Iterative Loop Learning Combining Self-Training and Active Learning for Domain Adaptive Semantic Segmentation

Licong Guan[1] Xue Yuan[1] ✉
[1]School of Electronic and Information Engineering, Beijing Jiaotong University
{lcguan941,xyuan}@bjtu.edu.cn

Abstract

Recently, self-training and active learning have been proposed to alleviate this problem. Self-training can improve model accuracy with massive unlabeled data, but some pseudo labels containing noise would be generated with limited or imbalanced training data. And there will be suboptimal models if human guidance is absent. Active learning can select more effective data to intervene, while the model accuracy can not be improved because the massive unlabeled data are not used. And the probability of querying sub-optimal samples will increase when the domain difference is too large, increasing annotation cost. This paper proposes an iterative loop learning method combining Self-Training and Active Learning (STAL) for domain adaptive semantic segmentation. The method first uses self-training to learn massive unlabeled data to improve model accuracy and provide more accurate selection models for active learning. Secondly, combined with the sample selection strategy of active learning, manual intervention is used to correct the self-training learning. Iterative loop to achieve the best performance with minimal label cost. Extensive experiments show that our method establishes state-of-the-art performance on tasks of GTAV → Cityscapes, SYNTHIA → Cityscapes, improving by 4.9% mIoU and 5.2% mIoU, compared to the previous best method, respectively. Code will be available.

1. Introduction

Semantic segmentation can understand image scenes at the pixel level and is crucial for various real-world applications. Thanks to the rapid development of deep learning, many advanced segmentation methods have been proposed and achieved great breakthroughs in various tasks such as autonomous driving [24], scene parsing [13, 50], medical analysis [2], and human-computer interaction [47].

✉ Corresponding author

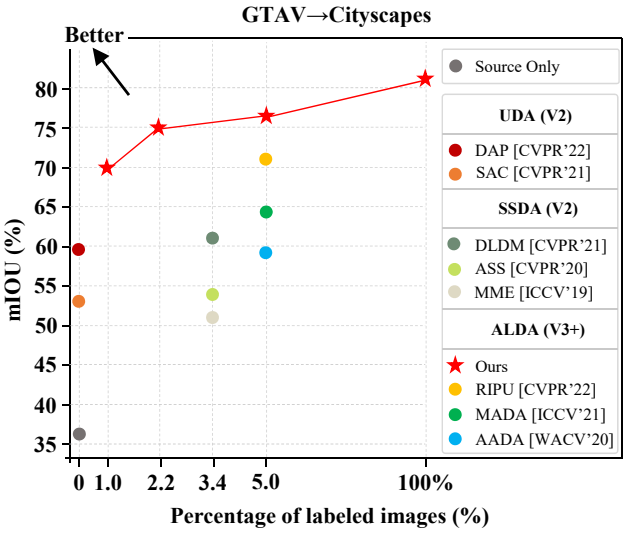


Figure 1. Performance comparison between our method and other methods on GTAV → Cityscapes. V2 and V3+ are based on DeepLab-v2 and DeepLab-V3+, respectively.

However, the traditional fully-supervised scenes [43, 76, 78] are very eager for images carefully annotated by human annotators, especially in aforementioned areas, a large number of annotated images are very expensive, time-consuming [13, 43], or even infeasible, which greatly hinders their widespread application. Therefore, it remains a major challenge to guarantee good generalization for different domain scenarios with minimal annotation cost.

Nowadays, many research works have proved that domain adaptation is one of the powerful means to address the above issues [23, 37, 44, 60, 63]. Among them, unsupervised domain adaptation (UDA) [26, 28, 33, 41, 69, 77] aims to solve this problem by leveraging the knowledge of label-rich data (source data) and transferring it to unlabeled data (target data) [52]. While it can avoid the intensive workload of manual annotation, the performance still lags far behind fully supervised models [56]. Furthermore, active learning can significantly improve performance on both classi-

fication and detection tasks by introducing a few additional manual annotations to a few selected samples from the target domain [57]. Nevertheless, active learning does not utilize the massive unlabeled data in the target domain, and only improves the accuracy through manual intervention, which makes the labeling cost difficult to control. And the probability of querying sub-optimal samples will increase when the domain difference is too large, increasing annotation cost. Moreover, in the process of semantic segmentation data labeling, the outline labeling of objects is usually completed by clicking with a mouse. Some previous studies on active learning domain adaptation are based on pixel selectors [56], which are difficult to generalize to real-world applications using a single image as the minimum selection unit and labeling based on object contours. Therefore, how to provide an accurate selection model for active learning and design a practical selection strategy is one the urgent problems to be solved.

Recently, self-training has greatly facilitated domain adaptation, as it can retrain the network with pseudo-labels generated from massive unlabeled data [11, 45, 65, 77, 79–81]. However, due to the limited and imbalanced training data, the pseudo-labels generated by self-training usually contain noise. This disadvantageous experience not only can not improve the accuracy of pseudo-labels but also further affect the performance of machine learning models without timely manual intervention and guidance. Therefore, how discovering and correcting false pseudo-labels in the self-training learning process is one of the problems to be solved urgently.

To solve the above problems, this paper proposes an iterative loop learning method combining self-training and active learning for domain-adaptive semantic segmentation. It learns massive unlabeled data through self-training to improve the accuracy in the target domain and provides an accurate selection model for active learning. Then, the active learning sample selection strategy is used to further correct the false pseudo labels in the self-training process through manual intervention. Self-training and active learning complement each other and help the model achieve the best performance with minimal annotation cost. (Fig. 1).

In a nutshell, our contributions can be summarized as:

- We propose a method combining self-training with active learning for domain-adaptive semantic segmentation, termed STAL. By complementing the advantages of self-training and active learning, we achieve the best performance for domain-adaptive semantic segmentation with minimal label cost.
- We propose an iterative loop learning strategy to optimize the performance of semantic segmentation models through three stages: warm-up learning, active selection, and incremental learning.

2. Related Work

Domain adaptation (DA) can transfer knowledge from a label-rich source domain to a label-scarce target domain, and recent work has achieved great success on a range of tasks. Such as classification [38, 40, 72], detection [10, 62], and segmentation [41, 42]. Most previous works have used adversarial learning [30, 39, 63, 68] in an attempt to reduce the domain gap between source and target features from the image level or feature level [34]. Recent work on domain-adaptive semantic segmentation can be mainly divided into two categories: adversarial training-based methods [63, 64] and self-training-based methods [31, 77, 79, 81]. For the first branch, most works tend to learn domain-invariant representations based on min-max adversarial optimization games by tricking the domain discriminator to obtain aligned feature distributions [63, 64]. The second branch focuses on how to generate high-quality pseudo-labels for target domain data for further model optimization [77, 79], which drives the development of self-training techniques.

