Unsupervised Domain Adaptation in the Real World: A Case Study in Sonar Video

Justin Kay* MIT
Suzanne Stathatos Caltech
Siqi Deng† Amazon
Erik Young Trout Unlimited
Pietro Perona Caltech
Sara Beery MIT
Grant Van Horn UMass Amherst

Abstract

In real world applications of machine learning, adaptation to new domains (e.g. new regions, new populations, new sensors, or new points in time) has been shown to be an ongoing challenge. In *unsupervised domain adaptation*, the assumption is that the user has access to a large labeled set of source domain data, and the goal is to adapt to a new target domain without the use of any labeled target data. The open question is how unlabeled samples from the target domain should be incorporated into the model training process. In this work we document our experiences applying recently proposed unsupervised domain adaption techniques for object detection to a novel application domain: **counting fish in sonar video**. We find that: (i) prior works that show progress on standard domain adaptation benchmark datasets do not necessarily translate to our domain, (ii) validation methods are often unrealistic in these prior works, and (iii) higher complexity (in terms of implementation and parameters) techniques work better. We aim for this work to be a useful guide for other practitioners looking to use unsupervised domain adaptation techniques in real world applications.

1 Introduction

Object detection and tracking have been shown to work well on benchmark datasets across diverse real world applications [16, 27, 26, 4, 23, 21], but these models often struggle to generalize out of domain [1, 9, 12]. The effort and expense required to develop training datasets and train models for new domains is a significant bottleneck to the reliable deployment of these models at scale. One approach to address this is *unsupervised domain adaptation* (UDA). UDA methods derive a training signal from a collection of unlabeled target domain data in order to adapt a model trained on data from one distribution (a "source" domain) to a new data distribution (the "target" domain).

Are UDA methods ready for real-world deployment? In this work we explore the practical challenges of utilizing UDA techniques beyond the standard computer vision benchmarks for which they were designed. We use detecting, tracking, and counting fish in sonar data as a case study representing a non-standard domain with real-world impact in sustainability. Sonar imaging provides a non-invasive way to monitor escapement—the number of salmon returning home each season to spawn—which helps inform sustainable fisheries management and supporting UN SDGs 2 (Zero Hunger), 13 (Climate Action), and 14 (Life Below Water) [13]. Automation using computer vision could enable current sonar-based monitoring programs to scale from a few locations to entire watersheds,

*Corresponding author: kayj@mit.edu
†Work done outside of Amazon

Computational Sustainability Workshop at NeurIPS 2023.

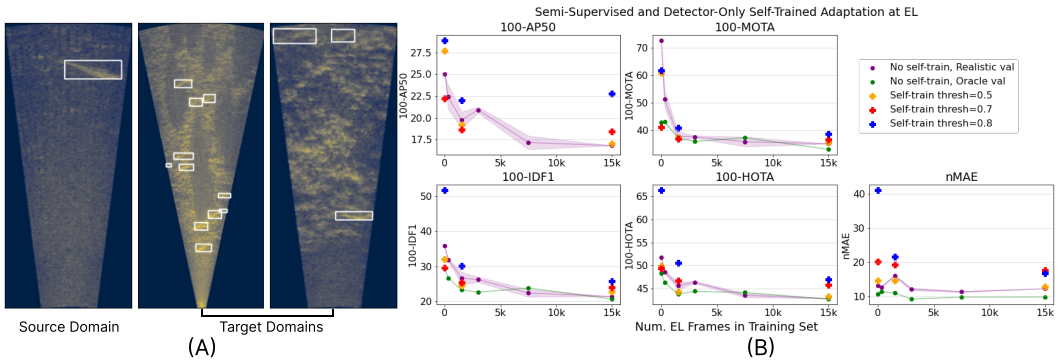

Figure 1: (A): Example imagery from the Caltech Fish Counting Domain Adaptation dataset. Models are trained on source imagery (left) and deployed on target imagery (right). (B): Purple and green lines show the effects of adding incrementally more target-domain supervision on detection (AP50), tracking (IDF1, MOTA, HOTA), and counting (nMAE) metrics. Green (oracle) uses the validation set to determine its optimal checkpoint, while purple uses the source set. All metrics are shown in terms of error, *i.e.* lower is better. The largest gains can be seen between 0 and 1500 additional annotated frames. The crosses indicate "vanilla" self-supervised performance after being trained on 2.6M unannotated images. We see there are there is no obvious best threshold, however, suboptimal settings can be quite harmful, since only one threshold shows an improvement for only one of the five metrics (red cross on MOTA).

