MADAv2: Advanced Multi-Anchor Based Active Domain Adaptation Segmentation

Munan Ning, Donghuan Lu, Yujia Xie, Dongdong Chen, Dong Wei, Yefeng Zheng, Yonghong Tian, Shuicheng Yan, Li Yuan

Abstract—Unsupervised domain adaption has been widely adopted in tasks with scarce annotated data. Unfortunately, mapping the target-domain distribution to the source-domain unconditionally may distort the essential structural information of the target-domain data, leading to inferior performance. To address this issue, we firstly propose to introduce active sample selection to assist domain adaptation regarding the semantic segmentation task. By innovatively adopting multiple anchors instead of a single centroid, both source and target domains can be better characterized as multimodal distributions, in which way more complementary and informative samples are selected from the target domain. With only a little workload to manually annotate these active samples, the distortion of the target-domain distribution can be effectively alleviated, achieving a large performance gain. In addition, a powerful semi-supervised domain adaptation strategy is proposed to alleviate the long-tail distribution problem and further improve the segmentation performance. Extensive experiments are conducted on public datasets, and the results demonstrate that the proposed approach outperforms state-of-the-art methods by large margins and achieves similar performance to the fully-supervised upperbound, *i.e.*, 71.4% mIoU on GTA5 and 71.8% mIoU on SYNTHIA. The effectiveness of each component is also verified by thorough ablation studies. Code is available at https://github.com/munanning/MADAv2.

Index Terms—Active learning, domain adaptation, semi-supervised learning, clustering, semantic segmentation

1 Introduction

As a fundamental task of computer vision, image segmentation has been long studied. In the recent decade, the rapid development of deep learning has brought great advances to the various tasks on top of image segmentation, such as autonomous driving [1], scene parsing [2], [3], object detection [4]–[7] and human-computer interaction [8]. However, it also leads to a hunger for huge training data, which is usually laborious and costly to obtain, especially for some expertise-demanding or complicated applications, *e.g.*, medical image segmentation [9]–[13] and auto-driving tasks [14]. This data insufficiency issue has greatly limited the application of automatic image segmentation in real-world scenarios.

One of the representative paradigms to solve this issue is unsupervised domain adaptation (UDA) [16]–[19]. UDA methods try to align the target-domain distribution towards the source-domain distribution, and apply the networks trained with the supervision of only the source data to the target data. Though the UDA methods have gained impressive achievements, they tend to undermine the latent structural pattern of the target domain as a result of them forcing representations of the target-domain to fit the distribution of the source-domain, which may lead to substantial performance degradation.

- *Munan Ning, Yonghong Tian, Li Yuan are with Peking University, School of Electronic and Computer Engineering, Shenzhen Graduate School, China, and also with PengCheng Laboratory. E-mail:{munanning, yhtian, yuanli-ece}@pku.edu.cn.*
- *Donghuan Lu, Dong Wei, Yefeng Zheng are with Jarvis Lab, Tencent Healthcare Co., China.*
- *Yujia Xie, Dongdong Chen are with the Microsoft Cloud AI, Redmond, WA 98052 USA.*
- *Shuicheng Yan is with Sea AI Lab. E-mail:yansc@sea.com.*

We provide in Fig. 1 an illustration of the distribution distortion of the target-domain features caused by applying UDA methods, *e.g.*, AdaptSeg [16], which is a typical adversarial training based UDA method. Fig. 1 (left) is the t-SNE [15] visualization of the latent representations. After applying AdaptSeg [16], most adapted target-domain features (red dots) are dragged away from target centers (yellow squares) and forced to align with the source centers (blue squares), as shown in region ②, or lose alignment as shown in region ③. We also show three exemplar images respectively from regions ①, ② and ③ with their corresponding segmentation by AdaptSeg in Fig. 1 (right). It can be seen that AdaptSeg only performs well in region ① where source and target samples share similar content, and generates error-prone segmentations in regions ② and ③. As demonstrated in Fig. 1, adversarial UDA methods tend to cause the generated target features to deviate from the real distribution (*i.e.*, to be distorted), thus losing some target-specific information and leading to bad segmentation.

To alleviate the above distribution distortion problem, a promising idea is to introduce the knowledge from the real target distribution, *i.e.*, selecting and annotating target samples and learning target-specific information from the annotated sample pairs. This can be done with active learning (AL) [20], which aims at getting better performance via a little annotation workload, with effectiveness well proved in the domain adaptation (DA) scenery for classification and detection tasks [21]. However, most previous AL methods [21] select samples only based on the target-domain distribution, which may not be optimal to training the network jointly with source-domain data as in common practice and lead to inferior performance.

Based on these observations, in our preliminary confer-

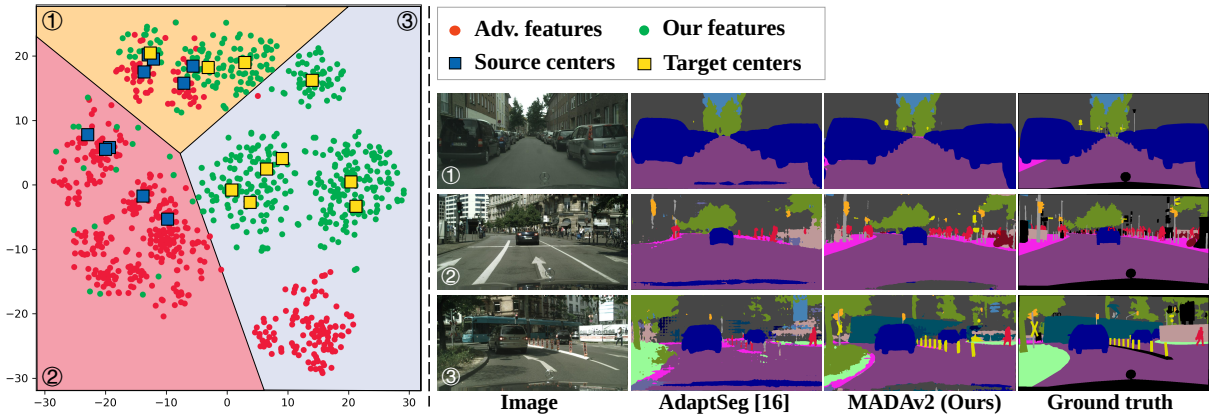

Fig. 1: Visualization (t-SNE [15]) of the target-domain distribution distortion problem in UDA. Left: Visualization of feature scatters by AdaptSeg [16] and MADAv2. The source cluster centers (blue squares) and the target cluster centers (yellow squares) can be very close, as shown in region ①, meaning samples here share similar content; the feature clusters may also show a source- and target-specific distribution, *i.e.*, containing only source or target centers, as shown in region ② and region ③, respectively. The features of AdaptSeg (red dots) are dragged away from the target centers, forced to align with the source centers (see region ②) or even lose alignment (see region ③). In contrast, our features are perfectly distributed around the target centers. Right: Exemplar samples and corresponding segmentation from AdaptSeg and our method MADAv2. Column 1 shows three samples respectively from regions ①, ② and ③; Columns 2 and 3 are segmentation results with AdaptSeg and our method, respectively. It can be seen that, AdaptSeg works well for the sample from regions ① as the target samples in region ① are similar to source samples, while for samples from region ② and ③, its segmentation is error-prone as the features do not follow the target centers. In contrast, our method can generate precise segmentation for all the samples. From these examples, we can see that the features of adversarial training tend to deviate from the real target distribution.

ence paper, we proposed a Multi-anchor Active Domain Adaptation (MADA) framework [22], which adopts the active learning strategy to assist DA regarding the semantic segmentation task, with a multi-anchor strategy to better characterize the source-domain and target-domain features. Specifically, the MADA framework consists of two stages. In the first stage, with the network pretrained in an adversarial UDA [16] manner, the most complementary samples are selected through the proposed multi-anchor strategy by exploiting the feature distribution across the source domain. Then in the second stage, the segmentation network is fine-tuned in a semi-supervised learning manner. The annotations of the source samples and the few selected target samples are jointly used for supervision, and all the available image information is additionally employed for optimization with a pseudo label loss and the proposed multi-anchor soft-alignment loss.

Though with noticeable improvement, our MADA method still shows an obvious gap with the fully-supervised upperbound, *i.e.*, 64.9% vs. 71.9%, a gap of 7.0% of mIoU on GTA5 [23] and 68.1% vs. 73.0%, 4.9% gap on SYNTHIA [24]. We arguably attribute such a gap to the sample selection method and the semi-supervised domain adaptation strategy. Hence in this work, we substantially extend our previously proposed MADA framework and propose an advanced multi-anchor based active domain adaptation segmentation framework (MADAv2). For sample selection, our previous MADA [22] adopts a single-domain selection method, which takes the distance between target samples and corresponding source anchors as the only metric, neglecting the possibility that selected samples may be *crowded* in the target domain and fail to provide sufficient knowledge of the whole target domain. In MADAv2, we argue that the selected samples should not only be complementary to the source domain, but also to most other target samples, such that more information of the target domain can be provided. Therefore, we extend the single-domain metric to a *Dual_Domain_Distance* active metric, *i.e.*, considering the distances from target samples to both source and target anchors comprehensively. Specifically, we select those target samples that are far away from the target anchors, which would bring better performance than the close ones (the result of *Dual_Domain_Distance* is 65.1% mIoU on GTA5, better than 62.6% of the close ones). The new sample selection metric leads to an improvement of 0.9% mIoU on GTA5, compared to the 64.2% result of MADA, as experimentally validated in Section 4.5. In addition, the previous compact semi-supervised phase in MADA cannot handle the long-tail distribution problem in the target domain [25], which is extremely severe due to the limited annotations. The tail classes are easily ignored with the cross-entropy loss because of their few pixels and the similarity between unlabeled target samples and corresponding pseudo labels. Therefore in MADAv2, we adopt the online hard example mining (OHEM) loss [26], which is designed to find hard examples in object detection tasks, to locate informative pixels and exploit the inconsistency between prediction and pseudo labels. Moreover, a combination of revised adaptive cutmix and copy-paste [25] is introduced as additional data augmentation to force the network to pay more attention to tail classes and alleviate the long-tail distribution problem.

