Towards Fewer Annotations: Active Learning via Region Impurity and Prediction Uncertainty for Domain Adaptive Semantic Segmentation

Binhui Xie[1]   Longhui Yuan[1]   Shuang Li[1]✉   Chi Harold Liu[1]   Xinjing Cheng[2,3]

[1]School of Computer Science and Technology, Beijing Institute of Technology
[2]School of Software, BNRist, Tsinghua University   [3]Inceptio Technology

{binhuixie,longhuiyuan,shuangli,chiliu}@bit.edu.cn   cnorbot@gmail.com

Abstract

Self-training has greatly facilitated domain adaptive semantic segmentation, which iteratively generates pseudo labels on unlabeled target data and retrains the network. However, realistic segmentation datasets are highly imbalanced, pseudo labels are typically biased to the majority classes and basically noisy, leading to an error-prone and suboptimal model. In this paper, we propose a simple region-based active learning approach for semantic segmentation under a domain shift, aiming to automatically query a small partition of image regions to be labeled while maximizing segmentation performance. Our algorithm, Region Impurity and Prediction Uncertainty (RIPU), introduces a new acquisition strategy characterizing the spatial adjacency of image regions along with the prediction confidence. We show that the proposed region-based selection strategy makes more efficient use of a limited budget than image-based or point-based counterparts. Further, we enforce local prediction consistency between a pixel and its nearest neighbors on a source image. Alongside, we develop a negative learning loss to make the features more discriminative. Extensive experiments demonstrate that our method only requires very few annotations to almost reach the supervised performance and substantially outperforms state-of-the-art methods. The code is available at https://github.com/BIT-DA/RIPU.

1. Introduction

Semantic segmentation, the task of comprehending an image at the pixel level, is the foundation for numerous applications such as autonomous driving [63, 80], robot manipulation [51, 67], and medical analysis [48, 62]. Learning of segmentation models, however, relies heavily on vast quantities of data with pixel-wise annotations, which

✉ Corresponding author

(a) Target image   (b) Image-based selection (100%)

(c) Point-based selection (2.2%)   (d) Region-based selection (2.2%)
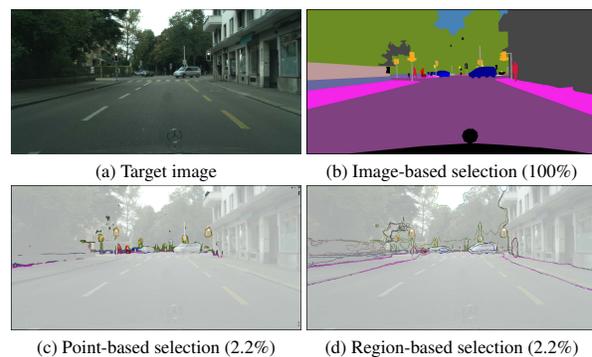
Figure 1. Illustration of different selection strategies. **Image-based selection** (e.g., MADA [41]) picks a few target samples and label the entire image, which is probably inefficient. **Point-based selection** (e.g., LabOR [54]) chooses scarce points about which the model is uncertain, while uncertainty estimation at point level is prone to lump pixels that come from particular categories. **Our region-based selection** asks for more annotations of regions with more categories as well as object boundaries in an effective way.

is onerous and prohibitively expensive [9, 35]. Further, it remains a major challenge to guarantee a good generalization to diverse testing situations. Various research efforts have been directed to address the above issues, with domain adaptation being promising methods [14, 29, 37, 64, 65, 72].

Recently, self-training has boosted domain adaptation, which retrains the network with the pseudo labels generated from confident predictions on the target domain [8, 40, 75, 83, 85–87]. Nevertheless, this competitive approach faces an inherent challenge: class unbalance is usually extreme. For instance, some classes e.g., "road" and "building", appear more frequently than others such as "rider" and "light". Thereby, pseudo labels are noisy and self-training would put heavy emphasis on classes with high frequency and sacrifice the performance on rare classes or small objects, resulting in undesired biases. Consequently, the performance lags far behind the supervised learning counterparts.

To overcome this obstacle and encourage maximizing

segmentation performance on the target domain, we show that a simple active learning strategy works well in adaptive semantic segmentation: annotating a small portion of image regions. Until recently, similar efforts have been made by Ning *et al*. [41] and Shin *et al*. [54]. The former uses multiple anchors to select representative target images to be labeled (Fig. 1b), which may be highly inefficient since it can waste the annotation budget on labeling redundant areas within objects. The latter utilizes an inconsistent mask of the bi-classifier predictions to query scarce points in each image for annotation (Fig. 1c). Although this process reduces human labor costs, under a severe domain shift, uncertainty estimation at the point level may be highly miscalibrated [60] or lead to sampling redundant points from certain categories. Moreover, both works are straightforward extensions of classification methods and weaken the significance of spatial proximity property in an image.

Driven by the above analysis, we present a simple, effective, and efficient active learning method, *Region Impurity and Prediction Uncertainty (RIPU)*, to assist domain adaptive semantic segmentation. A key design element of RIPU is to select the most diverse and uncertain regions in an image (Fig. 1d), eventually boosting the segmentation performance. To be concrete, we first generate the target pseudo labels from the model predictions and excavate all possible regions with the *k-square-neighbors* algorithm. Second, we take the entropy calculated on the percentage of internal pixels belonging to each distinct class as the region impurity score of each region. Finally, combining region impurity with the mean value of prediction uncertainty, i.e., the entropy of pixel prediction, a novel label acquisition strategy that jointly captures diversity and uncertainty is derived.

In this work, we introduce two labeling mechanisms for each target image, viz., "Region-based Annotating (RA)" (∼2.2% ground truth pixels) and "Pixel-based Annotating (PA)" (∼40 pixels). RA annotates every pixel in the selected regions—high annotation regime, while PA places its focus more on the labeling effort efficiency by selecting the center pixel within the region—low annotation regime. We further exploit local stability to enforce the prediction consistency between a certain pixel and its neighborhood pixels on the source domain and develop a negative learning loss to enhance the discriminative representation learning on the target domain. We demonstrate that our method can not only help the model to achieve near-supervised performance but also reduce human labeling costs dramatically.

In a nutshell, our contributions can be summarized as:

- We benchmark the performance of prior methods for active domain adaptation regarding semantic segmentation and uncover that methods using image-based or point-based selection strategies are not effective.

- We propose a region-based acquisition strategy for domain adaptive semantic segmentation, termed RIPU,

that utilizes region impurity and prediction uncertainty to identify image regions that are both diverse in spatial adjacency and uncertain in prediction output.

- We experimentally show that, with standard segmentation models, i.e., DeepLab-v2 and DeepLab-v3+, our method brings significant performance gains across two representative domain adaptation benchmarks, i.e., GTAV → Cityscapes, SYNTHIA → Cityscapes.

2. Related Work

Domain adaptation (DA) enables making predictions on an unlabeled target domain with the knowledge of a well-labeled source domain, which has been widely applied into an array of tasks such as classification [28, 30, 32, 37, 56, 65, 79], detection [7, 69] and segmentation [33, 34]. Initial studies minimize the discrepancy between source and target features to mitigate the domain gap [21, 36, 66]. As to semantic segmentation, most methods employ adversarial learning in three ways: appearance transfer [18, 31, 81], feature matching [22, 76, 84] and output space alignment [38, 64, 70].

