A Reproducible and Realistic Evaluation of Partial Domain Adaptation Methods

Anonymous authors
Paper under double-blind review

Abstract

Unsupervised Domain Adaptation (UDA) aims at classifying unlabeled target images leveraging source labeled ones. In the case of an extreme label shift scenario between the source and target domains, where we have extra source classes not present in the target domain, the UDA problem becomes a harder problem called Partial Domain Adaptation (PDA). While different methods have been developed to solve the PDA problem, most successful algorithms use model selection strategies that rely on target labels to find the best hyper-parameters and/or models along training. These strategies violate the main assumption in PDA: only unlabeled target domain samples are available. In addition, there are also experimental inconsistencies between developed methods - different architectures, hyper-parameter tuning, number of runs - yielding unfair comparisons. The main goal of this work is to provide a realistic evaluation of PDA methods under different model selection strategies and a consistent evaluation protocol. We evaluate 7 state-of-the-art PDA algorithms on 2 different real-world datasets using 7 different model selection strategies. Our two main findings are: *(i)* without target labels for model selection, the accuracy of the methods decreases up to 30 percentage points; *(ii)* only one method and model selection pair performs well on both datasets. Experiments were performed with our PyTorch framework, BenchmarkPDA, which we open source.

1 Introduction

Domain adaptation. Deep neural networks are highly successful in image recognition for in-distribution samples (He et al., 2016) with this success being intrinsically tied to the large number of labeled training data. However, they tend to not generalize as well on images with different backgrounds or colors not seen during training. Such shift in the samples is referred to as domain shift in the literature. Unfortunately, enriching the training set with new samples from different domains is challenging as labeling data is both an expensive and time-consuming task. Thus, researchers have focused on unsupervised domain adaptation (UDA) where we have access to unlabelled samples from a different domain, known as the target domain. The purpose of UDA is to classify these unlabeled samples by leveraging the knowledge given by the labeled samples from the source domain (Pan & Yang, 2010; Patel et al., 2015). In the standard UDA problem, the source and target domains are assumed to share the same classes. In this paper, we consider a more challenging variant of the problem called *partial domain adaptation* (PDA): the classes in the target domain $\mathcal{Y}_t$ form a subset of the classes in the source domain $\mathcal{Y}_s$ (Cao et al., 2018), *i.e.,* $\mathcal{Y}_t \subset \mathcal{Y}_s$. The number of target classes is unknown as we do not have access to the labels. The extra source classes, not present in the target domain, make the PDA problem more difficult: simply aligning the source and target domains forces a negative transfer where target samples are matched to outlier source-only labels.

Realistic evaluations. Most recent PDA methods report an increase of the target accuracy up to 15 percentage points on average when compared to the baseline approach that is trained only on source samples. While these successes constitute important breakthroughs in the DA research literature, target labels are used for model selection, violating the main UDA assumption. In their absence, the effectiveness of PDA methods remains unclear and model selection still constitutes an open problem as we show in this work. Moreover, the hyper-parameter tuning is either unknown or lacks details and sometimes requires labeled target data, which

Dataset	Model Selection	S. ONLY	PADA	SAFN	BA3US	AR	JUMBOT	MPOT
OFFICE-HOME	Worst (w/o target labels)	59.55 (-2.31)	52.72 (-11.00)	61.37 (-1.93)	62.25 (-13.73)	64.32 (-8.42)	61.28 (-15.87)	46.92 (-30.38)
	Best (w/o target labels)	60.73 (-1.14)	63.08 (-0.64)	62.59 (-0.71)	75.37 (-0.61)	70.58 (-2.16)	74.61 (-2.54)	66.24 (-11.07)
	ORACLE	61.87	63.72	63.30	75.98	72.73	77.15	77.31
VISDA	Worst (w/o target labels)	55.02 (-4.46)	32.32 (-22.26)	42.83 (-19.81)	51.07 (-16.60)	55.69 (-18.15)	59.86 (-24.15)	61.62 (-25.33)
	Best (w/o target labels)	55.24 (-4.24)	56.83 (2.26)	58.62 (-4.02)	65.58 (-2.09)	67.20 (-6.65)	77.69 (-6.31)	78.40 (-8.54)
	ORACLE	59.48	54.57	62.64	67.67	73.85	84.01	86.95

Table 1: Task accuracy average computed over three different seeds (2020, 2021, 2022) on Partial OFFICE-HOME and Partial-VISDA. For each dataset and PDA method, we display the results of the *worst and best performing model selection that do not use target labels* as well as the ORACLE model selection strategy. All results can be found in Table 6.

makes it challenging to apply PDA methods to new datasets. Recent work has highlighted the importance of model selection in the presence of domain shift. Gulrajani & Lopez-Paz (2021) showed that when evaluating domain generalization (DG) algorithms, whose goal is to generalize to a completely unseen domain, in a consistent and realistic setting no method outperforms the baseline ERM method by more than 1 percentage point. They argue that DG methods without a model selection strategy remain incomplete and should therefore be specified as part of the method. A similar recommendation was done by Saito et al. (2021) for domain adaptation.

