

Reducing Annotation Effort by Identifying and Labeling Contextually Diverse Classes for Semantic Segmentation Under Domain Shift

Sharat Agarwal¹, Saket Anand¹, and Chetan Arora²

¹IIT Delhi, India {sharata, anands}@iiitd.ac.in

²Indian Institute of Technology Delhi, India chetan@cse.iitd.ac.in

Abstract

In Active Domain Adaptation (ADA), one uses Active Learning (AL) to select a subset of images from the target domain, which are then annotated and used for supervised domain adaptation (DA). Given the large performance gap between supervised and unsupervised DA techniques, ADA allows for an excellent trade-off between annotation cost and performance. Prior art makes use of measures of uncertainty or disagreement of models to identify ‘regions’ to be annotated by the human oracle. However, these regions frequently comprise of pixels at object boundaries which are hard and tedious to annotate. Hence, even if the fraction of image pixels annotated reduces, the overall annotation time and the resulting cost still remain high. In this work, we propose an ADA strategy, which given a frame, identifies a set of classes that are hardest for the model to predict accurately, thereby recommending semantically meaningful regions to be annotated in a selected frame. We show that these set of ‘hard’ classes are context-dependent and typically vary across frames, and when annotated help the model generalize better. We propose two ADA techniques: the Anchor-based and Augmentation-based approaches to select complementary and diverse regions in the context of the current training set. Our approach achieves 66.6 mIoU on $GTA5 \rightarrow Cityscapes$ dataset with an annotation budget of 4.7% in comparison to 64.9 mIoU by MADA [22] using 5% of annotations. Our technique can also be used as a decorator for any existing frame-based AL technique, e.g., we report 1.5% performance improvement for CDAL [1] on *Cityscapes* using our approach.

1. Introduction

One of the major stumbling blocks for development of robust semantic segmentation techniques is the associated cost for obtaining annotated samples from a variety of tar-



Figure 1. Different annotation strategies for semantic segmentation in active learning. (a) **Target Image**. (b) **Frame-based** methods annotate the entire selected frame, which leads to annotating redundant pixels wasting annotation budget. (c) **Region-Based** methods select arbitrary regions based on uncertainty in each image across the unlabeled pool, which are often tedious and challenging for a human to annotate. We propose a novel and intuitive approach, (d) **Class-based** annotation, where contextually relevant classes that are complementary for model training are identified in each actively selected frame, thus reducing the annotation effort and simultaneously increasing the model performance.

get domains. For instance, it takes 90 minutes to annotate a single *Cityscapes* [6] image. One of the ways to mitigate data annotation effort/cost is by synthetically generating data with dense labels using 3D simulation platforms and game engines [28, 29]. However, Deep Neural Networks (DNNs) trained entirely on the synthetic data fail to generalize well in the real-world setting due to the domain shift. To address this issue several unsupervised [39, 42, 19, 14, 20, 24], semi-supervised [4, 43, 30] and self-training [21, 44, 47] based DA techniques were proposed.

However, despite the immense effort, the performance still lags far behind the fully supervised models.

To maximize model performance with minimal labeling effort, recently, ADA techniques [22, 33] have been proposed where the most *informative* samples from the unlabeled target domain are selected for the annotation. MADA [22], a frame selection approach, labels target samples that are most complementary to source domain anchors. LabOR [33] is a pixel-based approach which obtains labels for those uncertain pixels in an image where two classifiers disagree on their predictions. Both the approaches have shown significant performance improvement, but are highly inefficient in terms of annotation cost. On one hand, MADA wastes annotation budget by labeling redundant pixels in a selected frame (c.f. Fig. 1(b)). On the other hand, LabOR selects sparse pixels belonging to different classes (c.f. Fig. 1(c)), which are tedious and time-consuming to annotate by a human annotator. By choosing to work at a pixel level, where diversity is hard to compute, LabOR fails to consider annotated regions across the frames.

We argue that to gain cost efficiency, it is critical to select semantically meaningful target domain regions which are hard/novel for a model. Choosing the semantic regions rather than individual pixels maintains the simplicity of annotation task for a human oracle, and at the same time allows an automated algorithm to use dataset-wide criterion such as diversity, and novelty. E.g., if a model has only seen straight roads for the `road` label, it will likely struggle in images containing road with turns, thus making `road` as a *hard* label in those frames. It is important to point out the contextual nature of the *hard* classes, which may be quite different from the difficulty arising from the imbalanced classes. Consequently, we take an intuitive approach and define semantically meaningful regions as instances of the hard classes in a frame for the annotation.

Contributions: (1) We introduce two ADA techniques for selecting hard classes in the frames selected by any contemporary frame selection strategy such as CDAL [1]. Our *Anchor-based* approach (c.f. Section 3.2) selects class instances novel with respect to class-wise anchors chosen from the target dataset. This helps strengthen the per class representation in the feature space. Our *Augmentation-based* (c.f. Section 3.3) approach follows the self-supervised uncertainty estimation, and chooses class instances based on the disagreement in the prediction probabilities for the corresponding weakly and strongly corrupted samples. (2) To specifically understand our contribution of choosing semantic regions instead of pixels or frames in ADA, we ablate using an hypothetical *IoU-based* (Section 3.4) selection approach. Here, we select low confident classes in a frame based on the difference in their *IoU* values from the current model to a hypothetical fully supervised model. We show that us-

ing this approach one can achieve performance equal to a fully supervised model using only $\sim 7\%$ of the annotated data. This validates the significance of choosing semantic regions. (3) We compare with state-of-the-art (SOTA) UDA and ADA techniques, reducing the error margin (difference in *mIoU* from the fully supervised model) from 4.7 to 3 in `GTA5`→`Cityscapes` and from 5.7 to 3.2 in `Synthia`→`Cityscapes` at an annotation budget of merely 5%. Complete source-code and pre-trained models for this work will be publicly released post-acceptance.