We visualize the generated target features with our

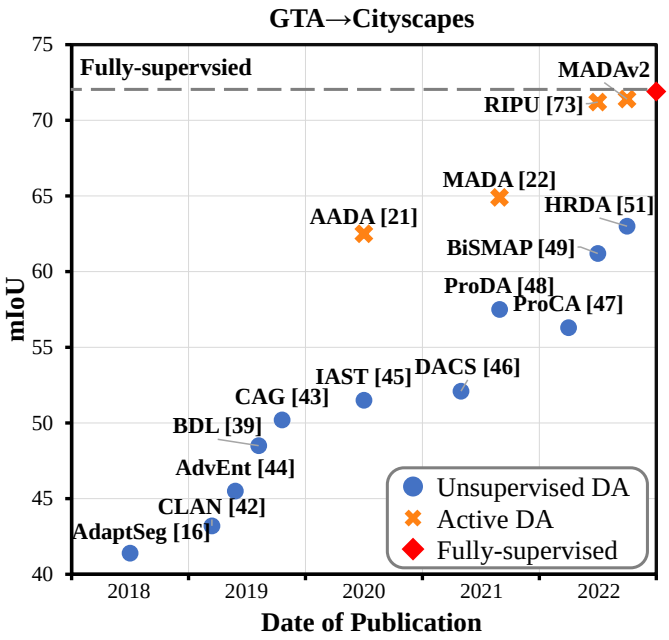

Fig. 2: The mIoU of state-of-the-art DA methods on Gta5 → Cityscapes adaptation. All the results of active DA methods are trained additionally with 5% annotation from the target domain. Our MADAv2 outperforms previous DA methods, and is very close to the fully-supervised upperbound.

MADAv2 using t-SNE in Fig. 1 as well as corresponding segmentation. It can be seen that MADAv2 effectively relieves the aforementioned distribution distortion problem and yields better segmentation. Also, as depicted in Fig. 2, it outperforms previous domain adaptation works. It achieves 71.4% and 71.8% mIoU on the segmentation of GTA5 and SYNTHIA datasets, respectively, improving upon the previous MADA method by 6.5% and 3.7% mIoU, and is close to the upperbound of 71.9% and 73.0%, on the GTA5 and SYNTHIA dataset respectively. In summary, this study makes the following contributions:

- To the best of our knowledge, our work is the first study to adopt active learning to assist domain adaptation regarding the semantic segmentation tasks. By annotating a few target-domain samples, the distortion of the target-domain feature distribution can be effectively alleviated and superior segmentation performance can be achieved.
- We propose a new *Dual_Domain_Distance* active metric, considering both the source and target domain distributions for selecting better active samples. The source and target anchors are obtained to characterize the multimodal distributions of both domains. Then we select samples far from source and target anchors simultaneously, which are the most complementary to the source domain and informative in the target domain.
- We propose to effectively tackle the long-tail distribution problem by introducing the OHEM loss along with the adaptive cutmix and copy-paste augmentation in the new MADAv2 method, and learn better latent representation through the multi-anchor soft-alignment loss.

2 Related Work

2.1 Unsupervised Domain Adaptation

Unsupervised domain adaptation (UDA) is aimed at addressing the domain shift problem in a wide variety of computer vision tasks including classification [27]–[30], detection [31], [32], and segmentation [16], [33].

Recent UDA methods can be roughly divided into two groups: maximum mean discrepancy (MMD) based [34]–[37] and adversarial learning based [16], [38]–[40]. The MMD kernel was first introduced in [34], which measures the discrepancy of features from different domains quantitatively. Subsequent studies proposed several improved MMD kernels for more accurate measurement of the domain discrepancy, including MK-MMD [34], JMMD [35], CMD [36] and CORAL [37]. These kernels minimize the discrepancy to force features from different domains to align with each other, thus addressing the domain shift problem. However, it is impractical to directly adopt the MMD-based methods in segmentation tasks, because these methods require complex computation in the high-dimension feature space.

In contrast, adversarial learning based methods are preferred for UDA of segmentation tasks, where the two domain distributions are drawn together via a domain discriminator. Among these methods, the classical appearance matching method CycleGAN [38] constructs two adversarial subnets to translate unpaired source and target images; BDL [39] leverages label consistency to improve the UDA performance; DISE [40] proposes a disentangled representation learning architecture [41] to preserve structural information during image translation. In addition, feature aligning methods such as CLAN [42] and CAG [43] utilize category-based distribution alignment to adapt the source and target domains in the feature and output spaces. Another work AdvEnt [44] designs a novel loss function to maximize the prediction certainty in the target domain to boost the UDA performance.

To address the instability of adversarial UDA methods, IAST [45] and DACS [46] apply a self-training (ST) framework to replace the adversarial training phase by learning from the refined pseudo labels. Based on the ST framework, ProCA [47], ProDA [48] and BiSMAP [49] introduce the representative prototypes to provide additional restrictions. Besides the training strategy, the Transformer [50] is also utilized to improve the UDA semantic segmentation performance. With its powerful representation learning ability, UDA methods [51], [52] based on a Transformer backbone show significantly better performance compared to those based on a ResNet backbone.

Despite the encouraging progress, UDA methods unconditionally force the distributions of the two domains to be similar, which may distort the underlying latent distribution of the target domain if it presents intrinsic difference from that of the source domain. In this work, we address such distortion with active learning (AL) [20], at the price of only a little annotation workload.

2.2 Active Learning and Domain Adaptation

AL aims at reaching the optimal performance at a low annotation cost by actively selecting a few samples that

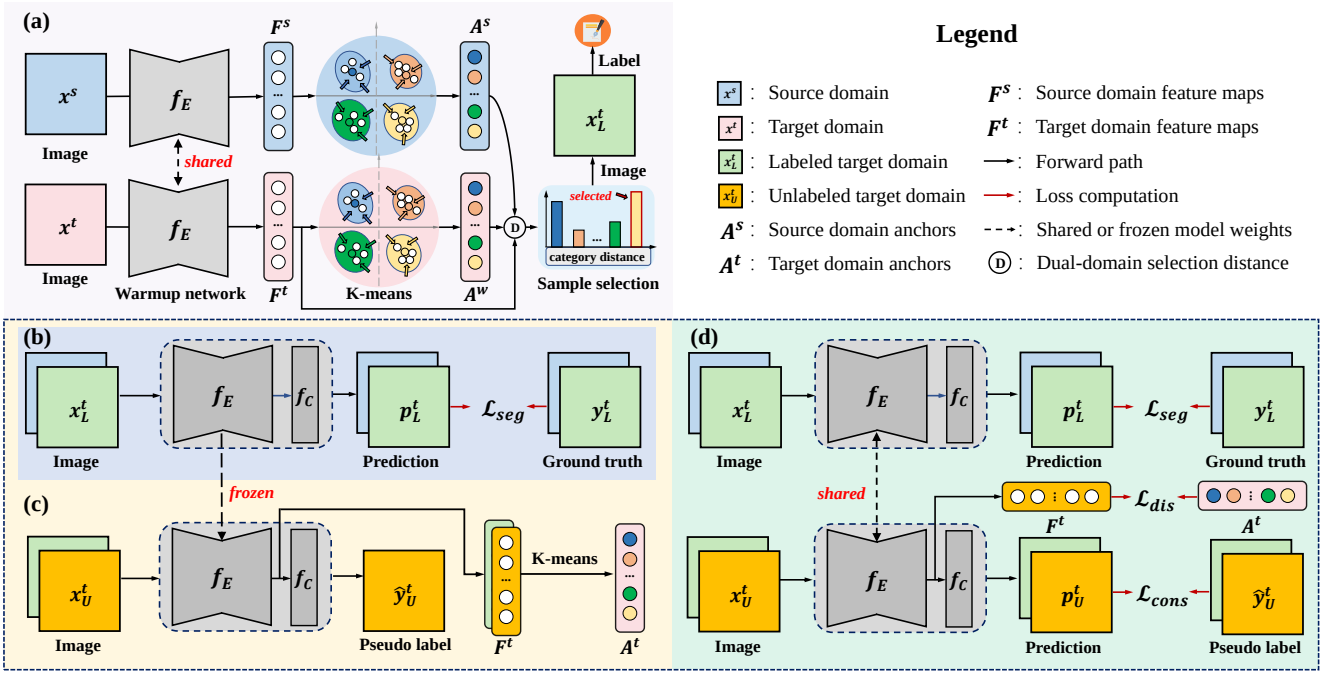

Fig. 3: Overview of the proposed MADAv2 framework.

are most helpful to performance improvement if labeled [53]. Over the past decade, several sample selection strategies have been proposed for AL, including uncertainty-based [54], [55], diversity-based [56], [57], representativeness-based [58]–[60], and expected model change based [61]–[63] strategies. These strategies have been successfully applied to various computer vision tasks, such as image classification [64], object detection [65]–[67], and image segmentation [68]. In this work, we argue that it is beneficial to introduce AL to the DA problem, to mitigate the distortion of the target-domain distribution. AL entails only a little annotation cost, which is acceptable in many scenarios considering the potential performance gain. Furthermore, with a proper sample selection strategy, AL can identify the samples most representative of the exclusive components in the target-domain distribution for annotation. Hence, how to select the AL samples becomes a critical issue.

