Active Domain Adaptation for Fish Segmentation

Simona Aksman
saksman@berkeley.edu

Abstract

Semantic segmentation is a promising approach for improving the sustainability and precision of deep sea fishing practices, but there are challenges to scaling adoption. We study two of these challenges: few ground-truth labels and a wide variety of marine habitats and fish species to adapt to. We experiment with applying active learning and domain adaptation approaches, first independently and then jointly, to handle these challenges.

1. Introduction

Human activity over the past several hundred years has inflicted significant damage on ocean ecosystems. Today, overfishing as a result of inefficient commercial trawling practices in deep sea ecosystems is a known cause of ecosystem collapse [7]. This practice occurs when unwanted fish species are caught up in fishing equipment. Over the past several decades, overfishing has become increasingly widespread. A 2020 report by the Food and Agriculture Organization of the United Nations estimated that 34% of the world's fisheries were overfished in 2017, and in some areas of the world, that number is estimated to be closer to 60% [3].

Within the past several years, advancements in computer vision have generated new opportunities to improve the sustainability and precision of commercial fisheries. In particular, semantic segmentation is a promising approach for addressing the problem of overfishing in ocean environments [4, 9]. Semantic segmentation is a computer vision task which predicts classes in images at the pixel level. When applied to fish, segmentation predictions can be used to estimate fish size, shape, and weight [9], statistics which can help to identify unwanted fish in commercial equipment.

In [9], Saleh et al. establish a set of benchmark models and an open-sourced dataset called DeepFish which provide a convenient starting point for the fish segmentation task. Saleh et al. present a strong out-of-sample semantic segmentation benchmark of 0.93 mIoU when training on 310 ground-truth labels. As a starting point, we nearly replicate the result, achieving 0.91 mIoU, and show the performance of this benchmark on several sample images in Figure 1.

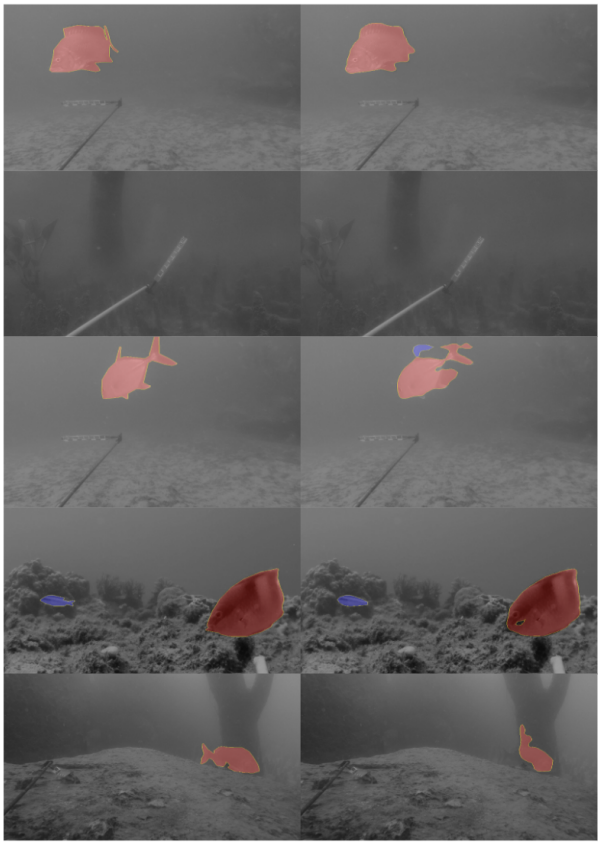

Figure 1. A sample of validation set results with 310 training labels. Ground-truth labels are in the **left** column, predicted labels in the **right** column.

With the availability of a strong benchmark result, we can turn our attention to the question of how to scale the fish segmentation task. There are several significant barriers to scaling, and in this paper we explore two of these barriers: a lack of ground-truth labels and a large variety of both marine environments and fish species to adapt to. This paper aims to address each of these barriers for fish segmentation by experimenting with the use of both domain

adaptation and active learning, for handling domain shifts and a lack of training labels, respectively. Specifically, we explore applications of one domain adaptation approach called AdaptSegNet [13] as well as two active learning approaches, Learned Loss for Active Learning (LL4AL) [15] and Active Adversarial Domain Adaptation (AADA) [12], on the DeepFish dataset. Finally, we evaluate the approach on another fish imagery dataset, QUT [2], for further out-of-sample validation.