2. Related Work

Domain Adaptation: Unsupervised Domain Adaptation (UDA) addresses the problem of domain shift between the labeled source and unlabeled target domains and has been extensively explored for image classification [7], object detection [5, 37] and semantic segmentation [39, 42, 20]. UDA is majorly categorized into two groups, based on (a) maximum mean discrepancy MMD [16, 17, 41] or (b) using adversarial learning [39, 42, 9]. Adversarial learning based approaches are more popular, and have been used to align source and target distributions at image [8, 13, 38], feature [39, 42], and output [40, 23] stages. Despite extensive interest, there is still a significant performance gap between supervised learning and UDA-based approaches [22]. To reduce the performance gap, semi-supervised learning (SSL) [4, 43, 30] based DA approaches have been proposed which utilize a small portion of randomly selected labeled target data for training. The implicit assumption is that the randomly selected set maintains the relationship between labeled and unlabeled data distribution [18].

Active Learning: Instead of labeling randomly selected samples, AL algorithms choose the most valuable samples to be labeled by a human annotator [32]. Since, annotating is far more expensive than collecting the data, several AL strategies have been proposed [27], based on ideas like membership query [10], stream-based sampling [11], and pool-based sampling [31, 1]. Problems of interest include image classification [31, 34], object detection [1] and semantic segmentation [34, 1]. Despite the enormous effort required in annotation for semantic segmentation, there has been a limited amount of work in this domain.

Active Domain Adaptation: Active Domain adaptation (ADA) techniques adopt active learning for the task of domain adaptation, where most valuable samples from the unlabeled target domain are labeled. Recently, ADA techniques have been proposed for image classification [25, 26]. Our focus in this paper is on semantic segmentation, where the SOTA techniques include MADA [22], and LabOR [33]. Whereas, MADA [22] annotates target domain frames most complementary to the anchors from the source domain, LabOR [33] annotates most uncertain regions based on the

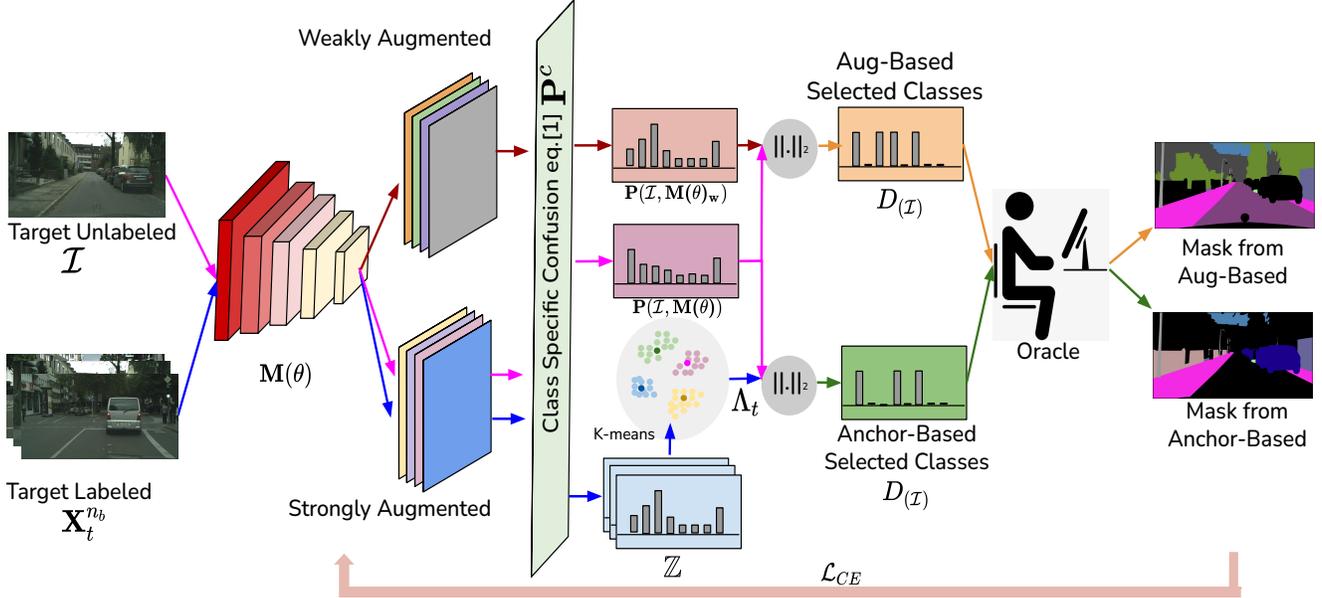


Figure 2. Overview of our proposed approach. Given small set of labeled target images $\mathbf{X}_t^{n_b}$ and an unlabeled image \mathcal{I} , Anchor-based and Aug-based select classes $D_{\mathcal{I}}$ to be labeled. In Anchor-based we select classes complementary to the target anchors exploiting labeled target class distribution. Whereas, in Aug-based we measure the dissimilarity in the class confusions at inference time from Strong and Weakly augmented model. (Best visible in color)

classifier disagreement in each image. As highlighted in Section 1 both the existing ADA approaches are highly inefficient in terms of annotation cost.

3. Methodology

In this section we firstly discuss the preliminaries and the problem setup. Then, we present the proposed class selection approaches, Aug-Based (3.3), Anchor-Based (3.2), the ground truth based skyline IoU-Based (3.4), and finally our training objective (3.5). Fig. 2 shows the overview of our proposed approach.

3.1. Problem Overview and Background

In UDA, given the source dataset $\mathcal{X}_s = \{x_s\}^{n_s}$ with pixel-level labels $\mathcal{Y}_s = \{y_s\}^{n_s}$, the goal is to learn a segmentation model $\mathbf{M}(\theta)$ which can correctly predict pixel-level labels for the target domain samples, $\mathcal{X}_t = \{x_t\}^{n_t}$ without using $\mathcal{Y}_t = \{y_t\}^{n_t}$, where \mathcal{Y}_s and \mathcal{Y}_t share the same label space of C classes and n_s and n_t are the number of images from the source and target domains. In Active Domain Adaptation (ADA) the task is to select a set of n_b images in each iteration, $\mathbf{S}_t^{n_b} \subset \mathcal{X}_t^{n_t}$ with $n_b \ll n_t$ as the annotation budget, to be labeled by a human oracle such that $\mathbf{M}(\theta)$ achieves good performance on the target domain with only a few annotations.