Semi-supervised learning (SSL) involves two typical paradigms: consistency regularization [3, 48, 73] and entropy minimization [5, 6, 27, 49]. Consistency regularization forces the model to produce stable and consistent predictions on the same unlabeled data under various perturbations [71]. On the other hand, entropy minimization, generalized by self-training pipelines [1, 12, 15], exploits unlabeled target domain data in a way that uses pseudo-labels for training. For example, Wang *et al*. [67] propose the Semantic-Level Shift (ASS) framework, which introduces an additional semantic-level adaptation module by adversarial training on the corresponding outputs of the source and target labeled inputs. However, adversarial loss makes training unstable due to weak supervision. Zou *et al*. [80] proposed an iterative learning strategy with class-balanced and spatial priors for target instances. Tranheden *et al*. [59] proposed a domain-mixed self-training pipeline to improve training stability. Wang *et al*. [66] exploit unreliable pixels by adding a contrastive learning loss on top of self-training. The above studies can improve the model accuracy by using massive unlabeled data, while the noise problem of pseudo-labels can not be effectively solved. Therefore, we incorporate a sample selection strategy of active learning and employ human intervention to correct self-training learning in this paper.

Active learning (AL) aims to maximize the model performance with the least labeling cost of the dataset. Query rules are the core content of active learning, and commonly used query strategies are divided into uncertainty-based methods [4, 16] and diversity-based methods [22]. In the field of active learning domain adaptation research, previous work has mainly focused on classification tasks [21, 51].

Ning *et al*. [46] and Shin *et al*. [56] were the first to adopt active learning domains for semantic segmentation tasks. Among them, [46] proposed a multi-anchor strategy to actively select image subsets, which can be inefficient. [56] proposed a more efficient point-based annotation. However, the selected points ignore the pixel spatial continuity of the image. Recently, Xie *et al*. [70] greatly improved the segmentation performance in the target domain by exploring the consistency of the image space and selecting the most diverse and uncertain image regions. However, most active learning research works ignore the utilization of massive unlabeled data in the target domain, resulting in high labeling costs. This paper uses self-training to learn massive unlabeled data to improve the accuracy of the model, provide an accurate selection model for active learning, and reduce the cost of labeling.

3. Approach

In this section, we establish our problem mathematically and first outline our proposed method in § 3.1. Our strategies for self-training and active learning are presented in § 3.2 and § 3.3, respectively.

3.1. Overview

In domain-adapted semantic segmentation, we have a set of labeled source domain data $\mathcal{D}_s = \{(\mathbf{x}_s, \mathbf{y}_s)\}$ and incompletely labeled target data $\mathcal{D}_t = \{(\mathbf{x}_t, \mathbf{y}_t)\}$, where $\mathbf{y}_s$ is the pixel label belonging to one of the C known classes in the label space $\mathcal{Y}$. Our goal is to learn a function $f \circ h : \mathbf{x} \rightarrow \mathbf{y}$ (a semantic segmentation network parameterized by θ) that achieves good segmentation performance on the target domain with a small amount of labeled target data and a large amount of labeled source data and unlabeled target data.

In general, the success of CNN-based methods benefits from a large amount of manually labeled data and the assumption of independent and identical data distributions between training and testing samples. However, when a model trained on the training set (source domain) is directly applied to an unseen test scene (target domain), the performance drops significantly. To transfer knowledge efficiently, recent advances employ self-training techniques and optimize the cross-entropy loss using target pseudo-labels $\widehat{\mathbf{y}}_t$. Due to limited training data and class imbalance, the pseudo-labels generated by self-training often contain noise, and when lacking human intervention, inferior experience during training can lead to sub-optimal models. Further, active learning can pick out more effective data to intervene. However, active learning does not use massive amounts of unlabeled data, and it relies too much on picking models. And the probability of querying sub-optimal samples will increase when the domain difference is too large, resulting in an increase in annotation cost.

To solve the above problems, we propose an iterative loop learning method that combines self-training and active learning. The proposed framework consists of three stages: the first stage is self-training learning (Fig. 2a). Use a very small amount of labeled data and a large amount of unlabeled data to perform self-training learning to obtain a warm-up model. The second stage is active selection (Fig. 2b). Use the current model to predict the unlabeled target domain data and send it to the acquisition function. The third stage is image labeling (Fig. 2c). The selected samples are manually labeled and added to the labeled dataset of the target domain, and the first stage of self-training is repeated to obtain the final model.

3.2. Self-Training Learning

STAL follows a typical self-training framework, which consists of a student model and a teacher model. The teacher model and the student model have the same schema. The two models differ only in updating their weights, the student's weight θ_s is updated by convention, while the teacher model's weight θ_t is updated by the exponential moving average (EMA) of the student model. For labeled images of source and target domains, we use standard cross-entropy loss on them. And for each unlabeled target domain image, we bring it into the teacher model for prediction and obtain the pseudo-label according to the pixel prediction entropy. Subsequently, the student model is trained on unlabeled target domain data and corresponding pseudo-labels. Our optimization objective is to minimize the overall loss, which can be expressed as:

$$\mathcal{L} = \mathcal{L}_s + \lambda_u \mathcal{L}_u + \lambda_c \mathcal{L}_c \,, \tag{1}$$

where $\mathcal{L}_s$ (Eq. 2) and $\mathcal{L}_u$ (Eq. 3) represent the supervised and unsupervised loss applied to labeled and unlabeled images, respectively. $\mathcal{L}_c$ (Eq. 4) is the contrastive learning loss [66]. λ_u is the weight of the unsupervised loss, λ_c is the weight of the contrastive loss, and $\mathcal{L}_s$ and $\mathcal{L}_u$ are both cross-entropy (CE) loss.