As far as the authors are aware of, only a few studies have attempted to apply AL to DA problems. An early work by Chattopadhyay *et al.* [69] uses the MMD distance between the source and target domains for active sample selection during the DA process. However, it is practically prohibitive to apply MMD distances for segmentation DA problems, as mentioned earlier. More recently, Huang *et al.* [70] proposed to fine-tune pre-trained models for classification tasks and involved additional active sample selection in every iteration. In comparison, our framework takes a step forward to make dense predictions for segmentation tasks, and simplifies the active learning process to a one-time sample selection. Being closely related to our work, Active Adversarial Domain Adaptation (AADA) [21] proposed AL for DA with the adversarial learning [71] strategy, which selects representative samples by jointly considering diversity and uncertainty criteria.

In addition to the sample-wise active learning, pixel-wise [72] and region-wise [73] active DA methods have also been proposed for semantic segmentation tasks. However, the annotators have to go through every image for annotating a few image areas, which costs much more time than only labeling a few images. Especially in some applications like medical images, the annotators need to analyze the whole image to determine the label of a specific region.

In this work, we propose to capture more comprehensive information from not only the target domain, but also the source domain. By modeling both the source and target distributions as multimodal (in contrast to the implicit unimodal assumption in previous works such as AADA), our method can achieve substantial performance improvement, as experimentally validated in Section 4.5.

3 Proposed Approach

In this section we elaborate on the proposed MADAv2 framework. An illustration of its overall structure is given in Fig. 3. MADAv2 consists of two main stages: active target sample selection based on multiple anchors of both source and target domains (Fig. 3(a)) in Section 3.2, and semi-supervised self-training aiming at exploiting information from both labeled and unlabeled data (Figs. 3(b), 3(c) and 3(d)) in Section 3.3. Below we first formally define our problem setting and then explain the two stages in detail.

3.1 Problem Setting

The goal of semantic segmentation is to train a model $\mathbf{M}$ to map a sample x in the image space X to a prediction y in the label space Y, where $x \in \mathbb{R}^{H \times W \times 3}$ with H denoting the height, W the width, and 3 the color channels, and

$y \in \{0,1\}^{H\times W\times C}$ with C denoting the number of segmentation categories. For DA, there are N_s image-label pairs $X^s = \{(x^s, y^s)\}$ in the source domain, and N_t unlabeled images $X^t = \{x^t\}$ in the target domain. For AL, N_a active samples are selected in the target domain for annotation, where $N_a \ll N_t$, so that the target-domain data consist of N_a image-label pairs $X_L^t = \{(x_L^t, y_L^t)\}$ and $N_t - N_a$ unlabeled images $X_U^t = \{x_U^t\}$. The target of this work is to optimize the segmentation performance of $\mathbf{M}$ in the target domain while keeping N_a small.

3.2 Multi-anchor based Active Sample Selection

Multiple Anchoring Mechanism. We propose an efficient anchoring mechanism to model the domain distributions, and shrink the gap between network predictions and the anchors by forming compact clusters around the anchors.

Previous works CAG [43] and ProDA [48] average all image-level features of the source domain to obtain a centroid representing the entire domain, which implicitly assume a unimodal distribution. In practice, however, the distribution of a domain often comprises more than a single mode [74]. Although different images may contain the same categories of objects (*e.g.*, road, car, human, and vegetable), they can be classified into various scenes (*e.g.*, highway, uptown, and suburb).

We then adopt *multiple anchors* instead of a single centroid to characterize the domain distribution. By concatenating the features of different categories into an image-level 'connected' vector, we perform clustering on them to estimate scene-specific representative distributions with cluster centers, denoted as 'anchors'. We then select the most complementary and informative samples based on both the source and the target anchors.

As a warm-up model, we employ the common adversarial training [16] strategy to narrow the gap between the source and target domains. Then we freeze the feature encoder f_E and calculate the feature map $F_c^s(x^s)$ of a source sample x^s for a certain category c by:

$$F_c^s(x^s) = \frac{1}{|\Lambda_c^s|} y_c^s \otimes f_E\left(x^s\right)\big|_c, \tag{1}$$

where y_c^s denotes the label map for category c, $f_E\left(x^s\right)\big|_c$ is the networks' output for category c, $\otimes$ denotes element-wise multiplication for extracting category-exclusive information, and $|\Lambda_c^s|$ is the number of pixels belonging to the specific category. The final feature vector $F^s(x^s)$ of the source image x^s is obtained by first flattening the $F_c^s(x)$ of each category into a vector followed by concatenating the vectors of all categories into a long vector. Then, we apply the K-means method [75] to feature vectors of all source images to group them into K clusters, by minimizing the following error:

$$\sum_{k=1}^{K} \sum_{x\in\mathcal{C}_k} \|F^s(x^s) - A_k^s\|_2^2, \tag{2}$$

where $\|\cdot\|_2^2$ denotes the $L2$ distance and A_k^s is the centroid of the cluster $\mathcal{C}_k$:

$$A_k^s = \frac{1}{|\mathcal{C}_k|} \sum_{x\in\mathcal{C}_k} F^s(x^s), \tag{3}$$

where $|\mathcal{C}_k|$ denotes the number of images belonging to $\mathcal{C}_k$. The centroids $\{A_k^s\}$ are used as the source-domain anchors, against which the target images will be compared for active sample selection. Note that the cluster number K is not the same as the number of segmentation categories C, and the impact of different values of K is explored in Section 4.8.

Active Target Sample Selection. For single-domain AL, uncertainty-based metrics have been extensively used to select the samples which are the most difficult to segment [76]. For multi-domain AL, however, we argue that the more dissimilar the target samples are to the source-domain, the more complementary they are to the segmentation network. Hence, we measure the dissimilarity by the distance between the target samples and the source anchors to assess the importance of unlabeled target samples for domain adaptation in our previous work [22]. However, the selected samples may be *crowded* in the target domain, as shown in Fig. 4 (a). In addition, with a sufficient amount of data, the knowledge of the samples close to the target anchors can be learned through the semi-supervised learning strategy. To avoid information redundancy, we propose to additionally consider the distance of the target samples to the target anchors, so that the selected samples can provide as much information as possible.

Specifically, we first calculate the per category feature map of a target-domain image x^t:

$$F_c^t(x^t) = \frac{1}{|\Lambda_c^t|} \hat{y}_c^t \otimes f_E\left(x^t\right)\big|_c, \tag{4}$$

where $\hat{y}_c^t$ is the predicted map for category c, and $|\Lambda_c^t|$ is the number of pixels belonging to the specific category according to $\hat{y}_c^t$. Then, we concatenate $F_c^t(x^t)$ of all categories to obtain the image-level feature vector $F^t(x^t)$.

Similar to the acquirement of source-domain anchors $\{A_k^s\}$ with Eq. (3), we can obtain the target-domain anchors $\{A_k^w\}$ with the warm-up model and K-means clustering. Eventually, we calculate the $L2$ distances from $F^t(x^t)$ to its nearest source anchor and target anchor, and define the sum of the two distances as the *Dual_Domain_Distance*, with which the distances from sample x^t to the source and target domains are considered comprehensively:

$$D(x^t) = \min_k \left\|F^t(x^t) - A_k^s\right\|_2^2 + \min_k \left\|F^t(x^t) - A_k^w\right\|_2^2. \tag{5}$$

Here, the first item $\min_k \|F^t(x^t) - A_k^s\|_2^2$ aims to find the samples which are most complementary to the source domain, while the second item $\min_k \|F^t(x^t) - A_k^w\|_2^2$ favors the '*outliers*' in the target domain. In this way, more information can be introduced for domain adaptation. Then, we select the samples with the largest *Dual_Domain_Distance* as active samples and annotate them for subsequent training, hoping to learn unique characteristics of the target-domain distribution from these active annotations. As shown in Fig. 4 (b), the samples selected in MADAv2 follow a better scatter in the target distribution.

