Multi-Anchor Active Domain Adaptation for Semantic Segmentation

Munan Ning*, Donghuan Lu*, Dong Wei†, Cheng Bian, Chenglang Yuan,
Shuang Yu, Kai Ma, Yefeng Zheng
Tencent Jarvis Lab, Shenzhen, China

Abstract

Unsupervised domain adaption has proven to be an effective approach for alleviating the intensive workload of manual annotation by aligning the synthetic source-domain data and the real-world target-domain samples. Unfortunately, mapping the target-domain distribution to the source-domain unconditionally may distort the essential structural information of the target-domain data. To this end, we firstly propose to introduce a novel multi-anchor based active learning strategy to assist domain adaptation regarding the semantic segmentation task. By innovatively adopting multiple anchors instead of a single centroid, the source domain can be better characterized as a multimodal distribution, thus more representative and complimentary samples are selected from the target domain. With little workload to manually annotate these active samples, the distortion of the target-domain distribution can be effectively alleviated, resulting in a large performance gain. The multi-anchor strategy is additionally employed to model the target-distribution. By regularizing the latent representation of the target samples compact around multiple anchors through a novel soft alignment loss, more precise segmentation can be achieved. Extensive experiments are conducted on public datasets to demonstrate that the proposed approach outperforms state-of-the-art methods significantly, along with thorough ablation study to verify the effectiveness of each component. The code will be released soon at https://github.com/munanning/MADA.

1. Introduction

Semantic segmentation has always been a fundamental task in computer vision. Benefiting from the rapid development of deep learning, many advanced segmentation methods have been proposed and achieved great breakthroughs with high accuracies for various tasks, such as autonomous driving [16], scene parsing [8, 44], object detection [26, 61] and human-computer interaction [43]. However, the requirement of large amount of data with accurate pixel-wise annotation limits their usage in many practical applications, *e.g.*, medical image segmentation [28, 40, 42, 41, 36] and auto-driving tasks [6].

To avoid the intensive workload of manual annotation, a lot of efforts have been made on unsupervised domain adaptation (UDA) [5, 20, 21, 55], which aims at aligning the target-domain distribution towards the source-domain distribution, so that networks trained with the supervision of only the synthetic source data can be applied to the real-world target data. However, forcing the target-domain features to fit the source-domain distribution may destroy the latent structural pattern of the target domain, resulting in inferior performance. As illustrated by the t-SNE [19] visualization in Fig. 1, the distributions of the source and target domains present both overlap (region ①) and obvious discrepancies (regions ② and ③). When the adapted target-domain features (red dots) obtained with a typical UDA method based on adversarial training [55]—despite generally aligned with the source domain distributions (blue squares)—show a clear distortion of the target-domain distribution in region ②, the adapted network presents unsatisfactory performance. Worse segmentation can be observed in region ③ when some specific targets are aligned neither with the source domain nor the target domain. A promising strategy to efficiently prevent such distortion of the target-domain distribution with minimal annotation workload is active learning (AL) [49]. By introducing little extra manual annotation for a few selected samples from the target domain, the performance can be significantly boosted regarding the classification and the detection tasks [51]. However, the sample selection methods in all previous active learning studies [51] assumed a unimodal source-domain distribution and neglected the potential multimodal distribution, resulting in sub-optimal active samples and inferior performance, as demonstrated in Table 4.

To address the above issues, we firstly propose to adopt the active learning strategy to assist domain adaptation (DA) regarding the semantic segmentation task, so that the essential structural pattern of the target domain can be maintained with minimal manual annotation workload. In ad-