Self-training has been gaining momentum as a competitive alternative, which trains the model with pseudo labels on the target domain [8, 40, 55, 75, 83, 85–87]. Popular as they are, the pseudo labels are noisy and rely primarily on a good initialization. Some efforts explore additional supervision to engage in this transfer. For example, Paul *et al*. [43] propose to use weak labels and Vu *et al*. [71] exploit dense depth information to perform adaptation. Another promising strategy to prevent such noise with minimal annotation workload is active learning, which we adopt in this work.

Active learning (AL) seeks to minimize labeling effort on an enormous dataset while maximizing performance of the model. Common strategies include uncertainty sampling [15, 20, 52] and representative sampling [1, 50, 58]. While label acquisition for dense prediction tasks such as segmentation is more expensive and laborious than image classification, there has been considerably less work [2, 4, 23, 53, 57, 77]. A recent example in [4] proposes to actively select image regions based on reinforcement learning, which is a more efficient way than labeling entire images.

Up to now, rather little work has been done to consider transferring annotations from a model trained on a given domain (a synthetic dataset) to a different domain (a real-world dataset) due to the dataset shift. However, it occurs frequently in practice but is not adequately addressed. In this work, we take a step forward to deal with this problem.

Active domain adaptation (ADA). Existing works mainly focus on image classification [13, 44–46, 61, 78]. To name a few, Prabhu *et al*. [44] combine the uncertainty and diversity into an acquisition round and integrate semi-supervised domain adaptation into a unified framework. Lately, Ning *et al*. [41] and Shin *et al*. [54] are among the first to study

the task of ADA applied to semantic segmentation, which greatly enhances the segmentation performance on the target domain. Ning *et al.* [41] put forward a multi-anchor strategy to actively select a subset of images and annotate the entire image, which is probably inefficient. While Shin *et al.* [54] present a more efficient point-based annotation with an adaptive pixel selector. But, the selected points are individual and discrete, neglecting the contextual structures of an image and pixel spatial contiguity within a region.

Though impressive, these methods neglect the value of spatial adjacency property within a region, and we argue that an effective and efficient region-based selection strategy is essential for lowering human labor costs while preserving model performance. In this work, we explore the spatial coherency of an image and favor the selection of the most diverse and uncertain image regions, promising high information content and low labeling costs.

3. Approach

3.1. Preliminaries and Motivation

Formally, in ADA semantic segmentation, we have a set of labeled source data $\mathcal{S} = \{(\mathbf{I}_s, \mathbf{Y}_s)\}$ and unlabeled target data $\mathcal{T} = \{(\mathbf{I}_t, \widetilde{\mathbf{Y}}_t)\}$, where $\mathbf{Y}_s$ is the pixel-wise label belonging to one of the C known classes in label space $\mathcal{Y}$ and $\widetilde{\mathbf{Y}}_t$ is the target active label that is initialized as $\emptyset$. The goal is to learn a function $h : \mathbf{I} \to \mathbf{Y}$ (a segmentation network parameterized by Θ) that achieves good segmentation performance on the target domain, with a few annotations.

Generally, the networks trained on the source domain generalize poorly on the target domain due to domain shift. To effectively transfer knowledge, recent advances resort to self-training techniques [40, 83, 87] and optimizes the cross-entropy loss with target pseudo labels $\widehat{\mathbf{Y}}_t$. But, the performance still far under-performs a fully supervised model. We hypothesis that pseudo labels are noisy, thus only the pixels whose prediction confidence is higher than a given threshold are accounted for retraining. In this way, the network training on target images is bootstrapped by the pixels that the model itself is confident in. To address this issue, we propose a simple yet effective active learning approach, Region Impurity and Prediction Uncertainty (RIPU), to assist domain adaptation by selecting a few informative image regions. The overall framework is illustrated in Fig. 2.

3.2. Region Generation

Traditional region-based active learning approaches in semantic segmentation simply divide an image into non-overlapping rectangles [4, 23] or employ superpixels algorithms such as SEEDS [11] to maintain object boundaries [2, 57]. However, we believe that these fixed regions are not flexible or suitable for a region-based selection strategy. The principal reason for this is that the model predic-

tions for adjacent regions should also be considered.

In this work, we consider *k-square-neighbors* of a pixel as a region, i.e., a regularly-shaped square of size $(2k + 1, 2k + 1)$ is treated as a region centered on each pixel. Formally, for any pixel $(i, j) \in \mathbb{R}^{H \times W}$ in an image $\mathbf{I}_t$ with H denoting the heigh and W for width, a region is denoted as

$$\mathcal{N}_k(i, j) = \{(u, v) \mid |u - i| \leq k, |v - j| \leq k\}. \quad (1)$$

Note that our method is a general one and any other division of region can be employed. In § 4.4, we further analyze the effectiveness of the shape and size of a region.

Discussion. *The concept of region in our work is different from that in recent work* [19]. RegionContrast [19] is proposed in a supervised manner and "region" denotes all pixel features belonging to one class, which is a semantic concept. In contrast, we aim to query informative regions for active domain adaptation and consider a regularly-shaped square centered on each pixel of image as a "region".

3.3. Region Impurity and Prediction Uncertainty

We notice that in practice, semantic segmentation often faces class imbalance problem as some events or objects are naturally rare in quantity. Consequently, the performance of the minority classes significantly degrades due to insufficient training samples. One can solve this during the training process through class re-balancing [73] or augmentation [17]. In contrast, we show that active learning can implicitly solve this by iteratively selecting a batch of samples to annotate at the data collection stage.

Given the pre-defined image regions, we describe our acquisition strategy to implement two different labeling mechanisms for each target image, viz., "Region-based Annotating (RA)" and "Pixel-based Annotating (PA)" as follows: at selection iteration n, we denote the model as Θ^n, a target image as $\mathbf{I}_t$ with corresponding active label $\widetilde{\mathbf{Y}}_t$, and an acquisition function $\mathcal{A}(\mathbf{I}_t; \Theta^n)$ is a function that the active learning system uses to query:

$$\mathcal{S}^* = \begin{cases} \mathcal{N}_k(i, j) \text{ if RA} \\ (i, j) \text{ else PA} \end{cases}, (i, j) = \underset{(u,v) \notin \widetilde{\mathbf{Y}}_t}{\arg \max} \mathcal{A}(\mathbf{I}_t; \Theta^n)^{(u,v)}. \quad (2)$$

In what follows, we propose an effective acquisition function that jointly captures region impurity and prediction uncertainty, favoring regions that are both diverse in spatial adjacency and uncertain in prediction output.