3.3 Semi-supervised Domain Adaptation

Step-1: Injecting Target-domain Specific Knowledge. The actively selected and annotated target-domain samples are added to the training process to learn information exclusive

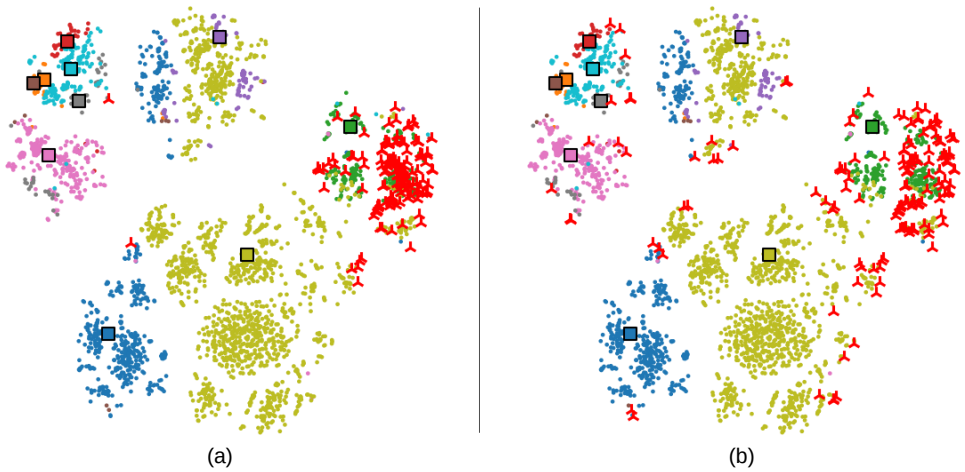

Fig. 4: Visualization (t-SNE [15]) of different sample selection methods. In the scatters, dots of different colors denote the target samples of different clusters, squares represent corresponding cluster centroids, and *red triangles* denote the selected AL samples. (a) The scatter of MADA [22], where we can find the active samples are crowded in the target domain. (b) The scatter of MADAv2, which considers the distance to both the source and target centers comprehensively.

to the target domain (Fig. 3(b)). Training data in this step consist of two parts: the labeled source samples X^s and the active target samples X_L^t, and the model f_E is fine-tuned with typical cross-entropy based segmentation losses:

$$\mathcal{L}_{seg} = \mathcal{L}_{ce}\left(x^s, y^s\right) + \mathcal{L}_{ce}\left(x_L^t, y_L^t\right), \tag{6}$$

where the cross-entropy loss $\mathcal{L}_{ce}$ is defined as:

$$\mathcal{L}_{ce} = -\frac{1}{HW}\sum_{i=1}^{H\times W}\sum_{c=1}^{C} y_{i,c}\log\left(p_{i,c}\right), \tag{7}$$

where y_i denotes the label for pixel i and p_i is the probability predicted by the classifier f_C. As experimentally validated in Section 4.5, our multi-anchor based active sample selecting strategy is superior to previous strategies, and the model gets a steady improvement in performance with the actively selected samples.

Step-2: Constructing Target-domain Anchors and Pseudo Labels. To fully utilize the unlabeled target data X_U^t, we use the fine-tuned f_E in step-1 to compute pseudo labels $\{\hat{y}_U^t\}$ for unlabeled target-domain samples as well as target-domain anchors $\{A_v^t\}_{v=1}^V$ (Fig. 3(c)), where V represents the number of target-domain anchors. Notably, as the target-domain anchors are a potentially biased estimation of the actual target-domain distribution, we naturally consider correcting them dynamically. As indicated by Xie *et al.* [77], re-clustering at each epoch may lead to the collapse of the training process due to jumps in cluster centroids between epochs. Therefore, we treat the target-domain anchors as a memory bank, and employ the exponential moving average (EMA) [78] to progressively update each anchor in a smooth manner:

$$A_v^t = \alpha A_v^t + (1-\alpha)F^t(x^t), \tag{8}$$

where α is set to 0.999 following [78], and $F^t(x^t)$ is utilized to update the closest anchor. With both $\{\hat{y}_U^t\}$ and $\{A_v^t\}$ computed, we proceed to the next step for semi-supervised domain adaptation.

Algorithm 1: Advanced Multi-Anchor Based Active Domain Adaptation Segmentation (MADAv2)

Notation: Source-domain set $\{(x^s, y^s)\}$, selected active sample set $\{(x_L^t, y_L^t)\}$, unlabeled target-domain set $\{x_U^t\}$, encoder f_E, feature vector set of the source domain $\{F^s(x^s)\}$ and feature vector set of the target domain $\{F^t(x^t)\}$, and number of iterations N.

Stage 1:

Warm-up f_E with adversarial training [16] to obtain $\{F^s(x^s)\}$ and $\{F^t(x^t)\}$.
Apply K-means on $\{F^s(x^s)\}$ and $\{F^t(x^t)\}$ to group the source-domain features and warm-up target-domain features into K clusters;
Compute the centroids $\{A_k^s\}$ and $\{A_k^w\}$ of the clusters (Eq. (3)) to serve as the anchors of the source and target domains, respectively;
Calculate the distance from each target-domain sample to both $\{A_k^s\}$ and $\{A_k^w\}$ (Eq. (5));
Select 5% target-domain samples with the smallest distances as active samples for annotation, resulting in set $\{(x_L^t, y_L^t)\}$.

Stage 2:

Fine-tune f_E with both $\{(x^s, y^s)\}$ and $\{(x_L^t, y_L^t)\}$ by minimizing $\mathcal{L}_{seg}$ (Eq. (6)), and re-obtain $\{F^t(x^t)\}$ with fine-tuned f_E;
Initialize $\{A_v^t\}$ with K-means clustering on $\{F^t(x^t)\}$;
for $i = 1, ..., N$ **do**
  Calculate $\mathcal{L}_{seg}$ (Eq. (6)) with $\{(x^s, y^s)\}$ and $\{(x_L^t, y_L^t)\}$;
  Calculate $\mathcal{L}_{cons}$ (Eq. (13)) and $\mathcal{L}_{dis}^t$ (Eq. (14)) with $\{x^t\}$;
  Update f_E by gradient descending $\nabla(\mathcal{L}_{seg} + \mathcal{L}_{cons} + \mathcal{L}_{dis}^t)$ (Eq. (15));
  Update A_v^t with EMA (Eq. (8));
end for

Step-3: Semi-supervised Adaptation. Lastly, we combine the source data X^s, labeled target samples X_L^t, and unlabeled target samples X_U^t for a semi-supervised training (*i.e.*, a further fine-tuning of f_E) for domain adaptation (Fig. 3(d)).

In MADA [22], we adopt a compact and efficient semi-

supervised phase, but it cannot fully exploit the knowledge from unlabeled samples. Due to the limited number of selected samples, the long-tail problem is much severer than the case in supervised semantic segmentation because of the small number of annotated pixels from tail classes. Inspired by ST based semi-supervised methods [25], [79], we build a more powerful semi-supervised domain adaptation framework to alleviate the problem by focusing more on the difficult regions and pixels.

First, we introduce additional data augmentations, adaptive cutmix (ACM) and adaptive copy-paste (ACP) [25], to achieve region- and pixel-level augmentations. To be specific, given an unlabeled target sample X_U^t and the corresponding pseudo label $\hat{y}_U^t$, the ACM selects less confidence pair $\left\{x_U^{t'}, \hat{y}_U^{t'}\right\}$ to replace some random regions of $\{x_U^t, \hat{y}_U^t\}$. It can be formulated as:

$$\left\{\widehat{x_U^t}, \widehat{\hat{y}_U^t}\right\} = \text{Cutmix}\left(\text{Crop}\left(\left\{x_U^{t'}, \hat{y}_U^{t'}\right\}\right), \{x_U^t, \hat{y}_U^t\}\right), \quad (9)$$

where $\{x_U^t, \hat{y}_U^t\}$ denotes a randomly chosen unlabeled target image and its corresponding pseudo label, $\left\{x_U^{t'}, \hat{y}_U^{t'}\right\}$ is elaborately chosen to improve the distribution of classes based on the confidence, and $\left\{\widehat{x_U^t}, \widehat{\hat{y}_U^t}\right\}$ is generated by cropping a random region from $\left\{x_U^{t'}, \hat{y}_U^{t'}\right\}$ and covering the corresponding area in $\{x_U^t, \hat{y}_U^t\}$. Specifically, we first calculate the confidence of unlabeled samples as:

$$\text{Conf}_c = \frac{1}{|\Lambda_c^t|} \sum_{i=1}^{H \times W} \hat{y}_{i,c}^t \log\left(p_{i,c}\right), \quad c \in \{1, \ldots, C\}, \quad (10)$$

where C is the number of categories, $|\Lambda_c^t|$ denotes the number of pixels belonging to category c according to its pseudo label $\hat{y}_i^t$, and $log\left(p_{i,c}\right)$ denotes the c-th channel prediction of the i-th pixel.