PDA methods have been designed using target labels at test time to select the best models. Related work (Saito et al., 2021; You et al., 2019) on model selection strategies for domain adaptation claimed to select the best models without using target labels. However, a realistic empirical study of these strategies in PDA is still lacking. In this work, we conduct extensive experiments to study the impact of model selection strategies on the performance of partial domain adaptation methods. We evaluate 7 different PDA methods over 7 different model selection strategies, 4 of which do not use target labels, and 2 different datasets under the same experimental protocol for a fair comparison. We list below our major findings:

- The accuracy attained by models selected without target labels can decrease up to 30 percentage points compared to the one reported using target labels (See Table 1 for a summary of results).
- Only 1 pair of PDA methods and target label-free model selection strategies achieve comparable accuracies to when target labels are used, while still improving over a source only baseline.
- Random seed plays an important role in the selection of hyper-parameters. Selected parameters are not stable across different seeds and the standard deviation between accuracies on the same task can be up to 8.4% even when relying on target labels for model selection.
- Under a more realistic scenario where some target labels are available, 100 random samples is enough to see only a drop of 1 percentage point in accuracy (when compared to using all target samples). However, the extreme case of using only one labeled target sample per class leads to a significant drop in performance.

Outline. In Section 2, we provide an overview of the different model selection strategies considered in this work. Then in Section 3, we discuss the PDA methods that we consider. In Section 4 we describe the training procedures, hyper-parameter tuning and evaluation protocols used to evaluate all methods fairly. In Section 5, we discuss the results of the different benchmarked methods and the performance of the different model selection strategies. Finally in Section 6, we give some recommendations for future work in partial domain adaptation.

2 Model Selection Strategies: An Overview

Model selection (choosing hyper-parameters, training checkpoints, neural network architectures) is a crucial part of training neural networks. In the supervised learning setting, a validation set is used to estimate the

model's accuracy. However, in UDA such approach is not possible as we have unlabeled target samples. Several strategies have been designed to address this issue. Below, we discuss the ones used in this work.

Source Accuracy (s-acc). Ganin & Lempitsky (2015) used the accuracy estimated on a small validation set from the source domain to perform the model selection. While the source and target accuracies are related, there are no theoretical guarantees. You et al. (2019) showed that when the domain gap is large this approach fails to select competitive models.

Deep Embedded Validation (dev). Sugiyama et al. (2007) and Long et al. (2018) perform model selection through Importance-Weighted Cross-Validation (IWCV). Under the assumption that the source and target domain follow a covariate shift, the target risk can be estimated from the source risk through importance weights that give increased importance to source samples that are closer to target samples. These importance weights correspond to the ratio of the target and source densities and are estimated using Gaussian kernels. Recently, You et al. (2019) proposed an improved variant, Deep Embedded Validation (DEV), that controls the variance of the estimator and estimates the importance weights with a discriminative model that distinguish source samples from target samples leading to a more stable and effective method.

Entropy (ent). While minimizing the entropy of the target samples has been used in domain adaptation to improve accuracy by promoting tighter clusters, Morerio et al. (2018) showed that it can also be used for model selection. The intuition is that a lower entropy model corresponds to a highly confident model with discriminative target features and therefore reliable predictions.

Soft Neighborhood Density (snd). Saito et al. (2021) argue that a good UDA model will have a cluster structure where nearby target samples are in the same class. They claim that entropy is not able to capture this property and propose the Soft Neighborhood Density (SND) score to address it.

Target Accuracy (oracle). We consider as well the target accuracy on all target samples. While we emphasize once again its use is not realistic in unsupervised domain adaptation (hence why we will refer to it as ORACLE), it has nonetheless been used to report the best accuracy achieved by the model along training in several previous works (Cao et al., 2018; Xu et al., 2019; Jian et al., 2020; Gu et al., 2021; Nguyen et al., 2022). Here, we use it as an upper bound for all the other model selection strategies and to check the reproducibility of previous works.

Small Labeled Target Set (1-shot and 100-rnd). For real-world applications in an industry setting, it is unlikely that a model will be deployed without an estimation of its performance on the target domain. Therefore, one can imagine a situation where a PDA method is used and a small set of target samples is available. Thus, we will compute the target accuracy with 1 labeled sample per class (1-SHOT) and 100 random labeled target samples (100-RND) as model selection strategies. One could argue that the 100 random samples could have been used in the training with semi-supervised domain adaptation methods. However, note that we do not know how many classes we have on the target domain so it is hard to form a split when we have uncertainty of classes. For instance, 100-RND represents possibly less than 2 samples per class for one of our real-world dataset, as we do not know the number of classes, making a potential split between a train and validation target sets not possible.

3 Partial Domain Adaptation Methods

In this section, we give a brief description of the PDA methods considered in our study. They can be grouped into two families: adversarial training and divergence minimization.

Adversarial training. To solve the UDA problem, Ganin et al. (2016) aligned the source and target domains with the help of a domain discriminator trained adversarially to be able to distinguish the samples from the two domains. However, when applied to the PDA problem this strategy leads to negative transfer and the model performs worse than a model trained only on source data. Cao et al. (2018) proposed PADA that introduces a PDA specific solution to adversarial domain adaptation: the contribution of the source-only class samples to the training of both the source classifier and the domain adversarial network is decreased. This is achieved through class weights that are calculated by simply averaging the classifier prediction on all target samples. As the source-only classes should not be predicted in the target domain, they should have

PDA Methods	PADA, SAFN, BA3US AR, JUMBOT, MPOT
Model Selection Strategies	S-ACC, ENT, DEV, SND, 1-SHOT, 100-RND, ORACLE
Architecture	ResNet50 backbone $\oplus$ linear bottleneck $\oplus$ linear classification head
Experimental protocol	3 seeds on the 12 tasks of OFFICE-HOME and 2 tasks of VISDA

Table 2: Summary of our considered methods, model selection strategies, architecture and datasets.