*Contributed equally.
†Correspondence: donwei@tencent.com

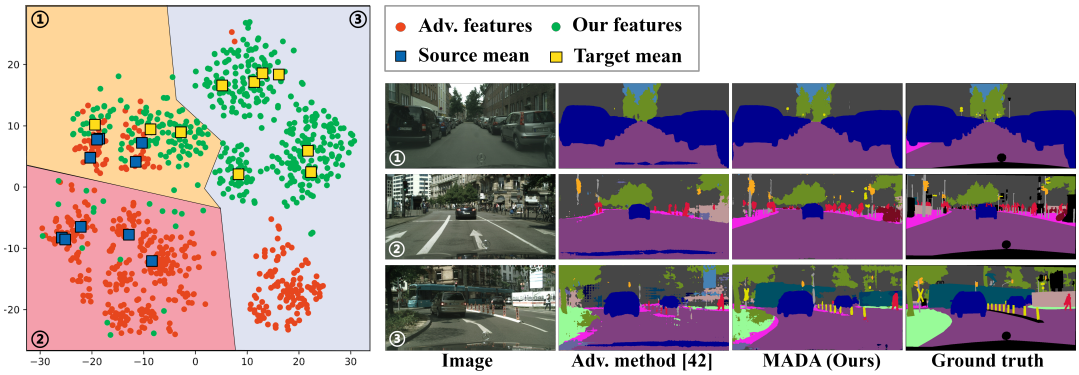

Figure 1. Visualization (t-SNE [19]) of the target-domain distribution distortion problem in UDA (left). The blue and yellow squares are the average latent representation of different category samples from the target-domain extracted by two networks trained with the source- and target-domain data, respectively, where we can observe little overlap (region ①) along with large discrepancy (regions ② and ③) between the two distributions. The red dots denote the adapted target-domain features by a typical adversarial (adv.) training based UDA method [55]. A general alignment to the source-domain distribution can be observed in region ①, while an obvious distortion of the target-domain distribution is displayed in region ② and ③, yielding unsatisfactory performance as presented on the right figures. By adopting active learning for the domain adaptation, such distortion is effectively alleviated, as demonstrated by the correctly distributed green dots.

dition, a multi-anchor strategy is proposed to better characterize the source-domain features as well as the target-domain features. Specifically, the proposed Multi-anchor Active Domain Adaptation (MADA) framework consists of two stages. In the first stage, with the network pretrained in an adversarial UDA [55] manner, a multi-anchor based active sample selection strategy is proposed to identify the most complementary and representative samples for manual annotation by exploiting the feature distributions across the target and source domains. 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 used for supervision, while all the available image information is additionally applied for optimization with a pseudo label loss and the proposed multi-anchor soft-alignment loss. In summary, our paper makes the following contributions:

- To the best of our knowledge, our work is the first study to adopt active learning to assist the domain adaptation regarding the semantic segmentation tasks. With little manual annotation workload of few target-domain samples, the distortion of the target-domain feature distribution can be effectively prevented and superior segmentation performance can be achieved.
- Assuming a multimodal distribution in practical situations, we propose to adopt multiple anchors obtained via clustering-based method to characterize the feature distribution of the source-domain, so that the representative target-domain samples which are the most complimentary to the source-domain can be selected.
- The multi-anchor strategy is used further to model the target-domain feature distribution. With the proposed multi-anchor soft-alignment loss, we show that explicitly pushing the features of the target samples towards multiple anchors leads to better latent representation, thus notably improve the segmentation performance.
- We conduct extensive experiments to demonstrate the superiority of the proposed MADA framework, along with thorough ablation studies to evaluate the effectiveness of the multi-anchor strategy on modeling the feature distribution.

2. Related Work

2.1. Unsupervised Domain Adaptation

Unsupervised domain adaptation (UDA) has been proposed for years, aiming to address the domain shift problem in a wide variety of computer vision tasks including classification [17], detection [4], and segmentation [55]. Recent UDA methods can be roughly divided into two groups: maximum mean discrepancy (MMD) based and adversarial learning based. The MMD kernel was first introduced in [33], which measured 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 [33], JMMD [34], CMD [59] and CORAL [52]. Minimization of the discrepancy yielded by these kernels forced 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 required complex computation in the high-dimension feature space.

In contrast, adversarial learning based methods are preferred for UDA of segmentation tasks, where the two do-

main distributions are drawn together via a domain discriminator. The classical appearance matching method CycleGAN [63] constructed two adversarial subnets to translate unpaired source and target images. BDL [32] leveraged label consistency to improve the UDA performance. DISE [1] proposed a disentangled representation learning architecture [25] to preserve structural information during image translation. Feature aligning methods such as CLAN [35] and CAG [62] utilized category-based distribution alignment to adapt the source and target domains in the feature and output spaces. AdvEnt [57] designed a novel loss function to maximize the prediction certainty in the target domain to boost the UDA performance.

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. A promising strategy to prevent such distortion with minimal annotation workload is active learning (AL) [49], which we adopt in this work.

2.2. Active Learning and Domain Adaptation

AL aims at optimal performance at a low annotation cost, by actively selecting the few samples that are most helpful to performance improvement, if labeled [7]. Over the past decade, several sample selection strategies have been proposed for AL, including uncertainty-based [31, 48], diversity-based [12, 22], representativeness-based [23, 10, 39], and expected model change based [14, 29, 56]. These strategies have been successfully applied to various computer vision tasks, such as image classification [45], object detection [30, 27, 60], and image segmentation [53]. In this work, we argue that it is beneficial to introduce AL to the DA problem, to avoid distortion of the target-domain distribution. First, AL only entails minimal annotation cost, which is acceptable in many scenarios considering the potential performance gain. Second, 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 few studies attempted applying AL to DA problems. An early work by Chattopadhyay *et al.* [2] proposed to use 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.* [24] proposed to fine-tune pre-trained models for classification tasks and involved additional active sample selection in every iteration. In contrast, 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) [51] proposed AL for DA with the adversarial learning [15] strategy, where representative samples were selected by jointly considering diversity and uncertainty criteria. In this work, 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 captures more comprehensive information from both domains and can achieve substantial performance improvement (experimentally validated in Section 4.5).