With the obtained confidence, the probability of sample $x_U^{t'}$ to be chosen to provide crops for x_U^t is defined as:

$$\boldsymbol{r} = \sum_{c=1}^{C} \text{Softmax}(1 - \text{Conf}_c). \quad (11)$$

For labeled active samples X_L^t, we utilize the adaptive copy-paste method to copy the pixels of non-long-tail classes to long-tail classes (*e.g.*, copy cars to the road, and persons to the sidewalk), which can be formulated as:

$$\{\widehat{x_L^t}, \widehat{y_L^t}\} = \text{Paste}\left(\text{Copy}\left(\{x_L^{t'}, y_L^{t'}\}\right), \{x_L^t, y_L^t\}\right), \quad (12)$$

where $\{\widetilde{x_L^{t'}}, \widetilde{y_L^{t'}}\}$ and $\{x_L^t, y_L^t\}$ denote active image-label pairs, $\{\widehat{x_L^t}, \widehat{y_L^t}\}$ is the augmented image. The procedure is similar to ACM, except that precise pixels instead of random regions are selected to replace the ones in the corresponding pairs. Based on the augmented training pairs, the consistency loss can be defined as:

$$\mathcal{L}_{cons} = \mathcal{L}_{ohem}(\widehat{x_L^t}, \widehat{y_L^t}) + \mathcal{L}_{ohem}(\widehat{x_U^t}, \widehat{\hat{y}_U^t}). \quad (13)$$

The online hard example mining (OHEM) loss is a masked cross-entropy loss, where only the pixels with low probabilities are considered. For more details regarding the pixel selection, please refer to the original study [26].

Beyond the self-training period, we propose a novel soft alignment loss to explicitly shrink the gap between the sample features and the anchors in the target domain:

$$\mathcal{L}_{dis}^t = V \Big/ \sum_{v=1}^{V} \frac{1}{\|F^t(x^t) - A_v^t\|_2^2}. \quad (14)$$

Intuitively, by minimizing the soft alignment loss, features of the target-domain samples output by the model are drawn towards the target-domain anchors, encouraging a more faithful learning of the underlying target-domain distribution represented by these anchors.

Thus, the overall loss function for the semi-supervised learning can be formulated as:

$$\mathcal{L}_{semi} = \mathcal{L}_{seg} + \mathcal{L}_{cons} + \mathcal{L}_{dis}^t. \quad (15)$$

The entire training pipeline is summarized in Algorithm 1.

4 Experiments

4.1 Datasets

To demonstrate the superiority of our proposed method, two challenging *synthetic-to-real* adaptation tasks, *i.e.*, GTA5 [23] $\rightarrow$ Cityscapes [2] and SYNTHIA [24] $\rightarrow$ Cityscapes are applied for evaluation. To be specific:

- GTA5 $\rightarrow$ Cityscapes: The GTA5 dataset consists of 24,966 synthetic images with 19-class segmentation, which is consistent with the Cityscapses dataset.
- SYNTHIA $\rightarrow$ Cityscapes: Following the previous study [39], the SYNTHIA-RAND-CITYSCAPES set with 9,400 synthetic images containing 16-class segmentation is utilized for training.

In both settings, Cityscapes serves as the target domain, with 2,975 images for training and 500 images for evaluation. The segmentation performance is measured with the mean-Intersection-over-Union (mIoU) [80] metric.

4.2 Implementation Details

We employ backbone DeepLab v3+ [81] as the feature extractor f_E, which is composed of the ResNet-101 [82] pretrained on ImageNet [83] and the Atrous Spatial Pyramid Pooling (ASPP) module. The classifier f_C is a typical convolutional layer with C channels and 1×1 kernel size to transform the latent representation to semantic segmentation. During the warm-up, the discriminator f_D consists of 5 convolutional layers of kernel size 3×3 and stride 2 with numbers of filters set to $\{64, 128, 256, 512, 1\}$. The first three convolutional layers are followed with a Rectified Linear Unit (ReLU) layer, while the fourth one is followed by a leaky ReLU [84] parameterized by 0.2. The proposed method is implemented on PyTorch with an NVIDIA Tesla V100 GPU. The input images are randomly resized with a ratio in $[0.5, 1.5]$ and then randomly cropped to 896×512 pixels.

For warm-up, we train the model for 20 epochs in an adversarial manner with a cross entropy loss and an adversarial loss weighted by 0.01. For the semi-supervised domain adaptation stage, we use the SGD optimizer to train our model for 200 epochs. The learning rate is initially set to 2.5×10^{-4} and decayed by the poly learning rate policy with a power of 0.9.

TABLE 1: Comparison with other DA methods on the GTA5 to Cityscapes adaptation task. Best results are shown in **bold**.

GTA5 → Cityscapes																				
Method	road	sidewalk	building	wall	fence	pole	light	sign	veg	terrain	sky	person	rider	car	truck	bus	train	mbike	bicycle	mIoU
AdaptSeg [16]	86.5	25.9	79.8	22.1	20.0	23.6	33.1	21.8	81.8	25.9	75.9	57.3	26.2	76.3	29.8	32.1	7.2	29.5	32.5	41.4
CLAN [42]	87.0	27.1	79.6	27.3	23.3	28.3	35.5	24.2	83.6	27.4	74.2	58.6	28.0	76.2	33.1	36.7	6.7	31.9	31.4	43.2
AdvEnt [44]	89.4	33.1	81.0	26.6	26.8	27.2	33.5	24.7	83.9	36.7	78.8	58.7	30.5	84.8	38.5	44.5	1.7	31.6	32.4	45.5
BDL [39]	91.0	44.7	84.2	34.6	27.6	30.2	36.0	36.0	85.0	43.6	83.0	58.6	31.6	83.3	35.3	49.7	3.3	28.8	35.6	48.5
CAG [43]	90.4	51.6	83.8	34.2	27.8	38.4	25.3	48.4	85.4	38.2	78.1	58.6	34.6	84.7	21.9	42.7	41.1	29.3	37.2	50.2
IAST [45]	93.8	57.8	85.1	39.5	26.7	26.2	43.1	34.7	84.9	32.9	88.0	62.6	29.0	87.3	39.2	49.6	23.2	34.7	39.6	51.5
DACS [46]	89.9	39.7	87.9	30.7	39.5	38.5	46.4	52.8	88.0	44.0	88.8	67.2	35.8	84.5	45.7	50.2	0.0	27.3	34.0	52.1
ProCA [47]	91.9	48.4	87.3	41.5	31.8	41.9	47.9	36.7	86.5	42.3	84.7	68.4	43.1	88.1	39.6	48.8	40.6	43.6	56.9	56.3
ProDA [48]	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
BiSMAP [49]	89.2	54.9	84.4	44.1	39.3	41.6	53.9	53.5	88.4	45.1	82.3	69.4	41.8	90.4	56.4	68.8	51.2	47.8	60.4	61.2
HRDA [51]	96.2	73.1	89.7	43.2	39.9	47.5	60.0	60.0	89.9	47.1	90.2	75.9	49.0	91.8	61.9	59.3	10.2	47.0	65.3	63.0
AADA (5%) [21]	94.1	66.7	87.7	43.0	49.9	49.4	54.8	59.8	89.3	47.8	89.7	72.5	43.0	90.7	51.8	48.4	41.8	41.5	66.3	62.5
MADA (5%) [22]	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%) [73]	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
MADAv2 (5%)	**97.4**	**79.8**	90.3	46.6	52.5	**54.3**	**62.3**	**71.7**	**91.2**	50.2	**94.0**	**77.8**	**56.8**	**93.2**	**74.7**	**80.3**	52.4	**57.5**	**73.6**	**71.4**
Fully-supervised	97.7	82.0	90.7	55.2	56.6	53.1	58.6	67.5	91.3	60.4	93.8	75.7	52.7	93.1	76.1	78.5	58.4	55.4	69.5	71.9

TABLE 2: Comparison with other DA methods on the SYNTHIA to Cityscapes adaptation task. Best results are shown in **bold**. The 'mIoU' and 'mIoU*' denote the average mIoU with 16 or 13 classes respectively.

SYNTHIA → Cityscapes																		
Method	road	sidewalk	building	wall	fence	pole	light	sign	veg	sky	person	rider	car	bus	mbike	bicycle	mIoU	mIoU*
AdaptSeg [16]	79.2	37.2	78.8	-	-	-	9.9	10.5	78.2	80.5	53.5	19.6	67.0	29.5	21.6	31.3	-	45.9
CLAN [42]	81.3	37.0	80.1	-	-	-	16.1	13.7	78.2	81.5	53.4	21.2	73.0	32.9	22.6	30.7	-	47.8
AdvEnt [44]	85.6	42.2	79.7	8.7	0.4	25.9	5.4	8.1	80.4	84.1	57.9	23.8	73.3	36.4	14.2	33.0	41.2	48.0
BDL [39]	86.0	46.7	80.3	-	-	-	14.1	11.6	79.2	81.3	54.1	27.9	73.7	42.2	25.7	45.3	-	51.4
CAG [43]	84.7	40.8	81.7	7.8	0.0	35.1	13.3	22.7	84.5	77.6	64.2	27.8	80.9	19.7	22.7	48.3	44.5	50.9
IAST [45]	81.9	41.5	83.3	17.7	4.6	32.3	30.9	28.8	83.4	85.0	65.5	30.8	86.5	38.2	33.1	52.7	49.8	57.0
DACS [46]	80.6	25.1	81.9	21.5	2.9	37.2	22.7	24.0	83.7	90.8	67.6	38.3	82.9	38.9	28.5	47.6	48.3	54.8
ProCA [47]	90.5	52.1	84.6	29.2	3.3	40.3	37.4	27.3	86.4	85.9	69.8	28.7	88.7	53.7	14.8	54.8	53.0	59.6
ProDA [48]	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
BiSMAP [49]	81.9	39.8	84.2	-	-	-	41.7	46.1	83.4	88.7	69.2	39.3	80.7	51.0	51.2	58.8	-	62.8
HRDA [51]	85.8	47.3	87.3	27.3	1.4	50.5	57.8	61.0	87.4	89.1	**76.2**	48.5	87.3	49.3	55.0	68.2	61.2	69.2
AADA (5%) [21]	93.9	66.3	87.5	36.9	41.1	47.5	53.1	59.8	88.6	92.9	71.3	42.4	90.0	52.8	34.8	66.6	64.1	69.2
MADA (5%) [22]	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%) [73]	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
MADAv2 (5%)	**97.3**	**80.4**	**89.9**	**49.2**	42.6	**52.9**	**60.7**	**70.4**	90.8	**94.1**	75.2	51.6	92.3	73.3	**55.4**	**73.3**	**71.8**	**77.2**
Fully-supervised	97.7	82.3	90.8	53.5	57.3	52.8	58.9	67.4	91.4	93.4	75.7	53.4	92.6	75.9	55.1	70.6	73.0	77.3