Usually, traditional Active Learning approaches either annotate the entire frame or annotate regions within a frame based on measures like uncertainty. Such measures are

based on the model’s performance, and may lead to semantically inconsistent regions that straddle class or object boundaries and are therefore harder to annotate manually. Contrary to this approach, we propose to select semantically meaningful regions as instances of certain classes in the selected frames to be annotated. Thus our Anchor-based and Aug-based identifies classes in the frames to be labeled. For our experiments we have used CDAL [1], but we further analyze the effectiveness of class selection in other frame selection techniques.

Let the frame selection function be Δ , such that $\mathbf{S}_t^{n_b} = \Delta(\mathcal{X}_t^{n_t}, n_b)$, where n_b is the annotation budget for each active learning iteration. The labeled pool is initialized as the set $\mathbf{X}_t[0] = \phi$ and in the k^{th} iteration, is updated as $\mathbf{X}_t[k] = \{\mathbf{X}_t[k-1] \cup \mathbf{S}_t^{n_b}\}$. In the first iteration, we initialize the model, $\mathbf{M}(\theta)$ with the warm-up weights from [39], and fine-tune it using the fully annotated frames from $\mathbf{X}_t[1]$. For every subsequent AL iteration, i.e., $k = 2, 3, \dots$, Δ selects a fresh subset $\mathbf{S}_t^{n_b}$ from $\{\mathcal{X}_t^{n_t} \setminus \mathbf{X}_t[k]\}$, where \setminus denotes the set difference operation. Our proposed class selection methods aim to select the most diverse and informative classes in each image $\mathcal{I} \in \mathbf{S}_t^{n_b}$, given the model trained on the most recent labeled set $\mathbf{X}_t[k]$.

Given $\mathbf{M}(\theta)$ and image \mathcal{I} , we extract class specific confusion \mathbf{P}^c for class $c \in C$ as proposed by [1].

$$\mathbf{P}^c(\mathcal{I}, \mathbf{M}(\theta)) = \frac{1}{|N_c|} \sum_{i \in N_c} \left[\frac{w_i \times \mathbf{p}_i[\hat{y}|\mathcal{I}; \mathbf{M}(\theta)]}{\sum_{i \in N_c} w_i} \right] \quad (1)$$

where N_c is the set of pixels that have been classified as class c and,

$$w_i = - \sum_{c \in C} \mathbf{p}_i[\hat{y}|\mathcal{I}; \mathbf{M}(\theta)] \log_2 \mathbf{p}_i[\hat{y}|\mathcal{I}; \mathbf{M}(\theta)] \quad (2)$$

is the Shannon's entropy for the i^{th} pixel, and \mathbf{p}_i is the softmax probability vector as predicted by the model $\mathbf{M}(\theta)$, and \hat{y} is the random variable corresponding to the predicted class for a pixel i in image \mathcal{I} . The class-specific confusion vector $\mathbf{P}^c(\mathcal{I}, \mathbf{M}(\theta))$ was introduced in [1] as an entropy weighted mixture of softmax probabilities of the pixels predicted as class c . It can be interpreted as a weighted average of *soft pseudo-labels*, where more uncertain pixels are assigned larger weights and highly confident pixels carry minimal weights. This weighted averaging of softmax probabilities results in a probability mass function that amplifies the probabilities of classes competing with class c , thus effectively capturing class confusion. Now we discuss our Anchor-based and Aug-based techniques, followed by IoU-based and training objective.

3.2. Anchor-Based Class Selection

Based on the observation that in the feature vectors corresponding to pixels of the same class belong to same cluster [46], we compute class representative *anchors* at each AL iteration using the labeled target data annotated thus far. These class anchors capture the class-specific confusion in the model predictions, which in turn helps in selecting the most informative classes in the selected pool $\mathbf{S}_t^{n_b}$ which when included in the labeled set helps the model improve the overall class representation.

For computing the class representative anchors in the k^{th} AL iteration, we construct a set of feature vectors \mathbf{Z}^c using eq. (3), which stores the class-specific confusion for a class c in the labeled pool $\mathbf{X}_t[k]$

$$\mathbf{Z}^c = \bigcup_{x \in \mathbf{X}_t[k]} \mathbf{P}^c(x, \mathbf{M}(\theta)) \quad (3)$$

Once we have \mathbf{Z}^c , we compute Λ_t^c , a c^{th} class representative anchor, by computing average over \mathbf{Z}^c

$$\Lambda_t^c = \text{average}(\mathbf{Z}^c). \quad (4)$$

Each target class anchor Λ_t^c serves as a representative of the c^{th} class in the feature space. Now, in order to identify classes that are informative in the sense of class confusion in each unlabeled image \mathcal{I}

$$D_{(\mathcal{I})}^c = \|\Lambda_t^c - \mathbf{P}^c(\mathcal{I}, \mathbf{M}(\theta))\|_2; \mathcal{I} \in \mathbf{S}_t^{n_b}, c \in C \quad (5)$$

where, $\|\cdot\|_2$ denotes the L_2 norm of a vector. $D_{(\mathcal{I})}$ stores the disparity between the class anchors Λ_t^c and class specific confusion for image \mathcal{I} . We say that a class c in \mathcal{I} is selected to be labeled if $D_{\mathcal{I}}^c > \delta$, where δ is a threshold set to 0.5.