Method	Architecture (bottleneck)	Runs per task	Model Selection	
			Hyper-Parameters	Along Training
PADA	Linear	1	IWCV (lacks details)	ORACLE
SAFN	Non-Linear	3	Unknown	ORACLE
BA3US	Linear	3	Unknown	ORACLE
AR	Non-Linear	1	IWCV (lacks details)	ORACLE
JUMBOT	Linear	1	ORACLE	FINAL
MPOT	Linear	3	Unknown	ORACLE

Table 3: Summary of the experimental protocol used for SOTA partial domain adaptation methods. We refer to Appendix A.1 for additional details.

lower weights. More recently, Jian et al. (2020) proposed BA3US which augments the target mini-batch with source samples to transform the PDA problem into a vanilla DA problem. In addition, an adaptive weighted entropy objective is used to encourage incorrect classes to have uniform and low prediction scores.

Divergence minimization. Another standard direction to align the source and target distributions in the feature space of a neural network is to minimize a given divergence between distributions of domains. Xu et al. (2019) empirically found that target samples have low feature norm compared to source samples. Based on this insight, they proposed SAFN which progressively adapts the feature norms of the two domains by minimizing the Maximum Mean Feature Norm Discrepancy (Gretton et al., 2012). Other approaches are based on optimal transport (OT) (Bhushan Damodaran et al., 2018) with mini-batches (Peyré & Cuturi, 2019; Fatras et al., 2020; 2021b). For the PDA problem in specific, (Fatras et al., 2021a) developed JUMBOT, a mini-batch unbalanced optimal transport that learns a joint distribution of the embedded samples and labels. The use of unbalanced OT is critical for the PDA problem as it allows to transport only a portion of the mass limiting the negative transfer between distributions. Based on this work, (Nguyen et al., 2022) investigated the partial OT variant (Chapel et al., 2020), a particular case of unbalanced OT, proposing M-POT. Finally, another line of work is to use the Kantorovich-Rubenstein duality of optimal transport to perform the alignment similarly to WGAN (Arjovsky et al., 2017). This is precisely the work of Gu et al. (2021) that proposed, AR. In addition, source samples are reweighted in order to reduce the negative transfer from the source-only class samples. The Kantorovich-Rubenstein duality relies on a one Lipschitz function which is approximated using adversarial training like the PDA methods described above.

4 Experimental Protocol

In this section, we discuss our choices regarding the training details, datasets and neural network architecture. We then discuss the hyper-parameter tuning used in this work. We summarize the PDA methods, model selection strategies and experimental protocol used in this work in Table 2. The main differences in the experimental protocol of the different published state-of-the-art (SOTA) methods is summarized in Table 3. To perform our experiments we developed a PyTorch (Paszke et al., 2019) framework: BenchmarkPDA. We make it available for researchers to use and contribute with new algorithms and model selection strategies:

`https://anonymous.4open.science/r/BenchmarkPDA-7F73`

It is the standard in the literature when proposing a new method to report directly the results of its competitors from the original papers (Cao et al., 2018; Xu et al., 2019; Jian et al., 2020; Gu et al., 2021;

METHOD	A2C	A2P	A2R	C2A	C2P	C2R	P2A	P2C	P2R	R2A	R2C	R2P	Avg
S. ONLY†	46.33	67.51	75.87	59.14	59.94	62.73	58.22	41.79	74.88	67.40	48.18	74.17	61.35
S. ONLY (Ours)	45.43	68.91	79.53	55.59	57.42	65.23	59.32	40.80	75.80	69.88	47.20	77.31	61.87
PADA†	51.95	67.00	78.74	52.16	53.78	59.03	52.61	43.22	78.79	73.73	56.60	77.09	62.06
PADA (Ours)	50.53	67.45	80.14	57.30	54.47	64.55	61.07	40.94	79.55	73.09	54.63	80.93	63.72
SAFN†*	58.93	76.25	81.42	70.43	72.97	77.78	72.36	55.34	80.40	75.81	60.42	79.92	71.84
SAFN* (Ours)	59.98	79.85	85.18	72.02	73.73	78.54	76.09	59.32	83.25	80.04	64.20	84.44	74.72
SAFN (Ours)	49.57	68.55	78.26	57.91	59.29	66.81	59.87	45.29	75.98	69.08	51.68	77.29	63.30
BA3US†	60.62	83.16	88.39	71.75	72.79	83.40	75.45	61.59	86.53	79.25	62.80	86.05	75.98
BA3US (Ours)	63.26	82.75	89.16	69.91	71.93	77.58	75.73	59.94	86.89	80.93	66.77	86.93	75.98
AR†*	62.13	79.22	89.12	73.92	75.57	84.37	78.42	61.91	87.85	82.19	65.37	85.27	77.11
AR* (Ours)	62.75	81.55	89.07	71.63	73.41	82.94	75.88	61.03	85.70	79.86	62.93	85.30	76.00
AR (Ours)	57.33	79.61	86.31	69.45	71.88	79.94	70.28	53.57	83.78	77.26	59.68	83.72	72.73
JUMBOT†	62.70	77.50	84.40	76.00	73.30	80.50	74.70	60.80	85.10	80.20	66.50	83.90	75.47
JUMBOT (Ours)	61.87	78.19	88.11	77.69	76.75	84.15	76.83	63.72	84.80	81.79	64.70	87.17	77.15
MPOT†	64.60	80.62	87.17	76.43	77.61	83.58	77.07	63.74	87.63	81.42	68.50	87.38	77.98
MPOT (Ours)	64.48	80.88	86.78	76.22	77.95	82.59	75.18	64.60	84.87	80.59	67.04	86.52	77.31

Table 4: Comparison between reported (†) accuracies on partial OFFICE-HOME from published methods with our implementation using the ORACLE model selection strategy. * denotes different bottleneck architectures.