Region Impurity. Given a target image $\mathbf{I}_t$, we first pass it through the network Θ, and then get the softmax output, i.e., prediction $\mathbf{P}_t \in \mathbb{R}^{H \times W \times C}$. Target pseudo label can be directly derived via the maximum probability output, $\widehat{\mathbf{Y}}_t^{(i,j)} = \arg \max_{c \in \{1, \dots, C\}} \mathbf{P}_t^{(i,j,c)}$. With $\widehat{\mathbf{Y}}_t$, we divide a region $\mathcal{N}_k(i, j)$ of $\mathbf{I}_t$ into C subsets:

$$\mathcal{N}_k^c(i, j) = \{(u, v) \in \mathcal{N}_k(i, j) \mid \widehat{\mathbf{Y}}_t^{(u,v)} = c\}. \quad (3)$$
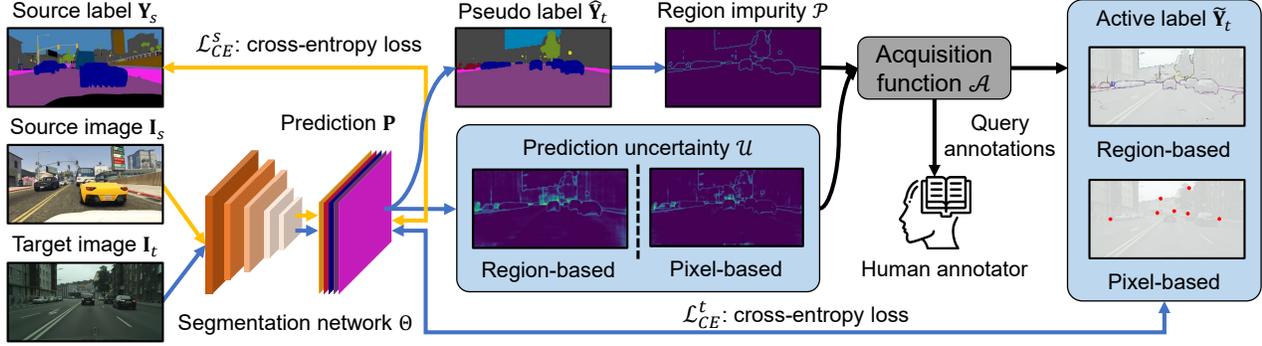
Figure 2. **The overview of the proposed RIPU.** At each round of active selection, we first utilize the current model to evaluate the region impurity $\mathcal{P}$ and prediction uncertainty $\mathcal{U}$ for each unlabeled target. Then we select a batch of regions (RA) or pixels (PA) using an acquisition function $\mathcal{A}$, and query annotations from a human annotator. Finally, the network is retrained using all the data labeled so far.

At the moment, we can collect statistical information about the categories in a region. If there are many objects in a region, we assume that it helps to train the network after being labeled. Mathematically, we introduce a novel region-based criterion, i.e., region impurity $\mathcal{P}$, to assess the significance of regions. Given a region $\mathcal{N}_k(i,j)$, its region impurity $\mathcal{P}^{(i,j)}$ is calculated as

$$\mathcal{P}^{(i,j)} = -\sum_{c=1}^{C} \frac{|\mathcal{N}_k^c(i,j)|}{|\mathcal{N}_k(i,j)|} \log \frac{|\mathcal{N}_k^c(i,j)|}{|\mathcal{N}_k(i,j)|}, \tag{4}$$

where $|\cdot|$ denotes the number of pixels in the set.

Prediction Uncertainty. As the predictions $\mathbf{P}$ carry the semantic relationship information, to measure uncertainty, we employ predictive entropy of each pixel $\mathcal{H}^{(i,j)}$. For C-way classification, $\mathcal{H}^{(i,j)} = -\sum_{c=1}^{C} \mathbf{P}_t^{(i,j,c)} \log \mathbf{P}_t^{(i,j,c)}$. For one thing, in region-based annotating (RA), we evaluate its prediction uncertainty by the average of all entropies of pixels within the region as follows,

$$\mathcal{U}^{(i,j)} = \frac{1}{|\mathcal{N}_k(i,j)|} \sum_{(u,v) \in \mathcal{N}_k(i,j)} \mathcal{H}^{(u,v)}. \tag{5}$$

For another, in pixel-based annotating (PA), the uncertainty of every pixel is the identity of entropy map, i.e., $\mathcal{U} = \mathcal{H}$.

Accordingly, we reckon the final acquisition function as

$$\mathcal{A}(\mathbf{I}_t; \Theta^n) = \mathcal{P} \odot \mathcal{U}, \tag{6}$$

where $\odot$ is the element-wise matrix multiplication. For RA/PA, we query regions/pixels with the highest scores, where the selected regions and the neighbors $\mathcal{N}_k$ of selected pixels are non-overlapping.

Discussion. *Why are region impurity and prediction uncertainty helpful for annotating informative pixels?* First, the region impurity prefers some categories with low occurrence frequency by exploring the spatial adjacency property. As shown in Fig. 4, this criterion exactly picks out areas of an image with more classes as well as object boundaries. Second, the prediction uncertainty is capable of finding the

---

Algorithm 1: Our proposed RIPU

1 **Require**: Labeled source data $(\mathbf{I}_s, \mathbf{Y}_s)$, unlabeled target data $\mathbf{I}_t$, total iteration N, selection rounds S, per-round budget b, and hyperparameters: τ, α_1, α_2.

2 **Define:** Target active label $\widetilde{\mathbf{Y}}_t = \emptyset$.

3 Pre-train the model Θ^0 on source data with Eq. (8).

4 **for** $n = 1$ *to* N **do**

5   Randomly select a batch source data and target and train the network Θ^n via Eq. (12).

6   **if** $n \in S$ **then**

7   Compute acquisition scores for all regions / pixels in $\mathbf{I}_t$ according to Eq. (6).

8   Iteratively sample regions/pixels using Eq.(2), and annotate by $\widetilde{\mathbf{Y}}_t[\mathcal{S}^*] = \mathbf{Y}_t[\mathcal{S}^*]$, until per-round b is exhausted.

9 **Return**: Final model parameters Θ^N.

---

regions whose model prediction is unsure as well. Finally, these two acquisition criteria gradually select the most diverse and uncertain regions for retraining the model, and in turn, make predictions closer to the ground-truth labels.

3.4. Training Objectives

With actively selected and annotated regions/pixels in target data, we can train the network to learn information exclusive to the target domain. Therefore, all labeled data from the source and target domain are used to fine-tune the network by optimizing the standard supervised loss:

$$\mathcal{L}_{sup} = \mathcal{L}_{CE}^s(\mathbf{I}_s, \mathbf{Y}_s) + \mathcal{L}_{CE}^t(\mathbf{I}_t, \widetilde{\mathbf{Y}}_t), \tag{7}$$

where $\mathcal{L}_{CE}$ is the categorical cross-entropy (CE) loss:

$$\mathcal{L}_{CE} = -\frac{1}{|\mathbf{I}|} \sum_{(i,j) \in \mathbf{I}} \sum_{c=1}^{C} \mathbf{Y}^{(i,j,c)} \log \mathbf{P}^{(i,j,c)}, \tag{8}$$

where $\mathbf{Y}^{(i,j,c)}$ denotes the label for pixel (i,j). Meanwhile, we enforce the prediction consistency between a certain

pixel and its neighborhood pixels in source images, which will promote the trained model to give more smooth predictions and avoid overfitting to the source data. Formally, this consistency regularization term is formulated as

$$\mathcal{L}_{cr}^s = \frac{1}{|\mathbf{I}_s|} \sum_{(i,j) \in \mathbf{I}_s} \left\| \mathbf{P}^{(i,j)} - \overline{\mathbf{P}}^{(i,j)} \right\|_1, \qquad (9)$$

where $\overline{\mathbf{P}}^{(i,j)}$ is the mean prediction of all pixels in a *1-square-neighbors*, i.e., the region size is 3×3, which can be calculated via $\overline{\mathbf{P}}^{(i,j)} = \frac{1}{|\mathcal{N}_1(i,j)|} \sum_{(u,v) \in \mathcal{N}_1(i,j)} \mathbf{P}^{(u,v)}$.

