UNIVERSITY OF CALIFORNIA
RIVERSIDE

Active Learning Guided Source-Free Domain Adaptive Semantic Segmentation and Applications in the Environmental Space

A Thesis submitted in partial satisfaction
of the requirements for the degree of

Master of Science

in

Computer Science

by

Sujith Kumar Sashikanth

March 2023

Thesis Committee:

Dr. Amit K. Roy-Chowdhury, Chairperson
Dr. Ahmed Eldawy
Dr. Konstantinos Karydis

The Thesis of Sujith Kumar Sashikanth is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

I am grateful to my advisor Professor Amit K. Roy-Chowdhury, without whose help, I would not have been here.

To my parents and brother for all the support.

ABSTRACT OF THE THESIS

Active Learning Guided Source-Free Domain Adaptive Semantic Segmentation and Applications in the Environmental Space

by

Sujith Kumar Sashikanth

Master of Science, Graduate Program in Computer Science
University of California, Riverside, March 2023
Dr. Amit K. Roy-Chowdhury, Chairperson

In most domain adaptation methods, concurrent access to source and target data is needed. In many real-life problem statements, it's highly likely that the access to the source data might not be possible during the adaptation process. Source-free domain adaptation methods, while very necessary, usually have significant performance gap when compared with those that assume access to the source data. To tackle this, we introduce a source free domain adaptation strategy for the semantic segmentation problem that internally uses active learning on the pixel level predictions (i.e., pseudo labels) to obtain a framework that can generalize well between the source and the target domain. Given a small budget for labeling, we select the pixels needed to be sent to an oracle for labeling. These labeled pixels are then used to boost performance of the the overall system. Our method does not assume any source data or any prior labels available on the target data. We observe that even without the source data and only by labeling a small percentage of pixels using active learning, we can still reach comparable performance with many standard unsupervised domain adaptive semantic segmentation methods. We also show a few brief example appli-

cations of domain adaptation on the environmental scenario. We apply domain adaptation to two different environmental applications, namely forest fire detection and also vegetation segmentation. In the forest fire detection project we localize the fires by predicting bounding boxes over them in the form of coordinates. In the vegetation segmentation project, we classify every pixel of the image into seven different classes. The overall domain adaptation operation is achieved through a simple self-learning approach.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Semantic segmentation involves the classification of every pixel in an image to the available semantic classes. It has numerous applications in autonomous driving[8, 53], medical imaging[72, 11], and robotics [35, 36]. Much of the success of existing semantic segmentation methods is owed to large-scale supervised training, which requires cumbersome labeling of each pixel of every image in large datasets [14]. However, in real-world scenarios, annotation efforts may be constrained by a pre-defined budget, and inference might need to be performed on images that incur a heavy domain shift with those used during training [27, 17]. Such domain shift is mainly due to the changes in the image environment and conditions leading to a shift in the data distribution [33, 48, 57, 75]. An example of domain shift is shown in Fig 1.1. If not addressed, domain shifts will result in a catastrophic drop in the performance of a semantic segmentation model. To address this issue researchers have proposed numerous domain adaptation approaches to make semantic segmentation models more practically applicable [42, 60, 59].

Figure 1.1: **Figure depicting the visual domain gap between the GTAV(Source) [45] and Cityscapes(Target) dataset [13]**. GTAV is a dataset of synthetic images and Cityscapes reflect real-world imagery. A model trained on synthetic images will fail to capture the semantic concept in real-world urban imagery.

A fully supervised domain adaptation (FSDA) task involves transferring knowledge learned from a labeled source domain to a labeled target domain. However, obtaining fully-annotated training data for any new domain is expensive. Additionally, concurrent access to source and target domain data does not reflect real-world scenarios where data privacy might deter a vendor from sharing source-side data with a client [25]. To address the issue of annotation unavailability, unsupervised domain adaptation (UDA) has been proposed in the literature [48, 42, 26, 47]. UDA approaches rely on pseudo labels based self-training principles to align source and target domains where the latter has no annotations [18, 62, 26, 59]. On the other hand, to preserve data privacy and make UDA semantic segmentation more practically viable, researchers have started to address UDA *without access to the source data*, formally known as source-free UDA (SF-UDA) [25, 67, 47, 29]. In such cases, only the source-trained models, which embed information about the source

dataset, are available to share.

UDA methods usually suffer significant performance drops compared to FSDA methods. Although self-training approaches for standard UDA have resulted in considerable performance gains over recent years, the noisy nature of pseudo labels acts as a bottleneck for such approaches to reach near FSDA performance. This is mainly due to the imbalance in class distributions whereby common entities, like "road" and "building", appear more frequently than others such as "rider" and "light", thereby making self-training methods heavily emphasize the high-frequency classes as opposed to the rare ones [67]. This issue is exacerbated for SF-UDA methods [25], where the unavailability of source domain data renders any domain-alignment regularization inapplicable [71, 67, 63, 66].

In recent years researchers have attempted to address the issue of self-training bias for standard UDA-based semantic segmentation by introducing human-in-the-loop approaches. These approaches focus on using a human annotator to label the smallest but most informative subset of the target domain training data, which suffices in boosting UDA performance to near full-supervision levels [67, 40, 68, 74]. The selection of the most informative subset of data during training cycles is formally known as active learning [30, 7, 32]. For semantic segmentation, the selection of the most informative pixels in each image [67, 50] is far more efficient than selective entire images to label [30, 74, 68]. However, the use of active learning for improving the performance of SF-UDA is unexplored. In this paper, we aim to address source-free UDA-based semantic segmentation with a human-in-the-loop active learning approach, which results in a more practically applicable model that adheres to pre-defined pixel-level annotation budgets while preserving data privacy.

Our method, **A**ctive **S**ource-free **D**omain **A**daptative semantic segmentation or **ASDA**, starts with a model pre-trained on the source domain. We utilize this model to perform self-training on a set of unlabeled target domain training images. During training, after each active learning cycle (generally composed of a few epochs) we query a human to select a budgeted set of informative pixels of the target side images based on a *selection criteria*. In the active learning method, the selection criteria chooses samples that the model is most confused about, thus harboring more information than the remaining training set. In ASDA, we utilize the predictive uncertainty associated with the pixel-level pseudo labels of the target domain images as the selection criteria. Quantifying the model's predictive uncertainty enables flagging pixels that tend to be from the rare semantic classes as the unavailability of data tends to increase model confusion [19, 21].

Additionally unlike prior active domain adaptation works [67, 41, 51] we do not assume the existence of an initial small labeled set for the target domain for the active learning to start with. This initial supervisory signal allows for improving the domain alignment especially if source data is available. In contrast, we start with a fully unlabeled set and show that we can still reach comparable performance with many standard UDA methods without the availability of source data.

The major contributions of our work are as follows:

- To the best of our knowledge, this is the first paper that incorporates a human-in-the-loop active learning approach to address SF-UDA for semantic segmentation.
- We show how predictive uncertainty guided active learning can bridge the gap between SF-UDA and standard UDA.

- Experimental results show that our method outperforms multiple prior UDA, as well as, SF-UDA based semantic segmentation methods.