3. Method

The proposed method consists of two main stages: active target sample selection based on multiple anchors of the source domain (Fig. 2(a)), and semi-supervised domain adaptation enhanced by a novel multi-anchor soft alignment loss (Figs. 2(b), 2(c) and 2(d)). Below we first formally define our problem setting, then elaborate the two stages.

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 for the width, and 3 for 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 $X^t_L = \{(x^t_L, y^t_L)\}$ and $N_t - N_a$ unlabeled images $X^t_U = \{x^t_U\}$. Given the scenario, 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. In this work, we propose an efficient anchoring mechanism to model the domain distributions, and close the gap between network predictions and the anchors by forming compact clusters around the anchors. Previously, CAG [62] averaged all image-level features of the source domain to obtain a centroid representing the entire domain, which implicity assumed a unimodal distribution. In practice, however, the distribution of a domain may actually comprise more than a single mode [9]. 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) based on their overall representative distributions. By concatenating the features of different categories into an image-level 'connected' vector, we perform clustering on them to estimate scene-specific representative distributions

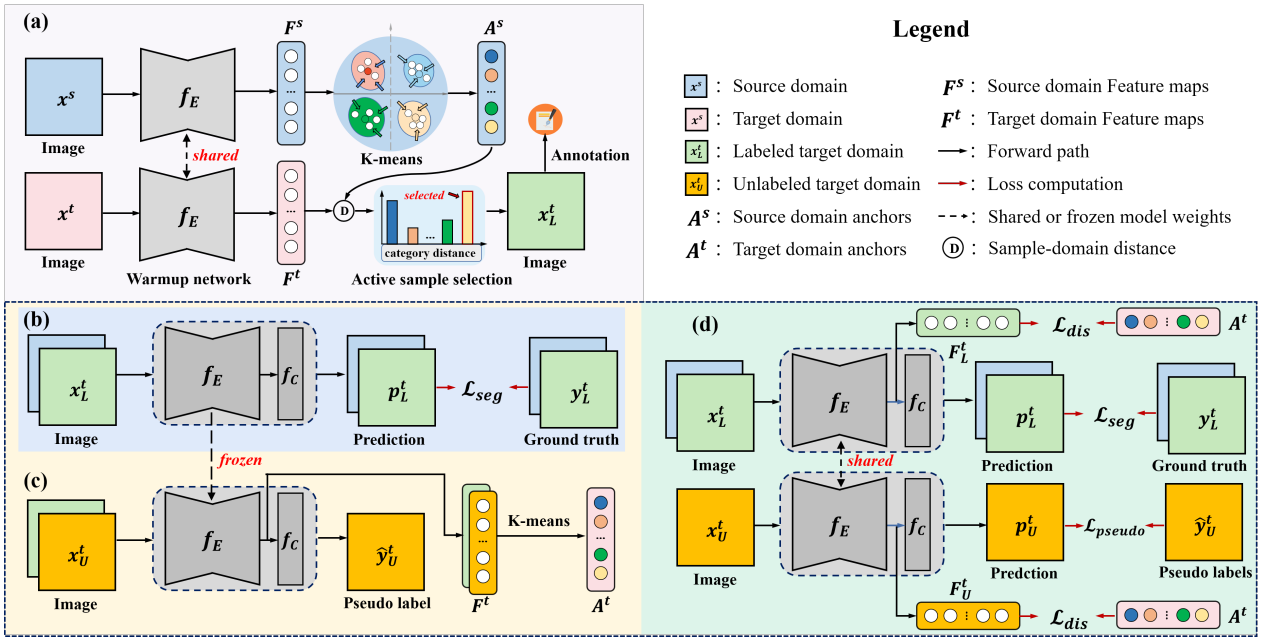

Figure 2. Overview of the proposed MADA framework.

with cluster centers, denoted as 'anchors'. We then measure the distance between each target sample and its nearest source anchor, and select the furthest samples. Below we first elaborate our multiple anchoring mechanism (demonstrated with the source domain), then describe how to use it for active target sample selection.