Nguyen et al., 2022). As a result some methods differ for instance in the neural network architecture implementation (AR (Gu et al., 2021), SAFN (Xu et al., 2019)) or evaluation protocol JUMBOT (Fatras et al., 2021a) with other methods. These changes often contribute to an increased performance of the newly proposed method leaving previous methods at a disadvantage. Therefore we chose to implement all methods with the same commonly used neural network architecture, optimizer, learning rate schedule and evaluation protocol. We discuss the details below.

4.1 Methods, Datasets, Training and Evaluation Details

Methods. We implemented 7 PDA methods by adapting the code from the Official GitHub repositories of each method: Source Only, PADA (Cao et al., 2018), SAFN (Xu et al., 2019), BA3US (Jian et al., 2020), AR (Gu et al., 2021), JUMBOT (Fatras et al., 2021a), MPOT (Nguyen et al., 2022). We provide the links to the different official repositories in Appendix A.1. A comparison with previous reported results can be found in Table 4 and we postpone the discussion to Section 5.

Datasets. We consider two standard real-world datasets used in DA. Our first dataset is OFFICE-HOME (Venkateswara et al., 2017). It is a difficult dataset for unsupervised domain adaptation (UDA), it has 15,500 images from four different domains: Art (A), Clipart (C), Product (P) and Real-World (R). For each domain, the dataset contains images of 65 object categories that are common in office and home scenarios. For the partial OFFICE-HOME setting, we follow Cao et al. (2018) and select the first 25 categories (in alphabetic order) in each domain as a partial target domain. We evaluate all methods in all 12 adaptation scenarios. VISDA (Peng et al., 2017) is a large-scale dataset for UDA. It has 152,397 synthetic images and 55,388 real-world images, where 12 object categories are shared by these two domains. For the partial VISDA setting, we follow Cao et al. (2018) and select the first 6 categories, taken in alphabetic order, in each domain as a partial target domain. We evaluate the models in the two possible scenarios. We highlight that we are the first to investigate the performance of JUMBOT and MPOT on partial VISDA.

Model Selection Strategies We consider the 7 different strategies for model selection described in Section 2: S-ACC, DEV, ENT, SND, ORACLE, 1-SHOT, 100-RND. We use them both for hyper-parameter tuning as well selecting the best model along training. Since S-ACC, DEV and SND require a source validation set, we divide the source samples into a training subset (80%) and validation subset (20%). Regardless of the model

Dataset	Variant	BA3US			JUMBOT			MPOT			SAFN		
		ENT	DEV	SND	ENT	DEV	SND	ENT	DEV	SND	ENT	DEV	SND
OFFICE-HOME	Naive	52.60	**63.10**	44.48	52.30	26.75	17.67	**49.01**	16.72	**30.63**	32.12	**49.67**	5.01
	Heuristic	**58.45**	**63.10**	**60.96**	**56.24**	**45.79**	**55.16**	**49.01**	**45.61**	**30.63**	**46.27**	**49.67**	**49.67**
VISDA	Naive	39.06	**36.99**	1.14	35.89	**54.53**	11.99	**75.04**	**55.33**	36.11	**52.82**	**53.26**	0.83
	Heuristic	**67.50**	34.94	**38.76**	**47.23**	**54.53**	**66.42**	**75.04**	**55.33**	**85.36**	**52.82**	**53.26**	**52.82**

Table 5: Comparison between the naive model selection strategy and our heuristic approach. Accuracy on AC task for OFFICE-HOME and SR task for VISDA. Best results in **bold**.

selection strategy used, all methods are trained using the source training subset. This is in contrast with previous work that uses all source samples, but necessary to ensure a fair comparison of the model selection strategies. We refer to Appendix A.2 for additional details.

Architecture. Our network is composed of a feature extractor with a linear classification layer on top of it. The feature extractor is a ResNet50 (He et al., 2016), pre-trained on ImageNet (Deng et al., 2009), with its last linear layer removed and replaced by a linear bottleneck layer of dimension 256.

Optimizer. We use the SGD (Robbins & Monro, 1951) algorithm with momentum of 0.9, a weight decay of $5e^{-4}$ and Nesterov acceleration. As the bottleneck and classifier layers are randomly initialized, we set their learning rates to be 10 times that of the pre-trained ResNet50 backbone. We schedule the learning rate with a strategy similar to the one in (Ganin et al., 2016): $\chi_p = \frac{\chi_0}{(1+\mu i)^{-\nu}}$, where i is the current iteration, $\chi_0 = 0.001$, $\gamma = 0.001$, $\nu = 0.75$. While this schedule is slightly different than the one reported in previous work, it is the one implemented in the different official code implementations. We elaborate in the Appendix A.3 on the differences and provide additional details. Finally, as for the mini-batch size, JUMBOT and M-POT were designed with a stratified sampling, *i.e.,* a balanced source mini-batch with the same number of samples per class. This allows to reduce the negative transfer between domains and is crucial to their success. On the other hand, it was shown that for some methods (e.g. BA3US) using a larger mini-batch, than what was reported, leads to a decreased performance (Fatras et al., 2021a). As a result, we used the default mini-batch strategies for each method. JUMBOT and M-POT use stratified mini-batches of size 65 for OFFICE-HOME and 36 for VISDA. All other methods use a random uniform sampling strategy with a mini-batch size of 36.