1.1 Related works

Domain Adaptation

The issue of domain shift can cause problems for models trying to generalize between different data distributions. To address this, several approaches have been proposed for semantic segmentation tasks. One such approach is class-balanced self-training, which formulates a loss minimization problem for end-to-end learning of both the classifier and domain-invariant features [80]. Another approach is the dual student network, which consists of two student networks that learn from each other through mutual supervision to achieve better feature learning and domain adaptation [20]. A structured output adaptation approach [58] uses adversarial learning to minimize the discrepancy between the two domains in the output space, and a curriculum-style approach [77] learns to estimate local and global label distributions to regularize the trained semantic segmentation network. These approaches aim to improve performance and overcome the difficulty of transferring between different domains.

As discussed earlier, domain shift could lead to a huge issue in terms of the model generalizing between two different domains.[80] proposes a class-balanced self-training method for domain adaptation in semantic segmentation tasks.The Self-training approach is formulated as a loss minimization problem that enables end-to-end learning of both the classifier and domain-invariant features. By creating pseudo-labels with a balanced class distribution,

a class-balanced self-training is proposed to address the imbalance problem of transferring difficulty among classes. [20] proposes a dual student network for semi-supervised semantic segmentation that surpasses the performance of a teacher-student network by a large margin. The dual student network consists of two student networks that learn from each other through mutual supervision, which allows for better feature learning and domain adaptation. They achieve this by introducing a novel training stabilization constraint called stable sample.[58] proposes a structured output adaptation approach for semantic segmentation that learns to adapt the output space between the source and target domains by minimizing the domain gap between the two domains. They use adversarial learning to minimize the discrepancy between the two domains in the output space. They also utilize a multi-level adversarial network to enhance the outputs from the trained model post the adaptation process. [77] suggest solving this problem using a curriculum-style approach. It proposes to learn to estimate the local label distributions of the landmark superpixels in the target domain as well as the global label distributions of the images. The authors believe this approach would give them an edge over considering it as a pixel level problem statement approach. They use this information to regularize the source trained semantic segmentation network.

Source Free Domain Adaptation

As explained earlier, in practical domain adaptation scenarios, it might not always be possible to have concurrent access to source and target data during training. Examples of these scenarios include decentralized or edge computing and medical image analysis. To address such cases, source-free domain adaptation methods have been proposed in literature.

These methods rely on self-training principles to distill knowledge from a source trained model to adapt to a new domain [31, 25]. Some methods also used entropy regularization [73, 16] to improve SF-UDA performance. In [31] an attention-based generative method is proposed to capture pixel-level information. In [24], the authors use a combination of multiple approaches such as online pseudolabel generation, self-attention and curriculum learning to achieve performance in the source-free scenario. Kundu et. al. [25] extend SF-UDA for the multi-source multi-domain scenario and use a Context prior enforcing autoencoder to refine the predicted pseudo labels. However, their approach relies on the assumption that such context encoders need to be trained on the source side. Additionally, since most of these methods rely on pseudo label based training, they are unable to compensate for the noisy nature of these pseudo labels. To alleviate this issue we introduce the concept of human-in-the-loop SF-UDA.

Active Learning

The goal of active learning is to minimize labeling effort on an enormous dataset while maximizing the performance of the model. Numerous methods have been proposed to address active learning for large scale image recognition [6, 9, 39, 28, 69]. In recent years active learning strategies have been extended to semantic segmentation. Common strategies include uncertainty based sampling [79, 7, 70] and representation/diversity based sampling [64, 54, 34]. In [69], classification difficulty is used to perform, whereby the error map between the prediction and ground truth is passed to a probabilistic attention module to predict pixel wise classification difficulty. A pixel wise attention module is used to fine tune the model on region's where there is difficulty in prediction.Here a custom acquisition

function is also devised so that examples that have high semantic difficulty levels are then selected on primary basis .This approach is particularly useful in regions where the segmentation model struggles.In [9], a reinforcement learning based approach is used to actively select pixels that are to be sent to an oracle for subsequent labelling. The Reinforcement learning system is basically used as an agent, where profitable selections in terms of segmentation metric turns into a reward for the agent and selections which hinder the segmentation metric are considered as a penalty for the agent. In [39],the authors propose a multi agent reinforcement learning system, where the internal function is in principle similar to the working of [9] but there also is an online retraining phase which is used to improvise the generated predictions. The authors use this overall system and integrate this into a robotic perception phase where the agents are placed in novel environments to navigate.

Active Domain Adaptation

It combines the principles of active learning with UDA for improving domain alignment. Due to its success in image recognition [51, 7, 74], researchers have extended active learning for more complex domain adaptation tasks like semantic segmentation. In [41], the authors introduce a multi-anchor strategy to select a subset of target domain images to label. Although effective but labeling all pixels in multiple images is very inefficient. Shin et al. [51] improved upon [41] by introducing a point-based selection strategy to efficiently label the most informative pixels. In [67], a region-based active learning approach was employed to select a minimal amount of regions to preserve neighborhood contextual information. Although these methods combine UDA with active learning they still require concurrent access to *labeled source data* to boast high performance. In comparison we cast

active domain adaptation as an SF-UDA problem making it more practically viable. In [76], similar to [41] the authors use a category based anchor centroid approach to match features in the target distribution space and generate the pseudolabels accordingly. The authors also employ an anchor based pixel level loss to refine the predictions.

There have been a few traditional methods in terms of unsuperivsed domain adaptation such as , in [16] it employs a task classifier to learn task-specific features and a domain classifier to learn domain-invariant features. [21] learns to reduce the domain shift between the source and target domains by utilizing an adversarial loss. In [14], By reducing the difference between the feature distributions of the source and target domains, [14] develops a non-linear mapping between them.Combining domain-specific and task-specific losses, the model is trained. In [23] maximizing the source domain's marginal distribution and reducing the target domain's conditional distribution, [22] aligns the joint distributions of the source and target domains. Combining domain-specific and task-specific losses, the model is trained. Also, domain shifts could lead to the need of adaptation techniques to solve this by building a feature map that generalizes between the multiple domains. To solve this problem of domain shift, domain adaptation methods have been proposed [56, 72, 27].

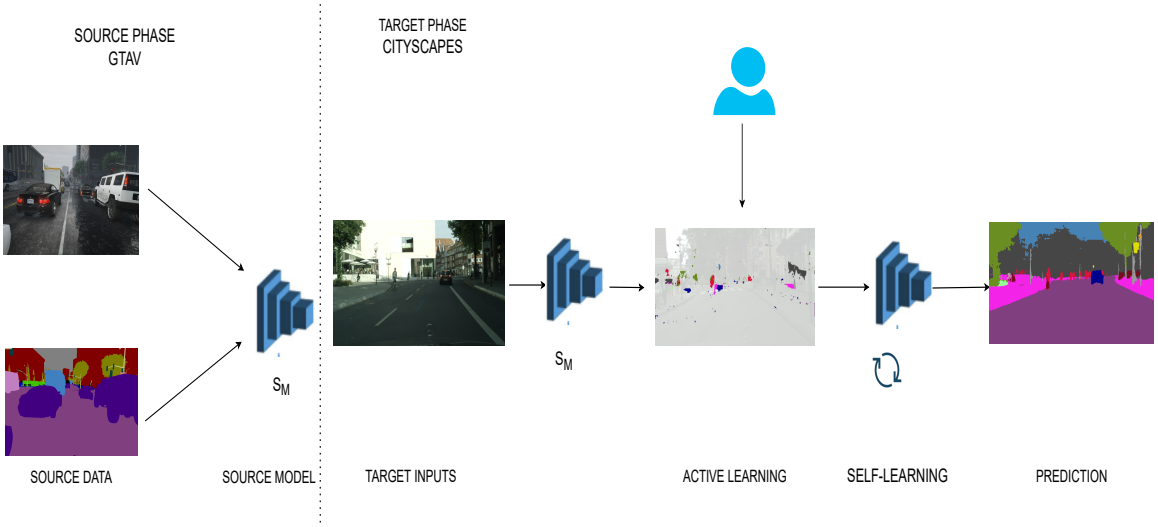


Figure 1.2: This is the overall figure that depicts the working of the ASDA system, to the left is the client side or the source side where the segmentation model is trained in conventional end-to-end manner. Following this, the source trained model is then handed over to the vendor side. The vendor obtains predictions on target images without access to the source data. A small percentage of these predicted pixels is then annotated by the oracle to provide the final prediction.