Except for the comparison study in Section 4.7, we select 5% target-domain samples as active samples for all experiments, which cost a little annotation workload but bring large performance gains.

4.3 Main Results

As presented in Table 1 and Table 2, the proposed framework is compared with a series of unsupervised (UDA) and active DA methods. For the UDA task, traditional adversarial-based [16], [39], [42]–[44], prototype-ST-based [47]–[49] and the ResNet-101 version of HRDA [51] (which is the current SOTA) UDA methods are included for comparison (Transformer based methods are not listed due to the different feature extraction ability and segmentation performance upperbound). For the active DA task, we compare with the sample-based [21] and region-based [73] approaches using the same amount of annotation. The results of MADA [22] are also listed. As expected, we observe substantial improvements of our proposed method MADAv2 over the compared UDA methods, which indicates that with elaborately selected active samples, a little manual annotation workload can lead to large performance gains.

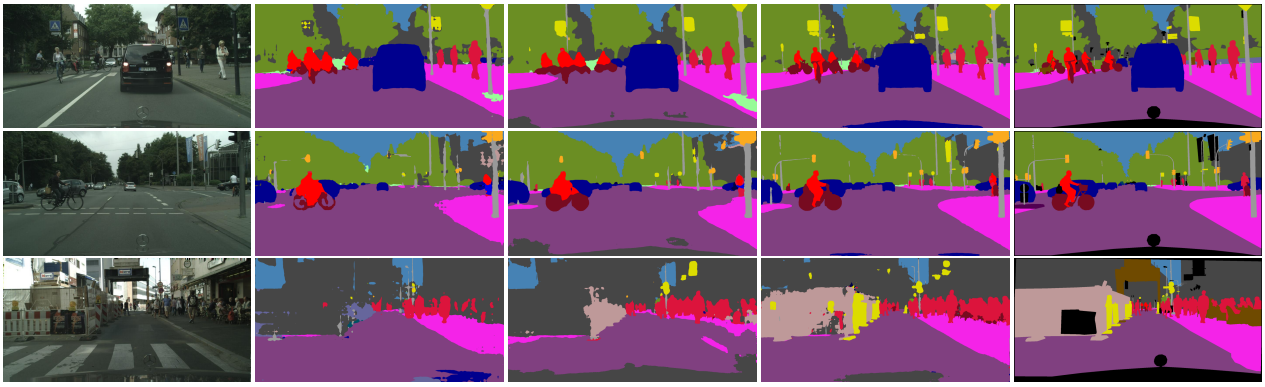

Fig. 5: Qualitative results of DA segmentation for GTA5 → Cityscapes. For each image, we show the results of a typical adversarial method (i.e., AdaptSeg [16]), typical unimodal prototype method (i.e., ProDA [48]) and MADAv2, respectively. The black region in 'Ground truth' is excluded from evaluation because it does not belong to any of the 19 classes.

In addition, MADAv2 outperforms another sample-based active DA method, *i.e.*, AADA [21], by a large margin (8.9% mIOU), demonstrating the effectiveness of the proposed multi-anchor strategy. And the improvement compared to MADA demonstrates that the new sample selection metric and ST based semi-supervised domain adaptation can effectively address the weaknesses of the previous method. In addition, our method consistently shows better performance than RIPU [73], which demonstrates that selecting a few images to annotate is better than labeling a few regions of each sample, despite the more time latter one costs.

The visualization of three example images is displayed in Fig. 5 for qualitative comparison. We can observe that by alleviating the distortion of target features, fewer segmentation errors as well as more precise boundaries can be obtained with the proposed MADAv2 method.

TABLE 3: Ablation study. G → C denotes the GTA5 → Cityscapes and S → C denotes the SYNTHIA → Cityscapes adaptation.

					G → C	S → C
Method	A	B	C	D	mIoU	mIoU
$\mathbf{M}^{(0)}$					42.5	42.9
$\mathbf{M}^{(1)}$	✓				65.1	65.2
$\mathbf{M}^{(2)}$	✓	✓			69.5	70.3
$\mathbf{M}^{(3)}$	✓	✓	✓		70.8	71.1
$\mathbf{M}^{(4)}$	✓	✓	✓	✓	71.4	71.8
$\mathbf{M}^{(u)}$					71.9	73.0

A: Training with active samples
B: Prediction consistency loss
C: ACM and ACP
D: Soft-anchor alignment loss

4.4 Ablation Study

To verify the effectiveness of each component, we perform ablation study with the following variants: $\mathbf{M}^{(0)}$: the baseline adversarial learning method [16] without any active annotation; $\mathbf{M}^{(1)}$: extending $\mathbf{M}^{(0)}$ by additionally introducing the active samples with cross entropy loss for training; $\mathbf{M}^{(2)}$: extending $\mathbf{M}^{(1)}$ by adding the prediction consistency loss and introducing the ST strategy; $\mathbf{M}^{(3)}$: extending $\mathbf{M}^{(2)}$ by augmenting target images with ACM and ACP; $\mathbf{M}^{(4)}$: adding the proposed multi-anchor soft alignment loss on target samples in addition to $\mathbf{M}^{(3)}$; $\mathbf{M}^{(u)}$: performing fully-supervised segmentation with the annotation of both the source and target datasets as the upper bound. As shown in Table 3, the consistent and notable improvements from $\mathbf{M}^{(0)}$ to $\mathbf{M}^{(4)}$ on two adaptation tasks demonstrate the effectiveness of each strategy. Furthermore, MADAv2 with only 5% of the target-domain samples actively annotated achieves comparable performance to that of the upper bound. This demonstrates that the proposed framework can select complementary samples to effectively shrink the performance gap between UDA and full supervision.

The visualization of the feature distribution with/without active learning is presented in Fig. 1. With the proposed MADAv2 framework, the target-specific information can be maintained as its original multimodal distribution.

4.5 Comparison of Sample Selection Methods

The performance of active learning depends heavily on the sample selection methods. In Table 4, we compare the proposed anchor-based method with the following popular sample selection approaches on the GTA5 to Cityscapes adaptation task.

Random Selection. Samples are randomly selected with equal probability from the target domain.

Entropy-based Uncertainty Method. The AdvEnt [44] is applied to obtain the prediction map entropy of each sample in the target domain and the ones with top 5% highest entropy are chosen for manual annotation:

$$E_{ent} = \frac{-1}{\log(C)} \sum_{c=1}^{C} \sum_{i=1}^{H \times W} p_{i,c}^{t} \log(p_{i,c}^{t}). \tag{16}$$

Adversarial-based Diversity Method. With the discriminator f_D trained in the warm-up stage as [16], we select the

TABLE 4: Experiments with different active sample selection methods. Best results are shown in **bold**.