2. Related Work

2.1 Semantic Segmentation. Semantic segmentation is a computer vision task that aims to classify objects and textures pixel-wise across entire images. State-of-the-art algorithms for semantic segmentation utilize deep learning due to the high quality results that these algorithms have produced in recent years [5]. In particular, the success of convolutional neural networks (CNNs) for image classification spurred research into semantic segmentation. The first semantic segmentation method was the fully-convolutional network (FCN) [11]. FCNs and other popular semantic segmentation approaches typically use CNN architectures, such as ResNet-50, as their backbone, and then perform upsampling to produce a pixel-wise map that is of the input image's dimensions.

2.2 Active Learning. In active learning, a learning algorithm can interactively query a human (often referred to as an oracle) to obtain ground-truth labels [10]. This approach is useful when there is a substantial quantity of unlabeled data and the task of labeling that data is expensive. Such is the case for the task of semantic segmentation, where ground-truth labels are defined at the pixel level, and it can therefore take many hours to annotate a batch of labels for training. In [9], it took Saleh et al. 25 hours to label the 310 images used to train the algorithm, with each label taking about 5 minutes to annotate and validate. Active learning is suggested in [9] as a potential next step given the labeling challenges the authors faced.

State-of-the-art active learning methods tend to use one of two methods: synthesized and pool-based [16]. Synthesized approaches use generative models, such as GANs or VAEs, to produce samples that are informative for training. Pool-based approaches, on the other hand, look for representative samples in the unlabeled data and suggest them to an oracle for labeling. Pool-based approaches tend to use uncertainty and diversity cues to determine how best to sample data, with many recent approaches utilizing both uncertainty and diversity cues [6]. We evaluate two pool-based approaches in this paper: Learned Loss for Active Learning (LL4AL) [15] and Active Adversarial Domain Adaptation (AADA) [12]. LL4AL is a task-agnostic active learning approach for deep learning models that appends a loss module to an existing neural network at every layer of the network. This module provides an estimate of the uncertainty of the network's predictions. Figure 2 gives a high-level overview of how this loss is learned. We provide details about AADA in the section about Active Domain Adaptation.

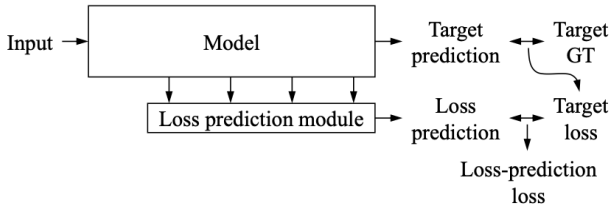

Figure 2. A loss prediction module used in [15] to quantify uncertainty in a network's prediction. This is then used to determine which samples to collect for active learning.

2.3 Domain Adaptation. The DeepFish benchmark dataset was carefully constructed to have similar training, validation and test sets. In a real-world setting, it may not be feasible to have such consistency between the data that an algorithm is trained on and the setting in which it is used. In such cases, domain adaptation is a useful approach. Domain adaptation (DA) is a technique which takes an algorithm trained on one domain, called the source domain, and optimizes its performance to a new target domain [14]. The target domain is typically unlabeled or partially labeled, and the target data distribution can be accessed during DA. Within the context of semantic segmentation, unsupervised DA is a popular approach as it does not require any target labels. Unsupervised DA techniques typically follow one of the following patterns: adversarial approaches using discriminative adversarial neural networks and self-training with pseudo-labels [17]. In this paper, we apply AdaptSegNet [13], a popular unsupervised DA algorithm which uses an adversarial approach. We provide more details about this approach in section 3.1.

2.4 Active Domain Adaptation. Recently, researchers have begun to combine active learning and domain adaptation for image classification and object detection tasks [6, 12]. This helps improve DA performance on challenging domain shifts and therefore allows DA to be more applicable within real-world settings. As was the case for DA, active DA approaches tend to follow one of two patterns: adversarial learning and pseudo-label cluster-based learning. Active DA has not yet been applied to the semantic segmentation task as far as we can tell. Our contribution in

this paper is to extend one particular active DA approach, AADA [12], to this task. This approach uses an adversarial DA approach which is similar to AdaptSegNet [13]. After the unsupervised DA step, it applies a pool-based sampling algorithm for active learning that estimates the uncertainty and diversity of the data from the output of the adversarial learning step. In the future, we also plan to extend ADA-CLUE [6] to semantic segmentation, as this method has been shown to outperform AADA on some classification benchmarks.