Chapter 2

Methodology and Experimental Results

2.1 Motivation and Overview

We consider two datasets from two different domains $\mathcal{D}_s = \{\mathbf{X}_s, \mathbf{Y}_s\}$ and $\mathcal{D}_t = \{\mathbf{X}_t\}$ where $\mathcal{D}_s$ is the source domain and $\mathcal{D}_t$ is the target domain. Both $\mathcal{D}_s$ and $\mathcal{D}_t$ share the same label space. In the standard UDA setup, one does not have access to the labels of $\mathbf{X}_t$ but has access to the entire labeled set $\mathcal{D}_s$. The final goal is to test on data from the target domain itself. However, in practical scenarios privacy issues can force a source side vendor to make $\mathcal{D}_s$ inaccessible to a client who aims to adapt to the target domain. In such an SF-UDA case a vendor shares a model ϕ trained on $\mathcal{D}_s$ to the client. The client aims at using only ϕ, which embeds information about $\mathcal{D}_s$, to adapt to $\mathcal{D}_t$. A common approach is to re-train ϕ on $\mathcal{D}_t$ using self-training based on pseudo labels, but as mentioned before, the noisy

nature of pseudo labels creates a bottleneck on SF-UDA performance [25, 67]. To alleviate this issue, we propose an uncertainty-guided active learning approach in our SF-UDA setup. The overview of ASDA is shown in Fig 1.2. Starting with ϕ, we start self-training on $\mathcal{D}_t$, and after a few cycles, we query a human to label the most informative subset of pixels in each image of $\mathcal{D}_t$. We then use the newly labeled set to provide a supervisory signal in the loss computation, enabling ASDA to better align ϕ to the target domain. It must be noted that compared to prior active domain adaptation works, [67, 74, 40] we start with a completely unlabeled set for $\mathcal{D}_t$.

2.2 Uncertainty based Active Learning

Given an image $\mathbf{X}_t \in R^{H \times W \times 3}$ from $\mathcal{D}_t$, we pass it through ϕ to obtain a pixel-wise probability mask $\mathbf{P}_t \in R^{H \times W \times C}$ where C is the number of semantic classes. Thus the pseudo label associated with pixel (i, j) is given as $\tilde{\mathbf{Y}}_t^{i,j} = arg\ max_{c \in \{1,\ldots,C\}} \mathbf{P}_t^{i,j,c}$.

To select the most informative subset of pixels for the human oracle to label we first compute the uncertainty associated with the pseudo labels of each pixel. We use entropy as the uncertainty measure, which for pixel (i, j) is shown below:

$$E(i,j) = -\sum_{c=1}^{C} \mathbf{P}_t^{i,j,c} \log \mathbf{P}_t^{i,j,c}. \tag{2.1}$$

Therefore, the entropy mask for $\mathbf{X}_t$ is $\mathbf{E}(\mathbf{X}_t) = \{E(i,j)\}_{i=1,j=1}^{H,W} \in R^{H \times W}$. $\mathbf{E}(\mathbf{X}_t)$ is flattened, and all its entries are sorted to find the pixels with the highest entropy values. Generally, the model tends to provide high confident pseudo labels for the pixels coming from high-frequent classes, thereby resulting in lower entropy values for such pixels. Therefore, the pixels that are harder to classify, possibly due to belonging to rare semantic classes,

will have higher entropy values. Based on a pre-defined cycle budget b, the most informative subset of pixels is chosen by taking the top b pixels in $\mathbf{X}_t$ with the highest $E(i,j)$ values.

The labeled pixel set for each image $\mathbf{X}_t$ is henceforth referred to as $\mathcal{S}_l^{x_t}$ and the remaining unlabeled pixels as $\mathcal{S}_{ul}^{x_t}$.

2.3 Training Methodology

After every active cycle, for each image, the labeled set of pixels is used to compute a supervised classification loss as shown below:

$$\mathcal{L}_{ce} = -\frac{1}{|\mathcal{S}_l^{x_t}|} \sum_{(i,j)\in\mathcal{S}_l^{x_t}} \sum_{c=1}^{C} \hat{\mathbf{Y}}^{i,j,c} \log \mathbf{P}_t^{i,j,c}, \tag{2.2}$$

where $\hat{\mathbf{Y}}^{i,j}$ is the queried label of $(i,j) \in$. Thus $\mathcal{L}_{ce}$ is the standard categorical cross-entropy loss. On the other hand, the remaining unlabeled set of pixels is used for self-training whereby the negative pseudo-label loss is computed as shown below:

$$\mathcal{L}_{nl} = -\frac{1}{\Gamma(S_{ul}^{x_t})} \sum_{(i,j)\in\mathcal{S}_{ul}} \sum_{c=1}^{C} \pi(\mathbf{P}_t^{i,j,c}) \log(1 - \mathbf{P}_t^{i,j,c}), \tag{2.3}$$

where $\pi(\mathbf{P}_t^{i,j,c})$ is the assigned negative pseudo label defined as follows:

$$\pi(\mathbf{P}_t^{i,j,c}) = \begin{cases} 1 & \text{if } \mathbf{P}_t^{i,j,c} < \tau \\ 0 & \text{otherwise} \end{cases} \tag{2.4}$$

with τ being a pre-defined threshold. $\Gamma(S_{ul}^{x_t})$ in $\mathcal{L}_{nl}$ is the set of all negative pseudo labels defined as

$$\Gamma(S_{ul}^{x_t}) = \sum_{(i,j)\in S_{ul}^{x_t}} \sum_{c=1}^{C} \pi(\mathbf{P}_t^{i,j,c}). \tag{2.5}$$

The negative pseudo-label loss has been shown to outperform the standard cross-entropic pseudo-label loss [23, 73, 67] by better handling ambiguous cases associated with low output probabilities in the initial stages of the learning [67].

The overall model ASDA is trained end-to-end with by minimizing the total loss $\mathcal{L}_{total} = \mathcal{L}_{ce} + \mathcal{L}_{nl}$. It must be noted that unlike prior active learning studies [67, 40, 74] we do not start with a small labeled target domain dataset $\mathcal{D}_t$ and instead, it is completely unlabeled. Therefore, before the first active cycle starts, the model is trained with $\mathcal{L}_{total} = \mathcal{L}_{nl}$ where the unlabeled set $\mathcal{S}_{ul}^{x_t}$ will be all the pixels in $\mathbf{X}_t$. The overall approach of ASDA is shown in Algorithm **??**.