Additionally, the lower output probabilities for target images in the early stages of training actually show particular absent classes, called negative pseudo labels [25, 74]. For instance, it is hard to judge which class a pixel with a predicted value of [0.49, 0.50, 0.01] belongs to, but we can clearly know that it does not belong to the class with a score of 0.01. Thus, we assign negative pseudo labels as below

$$\pi(\mathbf{P}_t^{(i,j,c)}) = \begin{cases} 1 & \text{if } \mathbf{P}_t^{(i,j,c)} < \tau, \\ 0 & \text{otherwise}, \end{cases} \qquad (10)$$

where τ is the negative threshold, and we use $\tau = 0.05$ in this work. Note that the negative pseudo labels are binary labels. Hence, the negative learning loss is formulated as

$$\mathcal{L}_{nl}^t = \frac{-1}{\Lambda(\mathbf{I}_t)} \sum_{(i,j) \in \mathbf{I}_t} \sum_{c=1}^C \pi(\mathbf{P}_t^{(i,j,c)}) \log(1 - \mathbf{P}_t^{(i,j,c)}), \quad (11)$$

where $\Lambda(\mathbf{I}_t)$ denotes all available negative pseudo labels and is calculated by $\Lambda(\mathbf{I}_t) = \sum_{(i,j) \in \mathbf{I}_t} \sum_{c=1}^C \pi(\mathbf{P}_t^{(i,j,c)})$.

Eventually, equipped with all the above losses, we train the network with the following total objective:

$$\min_{\Theta} \mathcal{L}_{sup} + \alpha_1 \mathcal{L}_{cr}^s + \alpha_2 \mathcal{L}_{nl}^t. \qquad (12)$$

By default, the weighting coefficients α_1 and α_2 both are set to 0.1 and 1.0 respectively in all experiments. To be clear, we summarize the overall algorithm in Algorithm 1.

4. Experiments

Dataset. For evaluation, we adapt the segmentation from synthetic images, GTAV [47] and SYNTHIA [49] datasets, to real scenes, the Cityscapes [9] dataset. **GTAV** contains 24,966 1914×1052 images, sharing 19 classes as Cityscapes. **SYNTHIA** contains 9,400 1280×760 images, sharing 16 classes. **Cityscapes** includes high-quality urban scene images with a resolution of 2048×1024. It is split into 2,975 training images and 500 images for validation.

Implementation details. All experiments are conducted on a Tesla V100 GPU with PyTorch [42]. We adopt DeepLab-v2 [5] and DeepLab-v3+ [6] architectures with ResNet-101 [16] pre-trained on ImageNet [26] as backbone. Regarding the training, we use the SGD optimizer with momentum of 0.9 and weight decay of 5×10^{-4}, and employ

the "poly" learning rate schedule with the initial learning rate at 2.5×10^{-4}. For all experiments, we train about 40K iterations with batch size of 2, and source data are resized into 1280×720 while target data are resized into 1280×640. Concerning the hyperparameters, we set $k = 1$ for RA and $k = 32$ for PA, $\tau = 0.05$, $\alpha_1 = 0.1$ and $\alpha_2 = 1.0$. Readers can refer to Appendix B for more details.

Evaluation metric. As a common practice [40, 41, 54, 64, 83, 85, 87], we report the mean Intersection-over-Union (mIoU) [12] on the Cityscapes validation set. Specifically, we report the mIoU on the shared 19 classes for GTAV → Cityscapes and report the results on 13 (mIoU*) and 16 (mIoU) common classes for SYNTHIA → Cityscapes.

Annotation budget. The selection process lasts for a total of 5 rounds. For region-based annotating (RA), we select 2.2% and 5.0% regions in total compared to LabOR [54] and MADA [41], respectively. Regarding pixel-based annotating (PA), we select 40 pixels per image like LabOR [54].

Fully Supervised baseline. We operate with the annotations of both the source and target data as the upper bound.

4.1. Comparisons with the state-of-the-arts

The results on GTAV → Cityscapes and SYNTHIA → Cityscapes are shown in Table 1 and Table 2, respectively. It can be seen that our method dramatically outperforms prior leading self-training methods. Even though we only use 40 pixels per target image, Ours (PA) also shows substantial improvements over the prior breaking record method, i.e., ProDA, which implies that active learning is a promising and complementary solution for domain adaptation.

For the task of GTAV → Cityscapes, compared to LabOR [54], the best baseline model, Ours (PA) obtains an improvement of 2.0 mIoU, in the meantime, Ours (RA) exceeds by 3.0 mIoU. Similarly, Ours (RA) is able to easily beat MADA [41] and AADA [61] when using the same backbone (DeepLab-v3+) and same annotation budget (5%). While for the task of SYNTHIA → Cityscapes, as expected, both Ours (PA) and Ours (RA), are superior to their corresponding state-of-the-art methods.

To sufficiently realize the capacity of our method, we also compare it with the Full Supervised model, which is trained on both source and target domain with all images labeled. It is noteworthy that our method even outperforms Full Supervised with respect to some specific categories such as "rider", "bus", and "motor", suggesting that the proposed method can select marvelous regions to surpass the performance of supervised counterpart.

In a nutshell, the results listed in Table 1 and Table 2 show our method performs favorably against existing ADA and competing DA methods regarding semantic segmentation, and performs comparable to Full supervised model, confirming the proposed method is effective and efficient.

Table 1. **Comparison with previous results on task GTAV → Cityscapes.** We report the mIoU and best results are shown in **bold**.

Method	road	side.	buil.	wall	fence	pole	light	sign	veg.	terr.	sky	pers.	rider	car	truck	bus	train	motor	bike	mIoU
Source Only	75.8	16.8	77.2	12.5	21.0	25.5	30.1	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
CBST [86]	91.8	53.5	80.5	32.7	21.0	34.0	28.9	20.4	83.9	34.2	80.9	53.1	24.0	82.7	30.3	35.9	16.0	25.9	42.8	45.9
MRKLD [87]	91.0	55.4	80.0	33.7	21.4	37.3	32.9	24.5	85.0	34.1	80.8	57.7	24.6	84.1	27.8	30.1	26.9	26.0	42.3	47.1
Seg-Uncertainty [85]	90.4	31.2	85.1	36.9	25.6	37.5	48.8	48.5	85.3	34.8	81.1	64.4	36.8	86.3	34.9	52.2	1.7	29.0	44.6	50.3
TPLD [55]	94.2	60.5	82.8	36.6	16.6	39.3	29.0	25.5	85.6	44.9	84.4	60.6	27.4	84.1	37.0	47.0	31.2	36.1	50.3	51.2
DPL-Dual [8]	92.8	54.4	86.2	41.6	32.7	36.4	49.0	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
ProDA [83]	87.8	56.0	79.7	46.3	44.8	45.6	53.5	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) [43]	94.0	62.7	86.3	36.5	32.8	38.4	44.9	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 (40 pixels) [54]	**96.1**	**71.8**	**88.8**	47.0	46.5	**42.2**	**53.1**	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
Ours (PA, 40 pixels)	95.5	69.2	88.2	**48.0**	**46.5**	36.9	45.2	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**
LabOR (2.2%) [54]	**96.6**	**77.0**	89.6	47.8	50.7	**48.0**	**56.6**	**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
Ours (RA, 2.2%)	96.5	74.1	**89.7**	**53.1**	**51.0**	43.8	53.4	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**
Fully Supervised (100%)	96.8	77.5	90.0	53.5	51.5	47.6	55.6	62.9	90.2	58.2	92.3	73.7	52.3	92.4	74.3	77.1	64.5	52.4	70.1	70.2
AADA (5%)♯ [61]	92.2	59.9	87.3	36.4	45.7	46.1	50.6	59.5	88.3	44.0	90.2	69.7	38.2	90.0	55.3	45.1	32.0	32.6	62.9	59.3
MADA (5%)♯ [41]	95.1	69.8	88.5	43.3	48.7	45.7	53.3	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
Ours (RA, 5%)♯	**97.0**	**77.3**	**90.4**	**54.6**	**53.2**	47.7	**55.9**	**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**
Fully Supervised (100%)♯	97.4	77.9	91.1	54.9	53.7	51.9	57.9	64.7	91.1	57.8	93.2	74.7	54.8	93.6	76.4	79.3	67.8	55.6	71.3	71.9

Methods with ♯ are based on DeepLab-v3+ [6] and others are based on DeepLab-v2 [5] for a fair comparison.