$$\mathcal{L}_s = \frac{1}{N_l} \sum_{i=1}^{N_l} \frac{1}{WH} \sum_{j=1}^{WH} \ell_{ce}(f \circ h(\mathbf{x}_{i,j}^l; \theta), \mathbf{y}_{i,j}^l) \,, \tag{2}$$

$$\mathcal{L}_u = \frac{1}{N_u} \sum_{i=1}^{N_u} \frac{1}{WH} \sum_{j=1}^{WH} \ell_{ce}(f \circ h(\mathbf{x}_{i,j}^u; \theta), \hat{\mathbf{y}}_{i,j}^u) \,, \tag{3}$$

$$\mathcal{L}_c = -\frac{1}{C \times M} \sum_{c=0}^{C-1} \sum_{i=1}^{M} \log \left[\frac{e^{\langle \mathbf{a}_{ci}, \mathbf{a}_{ci}^+ \rangle / \omega}}{e^{\langle \mathbf{a}_{ci}, \mathbf{a}_{ci}^+ \rangle / \omega} + \sum_{j=1}^{N} e^{\langle \mathbf{a}_{ci}, \mathbf{a}_{cij}^- \rangle / \omega}} \right] , \tag{4}$$

Among them, $f \circ h$ is the composition function of h and f, which means that the image $\mathbf{x}_{i,j}$ is first sent to h to extract features, and then sent to f to obtain segmentation results. $\mathbf{y}_{i,j}^l$ is the manually annotated mask label for the j-th pixel

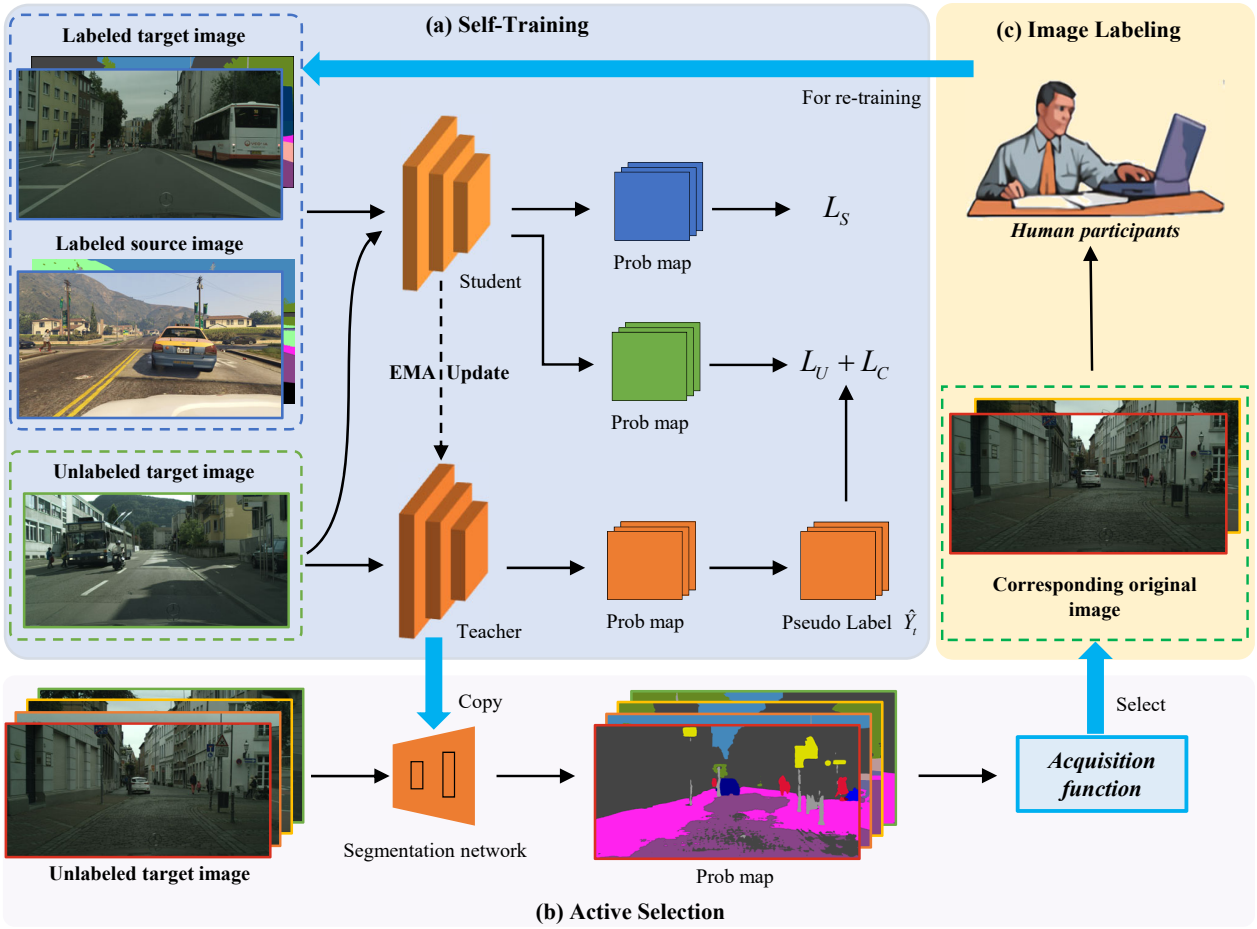


Figure 2. **The overview of the proposed STAL.** The proposed framework consists of three stages: the first stage (a) is self-training learning. Use a very small amount of labeled data and a large amount of unlabeled data to perform self-training learning to obtain a warm-up model. The second stage (b) is active selection. Use the current model to predict the unlabeled target domain data and send it to the acquisition function. The third stage (c) is image labeling. The selected samples are manually labeled and added to the labeled dataset of the target domain, and the first stage of self-training is repeated to obtain the final model.

in the i-th labeled image, $\hat{\mathbf{y}}^u_{i,j}$ is the pseudo-label for the j-th pixel in the i-th unlabeled image, N_l and N_u represents the number of labeled and unlabeled images in the training batch. W and H are the width and height of the input image. ℓ_{ce} is the standard cross-entropy loss. In the contrast loss Eq. 4, M is the total number of anchor pixels, and $\mathbf{a}_{ci}$ represents the representation of the i-th anchor of class c. The positive and negative samples corresponding to each anchor pixel are denoted as $\mathbf{a}^+_{ci}$ and $\mathbf{a}^-_{ci}$, $\langle \cdot, \cdot \rangle$ is the cosine similarity between features from two different pixels, whose range is limited between -1 to 1 according to ω. Following [66], we set $M = 50$, $N = 256$, and $\omega = 0.5$.

During training, some tail categories will introduce more pseudo-label noise due to insufficient training, which indirectly leads to the degradation of underperforming categories. We dynamically record the performance of each category during training by maintaining a confidence library, and for the underperforming categories, the confidence metric is shown in Eq. 5.

$$\text{Conf}^c = \frac{1}{N_l}\sum_{i=1}^{N_l}\frac{1}{N_i^c}\sum_{j=1}^{N_i^c}(f \circ h(x^c_{i,j};\theta)), c \in \{1,\ldots,C\}, \quad (5)$$

Among them, C is the category number, N_i^c indicates the number of pixels belonging to category c according to the ground truth label $y_{i,j}$, and $f \circ h(x^c_{i,j};\theta)$ indicates the c-th channel prediction result of the j-th pixel in the i-th image.