2.4 Experiments

Datasets. We use three different datasets to construct two domain adaptation tasks. The datasets are as follows:

- Cityscapes dataset [13] is used for real-scene evaluation. It contains high-quality urban road scenario images and is comprised of 2,975 training images and 500 validation images with a resolution of 2048x1024.
- GTAV dataset [45] has 24,966 images of size 1914x1052 with 19 classes common to Cityscapes. These images are synthetic and they are extracted from the game GTAV
- SYNTHIA dataset [46] has 9,400 images of size 1280x760 with 16 classes common to Cityscapes. This is also a synthetic urban scene dataset.

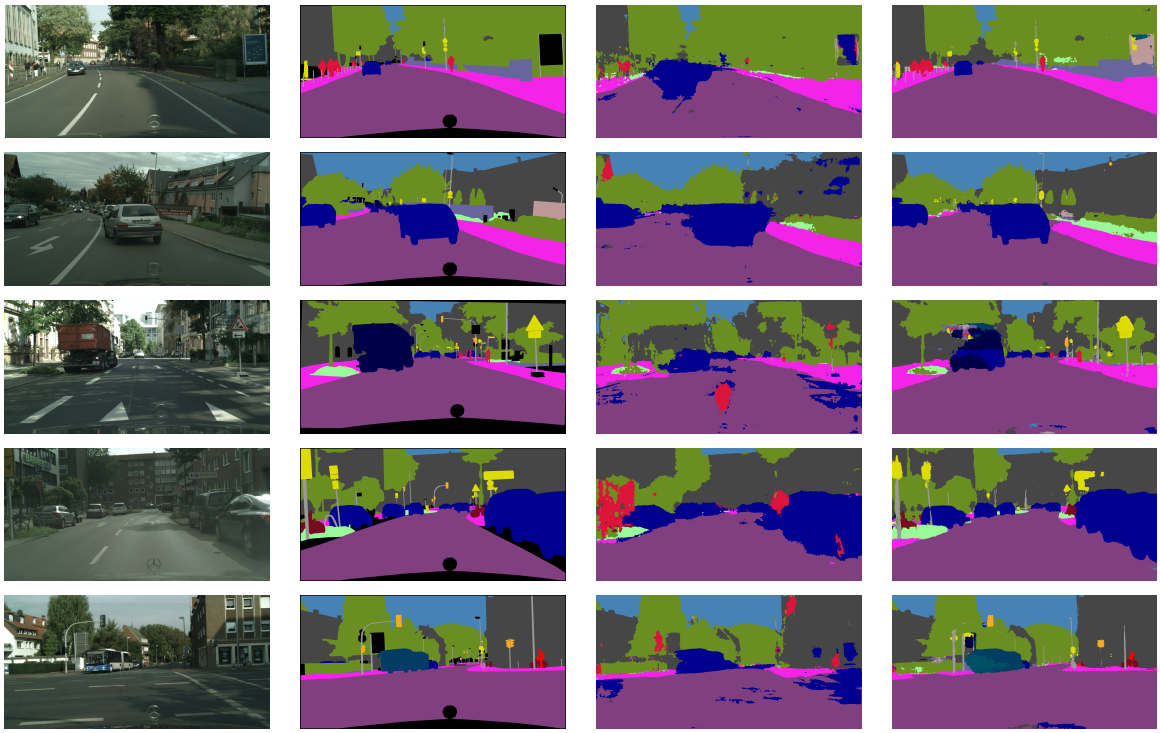


Figure 2.1: Qualitative visualization of segmentation results from GTAV→CITYSCAPES. From left to right: Original image, ground-truth label, semantic maps predicted by source-only model, and semantic maps predicted by ASDA

Implementation details. All of the experiments are carried out on Tesla A100 GPU. We use Deeplabv2 [10] with ResNet-101 pre-trained on ImageNet as the backbone. We use the SGD optimizer, with a weight decay of 1e-4 and a momentum of 0.9. Also, we employ the poly learning rate policy with the initial starting point of the curve at 1e-3. The target side images to 1280 × 640. Similar to [67] we perform active learning in 5 cycles.

Annotation Budget. We fix the annotation budget to 10% of the most informative pixels to be labeled in each image. We also run a study where we vary the budget to 5%, 7%, and 20%.

Table 2.1: **Comparison with prior domain UDA methods GTAV → Cityscapes.** All results are reported in terms of mIoU (%). We also report the class-wise performance of 16 out of 19 categories. $*$ refers to standard (with access to labeled source data) UDA methods, † refers to SF-UDA methods, and $\perp$ refers to active learning based UDA methods. ♯ refers to performance based on DeepLabV3. The best results w.r.t. standard UDA and SF-UDA methods are highlighted in bold.

[1.2pt] Method	road	side.	buil.	wall	fence	pole	light	sign	veg.	terr.	sky	pers.	rider	car	truck	bus	train	motor	bike	mIoU
Source Only	75.8	16.8	77.2	12.5	21.0	25.5	30.1	20.1	81.3	24.6	70.3	53.8	26.4	49.9	17.2	25.9	6.5	25.3	36.0	36.6
CBST* [80]	91.8	53.5	80.5	32.7	21.0	34.0	28.9	20.4	83.9	34.2	80.9	53.1	24.0	82.7	30.3	35.9	16.0	25.9	42.8	45.9
MRKLD*[81]	91.0	55.4	80.0	33.7	21.4	37.3	32.9	24.5	85.0	34.1	80.8	57.7	24.6	84.1	27.8	30.1	26.9	26.0	42.3	47.1
Seg-Uncertainty*[76]	90.4	31.2	**85.1**	36.9	25.6	37.5	48.8	48.5	85.3	34.8	81.1	64.4	36.8	86.3	34.9	52.2	1.7	29.0	44.6	50.3
TPLD*[52]	94.2	60.5	82.8	36.6	16.6	39.3	29.0	25.5	85.6	44.9	84.4	60.6	27.4	84.1	37.0	47.0	31.2	36.1	50.3	51.2
DPL-Dual*[12]	92.8	54.4	86.2	41.6	32.7	36.4	49.0	34.0	85.8	41.3	86.0	63.2	34.2	87.2	39.3	44.5	18.7	42.6	43.1	53.3
ProDA*[22]	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
URMA†[16]	92.3	55.2	81.6	30.8	18.8	37.1	17.7	12.1	84.2	35.9	83.8	57.7	24.1	81.7	27.5	44.3	6.9	24.1	40.4	45.1
SRDA†[5]	90.5	47.1	82.8	32.8	28.0	29.9	35.9	34.8	83.3	39.7	76.1	57.3	23.6	79.5	30.7	40.2	0.0	26.6	30.9	45.8
GND†[25]	90.9	48.6	85.5	35.3	31.7	36.9	34.7	34.8	86.2	**47.8**	**88.5**	61.7	32.6	85.9	46.9	50.4	0.0	38.9	52.4	51.6
ASDA(10%)	**94.9**	**69.4**	83.9	**50.8**	**50.3**	44.3	54.0	49.8	86.1	46.7	85.9	67.7	**40.4**	**90.1**	**48.5**	56.3	**49.4**	44.5	**63.7**	**62.9**
AADA (5%)$^{\perp,\sharp}$ [56]	92.2	59.9	87.3	36.4	45.7	46.1	50.6	59.5	88.3	44.0	90.2	69.7	38.2	90.0	55.3	45.1	32.0	32.6	62.9	59.3
MADA (5%)$^{\perp,\sharp}$ [41]	95.1	69.8	88.5	43.3	48.7	45.7	53.3	59.2	89.1	46.7	91.5	73.9	50.1	91.2	60.6	56.9	48.4	51.6	68.7	64.9
RIPU $^{\perp}$[67]	96.5	74.1	89.7	53.1	51.0	43.8	53.4	62.2	90.0	57.6	92.6	73.0	53.0	92.8	73.8	78.5	62.0	55.6	70.0	69.6
Fully Supervised (100%) [1.2pt]	96.8	77.5	90.0	53.5	51.5	47.6	55.6	62.9	90.2	58.2	92.3	73.7	52.3	92.4	74.3	77.1	64.5	52.4	70.1	70.2

Evaluation Metric. Like most prior related works [67, 25, 76, 69] we use the mean Intersection-over-Union (mIoU) metric for evaluating the performance of our method. For GTAV→Cityscapes the common 19 Classes between GTAV and CityScapes are considered and Synthia→Cityscpaes, the 16 common classes are considered.

