UDA-Bench: Revisiting Common Assumptions in Unsupervised Domain Adaptation Using a Standardized Framework

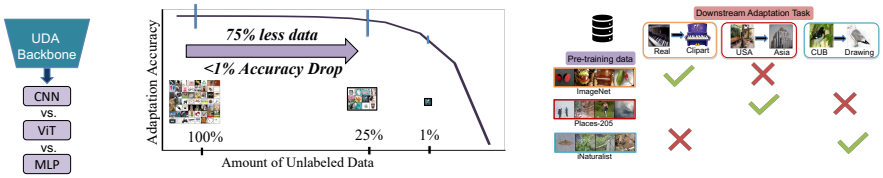
Tarun Kalluri, Sreyas Ravichandran, and Manmohan Chandraker

UC San Diego
https://github.com/ViLab-UCSD/UDABench_ECCV2024

Abstract. In this work, we take a deeper look into the diverse factors that influence the efficacy of modern unsupervised domain adaptation (UDA) methods using a large-scale, controlled empirical study. To facilitate our analysis, we first develop UDA-Bench, a novel PyTorch framework that standardizes training and evaluation for domain adaptation enabling fair comparisons across several UDA methods. Using UDA-Bench, our comprehensive empirical study into the impact of backbone architectures, unlabeled data quantity, and pre-training datasets reveals that: (i) the benefits of adaptation methods diminish with advanced backbones, (ii) current methods underutilize unlabeled data, and (iii) pre-training data significantly affects downstream adaptation in both supervised and self-supervised settings. In the context of unsupervised adaptation, these observations uncover several novel and surprising properties, while scientifically validating several others that were often considered empirical heuristics or practitioner intuitions in the absence of a standardized training and evaluation framework. The UDA-Bench framework and trained models are publicly available.

1 Introduction

Deep neural networks for image classification often suffer from dataset bias where accuracy significantly drops if the test-time data distribution does not match that of training, which often happens in real-world applications. To overcome the infeasibility of collecting labeled data from each application domain, a suite of methods have been recently proposed under the umbrella of unsupervised domain adaptation (UDA) [7,9,10,25,35,36,38,40–42,53,54,56,82,83,85,103,106,111,115] that allow training using only unlabeled data from the target domain of interest while leveraging supervision from a different source domain with abundant labels. These UDA methods have been greatly successful in improving the target accuracy on benchmark datasets under a variety of distribution shifts [11, 70, 71, 79, 100]. While literature in the area has predominantly focused on proposing new algorithms or loss functions, a holistic understanding of several fundamental assumptions that influence real-world effectiveness of domain adaptation has been lacking. In this paper, we address this through a large-scale empirical study of three major factors that potentially influence performance the most, namely,



(a) Backbone **(b)** Sample Efficiency of Target Data **(c)** Type of Pre-Training Data

Fig. 1: A summary of our contributions. We examine the effectiveness of SOTA UDA approaches using our proposed framework UDA-Bench by revisiting the role of backbone architectures (Fig. 1a, Sec. 4.1), unlabeled data (Fig. 1b, Sec. 4.2) and pre-training data (Fig. 1c, Sec. 4.3) with several useful observations.

1. **Choice of backbone architecture:** With recent advances in architecture designs such as vision transformers [23, 51, 93] and improved CNNs [52] we study which architectures suit domain transfer, and verify compatibility of existing adaptation methods with these backbones. 2. **Amount of unlabeled data:** Since the promise of unsupervised adaptation rests on its potential to leverage unlabeled target domain data, we study how much unlabeled data can really be digested by the adaptation methods. 3. **Nature of pre-training data:** We examine whether pre-training the backbone on similar data as the downstream adaptation task is more beneficial than commonly adopted ImageNet pre-training across several supervised and self-supervised pre-training strategies.