We use EMA to update the confidence by class at each training step, and the update criterion is shown in Eq. 6:

$$\text{Conf}^c_n \leftarrow \alpha \text{Conf}^c_{n-1} + (1-\alpha)\text{Conf}^c_n, \; c \in \{1,\ldots,C\}, \quad (6)$$

where n represents the n-th iteration and $\alpha \in [0, 1)$ is the momentum coefficient, which we set to 0.999 in our experiments. For underperforming categories, we improve by two data augmentation techniques, Copy_Paste [25] and Cutmix [75]. We calculate the sampling probability according

to Eq. 7, which converts the class confidence in the confidence base into the normalized sampling probability s.

$$\boldsymbol{s} = \text{Softmax}(1 - \text{Conf}). \tag{7}$$

3.3. Active Learning

Inferior experience during self-training can have an impact on the performance of machine learning models in the absence of human intervention. Therefore, we utilize active learning to correct self-training erroneous pseudo-labels.

Our sample acquisition strategy is as follows: Given an unlabeled target image $\mathbf{x}_t$ and a warm-up model θ_w, the acquisition function $\mathcal{A}$ is a function that the active learning system uses to query. First, the softmax output $\mathbf{P}_t$ of the unlabeled image $\mathbf{x}_t$ is obtained by warm-up model θ_w. Since the prediction $\mathbf{P}_t$ carries semantic relation information, we adopt the prediction entropy $\mathcal{H}$ of each pixel to measure the uncertainty. For a single image with a C classification, we evaluate the uncertainty $\widetilde{\mathbf{x}}_u$ by averaging all entropies of all pixels in the image. Calculated as follows:

$$\widetilde{\mathbf{x}}_u = -\frac{1}{WH}\sum_{j=1}^{WH}\sum_{c=1}^{C}\mathbf{P}_t^{i,j,c}\log\mathbf{P}_t^{i,j,c} \tag{8}$$

Among them, W and H are the width and height of the feature map, respectively, and $\mathbf{P}_t^{i,j,c}$ represents the c-th channel prediction result of the j-th pixel in the i-th image. After obtaining uncertainty results for all unlabeled images, we preferentially select the most uncertain images $\boldsymbol{S}$ for annotation according to Eq. 9.

$$\boldsymbol{S} = \text{argmax}\,\mathcal{A}(\widetilde{\mathbf{x}}_u) \tag{9}$$

4. Experiments

Dataset. To verify the effectiveness of the proposed method, we evaluate our method on two popular scenarios, transferring information from synthetic images GTAV [53] and SYNTHIA [54] to the real domain, the Cityscapes [13] dataset. **GTAV** is a synthetic image dataset containing 24,966 1914×1052 images, sharing 19 classes as Cityscapes. **SYNTHIA** is a synthetic urban scene dataset containing 9,400 1280×760 images, sharing 16 classes as Cityscapes. **Cityscapes** is an autonomous driving dataset of real urban scenes, containing 2,975 training images and 500 validation images, each with a resolution of 2048×1024.

Implementation details. All experiments are performed on NVIDIA A100 GPU with Pytorch. We adopt DeepLabv2 [7] and DeepLab-v3+ [8] architectures with ResNet-101 [29] pre-trained on ImageNet [14] as backbone. Regarding the training, we use the SGD optimizer with an initial learning rate of 0.0025, weight decay of 0.0001, and momentum of 0.9. For all experiments, we train about 200K iterations with batch size of 12, and data are resized into 769×769.

Evaluation metric. As a common practice [45, 46, 56, 70], we report the mean Intersection-over-Union (mIoU) [17] 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. We also add the 19-class evaluation for the SYNTHIA → Cityscapes task in § 4.4, and we believe that the extremely small amount of target domain data is sufficient to optimize the three classes missing from SYNTHIA.

Annotation budget. Previous active learning-based selection strategies use pixels or regions as selection units, and they select a fixed proportion (2.2% or 5%) of each sample in the dataset. While our sample selection strategy takes a single image as the smallest unit, for a fair comparison, we only select the corresponding percentage of images. The selection process is divided into two rounds, the first round we randomly select 1% (30 images) from the target domain dataset for self-training learning. In the second round, we selected 1.2% (35 images) or 4% (120 images). Therefore, we label 2.2% or 5% of the target domain data in total for self-training learning to obtain the final evaluation model.

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

Table 1 and Table 2 are the domain adaptation results of GTAV → Cityscapes and SYNTHIA → Cityscapes, respectively, and it can be seen that our method greatly outperforms the previous leading unsupervised domain adaptation and active learning domain adaptation methods. At the limit, our results using only 1% of the data are also substantial improvements over previous state-of-the-art unsupervised methods (DAP+ProDA [33]).

For the GTAV → Cityscapes task, based on using the same backbone (DeepLab-v3+), we can easily beat AADA [57] and MADA [46] with an annotation budget of 1%. Compared to the state-of-the-art model, using the same annotation budget (5%), our method achieves 4.9% mIoU improvement over RIPU [70]. Notably, our method significantly outperforms the contrastive methods in some specific categories, namely the tail category of Cityscapes (such as "traffic light", "traffic sign", "rider", "bus", "train", "motorcycle", and "bicycle"), which indicates that the proposed method can effectively alleviate the long-tailed distribution problem to outperform the adversary.

Our STAL is still competitive for the SYNTHIA → Cityscapes task. On the basis of using the same backbone (DeepLab-v3+), our method can beat all methods using 1% of the target data. Compared to the state-of-the-art model, our method achieves 5.2% mIoU improvement over RIPU [70] if the same annotation budget (5%) is used. Likewise, our method outperforms RIPU [70] on the tail categories of Cityscapes (such as "traffic light", "traffic sign", "rider", "bus", "motorcycle", and "bicycle").