3. Methodology

3.1 AdaptSegNet. AdaptSegNet [13] is an adversarial DA approach which trains a fully-convolutional discriminator network **D** to learn the difference between the source and target domains. The first step is to train the semantic segmentation network, which acts as the generator **G** within the adversarial learning setup. Next, **G**'s segmentation prediction is fed to **D**, which attempts to correctly classify whether it is coming from the source or target domain. The loss function for this joint learning process is:

$$L(I_s, I_t) = L_{seg}(I_s) + \lambda_{adv} L_{adv}(I_t) \tag{1}$$

where I_s and I_t are source and target images, respectively, L_{seg} is the cross-entropy loss learned by **G** in the source domain, and L_{adv} is the adversarial loss learned by **D**. λ_{adv} is a weight that balances the losses. In the original paper it is set to 0.01 based on a sensitivity analysis, but we reset it to 0.1 after some experimentation. Finally, the loss is optimized with a min-max objective:

$$\max_{\text{D}} \min_{\text{G}} L(I_s, I_t) \tag{2}$$

Given this objective, **G** will attempt to reduce the cross-entropy loss to improve predictions on the source images and fool **D** into classifying target predictions as source predictions. This has the effect of reducing the gap between the source and target domains.

3.2 AADA. Since AADA and AdaptSegNet both include an adversarial unsupervised DA step, AdaptSegNet can be used as a starting point for implementing AADA [12]. The output of the adversarial network is an input to the active learning sample selection step, which aims to find the most informative labels for sampling. The sampling criterion $s(x)$ for choosing labels from the unlabeled target dataset is defined as:

$$s(x) = \frac{1 - G_d(G_f(x))}{G_d(G_f(x))} \mathcal{H}(G_y(G_f(x))) \tag{3}$$

where $G_d(G_f(x))$ and $G_y(G_f(x))$ are the predictions made by discriminator **D** and generator **G**, respectively, and $\mathcal{H}(G_y(G_f(x)))$ indicates the entropy of **G**'s predictions. In this sampling strategy, $\frac{1-G_d(G_f(x))}{G_d(G_f(x))}$ is the diversity cue, and the entropy term is the uncertainty cue. This sampling strategy is applied per batch, and assumes that $G_d(G_f(x))$ and $G_y(G_f(x))$ are scalar values. This is not the case for semantic segmentation, where **G** and **D** both produce pixel predictions at the original image's dimensions. To compress these pixel maps to a set of scalar values, we compute channel-wise mean values of the pixel maps for each prediction, and then compute the mean for each image across channels. In the future, we plan to experiment with different schemes for compressing these pixel outputs for sampling, as well as test additional approaches for computing the sampling criteria. For instance, since the sampling criteria diversity cue is computed per batch, it likely does not pick up much signal in our experiments as our batch size is 2 (due to memory considerations). We will experiment with varying the batch size in future experiments.

4. Experiments

4.1 Experiment Setup. The DeepFish dataset was acquired by [9] by shooting underwater videos of 20 marine habitats in tropical Australia. Videos were recorded by cameras on metal frames and lowered into marine environments, where they were left for a period of time to collect footage. This footage was only collected during the day and within periods of reasonable visibility. Each image in this dataset is an RGB video frame of 1920×1080 pixel resolution. For modeling, images have been normalized on the population level by subtracting the mean and dividing by the standard deviation for each channel. This reduces the range and therefore variance of the data, which we found to improve model performance during training. We used the DeepFish open-sourced code [8] as a starting point for replication and our own experimentation. As such, we use many of the same modeling settings used in [9]. Our semenatic segmentation model is an FCN network with a ResNet-50 backbone that has been pretrained on ImageNet, since this pretraining step greatly improved performance in [9]. Models are optimized using the Adam optimizer with a learning rate of 10^{-3}, the batch size is set to 2 due to memory constraints, and experiments are run between 5-20 epochs, depending on the experiment. In the original DeepFish paper, experiments were run over 1,000 epochs, but we were able to nearly replicate the original benchmark results within 20 epochs (results are shown in Figure 1), so we kept our experiment runs within that range. Validation set mIoU (mean intersection over union) is our main metric. In general, mIoU values of 0.5 (50%) or greater are considered to be good scores. In this paper we aim for mIoU scores of 0.8 or higher given the strong benchmark and our goal for scaling this approach.