As a warm-up, we first employ the common adversarial training [55] strategy to narrow the gap between the source and target domains. After that, 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:

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

where y_c^s denotes the label map for category c, $f_E(x^s)|_c$ is the networks' output for category c, $\otimes$ denotes element-wise multiplication for extraction of 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 connecting the vectors of all categories into a long vector. Then, we apply the K-means method [38] 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 category C, and the impact of different K is explored in Section 4.6.

Active Target Sample Selection Against Source Anchors. For single-domain AL, uncertainty-based metrics were extensively used to select the samples which are the most difficult to segment [50]. For multi-domain AL, however, we argue that the more dissimilar the target samples are to the source-domain, the more complimentary they are to the segmentation network. Here, we measure the dissimilarity by the distance between the target-domain samples and the source-domain anchors to assess the importance of unlabeled target-domain samples to domain adaptation. Specifically, we first calculate the per category feature map of a target-domain image x^t:

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

where $\hat{y}_c^t$ is the predicted label 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 combine $F_c^t(x^t)$ of all categories to obtain the image-level feature vector $F^t(x^t)$. Eventually, we calculate the $L2$ distances from $F^t(x^t)$ to all source-

domain anchors, and define the smallest of them as the distance from the target-domain sample to the source domain:

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

Intuitively, this definition assigns the target-domain sample to the closest anchor of the source domain's, which corresponds to a mode in the multimodal source-domain distribution. Based on the distance, we can identify the target-domain samples that are far away from the entire source domain and thus are expected to contain target domain specific information. Therefore, we select them as active samples and annotate them for subsequent training, hoping to learn unique components of the target-domain distribution from these active annotations.

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 to the target domain (Fig. 2(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 model $f_C(f_E)$, and f_C is a classifier. As experimentally validated (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: Computing Target-domain Anchors and Pseudo Labels. To fully utilize the unlabeled target data X_U^t, we use the fine-tuned model to compute pseudo labels $\{\hat{y}^t\}$ for unlabeled target-domain samples as well as target-domain anchors $\{A_v^t\}_{v=1}^V$ (Fig. 2(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, it is natural to correct them dynamically. As indicated by Xie et al. [58], re-clustering at each epoch could 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) [54] 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 [54], and $F^t(x^t)$ is utilized to update the closest anchor. With both $\{\hat{y}^t\}$ and $\{A_v^t\}$ computed, we proceed to the next step for semi-supervised domain adaptation.

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. 2(d)). Notably, we propose a novel soft alignment loss to explicitly close the gap between the sample features and anchors in the target domain:

$$\mathcal{L}_{dis}^t = V \Big/ \sum_{v=1}^{V} \frac{1}{\|F^t(x^t) - A_v^t\|_2^2}. \tag{9}$$

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. Besides, to make a full use of X_U^t, we exploit the pseudo labels $\hat{y}^t$ to provide further supervision:

$$\mathcal{L}_{pseudo} = \mathcal{L}_{CE}\left(x_U^t, \hat{y}^t\right). \tag{10}$$

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

$$\mathcal{L}_{semi} = \mathcal{L}_{seg} + \mathcal{L}_{dis}^t + \mathcal{L}_{pseudo}. \tag{11}$$

The entire training pipeline is summarized in Algorithm 1.

4. Experiments

4.1. Datasets

To demonstrate the superiority of our proposed method, two challenging *synthia-2-real* adaptation tasks, *i.e.*, GTA5 [46] $\rightarrow$ Cityscapes [8] and SYNTHIA [47] $\rightarrow$ 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 [32], the SYNTHIA-RAND-CITYSCAPES set with 9,400 synthetic images containing 16-class segmentation is utilized for training.

In both sets, 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) [13] metric.

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 [55]	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 [35]	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 [57]	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 [32]	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 [62]	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
AADA [51]	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 (Ours)	**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**

Table 2. Comparison with other DA methods on the SYNTHIA to Cityscapes adaptation task. Best results are shown in **bold**.

SYNTHIA → Cityscapes																		
Method	road	sidewalk	building	wall	fence	pole	light	sign	veg	sky	person	rider	car	bus	mbike	bicycle	mIoU	mIoU*
AdaptSeg [55]	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 [35]	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 [57]	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	-
BDL [32]	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 [62]	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
AADA [51]	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 (Ours)	**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**

4.2. Implementation Details

We employ the DeepLab v3+ [3] as the feature extractor f_E, which is composed of the backbone ResNet-101 [18] pretrained on ImageNet [11] 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 [37] parameterized by 0.2. The proposed method is implemented on PyTorch with a TITAN Tesla V100 GPU. The input images are randomly resized with a ratio in $[0.5, 1.5]$ and then 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 fine-tuning in the second stage, we use the SGD optimizer to train our model for 50 epochs. The learning rate is initially set to 2.5×10^{-4} and decayed by poly learning rate policy with a power of 0.9.