Evaluation Protocol. For the hyper-parameters chosen with each model selection strategy, we run the methods for each task 3 times, each with a different seed (2020, 2021, 2022). We tried to control for the randomness across methods by setting the seeds at the beginning of training. Interestingly, as we discuss in more detail in Section 5, some methods demonstrated a non-negligible variance across the different seeds showing that some hyper-parameters and methods are not robust to randomness.

4.2 Hyper-Parameter Tuning

Previous works (Gulrajani & Lopez-Paz, 2021; Musgrave et al., 2021; 2022) perform random searches with the same number of runs for each method. In contrast, we perform hyper-parameter grid searches for each method. As a result, the hyper-parameter tuning budgets differs across the methods depending on the number of hyper-parameters and the chosen grid. While one can argue this leads to an unfair comparison of the methods, in practice in most real-world applications one will be interested in using the best method that our approach will precisely capture.

The hyper-parameter tuning needs to be performed for each task of each dataset, but that would require a significant computational resources without a clear added benefit. Instead for each dataset, we perform the hyper-parameter tuning on a single task: A2C for OFFICE-HOME and S2R for VISDA. This same strategy was adopted in (Fatras et al., 2021a) and the hyper-parameters were found to generalize to the remaining tasks in the dataset. We conjecture that this may be due to the fact that information regarding the number of target only classes is implicitly hidden in the hyper-parameters. See Appendix A.4 for more details regarding the hyper-parameters.

DATASET	METHOD	S-ACC	ENT	DEV	SND	1-SHOT	100-RND	ORACLE
OFFICE-HOME	S. ONLY	60.38±0.5	60.73±0.2	60.22±0.3	59.55±0.3	58.92±0.4	60.34±0.4	61.87±0.3
	PADA	63.08±0.3	59.74±0.5	52.72±2.8	62.36±0.4	62.00±0.5	63.22±0.1	63.72±0.3
	SAFN	62.09±0.2	61.37±0.3	62.03±0.4	62.59±0.1	49.30±0.7	62.36±0.2	63.30±0.2
	BA3US	68.32±1.1	73.36±0.6	62.25±7.1	75.37±0.8	65.56±7.6	75.19±0.4	75.98±0.3
	AR	65.68±0.3	70.58±0.4	64.32±0.9	70.25±0.2	70.56±0.7	70.34±0.2	72.73±0.3
	JUMBOT	62.89±0.2	74.61±0.8	61.28±0.1	72.29±0.2	74.95±0.1	75.74±0.3	77.15±0.4
	MPOT	66.24±0.1	64.46±0.1	61.37±0.2	46.92±0.4	68.28±0.2	73.06±0.3	77.31±0.5
VISDA	S. ONLY	55.15±2.4	55.24±3.2	55.07±1.2	55.02±2.9	55.72±2.2	58.16±0.6	59.48±0.4
	PADA	47.48±4.8	32.32±4.9	43.43±5.3	56.83±1.0	53.15±2.9	54.38±2.7	54.57±2.6
	SAFN	58.20±1.7	42.83±6.3	58.62±1.3	44.82±8.8	56.89±2.1	59.09±2.8	62.64±1.5
	BA3US	55.10±3.7	65.58±1.4	58.40±1.4	51.07±4.3	64.77±1.4	67.44±1.2	67.67±1.3
	AR	66.68±1.0	64.27±3.6	67.20±1.5	55.69±0.9	70.29±1.7	72.60±0.8	73.85±0.9
	JUMBOT	60.63±0.7	62.42±2.4	59.86±0.6	77.69±4.2	78.34±1.9	83.49±1.9	84.01±1.9
	MPOT	70.02±2.0	74.64±4.4	61.62±1.3	78.40±3.9	70.96±3.7	86.69±5.1	86.95±5.0

Table 6: Task accuracy average over seeds 2020, 2021, 2022 on Partial OFFICE-HOME and Partial VISDA for the PDA methods and model selection strategy. For each method, we highlight the best and worst label-free model selection strategies in green and red, respectively.

Several runs in our hyper-parameter search for JUMBOT, M-POT and BA3US were unsuccessful with the optimization reaching its end without the model being trained at all. This poses a challenge to DEV, SND and ENT and its one of the failures modes accounted for in (Saito et al., 2021). Following their recommendations, for JUMBOT, M-POT and BA3US, before applying the model selection strategy, we discard models whose source domain accuracy is below a certain threshold thr, which is set with the heuristic as $thr = 0.9 \cdot \text{Acc}$. Here Acc denotes the source domain accuracy of the Source-Only model. In our experiments, this leads to select models whom the source accuracy is at least of $thr = 69.01\%$ for the A2C task on OFFICE-HOME and $thr = 89.83\%$ for the S2R task on VISDA. We choose this heuristic because the ablation study of some methods showed that doing the adaptation decreased slightly the source accuracy (Bhushan Damodaran et al., 2018). Table 5 shows that our heuristic leads to improved results.

Lastly, when choosing the hyper-parameters, we only consider the model at the end of training, discarding the intermediate checkpoint models in order to select hyper-parameters which do not lead to overfitting at the end of training and better generalize to the other tasks. Following the above protocol, for each dataset we trained *468* models in total in order to find the best hyper-parameters. Then, to obtain the results with our neural network architecture on all tasks of each dataset, we trained an additional *1224* models for OFFICE-HOME and *156* models for VISDA. We additionally trained 231 models with the different neural network architectures for AR and SAFN. In total, *2547* models were trained to make this study and we present the different results in the next section.