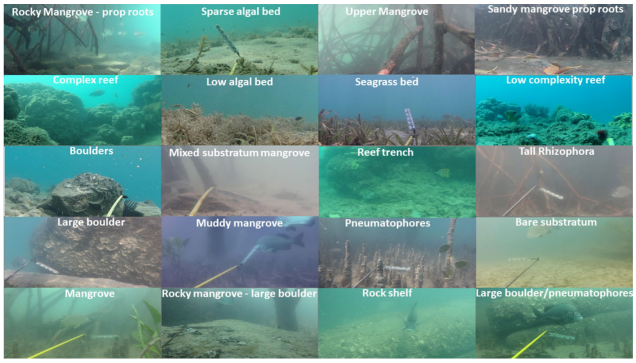

Figure 3. Sample images of each of the 20 habitats included in the DeepFish [9] dataset. A variety of habitats are represented here, with each containing unique fish species and displaying different environmental conditions. For instance, mangrove environments are distinguished from other habitats by their low visibility and the appearance of tree roots.

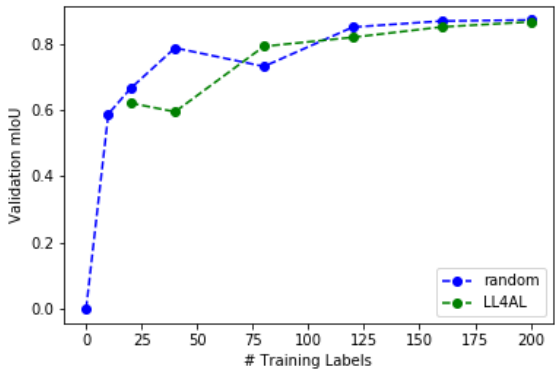

Figure 4. Active learning on benchmark training labels using LL4AL and random sampling. These methods show similar performance on this dataset. A random sampling strategy can nearly achieve 0.8 mIoU with only 40 training labels. Experiment results shown are averages over 5 trials.

4.2 Domain-shifted Dataset Creation. For the first set of active learning experiments, we use the same training and testing split used in [9], which includes all 20 habitats in each dataset, as we are not yet introducing a domain shift into the data. For experimenting with methods that address domain shifts, we construct a dataset that includes a domain shift. See Table 1 for details. In our domain-shifted dataset, the source domain is a reef habitat and the target domain is a mangrove habitat. The reef and mangrove habitats represent different domains because the environmental conditions between these two habitats are fairly different, with the mangrove environment containing visible tree roots and having much poorer visibility than the reef environment. In domain adaptation and fine-tuning experiments, the entire source dataset is used during training. Target data has been further split into training and validation sets, but target data is only used for training when applying active domain adaptation or fine-tuning.

Dataset	Habitats	Samples
Source	complex reef, low complexity reef, reef trench	144
Target	rocky mangrove - prop roots, sandy mangrove - prop roots, upper mangrove, mixed substratum mangrove, mangrove	179

Table 1. Details of the domain-shifted datasets we constructed.

4.3 Active Learning Experiments. The first question we set out to answer was: what is the minimum amount of ground-truth training labels that could be used to produce strong performance (0.8 mIoU) on the benchmark DeepFish dataset? We tested two methods on the original training dataset, LL4AL and random sampling. Figure 4 details the results. We ran experiments 5 times each and evaluated the average result across trials. It appears that 0.8 mIoU can nearly be achieved with as few as 40 training labels when using a random sampling strategy. Overall, LL4AL [1] appears to perform about as well as random sampling. We did not experiment with AADA in this section, as this approach only applies when the data contains a domain shift.

4.4 Unsupervised DA Experiments. Next we evaluated how the semantic segmentation model performs when a domain shift is introduced. Figure 5 shows the results of unsupervised DA with no target labels when there is a domain shift from a reef habitat to a mangrove habitat. As we increase the number of training epochs from 10-20 epochs, validation mIoU for the unsupervised DA method remains consistently higher than source-only training, but the size of the effect diminishes. In addition, although unsupervised DA improves validation set performance, it is still not able to reach an mIoU of 0.6, and therefore does not produce performance that would be strong enough to deploy in a real-world setting. Figure 6 provides a visual comparison of these results and confirms this intuition. While the unsupervised DA model seems to be less sensitive to misclassifying tree roots as fish when compared with the source-only method, it still struggles to identify fish effectively.