Figure 2.2: **Qualitative visualization of pixels annotated by the human oracle for GTAV→CITYSCAPES**. From left to right. Ground truth maps, pixels annotated by random selection, and pixels annotated by uncertainty based selection.

Table 2.2: Comparison with prior domain UDA methods SYNTHIA → Cityscapes. We report the mIoUs in terms of 13 classes (excluding the wall*, fence*, and pole*) and 16 classes which are correspondingly referred to as mIoU*(%) and mIoU(%) respectively. ∗ refers to standard (with access to labeled source data) UDA methods, † refers to SF-UDA methods, and ⊥ refers to active learning based UDA methods. ♯ refers to performance based on DeepLabV3. The best results w.r.t. standard UDA and SF-UDA methods are highlighted in bold.

[1.2pt] Method	road	side.	buil.	wall*	fence*	pole*	light	sign	veg.	sky	pers.	rider	car	bus	motor	bike	mIoU	mIoU*
Source Only	64.3	21.3	73.1	2.4	1.1	31.4	7.0	27.7	63.1	67.6	42.2	19.9	73.1	15.3	10.5	38.9	34.9	40.3
CBST* [80]	68.0	29.9	76.3	10.8	1.4	33.9	22.8	29.5	77.6	78.3	60.6	28.3	81.6	23.5	18.8	39.8	42.6	48.9
MRKLD *[81]	67.7	32.2	73.9	10.7	1.6	37.4	22.2	31.2	80.8	80.5	60.8	29.1	82.8	25.0	19.4	45.3	43.8	50.1
DPL-Dual *[12]	87.5	45.7	82.8	13.3	0.6	33.2	22.0	20.1	83.1	86.0	56.6	21.9	83.1	40.3	29.8	45.7	47.0	54.2
TPLD *[52]	80.9	44.3	82.2	19.9	0.3	40.6	20.5	30.1	77.2	80.9	60.6	25.5	84.8	41.1	24.7	43.7	47.3	53.5
Seg-Uncertainty *[76]	87.6	41.9	83.1	14.7	1.7	36.2	31.3	19.9	81.6	80.6	63.0	21.8	86.2	40.7	23.6	53.1	47.9	54.9
ProDA *[22]	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
URMA †[16]	59.3	24.6	77.0	14.0	1.8	31.5	18.3	32.0	83.1	80.4	46.3	17.8	76.7	17.0	18.5	34.6	39.6	45.0
GND †[25]	89.0	44.6	80.1	7.8	0.7	34.4	22.0	22.9	82.0	**86.5**	65.4	33.2	84.8	45.8	38.4	31.7	48.1	55.5
ASDA(10%)	**90.9**	**61.7**	83.4	31.7	**32.8**	33.1	36.1	**43.9**	**88.9**	**86.1**	52.9	**41.8**	87.3	40.2	32.4	**53.1**	**56.1**	**64.3**
RIPU $^{\perp,\sharp}$[67]	96.8	76.6	89.6	45.0	47.7	45.0	53.0	62.5	90.6	92.7	73.0	52.9	93.1	80.5	52.4	70.1	70.1	75.7
AADA (5%)$^{\perp,\sharp}$ [56]	91.3	57.6	86.9	37.6	48.3	45.0	50.4	58.5	88.2	90.3	69.4	37.9	89.9	44.5	32.8	62.5	61.9	66.2
MADA (5%)$^{\perp,\sharp}$ [41]	96.5	74.6	88.8	45.9	43.8	46.7	52.4	60.5	89.7	92.2	74.1	51.2	90.9	60.3	52.4	69.4	68.1	73.3
Fully Supervised (100%) [1.2pt]	96.7	77.8	90.2	40.1	49.8	52.2	58.5	67.6	91.7	93.8	74.9	52.0	92.6	70.5	50.6	70.2	70.6	75.9

2.5 Comparative Results

We compare ASDA with prior UDA, SF-UDA as well as active domain-adaptation based UDA methods. The comparative results are shown in Tables 2.1 and 2.2 for GTAV → Cityscapes and SYNTHIA → Cityscapes, respectively. The fully supervised baseline involves access to a fully-annotated source and target domain training set.

Comparison with standard UDA methods. In both Tables 2.1 and 2.2, the methods marked with ∗ reflect prior state-of-the-art standard (with labeled source data) UDA based semantic segmentation methods. It can be observed that with just 10% labeled pixels, ASDA significantly outperforms these approaches which highlights that with a human-in-the-loop approach, we can not only improve upon standard self-training approaches for UDA but also eliminate the need for concurrent access to both source and target data. Specifically for hard classes like fence and pole ASDA achieves very high performance gains. This makes

ASDA preserve source side data privacy making it more practically viable for real-world UDA tasks.

Comparison with SF-UDA methods. In both Tables 2.1 and 2.2, the methods marked with † reflect prior state-of-the-art SF-UDA based semantic segmentation methods. ASDA along with these prior methods, show considerable performance gains over just using the source only trained model for inference on target domain test data. Although the prior SF-UDA methods showed that eliminating source side data can still enable effective UDA, the noisy nature of pseudo label based self-training creates a bottleneck in the performance. With ASDA, we show that the performance of a complex task like SF-UDA can be significantly improved by enabling a human-in-the-loop training approach where we constrain labeling efforts by a pre-determined budget.

Comparison with Active Domain Adaptation methods. We also compare ASDA with prior active domain adaptation (marked with $\perp$ in tables 2.1 and 2.2) works like AADA [56], MADA [41] and RIPU [67]. These approaches are built on top of standard UDA methods which means they *assume access to a fully-labeled source dataset* for alignment regularization. Additionally, AADA [56] and MADA [41] label 5% of all images in the target domain which involves annotating all the pixels in these images which is very inefficient. In comparison, ASDA, only labels 10% of pixels which are of high uncertainty values. RIPU's selection strategy is much more efficient involving labeling 2.2% of 9×9 regions in each image, enabling it to achieve near full supervision performance. Although ASDA requires a higher annotation budget compared to RIPU, it must be noted that RIPU's performance

gains are largely attributed to access to fully labeled source side data for adaptation. Moreover, all three of these works *assume the availability of a small labeled subset of images* in the target domain to start the active learning cycles. Therefore, the assumption of a fully-unlabeled target domain training data is unmet. We relax this assumption and show that ASDA can achieve comparable performance (Tables 2.1 and 2.2) with these prior active domain adaptation tasks along with being a completely source-free setup. To summarize, ASDA is a truly source-free domain adaptive semantic segmentation method with no prior labels in the target domain, while the most competitive recent methods like AADA, MADA and RIPU assume (i) the existence of fully labeled source data, and (ii) small amount of labeled data on the target side.