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

4.3. Main Results

As presented in Table 1 and Table 2, the proposed framework is compared with five UDA methods [55, 35, 57, 32, 62] and an active DA approach [51]. As expected, we observe substantial improvements over the UDA methods, suggesting that with carefully selected active samples, little manual annotation workload can lead to large performance gains. In addition, the proposed method outperforms another active DA method, *i.e.*, AADA, by a large margin (5.6% mIOU), demonstrating the effectiveness of the proposed multi-anchor strategy. The visualization of three example images, which are the same as those in Fig. 1, is displayed in Fig. 3 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 MADA method.

4.4. Ablation Study

To verify the effectiveness of each component, we perform an ablation study with the following variants: $\mathbf{M}^{(0)}$: the baseline adversarial learning method [55] 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 proposed multi-anchor soft alignment loss on target samples for optimization; $\mathbf{M}^{(3)}$: extending $\mathbf{M}^{(2)}$ by progressively updating the target anchors with EMA; $\mathbf{M}^{(4)}$: adding the pseudo

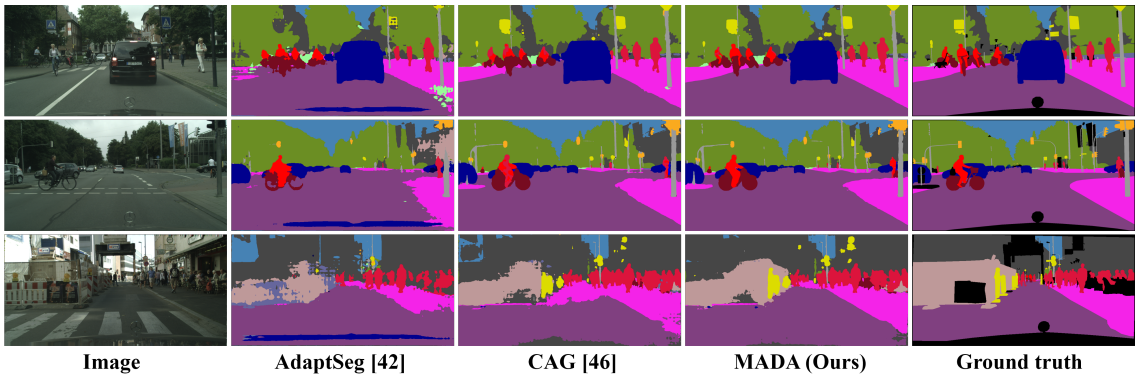

Figure 3. Qualitative results of DA segmentation for GTA5 → Cityscapes. For each image, we show the results of the typical adversarial method [55], state-of-the-art UDA method [62] and our proposed MADA, respectively. The black region in "Ground truth" is excluded from evaluation because it does not belong to any of the 19 classes.

Algorithm 1 Multi-anchor Active Domain Adaptation (MADA)

Notation: Source-domain set $\{(x^s, y^s)\}$, selected active sample set $\{(x_L^t, y_L^t)\}$, and 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 $\{F^t(x^t)\}$. Number of iterations N.

Stage 1:

1: Warm-up f_E with adversarial training [55] to obtain $\{F^s(x^s)\}$.
2: Apply K-means on $\{F^s(x^s)\}$ to group the source-domain samples into K clusters;
3: Compute the centroid A_k^s of the clusters (Eq. (3)) to serve as the source-domain anchors;
4: Calculate the distance from each target-domain sample to $\{A_k^s\}$ (Eq. (5));
5: Select 5% target-domain samples with the smallest distances as active samples for annotation, getting set $\{(x_L^t, y_L^t)\}$.

Stage 2:

6: 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 obtain $\{F^t(x^t)\}$;
7: Initialize A_v^t with K-means clustering on $\{F^t(x^t)\}$;
8: **for** $i = 1, ..., N$ **do**
9: Calculate $\mathcal{L}_{seg}$ (Eq. (6)) with $\{(x^s, y^s)\}$ and $\{(x_L^t, y_L^t)\}$;
10: Calculate $\mathcal{L}_{dis}^t$ (Eq. (9)) with $\{x^t\}$ and $\mathcal{L}_{pseudo}$ (Eq. (10)) with $\{x_U^t\}$;
11: Update f_E by gradient descending $\nabla(\mathcal{L}_{seg} + \mathcal{L}_{dis}^t + \mathcal{L}_{pseudo})$ (Eq. (11));
12: Update A_v^t with EMA (Eq. (8));
13: **end for**

label loss for optimization 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 public datasets demonstrate the effectiveness of each strategy. Furthermore, MADA with only 5% of the target-domain samples actively annotated achieves a comparable performance with that of the upper bound, suggesting that the proposed framework can select complimentary samples to effectively close the gap between UDA and full supervision.