5 Partial domain adaptation experiments

We start the results section by discussing the differences between our reproduced results and the published results from the different PDA methods. Then, we compare the performance of the different model selection strategies. Finally, we discuss the sensitivity of methods to the random seed.

5.1 Reproducibility Of Previous Results

We start by ensuring that our reimplementation of PDA methods was done correctly by comparing our reproduced results with previously reported results in Table 4. As such the model selection strategy used is ORACLE. On OFFICE-HOME, both PADA and JUMBOT achieved higher average task accuracy (1.6 and 1.7 percentage points, respectively) in our reimplementation, while for BA3US and MPOT we recover the reported accuracy in their respective papers. However, we saw a decrease in performance for both SAFN and AR of roughly 8 and 5 percentage points respectively. This is to be expected due to the differences in the neural

network architectures. While we use a linear bottleneck layer, SAFN uses a nonlinear bottleneck layer. As for AR, they make two significant changes: the linear classification head is replaced by a spherical logistic regression (SLR) layer (Gu et al., 2020) and the features are normalized (the 2-norm is set to a dataset dependent value, another hyper-parameter that requires tuning) before feeding them to the classification head. While we account for the first change by comparing to AR (w/ linear) results reported in (Gu et al., 2021), in our neural network architecture we do not normalize the features. These changes, nonlinear bottleneck layer for SAFN and feature normalization for AR, significantly boost the performance of both methods. When now comparing our reimplementation with the same neural network architectures, our SAFN reimplementation achieves a higher average task accuracy by 3 percentage points, while our AR reimplementation is now only 1 percentage points below. The fact that AR reported results are from only one run, while ours are averaged across 3 distinct seeds, justifies the small remaining gap. Moreover, we report higher accuracy or on par on 4 of the 12 tasks. Given all the above and further discussion of the VISDA dataset results in Appendix B, our reimplementations are trustworthy and give validity to the results we discuss in the next sections.

5.2 Results for Model Selection Strategies

Model Selection Strategies (w/ vs w/o target labels) All average accuracies on the OFFICE-HOME and VISDA datasets can be found in Table 6. For all methods on OFFICE-HOME, we can see that the results for model selection strategies which do not use target labels are below the results given by ORACLE. For some pairs, the drop of performance can be significant, leading some methods to perform on par with the S. ONLY method. That is the case on OFFICE-HOME when DEV is paired with either BA3US, JUMBOT and MPOT. Even worse is MPOT with SND as the average accuracy is more than 10 percentage points below that of S. ONLY with any model selection strategy. Overall on OFFICE-HOME, except for MPOT, all methods when paired with either ENT or SND give results that are at most 2 percentage points below compared to when paired with ORACLE.

A similar situation can be seen over the VISDA dataset where the accuracy without target labels can be down to 25 percentage points. Yet again, some model selection strategies can lead to scores even worse than S. ONLY. That is the case for PADA, SAFN and BA3US. Contrary to OFFICE-HOME, all model selection strategies without target labels lead to at least one method with results on par or worse in comparison to the S. ONLY method. More generally, no model selection strategy without target labels can lead to score on par to the ORACLE model selection strategy. Finally, PADA performs worse than S. ONLY for most model selection strategies, including the ones which use target labels. However, when combined with SND it performs better than with ORACLE on average, although still within the standard deviation. This is a consequence of the random seed dependence mentioned before on VISDA: as the hyper-parameters were chosen by performing just one run, we were simply "unlucky". In general, all of this confirms the standard assumption in the literature regarding the difficulty of the VISDA dataset.

Model Selection Strategies (w/ target labels) We recall that the ORACLE model selection strategy uses all the target samples to compute the accuracy while 1-SHOT and 100-RND use only subsets: 1-SHOT has only one sample per class for a total of 25 and 6 on OFFICE-HOME and VISDA, respectively, while 100-RND has 100 random target samples. Our results show that using only 100 random target labeled samples is enough to reasonably approximate the target accuracy leading to only a small accuracy drop (one percentage point in almost all cases) for both datasets. Not surprisingly, the gap between the 1-SHOT and ORACLE model selection strategies is even bigger, leading in some instances to worse results than with a model selection strategy that uses no target labels. This poor performance of the 1-SHOT model selection strategy also highlights that semi-supervised domain adaptation (SSDA) methods are not a straightforward alternative to the 100-RND model selection strategy. While one could argue that the target labels could be leveraged during training like in SSDA methods, one still needs labeled target data to perform model selection. However our results suggest that we would need at least 3 samples per class for SSDA methods. In addition, knowing that we have a certain number of labeled samples per class provides information regarding which classes are target only, one of the main assumptions in PDA. In that case, PDA methods could be tweaked. This warrants further study that we leave as future work. Finally, we have also investigated a smaller labeled target set of 50 random samples (50-RND) instead of 100 random samples. The accuracies of methods using