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

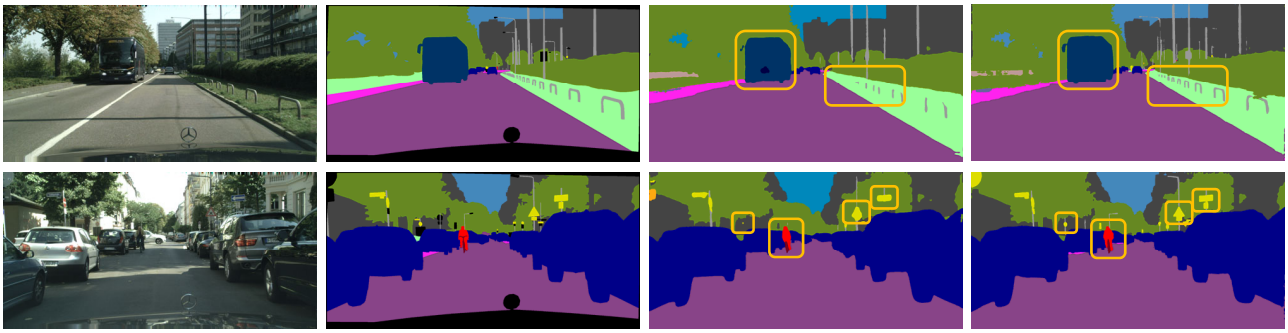
Method	Net.	road	side.	buil.	wall	fence	pole	light	sign	veg.	terr.	sky	pers.	rider	car	truck	bus	train	motor	bike	mIoU
Source Only	V2	75.8	16.8	77.2	12.5	21.0	25.5	30.1	20.1	81.3	24.6	70.3	53.8	26.4	49.9	17.2	25.9	6.5	25.3	36.0	36.6
CBST [80]	V2	91.8	53.5	80.5	32.7	21.0	34.0	28.9	20.4	83.9	34.2	80.9	53.1	24.0	82.7	30.3	35.9	16.0	25.9	42.8	45.9
MRKLD [81]	V2	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
SIM [68]	V2	90.6	44.7	84.8	34.3	28.7	31.6	35.0	37.6	84.7	43.3	85.3	57.0	31.5	83.8	42.6	48.5	1.9	30.4	39.0	49.2
SAC [1]	V2	90.4	53.9	86.6	42.4	27.3	45.1	48.5	42.7	87.4	40.1	86.1	67.5	29.7	88.5	49.1	54.6	9.8	26.6	45.3	53.8
ProDA [77]	V2	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
DAP+ProDA [33]	V2	94.5	63.1	89.1	29.8	47.5	50.4	56.7	58.7	89.5	50.2	87.0	73.6	38.6	91.3	50.2	52.9	0.0	50.2	63.5	59.8
LabOR (2.2%) [56]	V2	**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
RIPU (2.2%) [70]	V2	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
Ours (2.2%)	V2	96.4	74.6	**91.1**	45.9	**52.4**	**59.4**	**67.9**	**68.3**	**91.4**	50.0	**92.8**	**76.2**	**57.2**	**93.6**	**78.2**	**81.3**	**69.5**	**58.4**	**72.1**	**72.5**
AADA (5%) [57]	V3+	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%) [46]	V3+	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
RIPU (5%) [70]	V3+	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
Ours (1%)	V3+	95.2	67.0	90.9	47.4	49.6	60.9	68.2	67.5	90.9	44.6	91.5	81.3	60.5	93.9	67.2	76.6	47.9	54.7	74.8	70.0
Ours (2.2%)	V3+	96.5	75.6	91.2	46.7	53.6	62.1	70.3	76.0	91.4	52.1	94.1	82.0	60.8	94.4	**83.1**	86.4	71.9	61.2	75.8	75.0
Ours (5%)	V3+	**96.9**	**77.8**	**91.6**	46.7	**56.0**	**63.2**	**70.8**	**77.4**	**91.9**	54.9	**94.5**	**82.3**	**61.2**	**94.9**	79.3	**88.1**	**75.3**	**65.8**	**77.6**	**76.1**

Methods with V2 are based on DeepLab-v2 [7] and methods with V3+ are based on DeepLab-v3+ [8] 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	Net.	road	side.	buil.	wall*	fence*	pole*	light	sign	veg.	sky	pers.	rider	car	bus	motor	bike	mIoU	mIoU*
Source Only	V2	55.6	23.8	74.6	9.2	0.2	24.4	6.1	12.1	74.8	79.0	55.3	19.1	39.6	23.3	13.7	25.0	33.5	38.6
CBST [80]	V2	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 [81]	V2	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
SIM [68]	V2	83.0	44.0	80.3	-	-	-	17.1	15.8	80.5	81.8	59.9	33.1	70.2	37.3	28.5	45.8	-	52.1
SAC [1]	V2	89.3	47.2	85.5	26.5	1.3	43.0	45.5	32.0	87.1	89.3	63.6	25.4	86.9	35.6	30.4	53.0	52.6	59.3
ProDA [77]	V2	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
DAP+ProDA [33]	V2	84.2	46.5	82.5	35.1	0.2	46.7	53.6	45.7	89.3	87.5	75.7	34.6	91.7	73.5	49.4	60.5	59.8	64.3
RIPU (2.2%) [70]	V2	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
Ours (2.2%)	V2	**96.8**	76.1	**89.7**	**47.3**	**52.8**	**56.3**	**62.9**	**70.1**	**91.1**	**93.2**	**78.4**	**59.7**	**93.5**	78.2	**58.2**	**74.2**	**73.7**	**78.6**
AADA (5%) [57]	V3+	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%) [46]	V3+	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
RIPU (5%) [70]	V3+	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
Ours (1%)	V3+	96.8	74.8	90.0	34.0	46.3	60.9	68.0	74.8	90.2	92.5	81.1	58.2	93.0	72.3	63.4	75.6	73.2	79.3
Ours (2.2%)	V3+	96.8	76.3	90.9	48.1	54.2	62.4	69.0	77.3	91.0	93.7	82.2	**60.3**	94.2	80.0	63.8	76.0	76.0	80.9
Ours (5%)	V3+	**97.4**	**80.1**	**91.8**	38.6	**55.2**	**64.1**	**70.9**	**78.7**	**91.6**	**94.5**	**82.7**	60.1	**94.4**	**81.7**	**66.8**	**77.2**	**76.6**	**82.1**

Methods with V2 are based on DeepLab-v2 [7] and methods with V3+ are based on DeepLab-v3+ [8] for a fair comparison.