2.6 Qualitative Analysis

As shown in Fig 3.2 by utilizing a human-in-the-loop active learning approach ASDA significantly produces significantly better semantic maps compared to just using a source-only trained model. Fig 2.2 also shows that the uncertainty based selection strategy enables the labeling of more hard to classify pixels. In such cases, the standard pseudo-labeling approach would fail as it will propagate noisy predictions during the optimizing of the model.

2.7 Analysis of Annotation Budget

We run an ablation study on GTAV $\rightarrow$ Cityscapes by changing the annotation budget and as observed from Fig 2.3 the performance of ASDA increases with the rise in

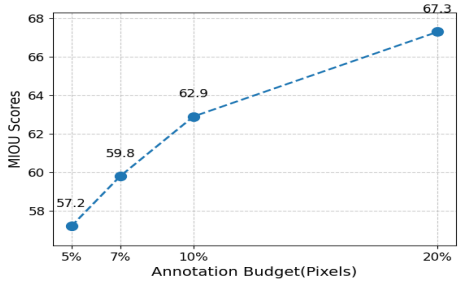


Figure 2.3: Performance variation of ASDA with varying pixel-level annotation budgets. Budget is capped at 20%

the budget. We constrain the budget to no more than 20% and can see that at 20% the performance is at par with fully-supervised setups even without access to source data during model adaptation.

2.8 Analysis of Pixel Selection Strategy

Table 2.3 shows the performance of ASDA in comparison to a random pixel selection. It can be observed that an uncertainty-guided pixel selection strategy is much more effective than random selection for all annotation budgets.

Table 2.3: **Comparison of our selection strategy with that of random pixel selection for different annotation budgets**. Results reported on GTAV → Cityscapes

Pixels	**Entropy MIOU**	**Random MIOU**
5%	57.2	45.1%
7%	59.8	47.4%
10%	62.9	49.7%
20%	67.3	50.3%

Chapter 3

Environmental Applications of Domain Adaptation

In this Chapter, we will see a couple of applications of domain adaptation applied to the environmental use cases. Here we explain how a simple yet effective strategy of self-learning can help bridge the gap between different domains. We apply this self-learning based adaptation system to the two of the following use cases:-

- Forest Fire Detection.
- Vegetation Segmentation from Satellite Imagery.

3.1 Forest Fire detection

Problem Statement:

The calamitous nature of forest fires warrants the development of autonomous systems for real-time detection of fires. In recent years, deep-learning based computer-vision frameworks have achieved tremendous success in the autonomous detection of everyday objects in images and videos. We aim to develop such learning models geared towards the detection of fires from image data at a near real-time speed, thereby allowing their deployment in drone-based environment monitoring systems.

Figure 3.1: **Visual results of the adapted YOLOV5s model on the Sparx project forest fire data**

Methodology:

We have developed methods for detection of forest fires from imagery and evaluated them on data collected by collaborators from the Sparx project, as well as publicly available opensource datasets. We start by developing a deep object detection model capable of detecting fires as objects in a given frame. For this we use the open source YOLOV5S[44] object detection architecture as the backbone for this system. We choose YOLOV5s[44] since it is an object detection framework known for its minimal latency and significantly quick inference time. To train the model we use two publicly available datasets FIRENET[1] and IEEE-Fire[49]. Individually the model gets an Average Precision (AP) of 81% and 94% on FIRENET[1] and IEEE-Fire[49], respectively. However, real world images of vegetation will have a significant domain shift wrt to such existing open-sourced datasets; hence,we also aim to develop frameworks that can adapt to images from different domains in order to make it generalizable to varying external environments and conditions. For example, a machine learning model trained on high resolution images captured by a mobile phone would underperform for the same task when deployed onto a CCTV camera. To address this issue, we develop a simple domain adaptation algorithm on top of the existing object detector. Basically we employ self learning as the fronte of the adaptation mechanism and to bridge the domain gap. Here we train the YOLOV5s[44] model initially on the source data until it attains a significantly good amount of performance. Once this is done, the model trained on the source data is then tested on the target dataset to generate pseudolabelled predictions. We then combine these weak pseudolabel predictions with the labelled source data to restart the training of the original model. To test its efficacy we perform a domain

Adaptation Type	Pseudolabels Generated	Ground Truth Used	mAP
SOURCE MODEL DIRECT	-	-	74%
SUPERVISED DA	-	FireNet & IEEE	91%
UNSUPERVISED DA	IEEE	ONLY FireNet	88%
SOURCE FREE DA	IEEE	-	81%

Table 3.1: Table depicting results on the FireNet $\rightarrow$ IEEE experiments. It is a Comparison of Mean Average Precision(mAP) scores for different adaptation strategies on the test dataset(IEEE) . Here acroos all of these experiments, FireNet is the source data and the IEEE dataset is the target dataset. Pseudolabels generated refers to the pseudolabels from which dataset used during the self-learning stage.Ground truth used basically means the ground truths from which dataset was added to the pseudolabels during the self-learning process.

adaptation test between FIRENET[1] and IEEE-Fire[49] where the former is used as the source domain and the latter the target domain on which the model is to be tested. The adapted model gets average precision(AP) of 88% on IEEE-Fire[49] which is significantly higher than the unadapted performance of just 74%. We apply our domain adapted model on some in-the-wild captured fire imagery obtained from our collaborators in the team, the qualitative results of which are shown below. The inference time per image is about 0.8ms, which is significantly faster than real-time processing. The detailed analytical results is given in Table 6.1. Some of the sample visual results is provided in Figure 6.1.

3.2 Vegetation Segmentation:

Problem Statement:

Similar to the Forest fire detection workflow, a vegetation segmentation pipeline was also built. Here, PSPNet[78] was used as a segmentation backbone to classify the Sentinel-2 data satellite images into seven different classes namely urban land,agriculutural

Adaptation Type	Pseudolabels Generated	Ground Truth Used	mIOU
SOURCE MODEL DIRECT	-	-	37%
SUPERVISED DA	-	DeepGlobe & LoveDA	51%
UNSUPERVISED DA	LoveDA	ONLY DeepGlobe	48%

Table 3.2: Table depicting results on the DeepGlobe → LoveDA. It is a Comparison of Mean Intersection over Union(mIOU) scores for different adaptation strategies on the test dataset(LoveDA) . Here across all of these experiments, Deepglobe is the source data and the LoveDA dataset is the target dataset. Pseudolabels generated refers to the pseudolabels from the target dataset(LoveDA) used during the self-learning stage. Ground truth used basically means the ground truths from which dataset that was added to the pseudolabels during the self-learning process. The first row depicts the results of a model trained on source data directly applied to the target, the second row depicts supervised DA where both the G.T from source and target are utilized. The third row depicts, unsupervised DA where Pseudolabels are generated on the Target and then self-learning is done