Algorithm 1 Algorithm for class selection

Given: Segmentation Model $\mathbf{M}(\theta)$, unlabeled target domain $\mathcal{X}_t^{n_t}$, frame selection function Δ , budget n_b , $\mathbf{X}_t[0] = \phi$,

Stage 1 (Active Labeling):

- 1: Warmup $\mathbf{M}(\theta)$ with adversarial UDA [39] weights
 - 2: $\mathbf{X}_t[1]^{n_b} = \Delta(\mathcal{X}_t, n_b)$
 - 3: Fine-tune $\mathbf{M}(\theta)$ using $\mathbf{X}_t[1]^{n_b}$
 - 4: for $k = 2:n$
 - 5: $\mathbf{S}_t^{n_b} = \Delta(\mathcal{X}_t^{n_t} \setminus \mathbf{X}_t[k-1])$;
 - 6: Selecting class to be labeled for an image $\mathcal{I} \in \mathbf{S}_t^{n_b}$
 - 7: if class selection == Anchor-based:
 - 8: $\mathbf{Z}^c = \bigcup_{x \in \mathbf{X}_t[k]} \mathbf{P}^c(x, \mathbf{M}(\theta))$
 - 9: $\Lambda_t^c = \text{average}(\mathbf{Z}^c)$.
 - 10: $D_{(\mathcal{I})}^c = \|\Lambda_t^c - \mathbf{P}^c(\mathcal{I}, \mathbf{M}(\theta))\|_2$;
 - 11: if class selection == Aug-based:
 - 12: $D_{(\mathcal{I})}^c = \|\mathbf{P}^c(\mathcal{I}_w, \mathbf{M}(\theta)) - \mathbf{P}^c(\mathcal{I}_s, \mathbf{M}(\theta))\|_2$
 - 13: if class selection == IoU-based:
 - 14: $D_{(\mathcal{I})}^c = (\text{IoU}_{\mathcal{I}}^c(\mathbf{M}(\theta)_{sup}) - \text{IoU}_{\mathcal{I}}^c(\mathbf{M}(\theta)))$
 - 15: A class c in \mathcal{I} is labeled if $D_{\mathcal{I}}^c > \delta(0.5)$
 - 16: $\mathcal{L}_{CE} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \sum_{c=1}^C \hat{Y}^c \log \mathbf{p}^c[i]$
 - 17: **Stage 2 (Self-training):**
 - 18: $\mathcal{L}_{pseudo} = \mathcal{L}_{CE}(\{\mathcal{X}_t^{n_t} \setminus \mathbf{X}_t[k]\}, \hat{y}_t)$
 - 19: $\mathcal{L}_{seg} = \mathcal{L}_{CE}(\mathbf{X}_t[k], y_t) + \mathcal{L}_{pseudo}$
-

3.3. Aug-Based Class Selection

In the proposed augmentation based method, the core idea is to capture the disagreement in the class confusion from model predictions over strong and weakly augmented data. Strong augmentations are intended to make the predictions harder, which helps in identifying classes which are more difficult for the model to learn. We use strong and weak augmentations at test time with the same model $\mathbf{M}(\theta)$. For weak augmentations, we use transforms like $\{\text{hflip}(0.5)\}$, while we use $\{\text{brightness}(0.3), \text{saturation}(0.1), \text{contrast}(0.3), \text{hflip}(0.5), \text{rotate}(0.2)\}$ for strong augmentations.

For an image $\mathcal{I} \in \mathbf{S}_t^{n_b}$, we extract class confusion using eq. (1), and compute the disparity amongst class confusion from weak and strong augmented models as follows

$$D_{(\mathcal{I})}^c = \|\mathbf{P}^c(\mathcal{I}_w, \mathbf{M}(\theta)) - \mathbf{P}^c(\mathcal{I}_s, \mathbf{M}(\theta))\|_2; c \in C \quad (6)$$

where \mathcal{I}_w and \mathcal{I}_s denote the weakly and strongly augmented versions of the image \mathcal{I} , respectively. A class c in \mathcal{I} is selected to be labeled if $D_{\mathcal{I}}^c > \delta$, where δ is a threshold set to 0.5.

3.4. IoU-Based Class Selection

We use IoU-based class selection as a setting to establish a skyline performance. Here, we assume that we



Figure 3. Representative RGB images with ground truth mask from the three datasets used: Cityscapes (first row), GTA5 (second row), Synthia (third row)

Table 1. Comparison with state-of-the-art DA techniques on GTA5→Cityscapes. Number in bracket represents % of annotated data.

		GTA5 → Cityscapes																		
Method	Road	Sidewalk	Building	Wall	Fence	Pole	T.light	T.sign	Vege	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	Motorbike	Bicycle	mIoU
AdaptNet[39]	86.5	36	79.9	23.4	23.3	23.9	35.2	14.8	83.4	33.3	75.6	58.5	37.6	73.7	32.5	35.4	3.9	30.1	28.1	42.4
AdvEnt[42]	89.4	33.1	81	26.6	26.8	27.2	33.5	24.7	83.9	34.7	78.8	58.7	30.5	84.8	38.5	44.5	1.7	31.6	32.4	45.5
CBST[48]	91.8	53.5	80.5	32.7	21.0	34.0	28.9	20.4	83.9	34.2	80.9	53.1	24.0	82.7	30.3	35.9	16.0	25.9	42.8	45.9
PyCDA[14]	90.5	36.3	84.4	32.4	28.7	34.6	36.4	31.5	86.8	37.9	78.5	62.3	21.5	85.6	27.9	34.8	18.0	22.9	49.3	47.4
IAST[20]	94.1	58.8	85.4	39.7	29.2	25.1	43.1	34.2	84.8	34.6	88.7	62.7	30.3	87.6	42.3	50.3	24.7	35.2	40.2	52.2
WDA[24]	94.0	62.7	86.3	36.5	32.8	38.4	44.9	51.0	86.1	43.4	87.7	66.4	36.5	87.9	44.1	58.8	23.2	35.6	55.9	56.4
ProDA[45]	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
CAG[46]	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[36] (5%)	92.2	59.9	87.3	36.4	45.7	46.1	50.6	59.5	88.3	44.0	90.2	59.7	38.2	90.0	55.3	45.1	32.0	32.6	52.9	49.3
MADA[22] (5%)	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
Aug-based (4.8%)	97.5	76.9	90.2	46.5	48.9	47.9	55.9	61.2	90.3	52.9	93.2	71.6	40.6	92.1	62.9	60.9	52.6	47.5	67.9	66.2
Anchor-based (4.7%)	96.9	76.6	88.8	47.9	49.0	47.1	55.2	61.6	89.8	55.3	92.9	70.9	39.3	92.0	63.3	63.7	58.6	48.7	66.7	66.6
IoU-based (4.5%)	96.9	78.1	89.4	47.6	50.0	49.1	55.8	61.4	89.9	54.4	93.6	70.9	47.0	91.6	70.6	67.9	49.6	47.8	67.2	67.4
Supervised	97.3	80.6	90.1	53.2	54.8	51.0	55.4	64.0	90.5	55.1	93.3	74.3	51.0	92.7	75.7	76.5	55.2	42.9	68.0	69.6