yet computer vision methods for this task struggle in part due to domain shift caused by changing environmental conditions [9]. The problem of domain shift is faced in many applications, and our conclusions may have relevance to the broader community of computer vision practitioners.

We focus on a popular framework for UDA called *teacher-student self-training* [24, 29, 30, 14] due to its prevalence in the literature. In these approaches, a "teacher" model is used to label a set of (unlabeled) target-domain images. These images and machine labels are subsequently used to train a "student" model, thereby adapting the student model to the target domain. Self-training is intuitive and ostensibly simple to implement, making it an attractive UDA option for practitioners.

However in our investigation we find that these approaches are not as simple and effective as they are made out to be, and do not offer consistent benefit. Our results illuminate some surprising pitfalls and promises in the deployability of UDA for real-world applications:

1. Many self-training approaches are not effective outside of standard benchmarks. Specifically, we show that using fixed pseudo-labels is not an effective technique even when some supervised target-domain data is used.

2. Validation methods in UDA for object detection are unrealistic and over-optimistic. We show that UDA methods do not offer consistent improvements when using realistic validation methods.

3. Recent work is promising, at the cost of more complexity. We find some positive results from using recent extensions of the mean teacher framework [25].

2 Related Work

"Vanilla" self-training for UDA In self-training, a model is first trained on the source domain and then used to generate predictions on the target domain. Then, these predictions become "pseudo-labels" that are used as ground truth to train a new model for the target domain. In this work we call this approach "vanilla" self-training as it involves no modifications to model architectures. Vanilla self-training has been reported to yield more accurate models in semi-supervised learning [28] as well as in our focus area of UDA in object detection [22].

Teacher-student frameworks for UDA In teacher-student approaches, a "teacher" model and a "student" model are trained in a way that enforces consistency between the target-domain predictions from both models. A popular approach, *mean teacher* [25] uses an exponential moving average of the student weights as the teacher model. As opposed to vanilla self-training, where the teacher's pseudo-labels are fixed at the beginning of training, in mean teacher the teacher is frequently updated

and can produce different pseudo-labels every epoch. Most recent state-of-the-art UDA methods in object detection are based on the mean teacher framework [15, 8, 6, 5].

Validation methods and hyperparameter selection in UDA Prior work notes that unrealistic methods for selecting best models and hyperparameters have led to over-optimistic results in UDA for image classification [18, 19]. This remains a problem for UDA in object detection, despite attempts to automatically select optimal confidence thresholds or to minimize prediction entropy [22].

3 Dataset and metrics

The Caltech Fish Counting Dataset (CFC) [9] contains over 1,500 videos where the goal is to detect, track, and count moving fish in low signal-to-noise sonar video. The visual conditions encountered in this application are starkly different from existing benchmarks in domain adaptive object detection (see Fig. 1A), enabling us to test whether existing algorithms translate to new applications.

We introduce an extension to the CFC dataset, deemed Caltech Fish Counting Domain Adaptation (CFC-DA), to allow for the study of using unsupervised and semi-supervised domain adaptation to improve OOD performance. Specifically, we curated 2.6M unsupervised video frames and 15,000 supervised video frames from the Elwha ("EL") test location in CFC. Supervision consists of multi-object tracking annotations where each target is enclosed in a bounding box and maintains a constant unique ID within each clip. This data is sourced from the same time period (July 2018) and the same camera hardware as the original test set.

Our problem involves counting fish, which we do through a detection and tracking framework. We monitor the following metrics: AP50 (detection); MOTA [2], HOTA [17], and IDF1 [20] (tracking); and nMAE (counting) [9].