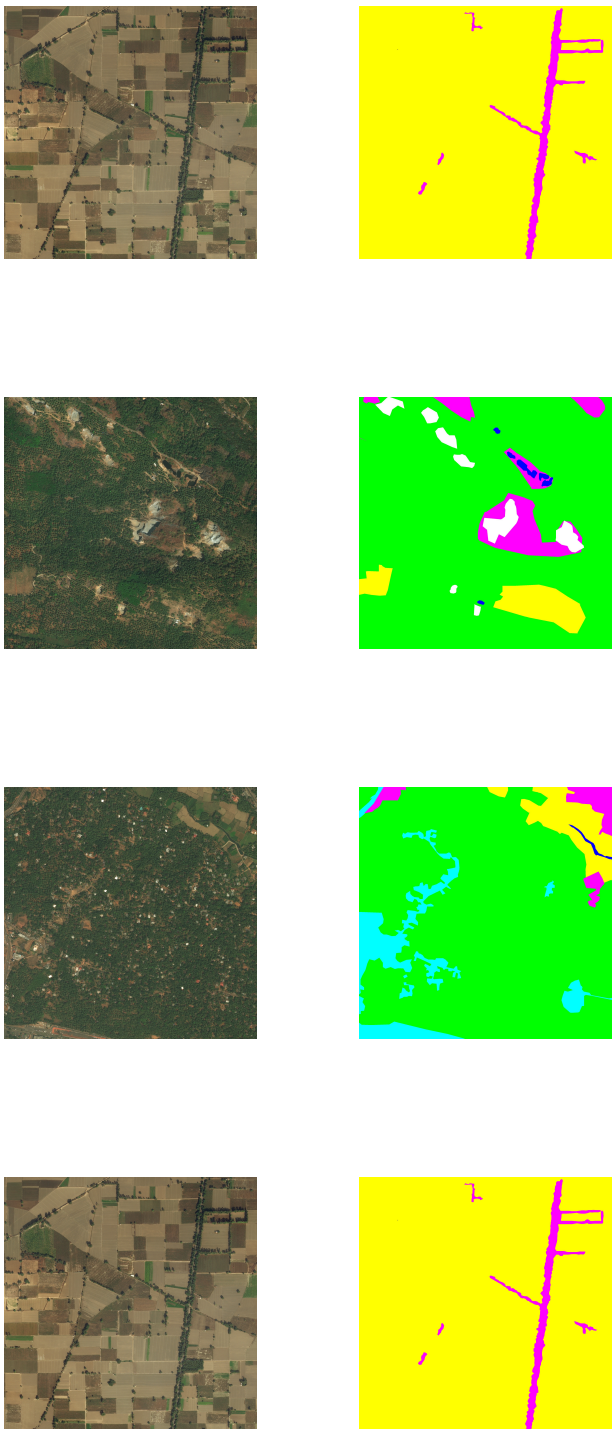
land, rangeland,forest land, water, barren land and unknown. PSPNet is a popular semantic segmentation network widely adopted in the industry due to its simplicity. A domain adaptation procedure was carried out between two different benchmark datasets called the DeepGlobe LandCover dataset[15] and the LoveDA dataset[61]. Self learning was once again used as a key element for the adaptation process. Similar to the fire detetcion project,we initially trained a PSPNet[78] model on the source data(DeepGlobe[15]), this trained model was then tested on the target datset to generate pseudolabel predictions. Thus by combining the generated pseudolabels and the source dataset, the original source trained model was then retrained to complete the self-learning system and thus bridging the domain gap. Some of the analytical results obtained have been furnished in Table 6.2. A few sample results on the Sentinel - 2 data has been shown in Figure 6.2

.

Summary:

Through this chapter,we were able to demonstrate that using a simple yet effective technique like self-learning, the domain gap within two different environmental use cases (i)Forest Fire Detection and (ii) Vegetation Segmentation was bridged. We had used popular architectures for this namely YOLOV5s[44] for detetction and PSPNet[78] for segmentation. We had also tested out these adapted models on a custom data such as the Sparx forest fire detection project data thus supporting the usefulness of these models in a real world environment.

Figure 3.2: **Visual Results of adapted PSPNet model predictions on Sentinel-2 data. On the left we have the input satelite images, and on the right we have the corresponding mask predictions as color maps**



Chapter 4

Conclusions

In this thesis, a source-free domain adaptive semantic segmentation method with a human in the loop ispresented. Given a pre-trained source model, but no source data, we adapt the model to the target domain, which is completely unlabeled. By analyzing the performance of the adaptation, we estimate which pixels in the segmentation task should be labeled to improve the performance of the overall system. This drives the selection of such pixels in the active learning strategy, based on a small labeling budget that is provided. We show that the resulting approach can achieve results that are comparable to many unsupervised domain adaptation methods for segmentation. In fact, with about 20% labeling budget, this method achieves the same performance as the fully supervised case. In the end, a few environmental applications of domain adaptation such as forest fire prediction and vegetation segmentation have also been demonstrated.

4.1 Acknowledgements:

This work has been partially supported by NSF grant 1724341 and the UC Lab Fees program through the UC Office of the President.

Bibliography

[1] FireNet OpenSource Dataset. https://github.com/OlafenwaMoses/FireNET/.

[2] Alvin Alpher. Frobnication. *Journal of Foo*, 12(1):234–778, 2002.

[3] Alvin Alpher and Ferris P. N. Fotheringham-Smythe. Frobnication revisited. *Journal of Foo*, 13(1):234–778, 2003.

[4] Alvin Alpher, Ferris P. N. Fotheringham-Smythe, and Gavin Gamow. Can a machine frobnicate? *Journal of Foo*, 14(1):234–778, 2004.

[5] Mathilde Bateson, Hoel Kervadec, Jose Dolz, Hervé Lombaert, and Ismail Ben Ayed. Source-relaxed domain adaptation for image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*, pages 490–499. Springer, 2020.

[6] Soufiane Belharbi, Ismail Ben Ayed, Luke McCaffrey, and Eric Granger. Deep active learning for joint classification & segmentation with weak annotator. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3338–3347, 2021.

[7] William H. Beluch, Tim Genewein, Andreas Nürnberger, and Jan M. Köhler. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[8] Hermann Blum, Paul-Edouard Sarlin, Juan Nieto, Roland Siegwart, and Cesar Cadena. Fishyscapes: A benchmark for safe semantic segmentation in autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, Oct 2019.

[9] Arantxa Casanova, Pedro O Pinheiro, Negar Rostamzadeh, and Christopher J Pal. Reinforced active learning for image segmentation. *arXiv preprint arXiv:2002.06583*, 2020.

[10] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous

convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[11] Zhang Chen, Zhiqiang Tian, Jihua Zhu, Ce Li, and Shaoyi Du. C-cam: Causal cam for weakly supervised semantic segmentation on medical image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11676–11685, June 2022.

[12] Yiting Cheng, Fangyun Wei, Jianmin Bao, Dong Chen, Fang Wen, and Wenqiang Zhang. Dual path learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9082–9091, 2021.

[13] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[14] 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 (CVPR)*, June 2016.

[15] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 172–181, 2018.

[16] Francois Fleuret et al. Uncertainty reduction for model adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9613–9623, 2021.

[17] Xiaoqing Guo, Jie Liu, Tongliang Liu, and Yixuan Yuan. Simt: Handling open-set noise for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7032–7041, June 2022.

[18] Xiaoqing Guo, Chen Yang, Baopu Li, and Yixuan Yuan. Metacorrection: Domain-aware meta loss correction for unsupervised domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3927–3936, June 2021.

[19] Yingsong Huang, Bing Bai, Shengwei Zhao, Kun Bai, and Fei Wang. Uncertainty-aware learning against label noise on imbalanced datasets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6960–6969, 2022.