have a model $M(\theta)_{sup}$ trained on the entire labeled target domain. The class selection is now based on the discrepancy between the class-wise IoU score between $M(\theta)$ (which is trained on $\mathbf{X}_t[k]$ in the k^{th} AL iteration) and that of $M(\theta)_{sup}$. Bigger the gap between the two class-wise IoU scores, harder is the class. For an image $\mathcal{I} \in \mathbf{S}_t^{nb}$, we construct $D_{\mathcal{I}}^c$, measuring the difference in IoU score for a class c when predicted using $M(\theta)$ and $M(\theta)_{sup}$

$$D_{\mathcal{I}}^c = (IoU_{\mathcal{I}}^c(M(\theta)_{sup}) - IoU_{\mathcal{I}}^c(M(\theta))); c \in C \quad (7)$$

A class c in \mathcal{I} is selected to be labeled if $D_{\mathcal{I}}^c > \delta$, where δ is a threshold set to 0.5.

3.5. Training Objective

Using all the actively labeled data, either by Anchor-based or Aug-based in the target domain, we can fine-tune the network to learn exclusive target domain information. Similar to MADA [22], our training process comprises of two stages in each AL iteration. In the first stage, we use the standard cross entropy (CE) loss to train the network over the labeled data. To further exploit the available unlabeled data, in the second stage we use a self-training approach using the *pseudo-labels*

obtained from the model trained in the first stage, such that $\hat{y}_t = \arg \max \mathbf{p}^c$ for the remaining unlabeled samples

$$\mathcal{L}_{pseudo} = \mathcal{L}_{CE}(\{\mathcal{X}_t^{n_t} \setminus \mathbf{X}_t[k]\}, \hat{y}_t) \quad (8)$$

Thus the overall loss function for segmentation model is given as

$$\mathcal{L}_{seg} = \mathcal{L}_{CE}(\mathbf{X}_t[k], y_t) + \mathcal{L}_{pseudo} \quad (9)$$

The overall training pipeline is summarized in Algo.1.

4. Dataset and Evaluation

Dataset: For evaluation we use two common “*synthetic-2-real*” segmentation setups as used in the contemporary SOTA approaches [22, 33], namely GTA5→Cityscapes and Synthia→Cityscapes. GTA5[28] contains 24966 (1914x1052) images, sharing 19 classes with Cityscapes [6]. Synthia[29] contains 9400 (1280x760) images, sharing 16 classes. Cityscapes includes high resolution real world images of 2048x1024, with a split of 2975 training and 500 validation images. Fig. 3 shows samples from the three datasets used to

Table 2. Comparison with state-of-the-art DA techniques on Synthia→Cityscapes. Number in bracket represents % of annotated data.

Method	Road	Sidewalk	Building	Wall*	Fence*	Pole*	T.light	T.sign	Vege	Sky	Person	Rider	Car	Bus	Motorbike	Bicycle	mIoU	mIoU*
AdaptNet[39]	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
AdvEnt[42]	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
CBST[48]	68.0	29.9	76.3	10.8	1.4	33.9	22.8	29.5	77.6	78.3	60.6	28.3	81.6	23.5	18.8	39.8	42.6	48.9
PyCDA[14]	75.5	30.9	83.3	20.8	0.7	32.7	27.3	33.5	84.7	85.0	64.1	25.4	85.0	45.2	21.2	32.0	46.7	53.3
IAST[20]	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
ProDA[45]	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
WDA[24]	94.9	63.2	85.0	27.3	24.2	34.9	37.3	50.8	84.4	88.2	60.6	36.3	86.4	43.2	36.5	61.3	57.2	63.7
CAG[46]	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[36] (5%)	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[22] (5%)	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
Aug Based (4.9%)	97.1	78.8	90.9	48.4	45.7	48.9	50.4	65.5	90.7	93.2	75.9	49.9	92.7	69.9	52.9	71.4	70.1	75.3
Anchor Based (4.7%)	97.2	79.6	90.5	45.5	50.8	48.7	55.4	67.1	90.2	93.2	76.1	53.2	90.1	73.3	53.9	70.4	70.9	76.1
IoU Based (5.1%)	97.7	80.9	90.8	49.1	56.1	52.3	59.1	68.5	90.8	93.1	75.8	54.1	93.1	78.4	56.7	71.1	72.9	77.7
Fully Supervised	97.6	81.3	91.1	49.8	57.6	53.8	59.6	69.1	91.2	94.4	76.7	55.6	93.3	79.9	57.7	72.2	73.8	78.4

illustrate the significant distribution shift between the datasets.

Implementation Details¹: We have followed the experimental setup of MADA [22], and have used DeepLabV3+ [3] with a ResNet-101 backbone for fair comparison. We have initialized the model with warm-up weights from AdaptNet [39], an adversarial unsupervised domain adaptation framework. For training we have used 50 epochs with a batch size of 4 across all the experiments. For evaluation we have used mIoU as a metric to measure model performance on Cityscapes validation set. We also report *error margin* for various techniques defined as the difference between the particular ADA approach (at a certain annotation budget), and a fully supervised technique using the same backbone.