GTA5 → Cityscapses																				
Method	road	sidewalk	building	wall	fence	pole	light	sign	veg	terrain	sky	person	rider	car	truck	bus	train	mbike	bicycle	mIoU
Random	94.8	71.0	86.6	39.0	43.5	47.1	55.2	59.1	88.7	43.5	87.1	72.2	35.8	90.1	51.1	49.6	29.8	50.1	66.2	61.1
Adversarial [16]	93.2	65.3	86.9	38.5	47.8	48.3	55.0	56.2	89.3	46.5	90.5	71.9	42.5	90.6	50.5	47.4	41.7	41.0	66.1	61.5
Entropy [44]	94.7	68.2	87.8	44.4	47.5	46.7	49.9	57.5	89.4	48.5	90.9	72.8	43.4	90.7	54.6	49.4	30.3	51.6	66.4	62.4
AADA [21]	94.1	66.7	87.7	43.0	49.9	49.4	54.8	59.8	89.3	47.8	89.7	72.5	43.0	90.7	51.8	48.4	41.8	41.5	66.3	62.5
Prototype [48]	95.6	71.5	87.9	36.6	47.6	49.4	53.5	60.9	88.5	48.9	91.0	74.1	48.4	91.3	51.7	57.4	22.5	52.4	68.9	63.1
MADA [22]	93.5	66.7	86.8	31.5	46.3	50.4	57.5	62.5	88.8	43.4	91.2	75.6	51.8	91.2	59.0	58.4	41.5	54.5	70.0	64.2
MADAv2	**97.4**	**79.8**	**90.3**	**46.6**	**52.5**	**54.3**	**62.3**	**71.7**	**91.2**	**50.2**	**94.0**	**77.8**	**56.8**	**93.2**	**74.7**	**80.3**	**52.4**	**57.5**	**73.6**	**65.1**

samples with least predicted probabilities, *i.e.*, the ones that are the most distinguishable from the source domain:

$$E_{adv} = \frac{1 - f_D(f_E(x^t)}{f_D(f_E(x^t))}. \tag{17}$$

AADA Method. In addition to the discriminator-based diversity, the AADA [21] method also takes the certainty of prediction into consideration:

$$E_{AADA} = E_{ent} E_{adv}. \tag{18}$$

Prototype Method. To verify the effectiveness of multi-modal distributions, we also compare with the unimodal prototype method, *i.e.*, ProDA [48]. If we define η as the single centroid, the samples with a large distance to the centorid are selected:

$$E_{Proto} = \left\| F^t(x^t) - \eta \right\|_2^2. \tag{19}$$

Note that for a fair comparison, all the comparison experiments are subject to the same experimental setup. The same percentage of active samples, 5%, are selected, and no unlabeled samples are used for optimization. From the results, we observe that the proposed *Dual_Domain_Distance* strategy delivers the best segmentation performance, demonstrating the superiority of the proposed strategy and the effectiveness of considering the distributions of both domains for sample selection.

4.6 Impact of Source Information

As demonstrated in Table 3, the performance is improved steadily with ST-based semi-supervised learning. One may wonder how much the source information contributes to the performance, or whether semi-supervised learning with only active samples can achieve similar results.

To investigate this question, we conduct an additional experiment by modifying our MADAv2 from two aspects: 1) only considering the distance from target samples to their nearest target anchor as the selection metric; 2) only training on target labeled or unlabeled samples. The final performance is significantly lower than MADAv2 (67.3% compared to 71.4% in mIoU), indicating that the source samples can provide effective information when the amount of target annotations is limited.

4.7 Impact of the Number of Active Samples

In order to verify the stability of our sample selection metric, comparative experiments regarding different percentages of active samples are conducted. To avoid the effect of unlabeled samples and semi-supervised learning method, only the source samples and labeled target samples are adopted to train the network in a fully supervised manner for this experiment. As shown in Table 5, as the percentage of samples increases from 1% to 20%, the mIoU increases steadily from 57.4% to 67.3%. We also gain the upper bound by optimizing with all target labels, and find a narrow gap of 6.8% in mIoU between using only 5% of target-domain data for AL and the upper bound, demonstrating that the proposed method can effectively exploit the information from active samples. We also find a trend that the performance rises sharply by introducing new AL samples when the quantity of samples is small, but improves slowly with further more samples introduced. This trend demonstrates that we can claim most of the benefit by selecting a few informative samples at the beginning, and 5% of samples is a trade-off between annotation cost and segmentation performance.

TABLE 5: Experiments with different numbers of active samples.

GTA5 → Cityscapse						
Percentage	1%	2%	5%	10%	20%	100%
mIoU	57.4	59.7	65.1	66.5	67.3	71.9
mIoU Gap	−14.5	−12.2	−6.8	−5.4	−4.6	-

4.8 Impact of the Number of Anchors

We evaluate the impact of different anchor numbers on modeling the source and target domains with the GTA5 to Cityscapes adaptation task, where the number of anchors varies from 1 to 100 in one domain while fixing the anchor number in the other domain to 10.. As shown in Fig. 6, for both domains, using multiple anchors performs consistently better than using a single centroid, and using 5-10 anchors stably yields superior performance. This might be because there are only limited types of scenarios in these datasets, and a few anchors are sufficient to representing their distributions. We therefore use 10 clusters considering the top performance in both domains.

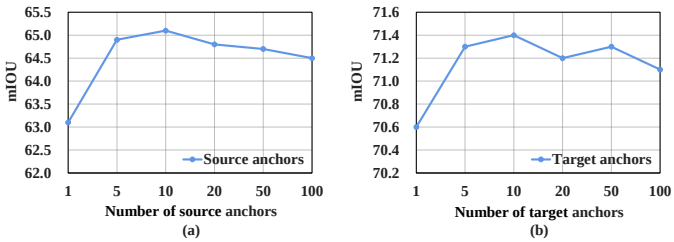

Fig. 6: Experiments on different numbers of anchors for the source domain (a) and the target domain (b).

5 Conclusion

In this paper, we proposed an advanced multi-anchor based active domain adaptation segmentation framework, namely MADAv2, to reduce the distortion of source-to-target domain adaptation for segmentation tasks at minimal annotation cost. MADAv2 performs anchor-based active sample selection in DA, which selects only a limited number of target-domain samples that are however most complementary to the source-domain distribution and meanwhile unique to the target-domain distribution. Adding active annotations of these selected target-domain samples to training can effectively alleviate the distortion of the target-domain distribution that could otherwise happen to typical UDA methods. Different from previous works which assume unimodal distributions for both the source and target domains, MADAv2 uses multiple anchors to realize multimodal distributions for both domains. On top of that, MADAv2 further introduces ACM, ACP and OHEM loss to address the long-tail issue. With the multi-anchor soft-alignment loss to explicitly push the target-domain features towards the anchors, the unlabeled target-domain samples can be fully utilized. Experimental results on two public benchmark datasets have demonstrated the effectiveness of 1) introducing AL into DA, 2) multiple anchors versus a single centroid, 3) introduction of ACM, ACP and OHEM, 4) adding the soft-alignment loss, as well as the superior performance of MADAv2 to existing state-of-the-art UDA and active DA methods.

References

[1] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3354–3361.

[2] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3213–3223.

[3] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, "Depth-induced multi-scale recurrent attention network for saliency detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7254–7263.

[4] W. Ji, J. Li, S. Yu, M. Zhang, Y. Piao, S. Yao, Q. Bi, K. Ma, Y. Zheng, H. Lu, and L. Cheng, "Calibrated RGB-D salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2021, pp. 9471–9481.

[5] M. Zhang, J. Li, W. Ji, Y. Piao, and H. Lu, "Memory-oriented decoder for light field salient object detection," in *Advances in Neural Information Processing Systems*, 2019, pp. 896–906.

[6] J. Li, W. Ji, Q. Bi, C. Yan, M. Zhang, Y. Piao, H. Lu *et al.*, "Joint semantic mining for weakly supervised RGB-D salient object detection," *Advances in Neural Information Processing Systems*, vol. 34, pp. 11 945–11 959, 2021.

[7] W. Ji, J. Li, Q. Bi, C. Guo, J. Liu, and L. Cheng, "Promoting saliency from depth: Deep unsupervised RGB-D saliency detection," in *International Conference on Learning Representations*, 2022.

[8] M. Oberweger, P. Wohlhart, and V. Lepetit, "Hands deep in deep learning for hand pose estimation," *arXiv preprint arXiv:1502.06807*, 2015.

[9] W. Ji, S. Yu, J. Wu, K. Ma, C. Bian, Q. Bi, J. Li, H. Liu, L. Cheng, and Y. Zheng, "Learning calibrated medical image segmentation via multi-rater agreement modeling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 341–12 351.

[10] M. Ning, C. Bian, D. Lu, H.-Y. Zhou, S. Yu, C. Yuan, Y. Guo, Y. Wang, K. Ma, and Y. Zheng, "A macro-micro weakly-supervised framework for AS-OCT tissue segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 725–734.

[11] M. Ning, C. Bian, C. Yuan, K. Ma, and Y. Zheng, "Ensembled ResUnet for anatomical brain barriers segmentation," *Segmentation, Classification, and Registration of Multi-modality Medical Imaging Data*, vol. 12587, p. 27, 2021.

[12] M. Ning, C. Bian, D. Wei, S. Yu, C. Yuan, Y. Wang, Y. Guo, K. Ma, and Y. Zheng, "A new bidirectional unsupervised domain adaptation segmentation framework," in *International Conference on Information Processing in Medical Imaging*. Springer, 2021, pp. 492–503.

[13] J. Ma, Y. Zhang, S. Gu, C. Zhu, C. Ge, Y. Zhang, X. An, C. Wang, Q. Wang, X. Liu *et al.*, "AbdomenCT-1K: Is abdominal organ segmentation a solved problem," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[14] S. Choi, J. T. Kim, and J. Choo, "Cars can't fly up in the sky: Improving urban-scene segmentation via height-driven attention networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9373–9383.

[15] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE." *Journal of Machine Learning Research*, vol. 9, no. 11, 2008.

[16] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7472–7481.

[17] Y.-H. Chen, W.-Y. Chen, Y.-T. Chen, B.-C. Tsai, Y.-C. Frank Wang, and M. Sun, "No more discrimination: Cross city adaptation of road scene segmenters," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1992–2001.