Table 2. **Comparisons with previous results on task SYNTHIA → Cityscapes.** We report the mIoUs in terms of 13 classes (excluding the "wall", "fence", and "pole") and 16 classes. Best results are shown in **bold**.

Method	road	side.	buil.	wall*	fence*	pole*	light	sign	veg.	sky	pers.	rider	car	bus	motor	bike	mIoU	mIoU*
Source Only	64.3	21.3	73.1	2.4	1.1	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
CBST [86]	68.0	29.9	76.3	10.8	1.4	33.9	22.8	29.5	77.6	78.3	60.6	28.3	81.6	23.5	18.8	39.8	42.6	48.9
MRKLD [87]	67.7	32.2	73.9	10.7	1.6	37.4	22.2	31.2	80.8	80.5	60.8	29.1	82.8	25.0	19.4	45.3	43.8	50.1
DPL-Dual [8]	87.5	45.7	82.8	13.3	0.6	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
TPLD [55]	80.9	44.3	82.2	19.9	0.3	40.6	20.5	30.1	77.2	80.9	60.6	25.5	84.8	41.1	24.7	43.7	47.3	53.5
Seg-Uncertainty [85]	87.6	41.9	83.1	14.7	1.7	36.2	31.3	19.9	81.6	80.6	63.0	21.8	86.2	40.7	23.6	53.1	47.9	54.9
ProDA [24]	87.8	45.7	84.6	37.1	0.6	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) [43]	94.9	63.2	85.0	27.3	24.2	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
Ours (PA, 40 pixels)	95.8	71.9	87.8	39.9	41.5	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 (RA, 2.2%)	96.8	76.6	89.6	45.0	47.7	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
Fully Supervised (100%)	96.7	77.8	90.2	40.1	49.8	52.2	58.5	67.6	91.7	93.8	74.9	52.0	92.6	70.5	50.6	70.2	70.6	75.9
AADA (5%)♯ [61]	91.3	57.6	86.9	37.6	48.3	45.0	50.4	58.5	88.2	90.3	69.4	37.9	89.9	44.5	32.8	62.5	61.9	66.2
MADA (5%)♯ [41]	96.5	74.6	88.8	45.9	43.8	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
Ours (RA, 5%)♯	**97.0**	**78.9**	**89.9**	**47.2**	**50.7**	**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**
Fully Supervised (100%)♯	97.5	81.4	90.9	48.5	51.3	53.6	59.4	68.1	91.7	93.4	75.6	51.9	93.2	75.6	52.0	71.2	72.2	77.1

Methods with ♯ are based on DeepLab-v3+ [6] and others are based on DeepLab-v2 [5] for a fair comparison.

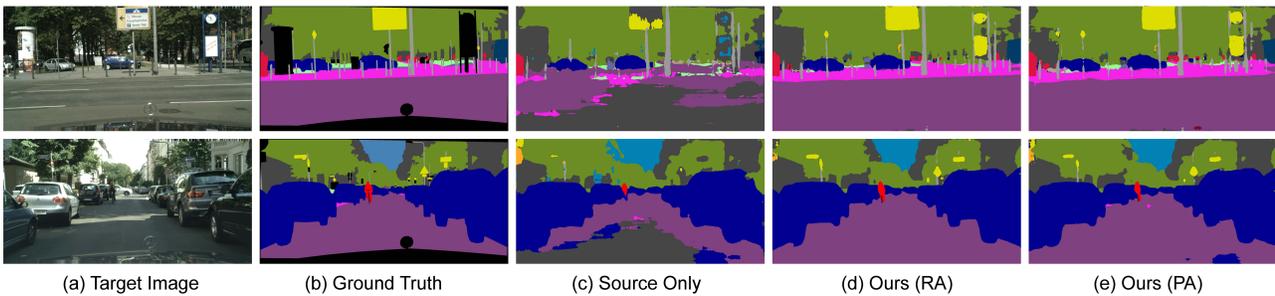| (a) Target Image | (b) Ground Truth | (c) Source Only | (d) Ours (RA) | (e) Ours (PA) |

Figure 3. **Visualization of segmentation results for the task GTAV → Cityscapes.** From left to right: original target image, ground-truth label, result predicted by Source Only model, result predicted by Ours (RA), and result predicted by Ours (PA) are shown one by one.

4.2. Qualitative results

We visualize the segmentation results predicted by RIPU and compare with those predicted by Source Only model in Fig. 3. The results predicted by RIPU are smoother and contain less spurious areas than those predicted by the Source Only model, showing that with RIPU, the performance has been largely improved, especially on hard classes.

Fig. 4 shows the selected regions for annotating from the RAND baseline, Uncertainty Only, Impurity Only, and Ours (RA). We observe that RAND uniformly picks image regions while Uncertainty Only and Impurity Only can cover

| (a) Ground Truth | (b) RAND | (c) Uncertainty Only | (d) Impurity Only | (e) Ours (RA) |

Figure 4. **Visualization of queried regions to annotate (2.2%) on GTAV→Cityscapes.** Compared to simple RAND baseline, Uncertainty Only, and Impurity Only, Ours (RA) is able to select the most diverse and uncertain regions of an image. Please zoom in to see the details.

Table 3. **Ablation study.** (a): use region impurity only as the selection criterion. (b): use prediction uncertainty only as the selection criterion. (c): combine impurity and uncertainty. (d): train with $\mathcal{L}_{cr}^s$ on source samples. (e): train with $\mathcal{L}_{nl}^t$ on target samples. (f): our full method RIPU for region-based annotating.

	Selection		Training		GTAV	SYNTHIA
Method	Impurity	Uncertainty	$\mathcal{L}_{cr}^s$	$\mathcal{L}_{nl}^t$	mIoU	mIoU
RAND: randomly selecting regions (2.2%)					63.8	64.7
Fully Supervised: all labeled source and target					70.2	70.6
(a)	✓				68.1	69.0
(b)		✓			66.2	67.9
(c)	✓	✓			68.5	69.2
(d)	✓	✓	✓		69.0	69.7
(e)	✓	✓		✓	69.2	69.8
(f)	✓	✓	✓	✓	**69.6**	**70.1**

a larger area of the entire image. However, uncertainty only tends to lump regions that are nearby together and Impurity Only focuses on regions gathered many categories. In contrast, Ours (RA) is also shown to pick diverse regions and not be grounded to a certain region of the image. More importantly, due to the limited budget, we adopt a relative small size to select regions, which might seem like an edge detection visually. Hence, we further operate a simple Canny algorithm [3] to select uncertain pixels from pre-detected edge and the results are reported in Appendix C.