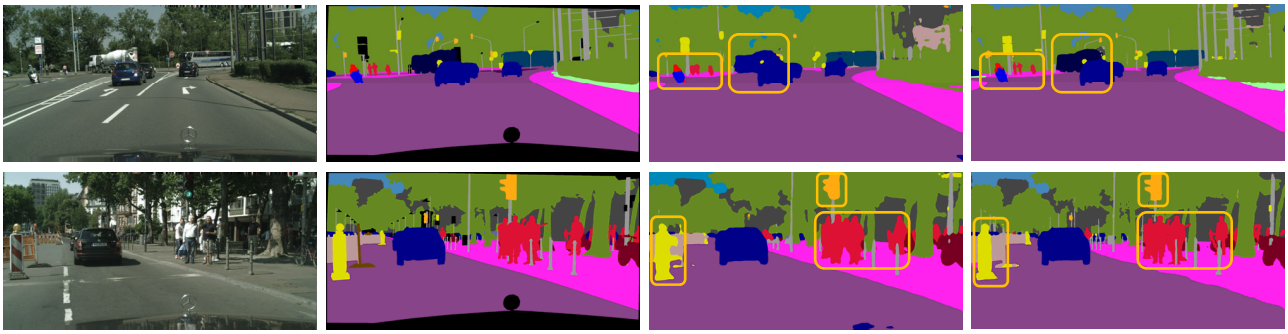


(a) Target Image **(b) Ground Truth** **(c) RIPU (5%)** **(d) Ours (5%)**

Figure 3. **Visualization of segmentation results for the task GTAV → Cityscapes.** From left to right: original target image, ground-truth label, result predicted by RIPU [70], and result predicted by Ours are shown one by one.

4.2. Qualitative results

We visualize the segmentation results predicted by STAL in GTAV → Cityscapes and compare them with the state-of-the-art RIPU [70] model prediction results. As can be seen from Fig. 3, our method has smoother prediction results for the head category (such as "pole"), and for Cityscapes tail categories (such as "bus", "rider", and "traffic sign") have a



(a) Target Image (b) Ground Truth (c) RIPU (5%) (d) Ours (5%)

Figure 4. **Visualization of segmentation results for the task SYNTHIA → Cityscapes.** From left to right: original target image, ground-truth label, result predicted by RIPU [70], and result predicted by Ours are shown one by one.

great improvement. The visualization of the segmentation results predicted by SYNTHIA → Cityscapes is shown in Fig. 4. Our method achieves accurate predictions for the missing class "truck" with only a small amount of target data. For the head class "person", STAL can achieve more detailed contour prediction than adversaries. For the tail categories of Cityscapes (such as "traffic sign" and "traffic light"), our predictions are greatly improved. Readers are referred to Appendix D for more details.

4.3. Ablation Study

To further investigate the efficacy of each component of our STAL, we perform ablation studies on GTAV → Cityscapes. We randomly select 5% of the data in the target domain to train in DeepLab-v3+ as the baseline. As shown in Table 3, satisfactory and consistent gains from the baseline to our full method demonstrate the effectiveness of each component. Compared to the baseline, $\mathcal{L}_u$ is able to leverage the unlabeled data of the target domain to improve performance by 10.38%. The performance is further improved by 0.9% and 1.63% after adding $\mathcal{L}_c$ and Data_aug in self-training, respectively. We applied both $\mathcal{L}_c$ and Data_aug strategies simultaneously and achieved a 4.32% improvement in model performance. Finally, by replacing the samples in the baseline with the samples picked by active learning, our performance reaches 76.11%, proving the effectiveness of the sample selection strategy based on prediction uncertainty.

4.4. Further Analysis

Extension to open set domain adaptive. SYNTHIA shares only 16 classes with Cityscapes, so previous methods only evaluate mIoU for 16 classes and 13 classes on this task. In active learning domain adaptation, we will report mIoU for 19 classes on the SYNTHIA → Cityscapes task due to the addition of data from the target domain. The evaluation results are shown in Table 4. Judging from the evaluation results of "terrain", "truck", and "train" missing three categories, we can still produce effects comparable to a large amount of labeled data under the premise of using a very small amount of target data.

Table 3. **Ablation study on the effectiveness of various components in our STAL**, including contrastive loss $\mathcal{L}_u$, $\mathcal{L}_c$, Data_aug, and Uncertainty.

Self-Training			Active Learning	GTAV
$\mathcal{L}_u$	$\mathcal{L}_c$	Data_aug	Uncertainty	mIoU
				59.33
✓				69.71
✓	✓			70.61
✓		✓		71.34
✓	✓	✓		74.03
✓	✓	✓	✓	**76.11**

Table 4. Experiments with **Open Set Domain Adaptation** on task SYNTHIA → Cityscapes.

Method	Net.	terrain	truck	train	mIoU
Ours (1%)	V3+	41.5	51.6	48.6	69.1
Ours (2.2%)		53.5	74.7	59.9	73.9

Comparison with SSDA and SSL methods. To better understand the performance improvement of strategies combining self-training with active learning, we compare our method with semi-supervised domain adaptation and semi-supervised learning methods. Among them, semi-supervised learning only uses the data of single-domain Cityscapes. The results are shown in Table 5.

Extension to source-free scenario. Due to data privacy and constraints on computing resources, domain adaptation sometimes fails to obtain source domain datasets. We further evaluate the generalization of STAL by extending to source-free domain adaptation (SFDA). Comparing methods include URMA [19], LD [74], SFDA [36], and