[18] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," in *International Conference on Machine Learning*, 2018, pp. 1989–1998.

[19] J. Hoffman, D. Wang, F. Yu, and T. Darrell, "FCNs in the wild: Pixel-level adversarial and constraint-based adaptation," *arXiv preprint arXiv:1612.02649*, 2016.

[20] B. Settles, "Active learning literature survey," Department of Computer Sciences, University of Wisconsin-Madison, Tech. Rep., 2009.

[21] J.-C. Su, Y.-H. Tsai, K. Sohn, B. Liu, S. Maji, and M. Chandraker, "Active adversarial domain adaptation," in *The IEEE Winter Conference on Applications of Computer Vision*, 2020, pp. 739–748.

[22] M. Ning, D. Lu, D. Wei, C. Bian, C. Yuan, S. Yu, K. Ma, and Y. Zheng, "Multi-anchor active domain adaptation for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9112–9122.

[23] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," in *European Conference on Computer Vision*. Springer, 2016, pp. 102–118.

[24] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The Synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3234–3243.

[25] H. Hu, F. Wei, H. Hu, Q. Ye, J. Cui, and L. Wang, "Semi-supervised semantic segmentation via adaptive equalization learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 22 106–22 118, 2021.

[26] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proceedings*

of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 761–769.

[27] X. Glorot, A. Bordes, and Y. Bengio, "Domain adaptation for large-scale sentiment classification: A deep learning approach," in *International Conference on Machine Learning*, 2011.

[28] S. Li, C. H. Liu, B. Xie, L. Su, Z. Ding, and G. Huang, "Joint adversarial domain adaptation," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 729–737.

[29] S. Li, M. Xie, K. Gong, C. H. Liu, Y. Wang, and W. Li, "Transferable semantic augmentation for domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 516–11 525.

[30] M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Conditional adversarial domain adaptation," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[31] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool, "Domain adaptive faster R-CNN for object detection in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3339–3348.

[32] V. Vs, V. Gupta, P. Oza, V. A. Sindagi, and V. M. Patel, "Mega-CDA: Memory guided attention for category-aware unsupervised domain adaptive object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4516–4526.

[33] Z. Liu, Z. Miao, X. Pan, X. Zhan, D. Lin, S. X. Yu, and B. Gong, "Open compound domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 406–12 415.

[34] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *International Conference on Machine Learning*, 2015, pp. 97–105.

[35] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in *International Conference on Machine Learning*, 2017, pp. 2208–2217.

[36] W. Zellinger, T. Grubinger, E. Lughofer, T. Natschläger, and S. Saminger-Platz, "Central moment discrepancy (CMD) for domain-invariant representation learning," *arXiv preprint arXiv:1702.08811*, 2017.

[37] B. Sun and K. Saenko, "Deep CORAL: Correlation alignment for deep domain adaptation," in *European Conference on Computer Vision*. Springer, 2016, pp. 443–450.

[38] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2223–2232.

[39] Y. Li, L. Yuan, and N. Vasconcelos, "Bidirectional learning for domain adaptation of semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6936–6945.

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

[41] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 172–189.

[42] Y. Luo, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2507–2516.

[43] Q. Zhang, J. Zhang, W. Liu, and D. Tao, "Category anchor-guided unsupervised domain adaptation for semantic segmentation," in *Advances in Neural Information Processing Systems*, 2019, pp. 435–445.

[44] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2517–2526.

[45] K. Mei, C. Zhu, J. Zou, and S. Zhang, "Instance adaptive self-training for unsupervised domain adaptation," in *European Conference on Computer Vision*. Springer, 2020, pp. 415–430.

[46] W. Tranheden, V. Olsson, J. Pinto, and L. Svensson, "DACS: Domain adaptation via cross-domain mixed sampling," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1379–1389.

[47] Z. Jiang, Y. Li, C. Yang, P. Gao, Y. Wang, Y. Tai, and C. Wang, "Prototypical contrast adaptation for domain adaptive semantic segmentation," *arXiv preprint arXiv:2207.06654*, 2022.

[48] P. Zhang, B. Zhang, T. Zhang, D. Chen, Y. Wang, and F. Wen, "Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 414–12 424.

[49] Y. Lu, Y. Luo, L. Zhang, Z. Li, Y. Yang, and J. Xiao, "Bidirectional self-training with multiple anisotropic prototypes for domain adaptive semantic segmentation," *arXiv preprint arXiv:2204.07730*, 2022.

[50] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical vision Transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[51] L. Hoyer, D. Dai, and L. Van Gool, "HRDA: Context-aware high-resolution domain-adaptive semantic segmentation," *arXiv preprint arXiv:2204.13132*, 2022.

[52] ——, "Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9924–9935.

[53] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, "Active learning with statistical models," *Journal of Artificial Intelligence Research*, vol. 4, pp. 129–145, 1996.

[54] D. D. Lewis and J. Catlett, "Heterogeneous uncertainty sampling for supervised learning," in *Proceedings of the International Conference on Machine Learning*. Elsevier, 1994, pp. 148–156.

[55] T. Scheffer, C. Decomain, and S. Wrobel, "Active hidden Markov models for information extraction," in *International Symposium on Intelligent Data Analysis*. Springer, 2001, pp. 309–318.

[56] S. Dutt Jain and K. Grauman, "Active image segmentation propagation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2864–2873.

[57] S. C. Hoi, R. Jin, J. Zhu, and M. R. Lyu, "Semisupervised SVM batch mode active learning with applications to image retrieval," *ACM Transactions on Information Systems*, vol. 27, no. 3, pp. 1–29, 2009.

[58] S.-J. Huang, R. Jin, and Z.-H. Zhou, "Active learning by querying informative and representative examples," *Advances in Neural Information Processing Systems*, vol. 23, pp. 892–900, 2010.

[59] S. Dasgupta and D. Hsu, "Hierarchical sampling for active learning," in *Proceedings of the 25th International Conference on Machine Learning*, 2008, pp. 208–215.

[60] H. T. Nguyen and A. Smeulders, "Active learning using pre-clustering," in *Proceedings of the twenty-first International Conference on Machine Learning*, 2004, p. 79.

[61] A. Freytag, E. Rodner, and J. Denzler, "Selecting influential examples: Active learning with expected model output changes," in *European Conference on Computer Vision*. Springer, 2014, pp. 562–577.

[62] C. Kading, A. Freytag, E. Rodner, P. Bodesheim, and J. Denzler, "Active learning and discovery of object categories in the presence of unnameable instances," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4343–4352.

[63] A. Vezhnevets, V. Ferrari, and J. M. Buhmann, "Weakly supervised structured output learning for semantic segmentation," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 845–852.

[64] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, and H.-J. Zhang, "Two-dimensional active learning for image classification," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

[65] C.-C. Kao, T.-Y. Lee, P. Sen, and M.-Y. Liu, "Localization-aware active learning for object detection," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 506–522.

[66] W. Ji, J. Li, M. Zhang, Y. Piao, and H. Lu, "Accurate RGB-D salient object detection via collaborative learning," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 52–69.

[67] M. Zhang, W. Ji, Y. Piao, J. Li, Y. Zhang, S. Xu, and H. Lu, "LFNet: Light field fusion network for salient object detection," *IEEE Transactions on Image Processing*, vol. 29, pp. 6276–6287, 2020.

[68] Q. Sun, A. Laddha, and D. Batra, "Active learning for structured probabilistic models with histogram approximation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3612–3621.

[69] R. Chattopadhyay, W. Fan, I. Davidson, S. Panchanathan, and J. Ye, "Joint transfer and batch-mode active learning," in *International Conference on Machine Learning*, 2013, pp. 253–261.
[70] S.-J. Huang, J.-W. Zhao, and Z.-Y. Liu, "Cost-effective training of deep CNNs with active model adaptation," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 1580–1588.
[71] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.
[72] I. Shin, D.-J. Kim, J. W. Cho, S. Woo, K. Park, and I. S. Kweon, "LabOR: Labeling only if required for domain adaptive semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8588–8598.
[73] B. Xie, L. Yuan, S. Li, C. H. Liu, and X. Cheng, "Towards fewer annotations: Active learning via region impurity and prediction uncertainty for domain adaptive semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8068–8078.
[74] H. Cui, D. Wei, K. Ma, S. Gu, and Y. Zheng, "A unified framework for generalized low-shot medical image segmentation with scarce data," *IEEE Transactions on Medical Imaging*, 2020.
[75] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, 1967, pp. 281–297.
[76] Y. Siddiqui, J. Valentin, and M. Nießner, "ViewAL: Active learning with viewpoint entropy for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9433–9443.
[77] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *International Conference on Machine Learning*, 2016, pp. 478–487.
[78] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *arXiv preprint arXiv:1703.01780*, 2017.
[79] Y. Wang, H. Wang, Y. Shen, J. Fei, W. Li, G. Jin, L. Wu, R. Zhao, and X. Le, "Semi-supervised semantic segmentation using unreliable pseudo-labels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4248–4257.
[80] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, 2015.
[81] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 801–818.
[82] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
[83] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.
[84] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proceedings of the International Conference on Machine Learning*, vol. 30, no. 1, 2013, p. 3.