5. Experiments and Results

Quantitative Results: We first show quantitative results on the two settings, GTA5→Cityscapes and Synthia→Cityscapes, in Tables 1 and 2 respectively. We compare our results with existing UDA [39, 42, 48, 45, 20], semi-supervised [46], weakly supervised [24] and frame based ADA [36, 22] techniques. We observe that both of our proposed approaches, *aug-based* and *anchor-based*, surpass the SOTA techniques, reducing the error margin from 4.7 to 3 in GTA5→Cityscapes and from 5.7 to 3.2 in Synthia→Cityscapes using merely 5% of the annotated data when compared to a fully supervised model.

Effectiveness of Proposed AL Strategies: In Table 3 we break down the results of the two stages used in our pipeline: active learning, and pseudo labeling, and re-

Table 3. Ablation study on GTA5→Cityscapes and Synthia→Cityscapes, after training using AL as well as pseudo-labels.

Method	Active Labels	Pseudo Labels	G→C	S→C
MADA [22]	✓		61.6	65.0
MADA [22]	✓	✓	64.1	68.1
Aug-Based	✓		65.5	75.1
Aug-Based	✓	✓	66.2	75.9
Anchor-Based	✓		66.1	76.0
Anchor-Based	✓	✓	66.6	76.5

port mIoU at each stage for both GTA5→Cityscapes and Synthia→Cityscapes setups. The purpose is to highlight the significant improvement of our active learning strategy over the MADA (row 1,3,5). In GTA5→Cityscapes we observe an mIoU improvement of 4.1 (61.6 to 66.1) and 3.9 (61.6 to 65.5) for *anchor-based* and *aug-based* approaches respectively. In future, we wish to work on improving our stage-2 performance by effectively using the pseudo labels complementary to our labeled samples.

Discussion: (1) It is noteworthy that our proposed approach helps in increasing the IoU for most of the class labels. We also wish to highlight the simplicity of our approach in comparison to MADA [22] which selects target samples based on source anchors and multiple add-on components in the training process such as soft-anchor alignment loss, and updating target anchors with EMA. **(2)** We observe reduction in annotation cost for both the proposed approaches, but *anchor-based* performs better. We speculate this is due to inability of *aug-based* approach in capturing contextual diversity. This is inline with the earlier works [1] emphasizing the role of contextual diversity in AL.

Result Visualization: In Fig. 4 we visualize the output gen-

¹Code:https://github.com/sharat29ag/contextual_class

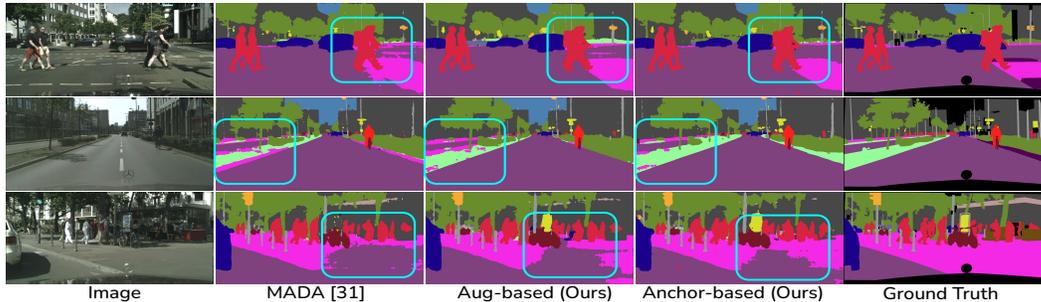


Figure 4. Qualitative results on Cityscapes Validation set after Domain Adaptation from GTA5→Cityscapes. We compare our results of Augmentation-based and Anchor-Based with MADA [22]. We can clearly see the improvement in the highlighted regions of each image.

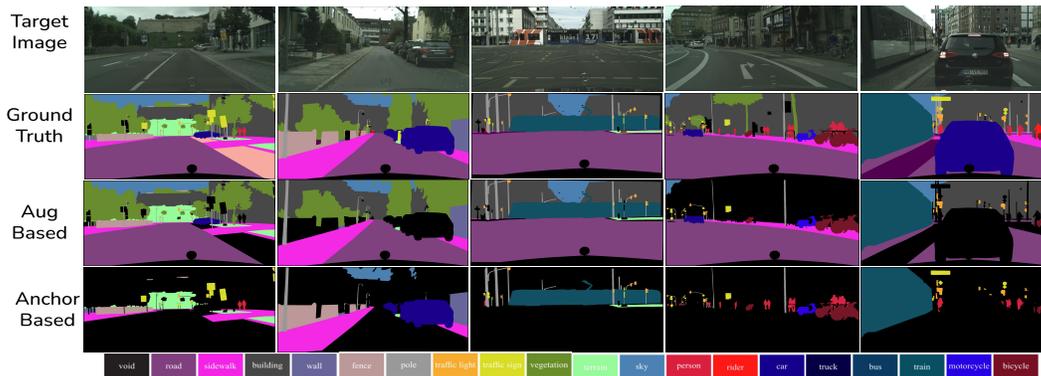


Figure 5. Qualitative samples of selecting classes using Aug-based (row 2) and Anchor-based (row 3) approaches. Both the approaches reduce the annotation cost by selecting contextually diverse and informative classes. Notice, how the road regions frequently observed in the target domain remains unlabeled in the frames selected by Anchor-based approach, but are labeled in Aug-based approaches due to the low confidence of the model.

erated by our technique and compare it with the results of MADA. We can see in the highlighted regions that the predicted masks are more accurate and smooth in the confusing regions.

Selection Visualization: In Fig. 5 we show the selections made by two of our approaches (black color shows the non-selected regions). Notice, how the road regions frequently observed in the target domain remains unlabeled in the frames selected by Anchor-based approach, but are labeled in Aug-based approaches due to the low confidence of the model because of appearance differences from the target. On the other hand, the label *fence* is less frequent in the target domain, and hence is selected by the Anchor-based approach.