Task	Method	S-ACC	ENT	DEV	SND	1-SHOT	100-RND	ORACLE
S2R	S. ONLY	46.96 ± 1.5	48.17 ± 3.9	49.00 ± 0.9	48.17 ± 3.9	49.43 ± 0.8	50.01 ± 1.6	51.86 ± 1.4
	PADA	44.56 ± 5.9	40.83 ± 11.3	41.04 ± 4.3	56.14 ± 9.7	52.94 ± 4.3	49.34 ± 8.4	49.34 ± 8.4
	SAFN	52.04 ± 3.5	29.86 ± 16.7	52.42 ± 2.9	28.46 ± 16.5	49.97 ± 3.3	47.83 ± 0.6	56.88 ± 2.1
	BA3US	44.21 ± 3.0	71.17 ± 1.9	48.78 ± 1.9	46.12 ± 7.8	66.79 ± 1.5	71.45 ± 0.8	71.77 ± 1.1
	AR	68.39 ± 1.3	75.28 ± 2.9	68.54 ± 1.3	57.61 ± 0.4	70.11 ± 1.4	75.09 ± 5.2	76.33 ± 4.5
	JUMBOT	55.23 ± 2.3	56.25 ± 2.1	54.35 ± 2.0	75.23 ± 8.4	81.27 ± 6.9	89.94 ± 1.1	90.55 ± 0.5
	MPOT	64.57 ± 2.9	82.10 ± 2.0	57.02 ± 1.5	84.45 ± 0.4	71.33 ± 4.4	87.20 ± 2.3	87.23 ± 2.3
R2S	S. ONLY	63.34 ± 3.4	62.32 ± 2.7	61.13 ± 3.3	61.88 ± 2.3	62.00 ± 3.9	66.30 ± 2.0	67.11 ± 2.1
	PADA	50.39 ± 3.8	23.80 ± 1.6	45.82 ± 9.2	57.53 ± 10.3	53.36 ± 1.7	59.43 ± 5.8	59.81 ± 6.2
	SAFN	64.37 ± 0.7	55.80 ± 5.2	64.82 ± 0.5	61.19 ± 3.3	63.82 ± 1.0	70.34 ± 5.8	68.40 ± 1.2
	BA3US	65.99 ± 4.6	59.99 ± 1.3	68.01 ± 1.9	56.01 ± 2.9	62.75 ± 2.6	63.44 ± 1.9	63.56 ± 1.8
	AR	64.97 ± 0.8	53.26 ± 9.7	65.86 ± 3.5	53.78 ± 2.1	70.46 ± 4.7	70.11 ± 5.0	71.36 ± 5.5
	JUMBOT	66.04 ± 1.0	68.59 ± 4.6	65.36 ± 0.8	80.16 ± 1.1	75.42 ± 4.8	77.03 ± 2.7	77.46 ± 3.3
	MPOT	75.47 ± 3.8	67.18 ± 9.1	66.21 ± 1.2	72.36 ± 7.4	70.58 ± 3.1	86.18 ± 8.1	86.67 ± 7.8
Avg	S. ONLY	55.15 ± 2.4	55.24 ± 3.2	55.07 ± 1.2	55.02 ± 2.9	55.72 ± 2.2	58.16 ± 0.6	59.48 ± 0.4
	PADA	47.48 ± 4.8	32.32 ± 4.9	43.43 ± 5.3	56.83 ± 1.0	53.15 ± 2.9	54.38 ± 2.7	54.57 ± 2.6
	SAFN	58.20 ± 1.7	42.83 ± 6.3	58.62 ± 1.3	44.82 ± 8.8	56.89 ± 2.1	59.09 ± 2.8	62.64 ± 1.5
	BA3US	55.10 ± 3.7	65.58 ± 1.4	58.40 ± 1.4	51.07 ± 4.3	64.77 ± 1.4	67.44 ± 1.2	67.67 ± 1.3
	AR	66.68 ± 1.0	64.27 ± 3.6	67.20 ± 1.5	55.69 ± 0.9	70.29 ± 1.7	72.60 ± 0.8	73.85 ± 0.9
	JUMBOT	60.63 ± 0.7	62.42 ± 2.4	59.86 ± 0.6	77.69 ± 4.2	78.34 ± 1.9	83.49 ± 1.9	84.01 ± 1.9
	MPOT	70.02 ± 2.0	74.64 ± 4.4	61.62 ± 1.3	78.40 ± 3.9	70.96 ± 3.7	86.69 ± 5.1	86.95 ± 5.0

Table 7: Accuracy of different PDA methods based on different model selection strategies on the 2 Partial VISDA tasks. Average is done over three seeds (2020, 2021, 2022). For each method, we highlight the best and worst label-free model selection strategies in green and red, respectively.

50-RND were not as good as when using 100-RND. All results of pairs of methods and 50-RND can be found in Appendix B. The smaller performance show that the size of the labeled target set is an important element and we suggest to use at least 100 random samples.

Model Selection Strategies (w/o target labels) Only the (BA3US, ENT) pair achieved average task accuracies within 3 percentage points of its ORACLE counterpart (*i.e.*, (BA3US, ORACLE)), while still improving over S. ONLY model. Our experiments show that there is no model selection strategy which performs well for all methods. That is why to deploy models in a real-world scenario, we advise to test selected models on a small labeled target set (*i.e.*, 100-RND)) to assess the performance of the models as model selection strategies without target labels can perform poorly.

Our conclusion is that the model selection for PDA methods is still an open problem. We conjecture that it is also the case for domain adaptation as the considered metrics were developed first for this setting. For future proposed methods, researchers should specify not only which model selection strategy should be used, but also which hyper-parameter search grid should be considered, to deploy them in a real-world scenario.