Table 3. Ablation study. G → C denotes the GTA5 → Cityscapes scenario and S → C denotes the SYNTHIA → Cityscapes scenario.

					G → C	S → C
Method	A	B	C	D	mIoU	mIoU
$\mathbf{M}^{(0)}$					42.5	42.9
$\mathbf{M}^{(1)}$	✓				61.6	65.0
$\mathbf{M}^{(2)}$	✓	✓			63.2	66.6
$\mathbf{M}^{(3)}$	✓	✓	✓		63.8	67.6
$\mathbf{M}^{(4)}$	✓	✓	✓	✓	64.9	68.1
$\mathbf{M}^{(u)}$					69.3	70.8

A: Training with active samples
B: Soft-anchor alignment loss
C: Updating target anchor with EMA
D: Pseudo training for unlabled target samples

The visualization of the feature distribution with/without active learning is presented in Fig 1. With the proposed MADA 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. On 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 [57] is applied to obtain the prediction map entropy of each sample in the target domain and the ones with top 5% entropy are chosen for manual annotation:

$$E_{ent} = \frac{-1}{\log(C)} \sum_{c=1}^{C} \sum_{i=1}^{H \times W} p_{i,c}^t \log(p_{i,c}^t). \quad (12)$$

Table 4. Experiments on 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	92.8	64.5	85.8	38.0	34.8	43.7	50.1	56.9	87.9	40.4	87.7	69.0	30.8	89.4	51.1	43.8	21.7	29.9	59.4	56.7
Entropy [57]	93.9	65.4	87.7	42.2	48.4	46.7	47.3	57.0	88.5	44.3	90.4	70.8	32.8	90.0	53.8	49.9	30.0	41.1	63.6	60.2
Adversarial [55]	91.8	59.2	87.5	37.8	45.2	45.5	51.5	56.9	88.5	43.0	90.3	69.0	37.1	89.9	54.5	46.1	35.9	28.1	61.3	58.9
AADA [51]	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
Proposed	**92.4**	**61.4**	**87.4**	**39.5**	**45.9**	45.2	**50.6**	57.5	87.8	42.4	89.2	**72.7**	**44.9**	**90.0**	54.7	**50.5**	**43.4**	**47.8**	**66.9**	**61.6**

Adversarial-based Diversity Method. With the discriminator f_D trained in the warm-up stage as [55], we select the samples with least predicted probabilities, *i.e.*, the ones that are most distinguishable from the source domain:

$$E_{adv} = \frac{1 - f_D(f_E(x^t)}{f_D(f_E(x^t))}. \tag{13}$$

AADA Method. In addition to the discriminator-based diversity, the AADA [51] method also takes the certainty of prediction into consideration:

$$E_{AADA} = E_{ent} E_{adv}. \tag{14}$$

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, while no unlabeled samples are used for optimization. We can observe that the proposed multi-anchor strategy delivers the best segmentation performance in mIoU, suggesting that better active samples are selected by our proposed strategy.

4.6. 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 both domains. As shown in Fig. 4, for both domains, using multiple anchors was consistently better than using a single centroid, and using 5–10 anchors stably yielded superior performance. This might be because there are only limited types of scenarios in these datasets, and a few anchors are sufficient to represent their distributions. We therefore use 10 clusters considering the top performance in both domains.

4.7. Impact of the Number of Active Samples

In order to verify the stability of our proposed method, comparative experiments for different percentages of active samples are conducted. As shown in Table 5, as the percentage of samples increases from 1% to 20%, the mIoU increases steadily from 56.7% to 64.1%. We also introduce the upper bound by optimizing with all target labels, the narrow gap of 7.7% in mIoU between using only 5% of target-domain data for AL and the upper bound demonstrates that the proposed method can effectively exploit the information from active samples.

Table 5. Experiments on different number of active samples.

GTA5 → Cityscapse						
Percentage	1%	2%	5%	10%	20%	100%
mIoU	56.7	59.1	61.6	62.7	64.1	69.3
mIoU Gap	−12.6	−10.2	−7.7	−6.6	−5.2	-

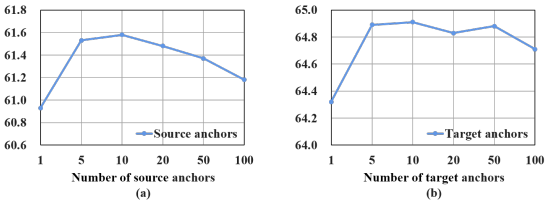

Figure 4. Experiments on different number of anchors for the source domain (a) and target domain (b).