Ablation Study for Effect of Frame Selection Strategy: As stated in Section 3.5, we have used CDAL [1] as our base frame selection technique over which we reduce the annotation effort by selecting informative classes. To understand the impact of frame selection strategy, we replace CDAL [1] with other frame selection techniques, such as Core-Set [31], MADA [22] and Random selection. Table 4 shows the results. We observe a significant improve-

ment in performance using both our approaches for each of the frame selection strategies.

Method	Frame Based		Aug-based		Anchor-based	
	Data(%)	mIoU	Data(%)	mIoU	Data(%)	mIoU
Random	3.4	57.8	2.8	60.2	2.1	60.5
CoreSet [31]	3.4	58.8	2.9	60.5	2.4	61.2
MADA [22]	3.4	60.3	2.6	62.5	2.3	62.9
CDAL [1]	3.4	62.04	2.8	63.8	2.2	63.5

ment in performance using both our approaches for each of the frame selection strategies.

Improvements Obtained at Various Annotation Budgets: We measure the impact of using various AL strategies at various annotation budget levels for the ADA problem on the GTA5→Cityscapes experimental setup. The challenge for each technique is to reach the performance supervised model, 69.6, with minimum annotation budget. For frame based techniques we increase the budget of 50 frames (1.7%) at each AL step. Similarly, for our approaches we select 50 frames at each active cycle using CDAL and annotate classes either using one anchor-based,

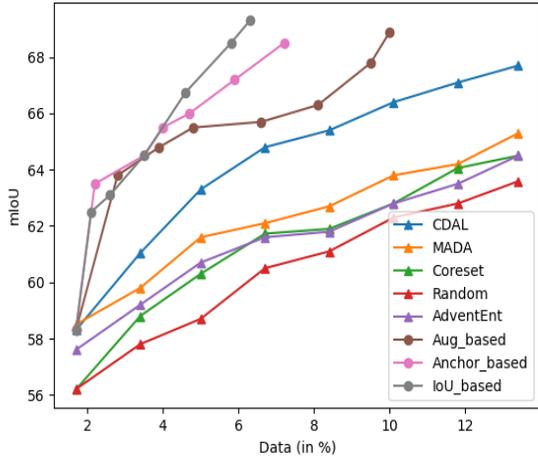


Figure 6. Comparison of state-of-the-art ADA, and AL techniques at different annotation budgets. We use IoU -based as a skyline since it uses whole supervised information. We observe significant improvement in mIoU using our techniques for all annotation budgets.

`aug-based` or `IoU-based` techniques. We note that it is difficult to control the exact annotation budget at each step in our approaches as we are annotating certain selected classes entirely. We now briefly discuss the baselines used in this experiment:

1. `Random-sampling`: For each active learning budgets, samples are randomly selected from the unlabeled pool.
2. `Coreset` [31]: A subset selection approach, using K-center greedy algorithm for selecting diverse samples.
3. `AdvEnt` [42]: Samples were selected using the entropy maps of each samples predicted using [42] in the target domain.
4. `CDAL` [1]: Selects contextually diverse samples exploiting the contextual information among the frames.
5. `MADA` [22]: Selects samples complementary to the source anchors.

Fig. 6 shows the results. Both of our proposed approaches, `aug-based` and `anchor-based`, surpass the baselines with a significant margin. We also note that we are very close to `IoU-based` selection at low annotation budgets. We also observe that using only 10% annotation our `anchor-based` approach is able to achieve the performance of full supervised moodel.

Source Free Domain Adaptation: To show the extended utility of our approach beyond ADA, we also compare our `Anchor-based` and `Aug-based` approaches in a Source-Free Domain Adaptation (SFDA) setting. In SFDA,

Table 5. Results on the Source Free domain adaptation setting, where source data is unavailable (due to privacy or other such constraints), but we have a model trained on the source data. Note that existing ADA approaches like `MADA` [22] fail in this setting due to critical dependency on the source data.

Method	Data(%)	mIoU
URMA [35]	-	45.1
SFDA [15]	-	43.16
SRDA [2]	-	45.8
SFDASS [12]	-	53.4
<code>Aug-based</code>	3.3	60.9
<code>Anchor-Based</code>	3.0	61.7

the source dataset is unavailable due to privacy issues, but we have a segmentation model trained on the source dataset. The existing ADA technique, `MADA` [22] fails to adapt in the SFDA setting due to its dependency on the source data for computing source anchors to select the samples from target dataset. Table 5 shows the results comparing with state-of-the-art SFDA approaches like [35, 15, 2, 12]. We observe an improvement of 7.5 and 8.3 using our two approaches, over the baseline of 53.4, using only 3.3% and 3.0% of the annotated data.

6. Discussion

In this paper, we have proposed a novel and intuitive approach to reduce annotation effort by labeling certain informative classes in a frame instead of wasting the annotation budget by labeling redundant regions. Through extensive experiments and comparison with different ADA and Active Learning baselines, we highlight the improvement in performance using both proposed `Aug-based` and `Anchor-based` approaches. We also validate that our approaches can be used as a decorator for any frame-based active learning approaches, which helps reduce annotation cost and increases the model performance beyond the existing state-of-the-art. While our work can effectively utilize a given annotation budget by selecting most informative samples, we have used some off-the-shelf techniques for generating pseudo-labels. In future, we would like to explore generating pseudo labels complementary to the labeled data from the target domain, as well as exploit label distribution and extracting useful information from the abundantly available source labeled data.

Acknowledgement: S. Agarwal was partially supported by the Infosys Center for Artificial Intelligence at IIT-Delhi. S. Anand was partially supported by MeitY, Govt. of India (No. 25(13)/2019-ESDA) and C. Arora acknowledges support received from DST through the IMPRINT program (IMP/2019/000250).