5.3 Random Seed Dependence

Ideally, PDA methods should be robust to the choice of random seed. This is of particular importance when performing hyper-parameter tuning since typically only one run per set of hyper-parameters is done (that was the case in our work as well). We investigate this robustness by averaging all the results presented over three different seeds (2020, 2021 and 2022) and reporting the standard deviations. This is in contrast with previous work where only a single run is reported (Fatras et al., 2021a; Gu et al., 2021). Other works (Cao et al., 2018; Xu et al., 2019; Jian et al., 2020) that report standard deviations do not specify if the random seed is different across runs. Results for all tasks on VISDA dataset are in Table 7 and on OFFICE-HOME in Appendix B due to space constraints.

Our experiments show that some methods express a non-negligible instabilities over randomness with respect to any model selection methods. This is particularly true for BA3US when paired with DEV and 1-SHOT as model selection strategies: there are several tasks where the standard deviation is above 10%. While in this case this instability may stem from the poor performance of the model selection strategies, it is also visible when ORACLE is the model selection strategy used. For instance, the M-POT has a standard deviation of 3.3% on the AP task of OFFICE-HOME which corresponds to a variance of 11%. On VISDA this instability and seed dependence is even larger.

6 Conclusion

In this paper, we investigated how model selection strategies affect the performance of PDA methods. We performed a quantitative study with seven PDA methods and seven model selection strategies on two real-word datasets. Based on our findings, we provide the following recommendations:

i) Target label samples should be used to test models before using them in real-world scenario. While this breaks the main PDA assumption, it is impossible to confidently deploy PDA models selected without the use of target labels. Indeed, model selection strategies without target labels lead to a significant drop in performance in most cases in comparison to using a small validation set. We argue that the cost of labeling it outweighs the uncertainty in current model selection strategies.

ii) The robustness of new PDA method to randomness should be tested over at least three different seeds. We suggest to use the seeds (2020, 2021, 2022) to allow for a fair comparison with our results.

iii) An ablation study should be considered when a novel architecture is proposed to quantify the associated increase of performance.

As our work focus on a quantitative study of model selection methods and reproducibility of state-of-the-art partial domain adaptation methods, we do not see any potential ethical concern. Future work will investigate new model selection strategies which can achieve similar results as model selection strategies which use label target samples.

References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 214–223, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/arjovsky17a.html.

Bharath Bhushan Damodaran, Benjamin Kellenberger, Remi Flamary, Devis Tuia, and Nicolas Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *The European Conference on Computer Vision (ECCV)*, September 2018.

Zhangjie Cao, Lijia Ma, Mingsheng Long, and Jianmin Wang. Partial adversarial domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 135–150, 2018.

Laetitia Chapel, Mokhtar Z. Alaya, and Gilles Gasso. Partial optimal transport with applications on positive-unlabeled learning. In *Advances in Neural Information Processing Systems*, 2020.

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

Kilian Fatras, Younes Zine, Rémi Flamary, Remi Gribonval, and Nicolas Courty. Learning with minibatch wasserstein : asymptotic and gradient properties. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2131–2141, Online, 26–28 Aug 2020. PMLR. URL http://proceedings.mlr.press/v108/fatras20a.html.

Kilian Fatras, Thibault Sejourne, Rémi Flamary, and Nicolas Courty. Unbalanced minibatch optimal transport; applications to domain adaptation. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3186–3197. PMLR, 18–24 Jul 2021a. URL http://proceedings.mlr.press/v139/fatras21a.html.

Kilian Fatras, Younes Zine, Szymon Majewski, Rémi Flamary, Rémi Gribonval, and Nicolas Courty. Minibatch optimal transport distances; analysis and applications, 2021b.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. URL http://jmlr.org/papers/v22/20-451.html.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1180–1189, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/ganin15.html.

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

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL http://jmlr.org/papers/v13/gretton12a.html.

Xiang Gu, Jian Sun, and Zongben Xu. Spherical space domain adaptation with robust pseudo-label loss. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Xiang Gu, Xi Yu, Yan Yang, Jian Sun, and Zongben Xu. Adversarial reweighting for partial domain adaptation. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=f5liPryFRoA.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=lQdXeXDoWtI.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Liang Jian, Wang Yunbo, Hu Dapeng, He Ran, and Feng Jiashi. A balanced and uncertainty-aware approach for partial domain adaptation. In *European Conference on Computer Vision (ECCV)*, August 2020.

Mingsheng Long, ZHANGJIE CAO, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/ab88b15733f543179858600245108dd8-Paper.pdf.

Pietro Morerio, Jacopo Cavazza, and Vittorio Murino. Minimal-entropy correlation alignment for unsupervised deep domain adaptation. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJWechg0Z.

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

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

Khai Nguyen, Dang Nguyen, Tung Pham, and Nhat Ho. Improving mini-batch optimal transport via partial transportation. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. doi: 10.1109/TKDE.2009.191.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Vishal M Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE Signal Processing Magazine*, 32(3):53–69, 2015. doi: 10.1109/MSP.2014.2347059.

Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *CoRR*, abs/1710.06924, 2017.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 2019.

H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22: 400–407, 1951.

Kuniaki Saito, Donghyun Kim, Piotr Teterwak, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Tune it the right way: Unsupervised validation of domain adaptation via soft neighborhood density. *arXiv preprint arXiv:2108.10860*, 2021.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(35):985–1005, 2007. URL http://jmlr.org/papers/v8/sugiyama07a.html.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *(IEEE) Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Ruijia Xu, Guanbin Li, Jihan Yang, and Liang Lin. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

Kaichao You, Ximei Wang, Mingsheng Long, and Michael Jordan. Towards accurate model selection in deep unsupervised domain adaptation. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7124–7133. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/you19a.html.