5. Conclusion

In this paper, we proposed the Multi-anchor Active Domain Adaptation (MADA) framework, for distortion-free source-to-target domain adaptation of segmentation models at minimal annotation cost. MADA introduced anchor-based active sample selection into DA, for selection of limited target-domain samples that were most complementary to the source-domain distribution and meanwhile unique to the target-domain distribution. Adding active annotation of these selected target-domain samples for training can effectively prevent distortion of the target-domain distribution that could otherwise happen in typical UDA methods. Different from previous works which assumed unimodal distributions for both the source and target domains, MADA proposed to use multiple anchors to realize multimodal distributions for both domains. On top of that, MADA further proposed a multi-anchor soft-alignment loss to explicitly push the target-domain features towards these anchors, for full utilization of the unlabeled target-domain samples. Experimental results on two public benchmark datasets demonstrated the effectiveness of (i) introducing AL into DA, (ii) multiple anchors versus a single centroid, and (iii) adding the soft-alignment loss, as well as the superior performance of MADA towards existing state-of-the-art UDA and active DA methods.

References

[1] Wei-Lun Chang, Hui-Po Wang, Wen-Hsiao Peng, and Wei-Chen 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*, pages 1900–1909, 2019.

[2] Rita Chattopadhyay, Wei Fan, Ian Davidson, Sethuraman Panchanathan, and Jieping Ye. Joint transfer and batch-mode active learning. In *International Conference on Machine Learning*, pages 253–261, 2013.

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[4] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc 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*, pages 3339–3348, 2018.

[5] Yi-Hsin Chen, Wei-Yu Chen, Yu-Ting Chen, Bo-Cheng Tsai, Yu-Chiang Frank Wang, and Min Sun. No more discrimination: Cross city adaptation of road scene segmenters. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1992–2001, 2017.

[6] Sungha Choi, Joanne T Kim, and Jaegul 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*, pages 9373–9383, 2020.

[7] David A Cohn, Zoubin Ghahramani, and Michael I Jordan. Active learning with statistical models. *Journal of artificial intelligence research*, 4:129–145, 1996.

[8] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[9] Hengji Cui, Dong Wei, Kai Ma, Shi Gu, and Yefeng Zheng. A unified framework for generalized low-shot medical image segmentation with scarce data. *IEEE Transactions on Medical Imaging*, 2020.

[10] Sanjoy Dasgupta and Daniel Hsu. Hierarchical sampling for active learning. In *Proceedings of the 25th international conference on Machine learning*, pages 208–215, 2008.

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[12] Suyog Dutt Jain and Kristen Grauman. Active image segmentation propagation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2864–2873, 2016.

[13] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015.

[14] Alexander Freytag, Erik Rodner, and Joachim Denzler. Selecting influential examples: Active learning with expected model output changes. In *European Conference on Computer Vision*, pages 562–577. Springer, 2014.

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

[16] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012.

[17] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *ICML*, 2011.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[19] Geoffrey Hinton and Sam T Roweis. Stochastic neighbor embedding. In *NIPS*, volume 15, pages 833–840. Citeseer, 2002.

[20] 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 *International conference on machine learning*, pages 1989–1998. PMLR, 2018.

[21] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016.

[22] Steven CH Hoi, Rong Jin, Jianke Zhu, and Michael R Lyu. Semisupervised svm batch mode active learning with applications to image retrieval. *ACM Transactions on Information Systems (TOIS)*, 27(3):1–29, 2009.

[23] Sheng-Jun Huang, Rong Jin, and Zhi-Hua Zhou. Active learning by querying informative and representative examples. *Advances in neural information processing systems*, 23:892–900, 2010.

[24] Sheng-Jun Huang, Jia-Wei Zhao, and Zhao-Yang 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*, pages 1580–1588, 2018.

[25] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018.

[26] Wei Ji, Jingjing Li, Shuang Yu, Miao Zhang, Yongri Piao, Shunyu Yao, Qi Bi, Kai Ma, Yefeng Zheng, Huchuan Lu, and Li Cheng. Calibrated rgb-d salient object detection. In *CVPR*, pages 9471–9481, June 2021.

[27] Wei Ji, Jingjing Li, Miao Zhang, Yongri Piao, and Huchuan Lu. Accurate RGB-D salient object detection via collaborative learning. In *ECCV*, pages 52–69, 2020.

[28] Wei Ji, Shuang Yu, Junde Wu, Kai Ma, Cheng Bian, Qi Bi, Jingjing Li, Hanruo Liu, Li Cheng, and Yefeng Zheng. Learning calibrated medical image segmentation via multi-rater agreement modeling. In *CVPR*, pages 12341–12351, June 2021.