We believe that such insights into the behavior of UDA methods have been previously hindered due to varying choices of adaptation-independent factors like initialization, learning algorithm and batch sizes. To address this, we first propose UDA-Bench, a new PyTorch framework that standardizes these factors across multiple UDA methods and offers a unified training and evaluation platform for unsupervised adaptation. Using this framework, we study various UDA methods for image classification under different factors of variation. Among prior works which shared similar motivations as ours [44], the absence of standardized evaluation limits fair comparisons between UDA methods, where our distinction lies in establishing such a framework for consistent UDA training and evaluation. Through our analysis, we discover several new insights, while scientifically validating several phenomenon which were only considered empirical heuristics or practitioner intuitions due to the lack of a standardized approach. These are outlined in Fig. 1, and can be summarized as follows:

1. Recent advancements in vision transformers such as Swin [50] and DeiT [94] exhibit superior robustness against diverse domain shifts when compared to the conventional choice of ResNet-50 (see Tab. 1). However, incorporating these advancements into current UDA methods tends to diminish their benefits, leading to significant changes to the relative ranking among the methods. As a result, *older and simpler UDA methods often achieve comparable or even superior accuracies compared to more recent methods* (see Fig. 3 and Sec. 4.1).
2. Reducing the amount of unlabeled target data by up to 75% resulted in only a 1% decrease in target accuracy across all UDA methods studied (see Fig. 4), suggesting that *that current UDA methods saturate quickly, and are*

not well-equipped to exploit the increasing availability of inexpensive unlabeled data (see Sec. 4.2). This observation also contradicts the prevailing theory underpinning modern UDA research proposed in Ben-David et. al. [5], which suggests an inverse relation between the amount of unlabeled target data and target error, highlighting the discrepency between theory and practice.

3. Pre-training data matters for downstream adaptation, but in different ways for supervised and self-supervised pre-training. In supervised setting, *pre-training on similar data as the downstream adaptation task significantly improves the accuracy* compared to standard ImageNet pre-training (see Tab. 2).
4. In self-supervised setting, *object-centric pre-training datasets enhance accuracy for object-centric adaptation*, while scene-centric pre-training datasets are better suited for scene-centric tasks (see Tab. 3). This trend holds across different types of pre-text tasks in self-supervised pre-training (see Sec. 4.3).

Through a comprehensive analysis using our unified training and evaluation framework, our recommendations serve a dual purpose - enabling researchers in identifying future opportunities for developing more effective adaptation algorithms with fair comparisons, as well as guiding practitioners in maximizing the benefits derived from current UDA methods. Our framework is publicly available to continue improving our understanding of UDA methods.

2 Related Works

Unsupervised Domain Adaptation A majority of works in unsupervised adaptation aim to minimize some notion of divergence between the source and target domains estimated using unlabeled samples [5, 6]. Prior works studied various divergence metrics like MMD distance [4, 37, 53, 55, 56, 67, 98, 108], higher-order correlations [38, 59, 87, 88] or optimal-transport [20, 22, 74], but adversarial discriminative approaches [13, 25, 54, 83, 96–98, 105] have been the most popular. More recent works address the issue of noisy alignment with global domain discrimination [45] using category-level [21, 24, 42, 63, 69, 72, 81, 103], instance-level [41, 85], consistency-based [7], language-guided [40] or cross-attention [107, 115] based techniques. The primary focus of most of these works is on algorithmic innovations to improve adaptation. Instead, our emphasis in this paper lies in identifying several key method-agnostic factors that impact performance of UDA methods, and conducting a comprehensive empirical study along these factors for a better understanding of these methods. While domain adaptive semantic segmentation is also popular [36, 39, 46, 95, 101], we restrict focus on adaptation methods for image-classification in this paper.

Comparative Studies and Benchmarks Many recent works aim to enhance our understanding of the factors impacting the success of state-of-the-art methods through carefully crafted empirical analysis. A common theme in these works is to keep the algorithm itself fixed, but study several other factors which hold non-trivial importance in determining the performance of the algorithm. Within computer vision, these works span the areas of semi-supervised learning [66],

SLAM [64], metric learning [60, 77], transfer learning [58], domain generalization [30], optimization algorithms [18], few-shot learning [15], contrastive learning [19], GANs [57], fairness [28] and self-supervised learning [27, 29, 65]. Prior works also established standardized benchmarks to facilitate fair comparisons and quick prototyping [60, 64, 91]. Our work follows suit, where we develop a unified framework for UDA methods, and devise a controlled empirical study to revisit several standard training choices in unsupervised adaptation.