4 Experimental results

Validation settings. We compare two approaches: *"realistic" validation*, where hyperparameters and early stopping points are chosen based on model performance on a source-domain validation set, and *"oracle" validation*, where parameters are chosen based on a target-domain test set. We emphasize that *in true UDA settings, oracle validation is not possible*.

Incremental supervision as upper bounds. What if we could collect some target domain annotations? As is common in the UDA literature we consider this an upper bound for what unsupervised methods could achieve. We test incremental levels of supervision by progressively sampling up to 50 *annotated* target-domain video clips, 300 frames each, and train on the union of these clips and the CFC source-domain training set. In Fig. 1B we show the effect of this incremental supervision on detection, tracking, and counting performance when using 1, 5, 10, 25, and 50 annotated video clips. We compare using both realistic (purple) and oracle (green) validation settings, where we choose detector settings based on AP50 and tracker settings based on nMAE.

Our key takeaways are: (1) For object detection and tracking, adding annotated frames from the target domain improves performance roughly monotonically, with diminishing returns after a few thousand labeled images; (2) Validating on target-domain data has a larger effect on performance than training on target-domain data. For example, oracle validation with 0 supervised target-domain frames achieves better nMAE than realistic validation with 14k supervised target domain frames. This is worth emphasizing: *even if perfect target-domain labels are available during self-training (e.g., from a perfect teacher model), this still may not lead to improved performance under realistic (source-only) validation conditions*. Therefore, if possible, practitioners should allocate effort toward creating target-domain validation sets instead of solely focusing on training data.

Detector-only self-training Next we investigate the efficacy of vanilla self-training on the CFC-DA dataset as a method of closing the performance gap between source-supervised and target-supervised models. We run inference using our baseline detector (the "teacher") on the unsupervised portion of CFC-DA from the target domain (2.6M frames), and use all detections with a confidence score $s \geq \alpha$ as pseudo-labels for self-training, where α is a hyperparameter that we sweep from 0.5 to 0.8. We then train a second detector (the "student") on the union of the CFC training set and the target-domain pseudo-labels. We also test self-training in the semi-supervised case. Specifically, we train several teacher models using various amounts of target-domain data. Each teacher predicts pseudo-labels on the same 2.6M image dataset, and self-training proceeds as normal.

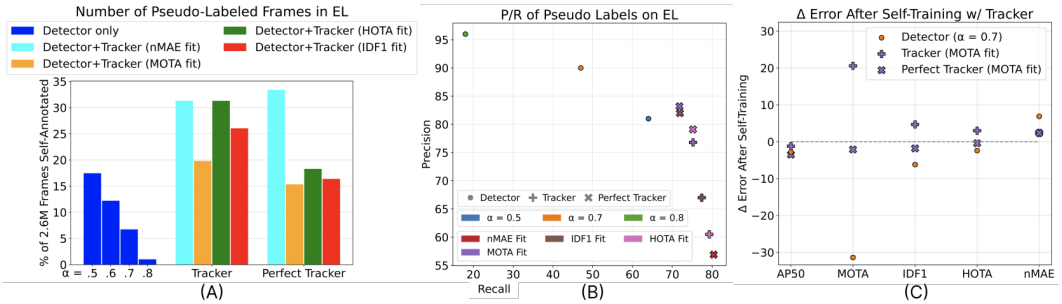

Figure 2: **Pseudo-label refinement using an object tracker introduces additional hyperparameters that are nontrivial to set.** Depending on which metric the tracker is optimized for, the resulting pseudo-labels vary greatly in quantity (A) and quality (B). We found that fitting tracker parameters to validation MOTA gave the best results. However, disappointingly, when using realistic validation adding a tracker made results worse on all metrics except AP50, where improvement was marginal (C).

We show results from detector-only self-training in Fig. 1B. Our key takeaways are: (1) Vanilla self-training does not consistently lead to performance improvements, and can dramatically hurt performance if the wrong α is chosen. (2) First training on target-domain data, and then pseudo-labeling (all points in Fig. 1B where Num EL Frames > 0) is not guaranteed to lead to effective self-training either. (3) Annotating about 1000 target-domain images by hand is better than self-training on millions of images using a source-only teacher.