4.5 Active Domain Adaptation Experiments. We now shift our attention to jointly evaluating active learning and unsupervised DA. The question we aim to answer is: what is the minimum quantity of training labels needed to produce strong mIoU results (0.8 or higher) when shifting to a new domain?

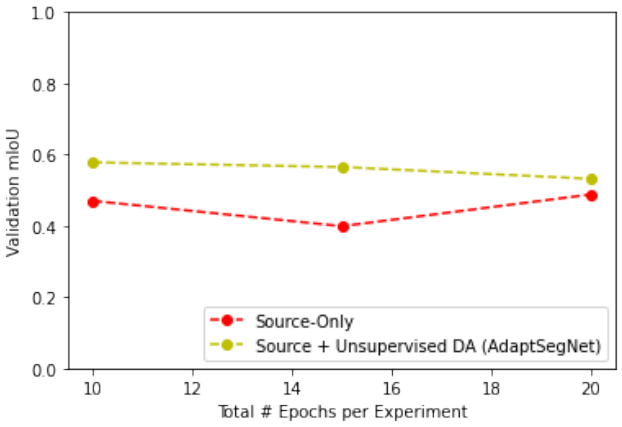

Figure 5. A comparison of training a semantic segmentation network with and without unsupervised domain adaptation (AdaptSegNet) as the number of training epochs varies.

Training Method	AL Strategy	Num Target Labels	mIoU
Source-only	None	None	0.40
Source + DA (AdaptSegNet)	None	None	**0.57**
Source + fine-tuning on target	Random	10	0.70
Source + DA (AdaptSegNet)	Random	10	**0.75**
Source + DA (AdaptSegNet)	AADA	10	0.63
Source + fine-tuning on target	Random	20	**0.81**
Source + DA (AdaptSegNet)	Random	20	0.80
Source + DA (AdaptSegNet)	AADA	20	0.66
Source + fine-tuning on target	Random	40	**0.83**
Source + DA (AdaptSegNet)	Random	40	0.79
Source + DA (AdaptSegNet)	AADA	40	0.80
Source + fine-tuning on target	Random	80	**0.90**
Source + DA (AdaptSegNet)	Random	80	0.89
Source + DA (AdaptSegNet)	AADA	80	**0.90**

Table 2. Reef → mangrove domain shift: results of training on images of reef habitats and validating on images of mangrove habitats. Each training component was run for 15 epochs.

Using previous active DA experimental setups as a guideline, [6, 12] we compare active DA approaches to a baseline of training on source and then training again (fine-tuning) on target with randomly-selected labels.

The training process for active DA is as follows:

- Iteratively train supervised semantic segmentation model on labeled source training set (144 labels) and run unsupervised DA (AdaptSegNet) [13] over 15 epochs. Train on labeled source only and run DA on both labeled source and unlabeled target training sets.
- Active learning step: apply a sampling strategy, either random sampling or AADA [12] to select a set of N training labels from the unlabeled target training set.
- Obtain N chosen labels for target training set.
- Train semantic segmentation model with new target set labels.

Results of active DA experiments are detailed in Table 2. All results are evaluated on the target validation set. We are able to achieve an mIoU of 0.81 when training with 20 target labels as well as all source labels and then randomly fine-tuning on the target. From these results, it appears that active DA does not outperform random fine-tuning on this task. In addition, random fine-tuning is the simplest method to implement. Figure 6 displays the results of the best models for each method. Interestingly, although random fine-tuning performs the best in terms of validation set mIoU, it does not seem to yield the best visual results.

4.6 Evaluating on the QUT dataset. Finally, we test the best-performing models on another fish dataset, QUT [2]. The results of this experiment are show in Figure 7. For this comparison, the benchmark model is our replication of the original DeepFish semantic segmentation model (trained with 310 samples). AADA and Random Fine-Tuning are both trained on the domain-shifted dataset (144 source samples and 80 target samples). The benchmark model appears to perform slightly better when fish are in a higher contrast setting and worse when fish are in a lower contrast setting. AADA and Random Fine-Tuning approaches have had to adjust to a domain shift within the hazy and low visibility mangrove environment, which may explain why these approaches could adapt better to low contrast settings. Additional out-of-sample validation would help strengthen this finding.