The works closest to ours in domain adaptation are [44], which carries UDA study but without a unified training framework, [61, 62], which study UDA methods through fair validation methods and [43] which studies adaptation for video segmentation. Different from these, our work lays emphasis on several other key factors that impact adaptation such as architectures, quantity of unlabeled data and nature of pretext data used in pre-training through design of a new standardized evaluation framework.

3 Analysis Setup

The task of unsupervised domain adaptation (UDA) aims to improve performance on a certain target domain with only unlabeled samples (D_t=$\{X_t\}$) by leveraging supervision from a different labeled source domain D_s=$\{X_s, y_s\}$. We assume that the source images are drawn from $X_s \sim P_s$, and target images from $X_t \sim P_t$. We assume a covariate shift [5] between the domains, which arises when $P_s \neq P_t$, although other forms of shift have also been studied in literature [1, 2, 26, 90]. The task of UDA is then to learn a predictive model using $\{X_s, X_t, y_s\}$ to improve performance on test samples from the target domain P_t. While recent literature focuses on novel training algorithms or loss functions to improve transfer, this paper aims to study their effectiveness under several important but often overlooked axes of variations pertaining to backbone architectures, unlabeled data quantity and backbone pre-training strategies.

The Need for UDA-Bench Framework Ensuring fair comparisons between different UDA methods necessitates controlling algorithm-independent factors during training and inference. However, we identify a problematic practice in most UDA methods where they are trained on different frameworks with different choices in various training hyper-parameters and settings, making fair comparison across these works difficult. To highlight this issue, we compute the plain source-only accuracy using original code-bases of various UDA algorithms in Fig. 2 (the links to the open-source code for each of these methods are given in the supplementary). Essentially, we take the open-source code base for the methods, switch off all the adaptation losses, and train the model only on the source dataset to compute the target accuracy. Ideally, this accuracy, which acts as the baseline, should be the same across all the methods since it is independent of any adaptation. In practice, however, we observe that this baseline accuracy varies significantly between various UDA codebases, pointing to an underlying discrepency in various training choices adopted by these works unrelated to the adaptation algorithm itself. For example, unique to the respective methods,

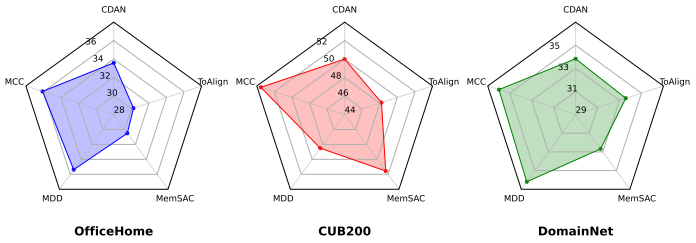


Fig. 2: Need for UDA-Bench. We illustrate the disparity between various codebases proposed for prior UDA methods by highlighting the different accuracy numbers obtained for a plain source only model. Computed without any adaptation, it should ideally match across implementations which is clearly not the case. To enable fair comparisons across UDA methods, we propose UDA-Bench, a new PyTorch framework to standardize training and evaluation across various methods.

MDD [111] uses a deeper MLP as a classifier, MCC [38] uses batchnorm layers in the bottleneck layer, CDAN [54] uses 10-crop evaluation and AdaMatch [7] uses stronger augmentation on source data.