[20] Zhanghan Ke, Daoye Wang, Qiong Yan, Jimmy Ren, and Rynson WH Lau. Dual student: Breaking the limits of the teacher in semi-supervised learning. In *Proceedings*

of the IEEE/CVF International Conference on Computer Vision, pages 6728–6736, 2019.

[21] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

[22] Kwanyoung Kim, Dongwon Park, Kwang In Kim, and Se Young Chun. Task-aware variational adversarial active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8166–8175, 2021.

[23] Youngdong Kim, Junho Yim, Juseung Yun, and Junmo Kim. Nlnl: Negative learning for noisy labels. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 101–110, 2019.

[24] Divya Kothandaraman, Rohan Chandra, and Dinesh Manocha. Ss-sfda: Self-supervised source-free domain adaptation for road segmentation in hazardous environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 3049–3059, October 2021.

[25] 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 (ICCV)*, pages 7046–7056, October 2021.

[26] Suhyeon Lee, Junhyuk Hyun, Hongje Seong, and Euntai Kim. Unsupervised domain adaptation for semantic segmentation by content transfer. 35:8306–8315, May 2021.

[27] Guangrui Li, Guoliang Kang, Wu Liu, Yunchao Wei, and Yi Yang. Content-consistent matching for domain adaptive semantic segmentation. In *Computer Vision – ECCV 2020*, pages 440–456, 2020.

[28] Jun Li, José M Bioucas-Dias, and Antonio Plaza. Hyperspectral image segmentation using a new bayesian approach with active learning. pages 3947–3960, 2011.

[29] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[30] Xin Li and Yuhong Guo. Adaptive active learning for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

[31] 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.

[32] Zimo Liu, Jingya Wang, Shaogang Gong, Huchuan Lu, and Dacheng Tao. Deep reinforcement active learning for human-in-the-loop person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[33] 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 (CVPR)*, June 2019.

[34] Prem Melville and Raymond J Mooney. Diverse ensembles for active learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 74, 2004.

[35] Andres Milioto, Philipp Lottes, and Cyrill Stachniss. Real-time semantic segmentation of crop and weed for precision agriculture robots leveraging background knowledge in cnns. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2229–2235, 2018.

[36] Andres Milioto and Cyrill Stachniss. Bonnet: An open-source training and deployment framework for semantic segmentation in robotics using cnns. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 7094–7100, 2019.

[37] Full Author Name. The frobnicatable foo filter, 2014. Face and Gesture submission ID 324. Supplied as additional material `fg324.pdf`.

[38] Full Author Name. Frobnication tutorial, 2014. Supplied as additional material `tr.pdf`.

[39] David Nilsson, Aleksis Pirinen, Erik Gärtner, and Cristian Sminchisescu. Embodied visual active learning for semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2373–2383, 2021.

[40] 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 (ICCV)*, pages 9112–9122, October 2021.

[41] Munan Ning, Donghuan Lu, Dong Wei, Cheng Bian, Chenglang Yuan, Shuang Yu, Kai Ma, and Yefeng Zheng. Multi-anchor active domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9112–9122, 2021.

[42] 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 (CVPR)*, June 2020.

[43] Pedro O. Pinheiro. Unsupervised domain adaptation with similarity learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[44] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[45] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 102–118. Springer, 2016.

[46] 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.

[47] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[48] Ozan Sener, Hyun Oh Song, Ashutosh Saxena, and Silvio Savarese. Learning transferrable representations for unsupervised domain adaptation. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, 2016.

[49] Alireza Shamsoshoara, Fatemeh Afghah, Abolfazl Razi, Liming Zheng, Peter Fulé, and Erik Blasch. The flame dataset: Aerial imagery pile burn detection using drones (uavs), 2020.

[50] Gyungin Shin, Weidi Xie, and Samuel Albanie. All you need are a few pixels: Semantic segmentation with pixelpick. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 1687–1697, October 2021.

[51] 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.

[52] Inkyu Shin, Sanghyun Woo, Fei Pan, and In So Kweon. Two-phase pseudo label densification for self-training based domain adaptation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 532–548. Springer, 2020.

[53] Mennatullah Siam, Mostafa Gamal, Moemen Abdel-Razek, Senthil Yogamani, Martin Jagersand, and Hong Zhang. A comparative study of real-time semantic segmentation for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[54] 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 (CVPR)*, June 2020.

[55] Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[56] 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.

[57] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. Mar. 2016.

[58] 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.

[59] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Perez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[60] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Perez. Dada: Depth-aware domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[61] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv preprint arXiv:2110.08733*, 2021.

[62] Zhonghao Wang, Mo Yu, Yunchao Wei, Rogerio Feris, Jinjun Xiong, Wen-mei Hwu, Thomas S. Huang, and Honghui Shi. Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[63] Guoqiang Wei, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Metaalign: Coordinating domain alignment and classification for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16643–16653, June 2021.

[64] Jiaxi Wu, Jiaxin Chen, and Di Huang. Entropy-based active learning for object detection with progressive diversity constraint. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9397–9406, 2022.

[65] Tsung-Han Wu, Yueh-Cheng Liu, Yu-Kai Huang, Hsin-Ying Lee, Hung-Ting Su, Ping-Chia Huang, and Winston H. Hsu. Redal: Region-based and diversity-aware active learning for point cloud semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15510–15519, October 2021.

[66] Ni Xiao and Lei Zhang. Dynamic weighted learning for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15242–15251, June 2021.

[67] Binhui Xie, Longhui Yuan, Shuang Li, Chi Harold Liu, and Xinjing Cheng. Towards fewer annotations: Active learning via region impurity and prediction uncertainty for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8068–8078, June 2022.

[68] Binhui Xie, Longhui Yuan, Shuang Li, Chi Harold Liu, Xinjing Cheng, and Guoren Wang. Active learning for domain adaptation: An energy-based approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8708–8716, 2022.

[69] Shuai Xie, Zunlei Feng, Ying Chen, Songtao Sun, Chao Ma, and Mingli Song. Deal: Difficulty-aware active learning for semantic segmentation. In *Proceedings of the Asian conference on computer vision*, 2020.

[70] Shuai Xie, Zunlei Feng, Ying Chen, Songtao Sun, Chao Ma, and Mingli Song. Deal: Difficulty-aware active learning for semantic segmentation. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, November 2020.

[71] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[72] Zhao Yi, Antonio Criminisi, Jamie Shotton, and Andrew Blake. Discriminative, semantic segmentation of brain tissue in mr images. In Guang-Zhong Yang, David Hawkes, Daniel Rueckert, Alison Noble, and Chris Taylor, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2009*, pages 558–565, 2009.

[73] Fuming You, Jingjing Li, Lei Zhu, Zhi Chen, and Zi Huang. Domain adaptive semantic segmentation without source data. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3293–3302, 2021.

[74] Hao Zhang and Ruimao Zhang. Active domain adaptation with multi-level contrastive units for semantic segmentation. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 1640–1657, December 2022.

[75] Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *Proceedings of the 30th International Conference on Machine Learning*, pages 819–827, 17–19 Jun 2013.

[76] 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.

[77] Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[78] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[79] Jingbo Zhu, Huizhen Wang, Benjamin K Tsou, and Matthew Ma. Active learning with sampling by uncertainty and density for data annotations. *IEEE Transactions on audio, speech, and language processing*, 18(6):1323–1331, 2009.

[80] Yang Zou, Zhiding Yu, B.V.K. Vijaya 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)*, September 2018.

[81] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5982–5991, 2019.