4.3. Ablation Study

To further investigate the efficacy of each component of our RIPU, we perform ablation studies on both GTAV → Cityscapes and SYNTHIA → Cityscapes. As shown in Table 3, satisfactory and consistent gains from RAND baseline to our full method prove the effectiveness of each factor.

Effect of region impurity and prediction uncertainty. In Table 3, both (a) and (b) achieve a clear improvement compared to RAND. This indicates that the information from either region impurity or prediction uncertainty helps to identify beneficial image regions. Among them, the advantages of the region impurity are more prominent since the spatial adjacency property within an image can alleviate the issue

Table 4. Experiments on different **active selection methods**.

Method	Budget	mIoU	Budget	mIoU
RAND	40 pixels	60.3	2.2%	63.8
ENT [52]	40 pixels	55.0	2.2%	66.2
SCONF [10]	40 pixels	59.1	2.2%	66.5
Ours	PA, 40 pixels	**64.9**	RA, 2.2%	**68.5**

of class imbalance. Furthermore, we can see that (c) is better than both (a) and (b), suggesting that integrating the two into a unified selection criterion facilitates the sampling of more diverse and uncertain regions. riping the benefits of diverse region selection with both impurity and uncertainty

Effect of $\mathcal{L}_{cr}^s$ and $\mathcal{L}_{nl}^t$. For the training losses, we ablate $\mathcal{L}_{cr}^s$ and $\mathcal{L}_{nl}^t$ one by one. From the bottom half of Table 3, we can notice that (e) provides 0.7 mIoU gain on GTAV → Cityscapes and 0.6 on SYNTHIA → Cityscapes compared to (c). This demonstrates that $\mathcal{L}_{cr}^s$ does help to learn local consistent prediction and to avoid overfitting to the source data, which is a complementary factor to the impurity criterion. Similarly, $\mathcal{L}_{nl}^t$ on target samples brings a comparable improvement compared to (c). Further, our full RIPU obtains the best results, which indicate the importance and complementary of the proposed losses.

4.4. Further Analysis

Comparison of different active selection methods. To better understand the performance gains from our region-based selection strategy, we compare Ours (RA) and Ours (PA), with other common selection methods such as Random selection (RAND), entropy (ENT) [52] and softmax confidence (SCONF) [10], without $\mathcal{L}_{cr}^s$ and $\mathcal{L}_{nl}^t$. The results on the task GTAV → Cityscapes are reported in Table 4. Readers can refer to Appendix D for more details.

Due to a very limited budget of 40 pixels, both ENT and SCONF are severely compromised by the aggregation of selected pixels, and even fall behind the RAND baseline. By comparison, Ours (PA) achieves a significant gain over RAND by 4.6 mIoU, proving that our method properly mitigates the effect of redundant pixels to capitalize on a limited budget. Concerning the budget of 2.2% regions, consistent
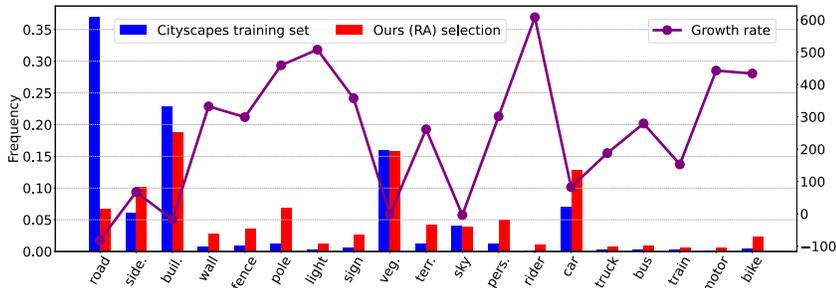
Figure 5. **Class frequencies (%)** in Cityscapes training set for the selected regions via our acquisition strategy. Additionally, we plot the growth rate between the above two.
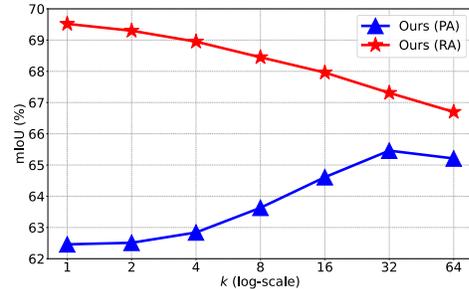

Figure 6. **Effect of region size** in the process of region generation on Cityscapes training set.

Table 5. Experiments on **source-free (SF) scenario**.

Method	Budget	GTAV mIoU	SYNTHIA mIoU	mIoU*
URMA [59]	-	45.1	39.6	45.0
LD [82]	-	45.5	42.6	50.1
SFDA [27]	-	53.4	52.0	60.1
Ours (RA)	2.2%	**67.1**	**68.7**	**74.1**

increases confirm the superiority and advantage.

Extension to source-free scenario. Considering the data sensitivity and security issues, we further evaluate the generalization of RIPU by extending it to source-free (SF) scenario [32]. Results in Table 5 validate the effectiveness of RIPU for this challenging DA task. More details about how RIPU works well are provided in Appendix E.

Class frequencies. Fig. 5 shows a detailed plot of the class frequencies of the 2.2% regions selected by Ours (RA) and true class frequencies of the Cityscape training set. As expected, we clearly see that the true distribution is exactly a long-tail distribution while our method is able to pick out more regions that contain rare classes. Particularly, it asks labels for more annotations of "light", "sign", "rider", "bus", "train" and "motor". This, along with the apparent gains in these classes from Table 1 and Table 2, confirms the diversity and balance of samples selected by our method.

Effect of region size. To understand how the region size influences our acquisition strategy, we vary k-square-neighbors on the task of GTAV $\rightarrow$ Cityscapes in Fig. 6. We can observe two opposite k-dependent laws for Ours (RA) and Ours (PA). For RA, the best performance is obtained at a small size $k = 1$ since smaller regions have a finer use of the sampling budget while larger regions contain more well-predicted pixels, causing a waste of budget. On the other hand, the best performance of PA occurs at a large size $k = 32$. We suppose when we perform PA with a smaller size, we may repeatedly pick pixels in a certain area. To make full use of the limited annotation budget, one should select pixels that cover the whole image.

Table 6. Effect of **region shape**.

Fixed rectangle	Superpixels	k-square-neighbors (**Ours**)
67.7	66.8	**69.6**

Effect of region shape In § 3.2, we define k-square-neighbors as a region that includes all possible areas within an image, allowing us adaptively select the most diverse and uncertain regions. In Table 6, we compare k-square-neighbors with other shapes of regions such as Fixed rectangle and Superpixels (the off-the-shelf SEEDS [11] algorithm) on the task GTAV $\rightarrow$ Cityscapes. As the model is training the importance of each region varies, however, the generated regions of the other two methods are fixed and do not fit well in this case. Thus, we observe the performance degradation using other methods, demonstrating the proposed region generation centered on each pixel is beneficial for acquiring the most concerning part in an image.