A number of solutions have been proposed to address the shortcomings of vanilla self-training [22, 10, 22, 11, 11, 31]. We investigate one such approach in the next section.

Pseudo-label refinement via tracking. CFC-DA contains video frames, offering the ability to use temporal information to improve pseudo-labels. We investigate using an object tracker [3] to do so, as in [22]. We run object tracking on the pseudo-label outputs and use the boxes from the final tracks as pseudo-labels. Intuitively, tracking can help remove spurious false positives as well as recover false negatives; however we find that small changes to the tracker hyperparameters can lead to big changes in the quantity (Fig. 2A) and quality (Fig. 2B) of the resulting pseudo-labels. This again poses the challenge of how to properly set these hyperparameters. We show in Fig. 2C the effect of tuning the tracker hyperparameters to the MOTA metric with source validation data (realistic) and target validation data (oracle). Under realistic validation, adding a tracker resulted in worse performance than detector-only self-training for four out of five metrics. Under oracle validation, refinement with a tracker still resulted in worse performance than detector-only self-training for all metrics except AP50. For AP50, the improvement was marginal (around 1%). The takeaway: filtering psuedo-labels using a tracker is not advantageous.

Preliminary results from more complex approaches. After finding vanilla self-training (with and without a tracker) to be ineffective, we tested mean teacher for object detection [25] using the Adaptive Teacher [15, 7] implementation. We upgraded the codebase to use a more modern backbone and training recipe, though we note that this required using a different detector architecture than our previous experiments (Faster R-CNN vs. YOLOv5). We find that compared to a two-stage baseline source-only model, which achieved an AP50 of 75.8 on EL, mean teacher self-training achieved AP50=79.8 in the realistic validation setting and AP50=80.6 in the oracle validation setting. Though there is still a performance cost due to the validation constraints of real-world settings, these experiments demonstrate that improvements are possible and potentially promising under the realistic validation setting.

5 Conclusion

This work focused on a practical question: how should developers of computer vision models adapt their models to changing data distributions in real-world applications? We tried the simple and obvious approaches first—*i.e.*, those that required no modification to underlying architectures—as these are appealing to practitioners for their ease of implementation. Unfortunately, we find these "simple" approaches actually require careful tuning of parameters that have not been fully explored or transparently discussed in prior work. A key takeaway from our collection of negative results is that

validation methods will be critical if UDA algorithms are to have real-world impact. Our experiments emphasize that the ability to estimate our performance on target-domain data is as important as the adaptation algorithm itself.

In short, for our problem, we found vanilla self-training to be a dead-end even when used in combination with additional target-domain data and pseudo-label refinement strategies. However, we have observed promising initial results using more complex mean teacher approaches. But these techniques still involve a collection of hyperparameters and design choices that must be appropriately set when moving to a new application domain. Future work in this space must involve rigorous and thorough analysis of these approaches to confirm our initial findings.

References

[1] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018.

[2] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008.

[3] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. IEEE, 2016.

[4] Elizabeth Bondi, Fei Fang, Mark Hamilton, Debarun Kar, Donnabell Dmello, Jongmoo Choi, Robert Hannaford, Arvind Iyer, Lucas Joppa, Milind Tambe, et al. Spot poachers in action: Augmenting conservation drones with automatic detection in near real time. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[5] Meilin Chen, Weijie Chen, Shicai Yang, Jie Song, Xinchao Wang, Lei Zhang, Yunfeng Yan, Donglian Qi, Yueting Zhuang, Di Xie, et al. Learning domain adaptive object detection with probabilistic teacher. *arXiv preprint arXiv:2206.06293*, 2022.

[6] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4091–4101, 2021.

[7] Ross Girshick, Ilija Radosavovic, Georgia Gkioxari, Piotrár Doll, and Kaiming He. Detectron. `https://github.com/facebookresearch/detectron`, 2018.