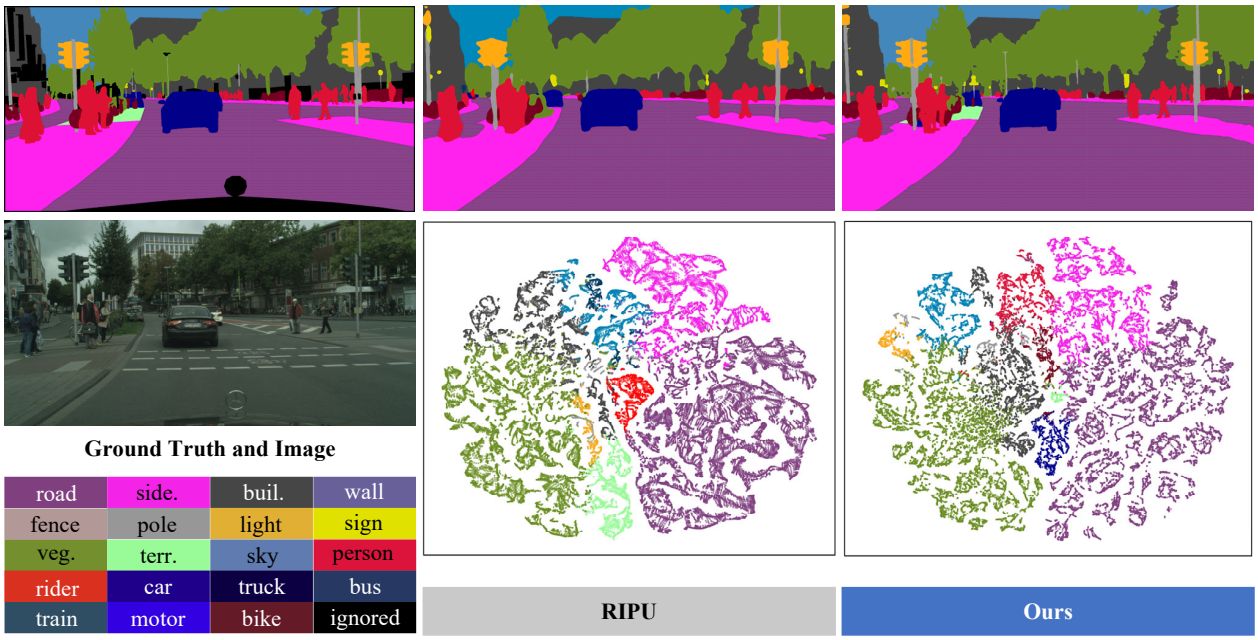


Figure 5. t-SNE analysis [61] of active learning method RIPU [70] and our method. The visualization of embedded features further demonstrates that our method can exhibit the clearest clustering.

Table 5. Comparisons with the **SSDA and SSL methods** on task GTAV → Cityscapes, SYNTHIA → Cityscapes. We report the mIoUs in terms of 19 classes and 13 classes.

Type	Method	Net.	GTAV mIoU	SYNTHIA mIoU*
SSDA	MME (3.4%) [55]	V2	52.6	59.6
	ASS (3.4%) [67]		54.2	62.1
	DIDM (3.4%) [9]		61.2	68.4
SSL	Cutmix (3.4%) [20]	V2	50.8	61.3
	DST-CBC (3.4%) [18]		48.7	59.7
ALDA	**Ours (2.2%)**	V2	**72.5**	**78.6**
SSL	GCT (3.1%) [35]	V3+	63.2	-
	MT (3.1%) [58]		64.1	-
	CCT (3.1%) [48]		66.4	-
	Cutmix (3.1%) [20]		69.1	-
	AEL (3.1%) [32]		74.3	-
	U2PL+AEL (6.3%) [9]		74.9	-
ALDA	**Ours (2.2%)**	V3+	75.0	80.9
	Ours (5.0%)		**76.1**	**82.1**

RIPU [70]. Results in Table 6 validate the effectiveness of STAL for this challenging DA task. More details about how STAL works well are provided in Appendix B.

t-SNE Visualization. To better develop intuition, we draw t-SNE visualization [61] of the learned feature representations for contrast methods (RIPU [70]) and ours STAL in Fig. 5. We randomly select an image from the target domain and then map its high-dimensional latent feature representation into 2D space. In Fig. 5, our method is able to separate the features between different categories better, and the decision boundary is clearer than other methods. This shows the discriminative power of STAL contrasting adaptations.

Table 6. Experiments on **source-free domain adaptation scenario (SFDA)**.

Method	Net.	Budget	GTAV mIoU	SYNTHIA mIoU	SYNTHIA mIoU*
URMA [19]	V2	-	45.1	39.6	45.0
LD [74]		-	45.5	42.6	50.1
SFDA [36]		-	53.4	52.0	60.1
RIPU [70]		2.2%	67.1	68.7	74.1
Ours	V2	2.2%	**70.4**	**72.0**	**78.1**

5. Conclusion

We propose an iterative loop learning algorithm that combines self-training and active learning. It successfully enhances the potential of combining the self-training paradigm with active learning to achieve the best performance for domain-adaptive semantic segmentation with minimal label cost. STAL leverages the ability of self-training to learn from massive unlabeled data to improve accuracy in the target domain and provide an accurate selection model for active learning. Then, the disadvantaged experience in the self-training is further corrected through active learning. The effectiveness of the proposed method is validated through extensive experiments and ablation studies, and our STAL achieves new state-of-the-art results.

References

[1] Nikita Araslanov and Stefan Roth. Self-supervised augmentation consistency for adapting semantic segmentation. In *CVPR*, pages 15384–15394, 2021. 2, 6

[2] 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

[3] Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. *NeurIPS*, 27, 2014. 2

[4] William H Beluch, Tim Genewein, Andreas Nürnberger, and Jan M Köhler. The power of ensembles for active learning in image classification. In *CVPR*, pages 9368–9377, 2018. 2

[5] Paola Cascante-Bonilla, Fuwen Tan, Yanjun Qi, and Vicente Ordonez. Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. In *AAAI*, volume 35, pages 6912–6920, 2021. 2

[6] Huaian Chen, Yi Jin, Guoqiang Jin, Changan Zhu, and Enhong Chen. Semisupervised semantic segmentation by improving prediction confidence. *IEEE Trans. Neural Networks Learn. Syst.*, 2021. 2

[7] 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 TPAMI*, 40(4):834–848, 2017. 5, 6

[8] 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 801–818, 2018. 5, 6

[9] Shuaijun Chen, Xu Jia, Jianzhong He, Yongjie Shi, and Jianzhuang Liu. Semi-supervised domain adaptation based on dual-level domain mixing for semantic segmentation. In *CVPR*, pages 11018–11027, 2021. 8

[10] 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

[11] 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. 2

[12] Charles Corbiere, Nicolas Thome, Antoine Saporta, Tuan-Hung Vu, Matthieu Cord, and Patrick Perez. Confidence estimation via auxiliary models. *IEEE TPAMI*, 2021. 2

[13] 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

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 5

[15] Jiahua Dong, Yang Cong, Gan Sun, Zhen Fang, and Zhengming Ding. Where and how to transfer: knowledge aggregation-induced transferability perception for unsupervised domain adaptation. *IEEE TPAMI*, 2021. 2

[16] Bo Du, Zengmao Wang, Lefei Zhang, Liangpei Zhang, and Dacheng Tao. Robust and discriminative labeling for multi-label active learning based on maximum correntropy criterion. *IEEE TIP*, 26(4):1694–1707, 2017. 2

[17] Mark Everingham, SM Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 111(1):98–136, 2015. 5

[18] Zhengyang Feng, Qianyu Zhou, Guangliang Cheng, Xin Tan, Jianping Shi, and Lizhuang Ma. Semi-supervised semantic segmentation via dynamic self-training and classbalanced curriculum. *arXiv preprint arXiv:2004.08514*, 1(2):5, 2020. 8