5. Discussion and Future Work

In this paper we explored active learning and domain adaptation as approaches for reducing the barriers to adopting fish segmentation in the wild. Through experimentation, we discovered that we only need about 13% of the original training dataset (40 labels v. 310 labels) to achieve strong validation set performance of 0.79 mIoU. In addition, when there is a domain shift, we can achieve 0.8 mIoU while training with a dataset that is about half of the size of the original training set (144 source labels and 20 target labels v. 310 labels). However, segmentation results visibly improve between 0.8 and 0.9 mIoU, and therefore it is likely preferable to collect slightly more labels to produce a more robust result. Random fine-tuning and AADA both achieve an mIoU of 0.90 with 224 total labels (144 source and 80 target), and appear to perform as well or even slightly better than the original benchmark model when evaluated on an outside dataset. AADA and random fine-tuning were trained in a more challenging, lower-visibility setting, which may be why these methods are slightly bet-

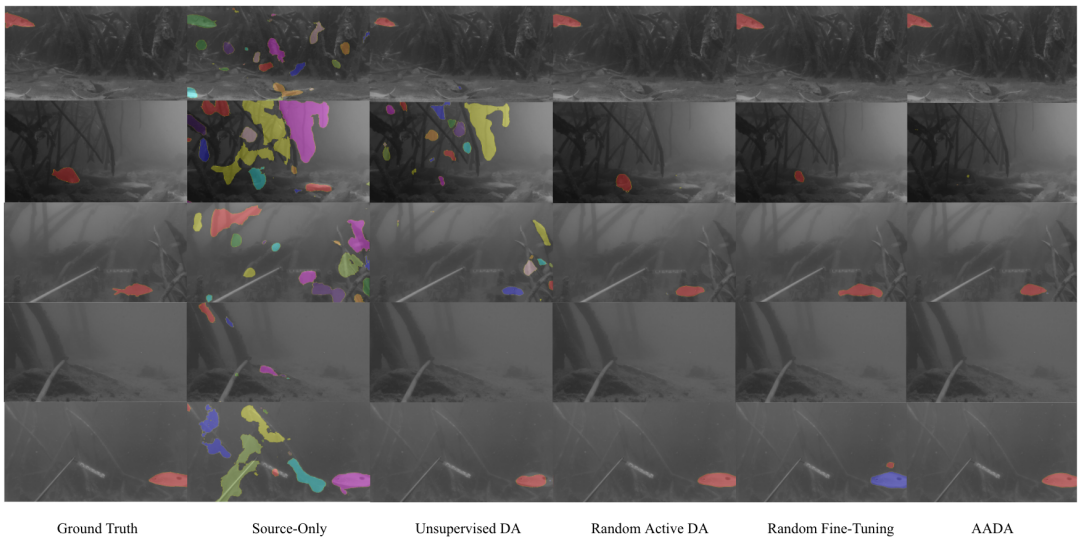

Figure 6. Results of various approaches when trained within a reef habitat and evaluated within a mangrove habitat. For each method, the best result in terms of validation mIoU is shown. See Table 2 for validation set results.

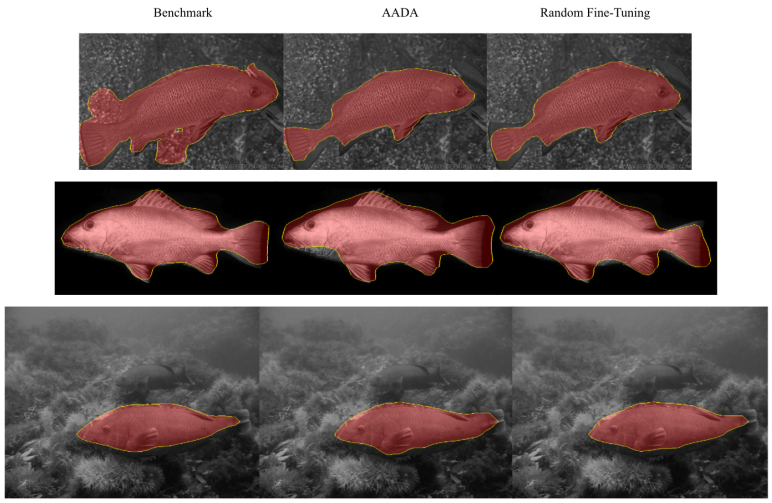

Figure 7. The best-performing models from our experiments when applied to another fish dataset, QUT [2].