[29] Christoph Kading, Alexander Freytag, Erik Rodner, Paul Bodesheim, and Joachim 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*, pages 4343–4352, 2015.

[30] Chieh-Chi Kao, Teng-Yok Lee, Pradeep Sen, and Ming-Yu Liu. Localization-aware active learning for object detection. In *Asian Conference on Computer Vision*, pages 506–522. Springer, 2018.

[31] David D Lewis and Jason Catlett. Heterogeneous uncertainty sampling for supervised learning. In *Machine learning proceedings 1994*, pages 148–156. Elsevier, 1994.

[32] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6936–6945, 2019.

[33] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015.

[34] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, pages 2208–2217. PMLR, 2017.

[35] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2507–2516, 2019.

[36] Jun Ma, Yao Zhang, Song Gu, Cheng Zhu, Cheng Ge, Yichi Zhang, Xingle An, Congcong Wang, Qiyuan Wang, Xin Liu, et al. Abdomenct-1k: Is abdominal organ segmentation a solved problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[37] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3. Citeseer, 2013.

[38] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[39] Hieu T Nguyen and Arnold Smeulders. Active learning using pre-clustering. In *Proceedings of the twenty-first international conference on Machine learning*, page 79, 2004.

[40] Munan Ning, Cheng Bian, Donghuan Lu, Hong-Yu Zhou, Shuang Yu, Chenglang Yuan, Yang Guo, Yaohua Wang, Kai Ma, and Yefeng Zheng. A macro-micro weakly-supervised framework for as-oct tissue segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 725–734. Springer, 2020.

[41] Munan Ning, Cheng Bian, Dong Wei, Shuang Yu, Chenglang Yuan, Yaohua Wang, Yang Guo, Kai Ma, and Yefeng Zheng. A new bidirectional unsupervised domain adaptation segmentation framework. In *International Conference on Information Processing in Medical Imaging*, pages 492–503. Springer, 2021.

[42] Munan Ning, Cheng Bian, Chenglang Yuan, Kai Ma, and Yefeng Zheng. Ensembled resunet for anatomical brain barriers segmentation. *Segmentation, Classification, and Registration of Multi-modality Medical Imaging Data*, 12587:27, 2021.

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

[44] 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.

[45] Guo-Jun Qi, Xian-Sheng Hua, Yong Rui, Jinhui Tang, and Hong-Jiang Zhang. Two-dimensional active learning for image classification. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[46] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European conference on computer vision*, pages 102–118. Springer, 2016.

[47] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016.

[48] Tobias Scheffer, Christian Decomain, and Stefan Wrobel. Active hidden markov models for information extraction. In *International Symposium on Intelligent Data Analysis*, pages 309–318. Springer, 2001.

[49] Burr Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

[50] Yawar Siddiqui, Julien Valentin, and Matthias Nießner. Viewal: Active learning with viewpoint entropy for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9433–9443, 2020.

[51] Jong-Chyi Su, Yi-Hsuan Tsai, Kihyuk Sohn, Buyu Liu, Subhransu Maji, and Manmohan Chandraker. Active adversarial domain adaptation. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 739–748, 2020.

[52] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer, 2016.

[53] Qing Sun, Ankit Laddha, and Dhruv Batra. Active learning for structured probabilistic models with histogram approximation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3612–3621, 2015.

[54] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv preprint arXiv:1703.01780*, 2017.

[55] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7472–7481, 2018.

[56] Alexander Vezhnevets, Vittorio Ferrari, and Joachim M Buhmann. Weakly supervised structured output learning for semantic segmentation. In *2012 IEEE conference on computer vision and pattern recognition*, pages 845–852. IEEE, 2012.

[57] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2517–2526, 2019.

[58] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pages 478–487. PMLR, 2016.

[59] Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. Central moment discrepancy (cmd) for domain-invariant representation learning. *arXiv preprint arXiv:1702.08811*, 2017.

[60] Miao Zhang, Wei Ji, Yongri Piao, Jingjing Li, Yu Zhang, Shuang Xu, and Huchuan Lu. LFNet: Light field fusion network for salient object detection. *IEEE Transactions on Image Processing*, 29:6276–6287, 2020.

[61] Miao Zhang, Jingjing Li, Wei Ji, Yongri Piao, and Huchuan Lu. Memory-oriented decoder for light field salient object detection. In *NeurIPS*, pages 896–906, 2019.

[62] Qiming Zhang, Jing Zhang, Wei Liu, and Dacheng Tao. Category anchor-guided unsupervised domain adaptation for semantic segmentation. In *Advances in Neural Information Processing Systems*, pages 435–445, 2019.

[63] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.