References

- [1] Sharat Agarwal, Himanshu Arora, Saket Anand, and Chetan Arora. Contextual diversity for active learning. In *European Conference on Computer Vision*, pages 137–153. Springer, 2020.
- [2] Mathilde Bateson, Hoel Kervadec, Jose Dolz, Hervé Lombaert, and Ismail Ben Ayed. Source-relaxed domain adaptation for image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 490–499. Springer, 2020.
- [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] Shuaijun Chen, Xu Jia, Jianzhong He, Yongjie Shi, and Jianzhuang Liu. Semi-supervised domain adaptation based on dual-level domain mixing for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11018–11027, 2021.
- [5] 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.
- [6] 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.
- [7] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.
- [8] 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.
- [9] Weixiang Hong, Zhenzhen Wang, Ming Yang, and Junsong Yuan. Conditional generative adversarial network for structured domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1335–1344, 2018.
- [10] Ross D King, Kenneth E Whelan, Ffion M Jones, Philip GK Reiser, Christopher H Bryant, Stephen H Muggleton, Douglas B Kell, and Stephen G Oliver. Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature*, 427(6971):247–252, 2004.
- [11] Vikram Krishnamurthy. Algorithms for optimal scheduling and management of hidden markov model sensors. *IEEE Transactions on Signal Processing*, 50(6):1382–1397, 2002.
- [12] Jogendra Nath Kundu, Akshay Kulkarni, Amit Singh, Varun Jampani, and R Venkatesh Babu. Generalize then adapt: Source-free domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7046–7056, 2021.
- [13] Peilun Li, Xiaodan Liang, Daoyuan Jia, and Eric P Xing. Semantic-aware grad-gan for virtual-to-real urban scene adaption. *arXiv preprint arXiv:1801.01726*, 2018.
- [14] Qing Lian, Fengmao Lv, Lixin Duan, and Boqing Gong. Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6758–6767, 2019.
- [15] Yuang Liu, Wei Zhang, and Jun Wang. Source-free domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1215–1224, 2021.
- [16] 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.
- [17] 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.
- [18] Tyler Tian Lu. Fundamental limitations of semi-supervised learning. Master’s thesis, University of Waterloo, 2009.
- [19] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 2507–2516, 2019.
- [20] Ke Mei, Chuang Zhu, Jiaqi Zou, and Shanghang Zhang. Instance adaptive self-training for unsupervised domain adaptation. In *European conference on computer vision*, pages 415–430. Springer, 2020.
- [21] Ke Mei, Chuang Zhu, Jiaqi Zou, and Shanghang Zhang. Instance adaptive self-training for unsupervised domain adaptation. In *European conference on computer vision*, pages 415–430. Springer, 2020.
- [22] Munan Ning, Donghuan Lu, Dong Wei, Cheng Bian, Chenglang Yuan, Shuang Yu, Kai Ma, and Yefeng Zheng. Multi-anchor active domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9112–9122, 2021.
- [23] Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3764–3773, 2020.
- [24] Sujoy Paul, Yi-Hsuan Tsai, Samuel Schuster, Amit K Roy-Chowdhury, and Manmohan Chandraker. Domain adaptive semantic segmentation using weak labels. In *European conference on computer vision*, pages 571–587. Springer, 2020.
- [25] Viraj Prabhu, Arjun Chandrasekaran, Kate Saenko, and Judy Hoffman. Active domain adaptation via clustering uncertainty-weighted embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8505–8514, 2021.
- [26] Harsh Rangwani, Arihant Jain, Sumukh K Aithal, and R Venkatesh Babu. S3vaada: Submodular subset selection for virtual adversarial active domain adaptation. In *Proceed-*

- ings of the *IEEE/CVF International Conference on Computer Vision*, pages 7516–7525, 2021.
- [27] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM Computing Surveys (CSUR)*, 54(9):1–40, 2021.
- [28] 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.
- [29] 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.
- [30] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8050–8058, 2019.
- [31] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018.
- [32] Burr Settles. Active learning literature survey. 2009.
- [33] Inkyu Shin, Dong-Jin Kim, Jae Won Cho, Sanghyun Woo, Kwanyong Park, and In So Kweon. Labor: Labeling only if required for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8588–8598, 2021.
- [34] Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5972–5981, 2019.
- [35] Prabhu Teja Sivaprasad and Francois Fleuret. Uncertainty reduction for model adaptation in semantic segmentation. In *2021 IEEE/CVF Conference On Computer Vision And Pattern Recognition, Cvpr 2021*, number CONF, pages 9608–9618. IEEE, 2021.
- [36] Jong-Chyi Su, Yi-Hsuan Tsai, Kihyuk Sohn, Buyu Liu, Subhransu Maji, and Manmohan Chandraker. Active adversarial domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 739–748, 2020.
- [37] Yuxing Tang, Josiah Wang, Boyang Gao, Emmanuel Delandrea, Robert Gaizauskas, and Liming Chen. Large scale semi-supervised object detection using visual and semantic knowledge transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2119–2128, 2016.
- [38] Marco Toldo, Umberto Michieli, Gianluca Agresti, and Pietro Zanuttigh. Unsupervised domain adaptation for mobile semantic segmentation based on cycle consistency and feature alignment. *Image and Vision Computing*, 95:103889, 2020.
- [39] 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.
- [40] Yi-Hsuan Tsai, Kihyuk Sohn, Samuel Schulter, and Manmohan Chandraker. Domain adaptation for structured output via discriminative patch representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1456–1465, 2019.
- [41] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.
- [42] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 2517–2526, 2019.
- [43] Zhonghao Wang, Yunchao Wei, Rogerio Feris, Jinjun Xiong, Wen-Mei Hwu, Thomas S Huang, and Honghui Shi. Alleviating semantic-level shift: A semi-supervised domain adaptation method for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 936–937, 2020.
- [44] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12414–12424, 2021.
- [45] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12414–12424, 2021.
- [46] Qiming Zhang, Jing Zhang, Wei Liu, and Dacheng Tao. Category anchor-guided unsupervised domain adaptation for semantic segmentation. *Advances in Neural Information Processing Systems*, 32, 2019.
- [47] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018.
- [48] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018.