To alleviate this issue, we create a new framework in PyTorch [68] for domain adaptation called *UDA-Bench* and implement several existing methods in this framework. Our framework standardizes different UDA methods with respect to adaptation-independent factors such as learning algorithm, network initialization and batch sizes while simultaneously allowing flexibility for incorporating algorithm-specific hyperparameters like loss coefficients and custom data loaders within a unified framework. All our comparisons and analyses in this paper are implemented using this framework, while using the same adaptation-specific hyperparameters proposed in the original papers in our re-implementation. We also verified that our re-implementations reproduced the original accuracies when using the hyper-parameters from the respective codebases. UDA-Bench, along with all our implementations, is publicly released to the research community to enable fair comparisons and fast prototyping of UDA methods in future works.

Axes of Variation We choose backbone architecture (Sec. 4.1), amount of unlabeled data in the target (Sec. 4.2) and the nature of data/algorithm used in pre-training the backbone (Sec. 4.3) as the different axes of variation in our study. The deliberate focus on backbone, data size, and pre-training factors is driven by the recognition that these factors hold the most potential to influence deep learning training in general and UDA algorithms in particular, while also being the most understudied in prior UDA literature. By analyzing these factors, we seek to offer insights into salient properties of UDA and provide practical guidance for enhancing accuracy through optimal design choices.

Adaptation Methods The selection of methods in our comparative study is not intended to be exhaustive of all the adaptation methods proposed in the literature thus far. Instead, we aim to provide a representative sample of works spanning a diverse range of model families from standard to state-of-the-art, although our inferences should readily transfer to any UDA method. In particular, the types

of UDA methods we study include *adversarial* (DANN [25], CDAN [54]), *non-adversarial* (MDD [111], MCC [38], DALN [14]), *consistency-based* (MemSAC [41], AdaMatch [7]), *alignment-based* (ToAlign [103]) and *pseudo-label based* [115] methods. In the supplementary, we show that the inferences made in our study also extend to several other adaptation methods (such as BSP [17], ILADA [85], AFN [106] and MCD [83]).

Adaptation Datasets Following popular choices in UDA literature, we use visDA [71], OfficeHome [100], DomainNet [70] and CUB200 [102] datasets in our analysis. VisDA studies synthetic to real transfer from 12 categories, OfficeHome contains 65 categories across four domains, DomainNet contains images from 345 categories from 6 domains while CUB200 is designed for fine-grained adaptation. In the supplementary, we also show results using adaptation on TinyImageNet [47] and variants [34].

Evaluation Metrics We report results using the accuracy on the test set of the target domain while correcting for a problematic practice in prior literature. In most prior works using OfficeHome and CUB200 datasets, the same set of data doubles up as the unlabeled target used in training as well as the target test set used to report the results. To avoid possible over-fitting to target unlabeled data, we create separate train and test sets for these datasets (using a 90%-10% ratio), and use images from train set as labeled or unlabeled data during training and report final numbers on the unused test images. While this could lead slightly different numbers from those reported in the original papers, it also leads to fair comparison with the source-only baseline.

Hyper-parameters In all our re-implementations of prior works, we use the default hyperparameters suggested by the original methods to keep the number of experiments manageable. Each method in the unlabeled data volume study (Sec. 4.2) takes about 24 hours to run on an NVIDIA A10 GPU, so 8 methods, across 4 settings, 6 data fractions and 3 random trials costs ∼14000 GPU hours. Likewise, the experiments in Sec. 4.1 cost 18640 GPU hours and Sec. 4.3 cost about 17356 GPU hours (including the pre-training). Incorporating experiments to seek optimal hyperparameters for several UDA methods on top of this would have incurred impractical levels of expenses.

4 Methodology and Evaluation

4.1 Which backbone architectures suit UDA best?

Motivation Although ResNet-50 [33] backbone is a widely adopted standard in domain adaptation research [7, 41, 54, 81, 83, 103], several recent architectures [23, 52, 93] have emerged as feasible alternatives with better performance. While a more recent method PMTrans [115] adopts a ViT backbone, all the prior methods were still compared using a ResNet-50 backbone. Therefore, we aim to study if the recent advances in vision transformers confer additional benefits to cross-domain transfer, and how ViT-specific methods [115] compare to classical methods while using a same backbone. While robustness properties of vision transformers to

Table 1: Comparison of domain robustness of various vision architectures on standard adaptation datasets. We use the source accuracy (λ_s) and the target accuracy (λ_t) of a model trained only on source data to calculate the relative drop in accuracy (σ_{st}=$100 * (\lambda_s - \lambda_t)/\lambda_s$, lower the better). Swin transformer shows consistently better robustness to domain shifts on several benchmarks.