[19] Francois Fleuret et al. Uncertainty reduction for model adaptation in semantic segmentation. In *CVPR*, pages 9613–9623, 2021. 7, 8

[20] Geoff French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. *arXiv preprint arXiv:1906.01916*, 2019. 8

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

[22] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *ICML*, pages 1183–1192. PMLR, 2017. 2

[23] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 17(1):2096–2030, 2016. 1

[24] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, pages 3354–3361. IEEE, 2012. 1

[25] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *CVPR*, pages 2918–2928, 2021. 4

[26] Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *ICCV*, pages 999–1006. IEEE, 2011. 1

[27] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. *NeurIPS*, 17, 2004. 2

[28] Xiaoqing Guo, Chen Yang, Baopu Li, and Yixuan Yuan. Metacorrection: Domain-aware meta loss correction for unsupervised domain adaptation in semantic segmentation. In *CVPR*, pages 3927–3936, 2021. 1

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

[30] 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. Pmlr, 2018. 2

[31] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *CVPR*, pages 9924–9935, 2022. 2

[32] Hanzhe Hu, Fangyun Wei, Han Hu, Qiwei Ye, Jinshi Cui, and Liwei Wang. Semi-supervised semantic segmentation via adaptive equalization learning. *NeurIPS*, 34:22106–22118, 2021. 8

[33] Xinyue Huo, Lingxi Xie, Hengtong Hu, Wengang Zhou, Houqiang Li, and Qi Tian. Domain-agnostic prior for transfer semantic segmentation. In *CVPR*, pages 7075–7085, 2022. 1, 5, 6

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

[35] Zhanghan Ke, Di Qiu, Kaican Li, Qiong Yan, and Rynson WH Lau. Guided collaborative training for pixel-wise semi-supervised learning. In *ECCV*, pages 429–445. Springer, 2020. 8

[36] 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. 7, 8

[37] Shuang Li, Binhui Xie, Qiuxia Lin, Chi Harold Liu, Gao Huang, and Guoren Wang. Generalized domain conditioned adaptation network. *IEEE TPAMI*, 2021. 1

[38] 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

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

[40] 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. PMLR, 2020. 2

[41] Yuang Liu, Wei Zhang, and Jun Wang. Source-free domain adaptation for semantic segmentation. In *CVPR*, pages 1215–1224, 2021. 1, 2

[42] Ziwei Liu, Zhongqi Miao, Xingang Pan, Xiaohang Zhan, Dahua Lin, Stella X Yu, and Boqing Gong. Open compound domain adaptation. In *CVPR*, pages 12406–12415, 2020. 2

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

[44] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *NeurIPS*, 31, 2018. 1

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

[46] 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. 3, 5, 6

[47] Markus Oberweger, Paul Wohlhart, and Vincent Lepetit. Hands deep in deep learning for hand pose estimation. *arXiv preprint arXiv:1502.06807*, 2015. 1

[48] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *CVPR*, pages 12674–12684, 2020. 2, 8

[49] Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V Le. Meta pseudo labels. In *CVPR*, pages 11557–11568, 2021. 2

[50] Yongri Piao, Wei Ji, Jingjing Li, Miao Zhang, and Huchuan Lu. Depth-induced multi-scale recurrent attention network for saliency detection. In *ICCV*, pages 7254–7263, 2019. 1

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

[52] Stephan R Richter, Zeeshan Hayder, and Vladlen Koltun. Playing for benchmarks. In *ICCV*, pages 2213–2222, 2017. 1

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

[54] 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

[55] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *ICCV*, pages 8050–8058, 2019. 8

[56] 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

[57] Jong-Chyi Su, Yi-Hsuan Tsai, Kihyuk Sohn, Buyu Liu, Subhransu Maji, and Manmohan Chandraker. Active adversarial domain adaptation. In *WACV*, pages 739–748, 2020. 2, 5, 6

[58] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *NeurIPS*, 30, 2017. 8

[59] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. In *WACV*, pages 1379–1389, 2021. 2

[60] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, pages 7167–7176, 2017. 1

[61] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(11), 2008. 8

[62] 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

[63] 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, 2

[64] Haoran Wang, Tong Shen, Wei Zhang, Ling-Yu Duan, and Tao Mei. Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation. In *ECCV*, pages 642–659. Springer, 2020. 2

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

[66] Yuchao Wang, Haochen Wang, Yujun Shen, Jingjing Fei, Wei Li, Guoqiang Jin, Liwei Wu, Rui Zhao, and Xinyi Le. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *CVPR*, pages 4248–4257, 2022. 2, 3, 4

[67] Zhonghao Wang, Yunchao Wei, Rogerio Feris, Jinjun Xiong, Wen-Mei Hwu, Thomas S Huang, and Honghui Shi. Alleviating semantic-level shift: A semi-supervised domain adaptation method for semantic segmentation. In *CVPR*, pages 936–937, 2020. 2, 8

[68] 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, 6

[69] Binhui Xie, Shuang Li, Mingjia Li, Chi Harold Liu, Gao Huang, and Guoren Wang. Sepico: Semantic-guided pixel contrast for domain adaptive semantic segmentation. *arXiv preprint arXiv:2204.08808*, 2022. 1

[70] Binhui Xie, Longhui Yuan, Shuang Li, Chi Harold Liu, and Xinjing Cheng. Towards fewer annotations: Active learning via region impurity and prediction uncertainty for domain adaptive semantic segmentation. In *CVPR*, pages 8068–8078, 2022. 3, 5, 6, 7, 8

[71] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *NeurIPS*, 33:6256–6268, 2020. 2

[72] 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

[73] Yi Xu, Lei Shang, Jinxing Ye, Qi Qian, Yu-Feng Li, Baigui Sun, Hao Li, and Rong Jin. Dash: Semi-supervised learning with dynamic thresholding. In *ICML*, pages 11525–11536. PMLR, 2021. 2

[74] Fuming You, Jingjing Li, Lei Zhu, Zhi Chen, and Zi Huang. Domain adaptive semantic segmentation without source data. In *ACM MM*, pages 3293–3302, 2021. 7, 8

[75] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, pages 6023–6032, 2019. 4

[76] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *CVPR*, pages 7151–7160, 2018. 1

[77] 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, 6

[78] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, pages 2881–2890, 2017. 1

[79] Zhedong Zheng and Yi Yang. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *IJCV*, 129(4):1106–1120, 2021. 2

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

[81] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *ICCV*, pages 5982–5991, 2019. 2, 6