5. Conclusion

This paper presents Region Impurity and Prediction Uncertainty (RIPU), an active learning algorithm to deal with performance limitations of domain adaptive semantic segmentation at minimal label cost. We propose a novel region-based acquisition strategy for the selection of limited target regions that are both diverse in spatial contiguity and uncertain under the model. Other than that, we further explore local consistent regularization on the source domain and negative learning on the target domain to advance the acquisition process. Extensive experiments and ablation studies are conducted to verify the effectiveness of the proposed method. Our RIPU achieves new state-of-the-art results and performs comparably to the supervised counterparts. We believe that this work will facilitate the development of stronger machine learning system, including active image segmentation [53] and universal domain adaptation [39].

References

[1] Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *ICLR*, 2020. 2

[2] Lile Cai, Xun Xu, Jun Hao Liew, and Chuan Sheng Foo. Revisiting superpixels for active learning in semantic segmentation with realistic annotation costs. In *CVPR*, pages 10988–10997, 2021. 2, 3

[3] John F. Canny. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8(6):679–698, 1986. 7, 12

[4] Arantxa Casanova, Pedro O. Pinheiro, Negar Rostamzadeh, and Christopher J. Pal. Reinforced active learning for image segmentation. In *ICLR*, 2020. 2, 3

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(4):834–848, 2018. 5, 6, 14

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, pages 833–851, 2018. 5, 6

[7] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster R-CNN for object detection in the wild. In *CVPR*, pages 3339–3348, 2018. 2

[8] Yiting Cheng, Fangyun Wei, Jianmin Bao, Dong Chen, Fang Wen, and Wenqiang Zhang. Dual path learning for domain adaptation of semantic segmentation. In *ICCV*, pages 9082–9091, 2021. 1, 2, 6

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016. 1, 5

[10] Aron Culotta and Andrew McCallum. Reducing labeling effort for structured prediction tasks. In *AAAI*, pages 746–751, 2005. 7, 13

[11] Michael Van den Bergh, Xavier Boix, Gemma Roig, and Luc Van Gool. SEEDS: superpixels extracted via energy-driven sampling. *Int. J. Comput. Vis.*, 111(3):298–314, 2015. 3, 8

[12] Mark Everingham, S M Eslami, Luc Van Gool, Christopher K I Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vis.*, 111(1):98–136, 2015. 5

[13] Bo Fu, Zhangjie Cao, Jianmin Wang, and Mingsheng Long. Transferable query selection for active domain adaptation. In *CVPR*, pages 7272–7281, 2021. 2

[14] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17:59:1–59:35, 2016. 1

[15] Steve Hanneke et al. Theory of disagreement-based active learning. *Found. Trends Mach. Learn.*, 7(2-3):131–309, 2014. 2

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 5

[17] Ruifei He, Jihan Yang, and Xiaojuan Qi. Re-distributing biased pseudo labels for semi-supervised semantic segmentation: A baseline investigation. In *ICCV*, pages 6930–6940, 2021. 3

[18] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. CyCADA: Cycle-consistent adversarial domain adaptation. In *ICML*, pages 1989–1998, 2018. 2

[19] Hanzhe Hu, Jinshi Cui, and Liwei Wang. Region-aware contrastive learning for semantic segmentation. In *ICCV*, pages 16271–16281, 2021. 3

[20] Ajay J. Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. Multi-class active learning for image classification. In *CVPR*, 2009. 2

[21] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G. Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *CVPR*, pages 4893–4902, 2019. 2

[22] Guoliang Kang, Yunchao Wei, Yi Yang, Yueting Zhuang, and Alexander G. Hauptmann. Pixel-level cycle association: A new perspective for domain adaptive semantic segmentation. In *NeurIPS*, 2020. 2

[23] Tejaswi Kasarla, Gattigorla Nagendar, Guruprasad M. Hegde, Vineeth Balasubramanian, and C. V. Jawahar. Region-based active learning for efficient labeling in semantic segmentation. In *WACV*, pages 1109–1117, 2019. 2, 3

[24] Kwanyoung Kim, Dongwon Park, Kwang In Kim, and Se Young Chun. Task-aware variational adversarial active learning. In *CVPR*, pages 8166–8175, 2021. 6

[25] Youngdong Kim, Junho Yim, Juseung Yun, and Junmo Kim. NLNL: negative learning for noisy labels. In *ICCV*, pages 101–110, 2019. 5

[26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, pages 1097–1105, 2012. 5

[27] Jogendra Nath Kundu, Akshay Kulkarni, Amit Singh, Varun Jampani, and R. Venkatesh Babu. Generalize then adapt: Source-free domain adaptive semantic segmentation. In *ICCV*, pages 7046–7056, 2021. 8, 14

[28] Shuang Li, Chi Harold Liu, Binhui Xie, Limin Su, Zhengming Ding, and Gao Huang. Joint adversarial domain adaptation. In *ACM Multimedia*, pages 729–737, 2019. 2

[29] Shuang Li, Binhui Xie, Qiuxia Lin, Chi Harold Liu, Gao Huang, and Guoren Wang. Generalized domain conditioned adaptation network. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 1–1, 2021. 1

[30] Shuang Li, Mixue Xie, Kaixiong Gong, Chi Harold Liu, Yulin Wang, and Wei Li. Transferable semantic augmentation for domain adaptation. In *CVPR*, pages 11516–11525, 2021. 2

[31] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *CVPR*, pages 6936–6945, 2019. 2

[32] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *ICML*, pages 6028–6039, 2020. 2, 8

[33] Yuang Liu, Wei Zhang, and Jun Wang. Source-free domain adaptation for semantic segmentation. In *CVPR*, pages 1215–1224, 2021. 2

[34] Ziwei Liu, Zhongqi Miao, Xingang Pan, Xiaohang Zhan, Dahua Lin, Stella X. Yu, and Boqing Gong. Open compound domain adaptation. In *CVPR*, pages 12403–12412, 2020. 2

[35] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. 1

[36] Mingsheng Long, Yue Cao, Zhangjie Cao, Jianmin Wang, and Michael I. Jordan. Transferable representation learning with deep adaptation networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(12):3071–3085, 2019. 2

[37] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *NeurIPS*, pages 1647–1657, 2018. 1, 2

[38] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *CVPR*, pages 2507–2516, 2019. 2

[39] Xinhong Ma, Junyu Gao, and Changsheng Xu. Active universal domain adaptation. In *ICCV*, pages 8968–8977, 2021. 8

[40] Ke Mei, Chuang Zhu, Jiaqi Zou, and Shanghang Zhang. Instance adaptive self-training for unsupervised domain adaptation. In *ECCV*, pages 415–430, 2020. 1, 2, 3, 5

[41] Munan Ning, Donghuan Lu, Dong Wei, Cheng Bian, Chenglang Yuan, Shuang Yu, Kai Ma, and Yefeng Zheng. Multi-anchor active domain adaptation for semantic segmentation. In *ICCV*, pages 9112–9122, 2021. 1, 2, 3, 5, 6

[42] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8024–8035, 2019. 5

[43] Sujoy Paul, Yi-Hsuan Tsai, Samuel Schulter, Amit K. Roy-Chowdhury, and Manmohan Chandraker. Domain adaptive semantic segmentation using weak labels. In *ECCV*, pages 571–587, 2020. 2, 6

[44] Viraj Prabhu, Arjun Chandrasekaran, Kate Saenko, and Judy Hoffman. Active domain adaptation via clustering uncertainty-weighted embeddings. In *ICCV*, pages 8505–8514, 2021. 2

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

[46] Harsh Rangwani, Arihant Jain, Sumukh K Aithal, and R. Venkatesh Babu. S3vaada: Submodular subset selection for virtual adversarial active domain adaptation. In *ICCV*, pages 7516–7525, 2021. 2