ter at distinguishing fish from their background environments in low-contrast settings as compared to the benchmark model. Overall, active domain adaptation did not outperform random fine-tuning for this dataset. This could possibly be due to our implementation of AADA, or the dataset. We should note that, in the original publication, some of the AADA [12] experiments did not outperform random fine-tuning, and it appeared to be dependent on both the dataset and number of training labels. A future step we will take is to experiment on standard benchmark datasets for semantic segmentation, such as the GTA5 and Cityscapes datasets. In addition, we plan to extend more active DA methods to semantic segmentation, such as ADA-CLUE [6], and make changes to the AADA method to further optimize it to the semantic segmentation task.

6. Acknowledgements

Thank you to my research advisor, Alberto Sangiovanni-Vincentelli, my research mentor, Xiangyu Yue, and my research collaborator, Mao Li, for discussing many of these

ideas with me over the past few months. In addition, thank you to Professor Alyosha Efros and Professor Angjoo Kanazawa for the helpful research direction and feedback.

References

[1] superannotateai: Active learning algorithms for classification, object detection, human pose estimation and semantic segmentation. https://github.com/superannotateai/active_learning/, 2019. 4

[2] K. Anantharajah, Z. Ge, C. McCool, S. Denman, C. Fookes, P.I. Corke, D. Tjondronegoro, and S. Sridharan. Local inter-session variability modelling for object classification. *IEEE Winter Conference on Applications of Computer Vision*, page 309–316, 2014. 2, 5, 6

[3] FAO. The State of World Fisheries and Aquaculture 2020. 2020. 1

[4] Rafael Garcia, Ricard Prados, Josep Quintana, Alexander Tempelaar, Nuno Gracias, Shale Rosen, Håvard Vågstøl, and Kristoffer Løvall. Automatic segmentation of fish using deep learning with application to fish size measurement. *ICES Journal of Marine Science*, pages 1354–1366, 2019. 1

[5] Alberto Garcia-Garcia, Sergio Orts-Escolano, Sergiu Oprea, Victor Villena-Martinez, and José García Rodríguez. A review on deep learning techniques applied to semantic segmentation. *CoRR*, abs/1704.06857, 2017. 2

[6] 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. 2, 3, 5, 6

[7] Antonio Pusceddu, Silvia Bianchelli, Jacobo Martín, Pere Puig, Albert Palanques, Pere Masqué, and Roberto Danovaro. Chronic and intensive bottom trawling impairs deep-sea biodiversity and ecosystem functioning. *Proceedings of the National Academy of Sciences*, pages 8861–8866, 2014. 1

[8] Alzayat Saleh, Issam H. Laradji, Dmitry A. Konovalov, Michael Bradley, David Vazquez, and Marcus Sheaves. Deepfish. https://github.com/alzayats/DeepFish, 2020. 3

[9] Alzayat Saleh, Issam H. Laradji, Dmitry A. Konovalov, Michael Bradley, David Vazquez, and Marcus Sheaves. A realistic fish-habitat dataset to evaluate algorithms for underwater visual analysis. *Nature Scientific Reports*, 2020. 1, 2, 3, 4

[10] Burr Settles. Active learning literature survey. 2009. 2

[11] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):640–651, 2017. 2

[12] Jong-Chyi Su, Yi-Hsuan Tsai, Kihyuk Sohn, Buyu Liu, Subhransu Maji, and Manmohan Chandraker. Active adversarial domain adaptation. *CoRR*, abs/1904.07848, 2019. 2, 3, 5, 6

[13] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker. Learning to adapt structured output space for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 3, 5

[14] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, and Tao Qin. Generalizing to Unseen Domains: A Survey on Domain Generalization. *CoRR*, abs/2103.03097, 2021. 2

[15] Donggeun Yoo and In So Kweon. Learning loss for active learning. *CoRR*, abs/1905.03677, 2019. 2

[16] Beichen Zhang, Liang Li, Shijie Yang, Shuhui Wang, Zheng-Jun Zha, and Qingming Huang. State-relabeling adversarial active learning, 2020. 2

[17] 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. *CoRR*, abs/2101.10979, 2021. 2