[8] Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. Mic: Masked image consistency for context-enhanced domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11721–11732, 2023.

[9] Justin Kay, Peter Kulits, Suzanne Stathatos, Siqi Deng, Erik Young, Sara Beery, Grant Van Horn, and Pietro Perona. The caltech fish counting dataset: A benchmark for multiple-object tracking and counting. In *European Conference on Computer Vision*, pages 290–311. Springer, 2022.

[10] Mehran Khodabandeh, Arash Vahdat, Mani Ranjbar, and William G Macready. A robust learning approach to domain adaptive object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 480–490, 2019.

[11] Seunghyeon Kim, Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6092–6101, 2019.

[12] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

[13] Bandy X Lee, Finn Kjaerulf, Shannon Turner, Larry Cohen, Peter D Donnelly, Robert Muggah, Rachel Davis, Anna Realini, Berit Kieselbach, Lori Snyder MacGregor, et al. Transforming our world: implementing the 2030 agenda through sustainable development goal indicators. *Journal of public health policy*, 37:13–31, 2016.

[14] Hengduo Li, Zuxuan Wu, Abhinav Shrivastava, and Larry S Davis. Rethinking pseudo labels for semi-supervised object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1314–1322, 2022.

[15] Yu-Jhe Li, Xiaoliang Dai, Chih-Yao Ma, Yen-Cheng Liu, Kan Chen, Bichen Wu, Zijian He, Kris Kitani, and Peter Vajda. Cross-domain adaptive teacher for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7581–7590, 2022.

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[17] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International journal of computer vision*, 129(2):548–578, 2021.

[18] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. Unsupervised domain adaptation: A reality check. *arXiv preprint arXiv:2111.15672*, 2021.

[19] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. Benchmarking validation methods for unsupervised domain adaptation. *arXiv preprint arXiv:2208.07360*, 2022.

[20] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*, pages 17–35. Springer, 2016.

[21] Mikel Rodriguez, Ivan Laptev, Josef Sivic, and Jean-Yves Audibert. Density-aware person detection and tracking in crowds. In *2011 International Conference on Computer Vision*, pages 2423–2430. IEEE, 2011.

[22] Aruni RoyChowdhury, Prithvijit Chakrabarty, Ashish Singh, SouYoung Jin, Huaizu Jiang, Liangliang Cao, and Erik Learned-Miller. Automatic adaptation of object detectors to new domains using self-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 780–790, 2019.

[23] Stefan Schneider, Graham W Taylor, and Stefan Kremer. Deep learning object detection methods for ecological camera trap data. In *2018 15th Conference on computer and robot vision (CRV)*, pages 321–328. IEEE, 2018.

[24] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*, 2020.

[25] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

[26] Ben G Weinstein, Lindsey Gardner, Vienna Saccomanno, Ashley Steinkraus, Andrew Ortega, Kristen Brush, Glenda Yenni, Ann E McKellar, Rowan Converse, Christopher Lippitt, et al. A general deep learning model for bird detection in high resolution airborne imagery. *bioRxiv*, 2021.

[27] Ben G Weinstein, Sarah J Graves, Sergio Marconi, Aditya Singh, Alina Zare, Dylan Stewart, Stephanie A Bohlman, and Ethan P White. A benchmark dataset for canopy crown detection and delineation in co-registered airborne rgb, lidar and hyperspectral imagery from the national ecological observation network. *PLoS computational biology*, 17(7):e1009180, 2021.

[28] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020.

[29] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3060–3069, 2021.

[30] Qize Yang, Xihan Wei, Biao Wang, Xian-Sheng Hua, and Lei Zhang. Interactive self-training with mean teachers for semi-supervised object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5941–5950, 2021.

[31] Fuxun Yu, Di Wang, Yinpeng Chen, Nikolaos Karianakis, Tong Shen, Pei Yu, Dimitrios Lymberopoulos, Sidi Lu, Weisong Shi, and Xiang Chen. Unsupervised domain adaptation for object detection via cross-domain semi-supervised learning. *arXiv preprint arXiv:1911.07158*, 2019.