[47] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *ECCV*, pages 102–118, 2016. 5

[48] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015. 1

[49] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*, pages 3234–3243, 2016. 5

[50] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *ICLR*, 2018. 2

[51] Rasha Sheikh, Andres Milioto, Philipp Lottes, Cyrill Stachniss, Maren Bennewitz, and Thomas Schultz. Gradient and log-based active learning for semantic segmentation of crop and weed for agricultural robots. In *ICRA*, pages 1350–1356, 2020. 1

[52] Yanyao Shen, Hyokun Yun, Zachary C. Lipton, Yakov Kronrod, and Animashree Anandkumar. Deep active learning for named entity recognition. In *ICLR*, 2018. 2, 7, 12, 13

[53] Gyungin Shin, Weidi Xie, and Samuel Albanie. All you need are a few pixels: semantic segmentation with pixelpick. *CoRR*, abs/2104.06394, 2021. 2, 8

[54] Inkyu Shin, Dong jin Kim, Jae Won Cho, Sanghyun Woo, Kwanyong Park, and In So Kweon. Labor: Labeling only if required for domain adaptive semantic segmentation. In *ICCV*, pages 8588–8598, 2021. 1, 2, 3, 5, 6

[55] Inkyu Shin, Sanghyun Woo, Fei Pan, and In So Kweon. Two-phase pseudo label densification for self-training based domain adaptation. In *ECCV*, pages 532–548, 2020. 2, 6

[56] Rui Shu, Hung H. Bui, Hirokazu Narui, and Stefano Ermon. A DIRT-T approach to unsupervised domain adaptation. In *ICLR*, 2018. 2

[57] Yawar Siddiqui, Julien Valentin, and Matthias Nießner. Viewal: Active learning with viewpoint entropy for semantic segmentation. In *CVPR*, pages 9430–9440, 2020. 2, 3

[58] Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *ICCV*, pages 5971–5980, 2019. 2

[59] Prabhu Teja Sivaprasad and François Fleuret. Uncertainty reduction for model adaptation in semantic segmentation. In *CVPR*, pages 9613–9623, 2021. 8, 14

[60] Jasper Snoek, Yaniv Ovadia, Emily Fertig, Balaji Lakshminarayanan, Sebastian Nowozin, D. Sculley, Joshua V. Dillon, Jie Ren, and Zachary Nado. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *NeurIPS*, pages 13969–13980, 2019. 2

[61] Jong-Chyi Su, Yi-Hsuan Tsai, Kihyuk Sohn, Buyu Liu, Subhransu Maji, and Manmohan Chandraker. Active adversarial domain adaptation. In *WACV*, pages 728–737, 2020. 2, 5, 6

[62] Saeid Asgari Taghanaki, Kumar Abhishek, Joseph Paul Cohen, Julien Cohen-Adad, and Ghassan Hamarneh. Deep semantic segmentation of natural and medical images: a review. *Artif. Intell. Rev.*, 54(1):137–178, 2021. 1

[63] Marvin Teichmann, Michael Weber, J. Marius Zöllner, Roberto Cipolla, and Raquel Urtasun. Multinet: Real-time joint semantic reasoning for autonomous driving. In *IV*, pages 1013–1020, 2018. 1

[64] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *CVPR*, pages 7472–7481, 2018. 1, 2, 5

[65] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, pages 2962–2971, 2017. 1, 2

[66] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *CoRR, abs/1412.3474*, 2014. 2

[67] Abhinav Valada, Johan Vertens, Ankit Dhall, and Wolfram Burgard. Adapnet: Adaptive semantic segmentation in adverse environmental conditions. In *ICRA*, pages 4644–4651, 2017. 1

[68] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *J. Mach. Learn. Res.*, 9(11), 2008. 13

[69] Vibashan VS, Vikram Gupta, Poojan Oza, Vishwanath A. Sindagi, and Vishal M. Patel. Mega-cda: Memory guided attention for category-aware unsupervised domain adaptive object detection. In *CVPR*, pages 4516–4526, 2021. 2

[70] Tuanhung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Perez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *CVPR*, pages 2517–2526, 2019. 2

[71] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. DADA: depth-aware domain adaptation in semantic segmentation. In *ICCV*, pages 7363–7372, 2019. 2

[72] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *CVPR*, pages 2517–2526, 2019. 1

[73] Tao Wang, Yu Li, Bingyi Kang, Junnan Li, Jun Hao Liew, Sheng Tang, Steven C. H. Hoi, and Jiashi Feng. The devil is in classification: A simple framework for long-tail instance segmentation. In *ECCV*, pages 728–744, 2020. 3

[74] Yuxi Wang, Jian Liang, and Zhaoxiang Zhang. Give me your trained model: Domain adaptive semantic segmentation without source data. *CoRR*, abs/2106.11653, 2021. 5

[75] Yuxi Wang, Junran Peng, and ZhaoXiang Zhang. Uncertainty-aware pseudo label refinery for domain adaptive semantic segmentation. In *ICCV*, pages 9092–9101, 2021. 1, 2

[76] Zhonghao Wang, Mo Yu, Yunchao Wei, Rogerio Feris, Jinjun Xiong, Wen-mei Hwu, Thomas S. Huang, and Honghui Shi. Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation. In *CVPR*, pages 12635–12644, 2020. 2

[77] Tsung-Han Wu, Yueh-Cheng Liu, Yu-Kai Huang, Hsin-Ying Lee, Hung-Ting Su, Ping-Chia Huang, and Winston H. Hsu. Redal: Region-based and diversity-aware active learning for point cloud semantic segmentation. In *ICCV*, pages 15510–15519, 2021. 2

[78] Binhui Xie, Longhui Yuan, Shuang Li, Chi Harold Liu, Xinjing Cheng, and Guoren Wang. Active learning for domain adaptation: An energy-based approach. *CoRR*, abs/2112.01406, 2021. 2

[79] Ruijia Xu, Guanbin Li, Jihan Yang, and Liang Lin. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *ICCV*, pages 1426–1435, 2019. 2

[80] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *CVPR*, pages 3684–3692, 2018. 1

[81] Yanchao Yang and Stefano Soatto. FDA: fourier domain adaptation for semantic segmentation. In *CVPR*, pages 4084–4094, 2020. 2

[82] Fuming You, Jingjing Li, Lei Zhu, Zhi Chen, and Zi Huang. Domain adaptive semantic segmentation without source data. In *ACM Multimedia*, pages 3293–3302, 2021. 8, 14

[83] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *CVPR*, pages 12414–12424, 2021. 1, 2, 3, 5, 6

[84] Qiming Zhang, Jing Zhang, Wei Liu, and Dacheng Tao. Category anchor-guided unsupervised domain adaptation for semantic segmentation. In *NeurIPS*, pages 433–443, 2019. 2

[85] Zhedong Zheng and Yi Yang. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *Int. J. Comput. Vis.*, 129(4):1106–1120, 2021. 1, 2, 5, 6

[86] Yang Zou, Zhiding Yu, BVK Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *ECCV*, pages 289–305, 2018. 1, 2, 6

[87] Yang Zou, Zhiding Yu, Xiaofeng Liu, B V K Vijaya Kumar, and Jinsong Wang. Confidence regularized self-training. In *ICCV*, pages 5982–5991, 2019. 1, 2, 3, 5, 6