Model	ResNet-50	Swin-V2-t	ConvNext-t	ResMLP-s	DeiT3-s	ResNet-50	Swin-V2-t	ConvNext-t	ResMLP-s	DeiT3-s
#Params	24.12 M	27.86 M	28.10 M	29.82 M	21.86 M	24.12 M	27.86 M	28.10 M	29.82 M	21.86 M
	DomainNet (R→C)					CUB200 (CUB→Draw)				
Source Accuracy (λ_s, ↑)	81.86	85.99	84.37	82.68	84.52	81.00	87.75	85.88	84.62	88.12
Target Accuracy (λ_t, ↑)	44.85	55.51	50.80	46.62	50.75	52.60	58.90	52.74	53.41	56.36
Relative Drop (σ_{st}, ↓)	45.21	**35.45**	39.78	43.61	39.95	35.0	**32.88**	38.50	36.88	36.05
Abs. Drop ($\lambda_s - \lambda_t$, ↓)	37.01	30.48	33.57	36.06	33.77	28.40	28.85	33.14	31.21	31.76
	OfficeHome (Ar→Pr)					GeoPlaces (USA→Asia)				
Source Accuracy (λ_s, ↑)	60.10	76.17	74.72	69.69	71.76	57.17	63.11	60.39	58.99	61.65
Target Accuracy (λ_t, ↑)	53.33	72.56	70.77	65.90	67.18	36.12	42.53	40.30	38.11	40.34
Relative Drop (σ_{st}, ↓)	11.26	**4.74**	5.29	5.44	6.38	36.82	**32.61**	33.27	35.40	34.57
Abs. Drop ($\lambda_s - \lambda_t$, ↓)	6.77	3.61	3.95	3.79	4.58	21.05	20.58	20.09	20.88	21.31

adversarial and out-of-context examples have been widely studied [3, 8, 84, 112, 114], our analysis differs from these by focusing on the *cross-domain robustness* properties of these architectures on standard UDA datasets and investigating their potential as an improved backbone for UDA methods.

Experimental Setup Along with ResNet-50, we choose four different vision architectures which showed great success on standard ImageNet classification benchmarks: DeiT [93], Swin [51], ResMLP [92], and ConvNext [52]. We use newer versions of DeiT (DeiT-III [93]) and Swin (Swin-V2 [51]) as they have better accuracy on ImageNet. We use the variants of these architectures which roughly have comparable number of parameters as ResNet-50, namely DeiT-small, Swin-tiny, ResMLP-small and ConvNext-tiny. All of them are pre-trained on ImageNet-1k, so their differences only arise from specific architectures. We use all pre-trained checkpoints from the timm library [104] and architecture-specific training details are provided in the supplementary.

Newer Architectures Show Better Domain Transfer For a model trained only on source-domain data (no adaptation), we use the accuracy on the source test-set (λ_s) and the accuracy on the target test-set (λ_t), to define relative cross-domain accuracy drop σ_{st}=$\frac{\lambda_s - \lambda_t}{\lambda_s} * 100$. While this metric is sensitive to the absolute value of the source accuracy (λ_s), we nevertheless find that it serves as a good indicator of cross-domain robustness. Additionally, we also show the absolute accuracy drop from source to target ($\lambda_s - \lambda_t$) to discount the effect of original source accuracy. From Tab. 1, vision transformer architectures have the least value of σ_{st} (least cross-domain drops) indicating better robustness properties compared to CNNs or MLPs. Specifically, Swin-V2-t pre-trained on ImageNet-1k showed least relative drop (σ_{st}) across all the datasets. Notably, on Real→Clipart from DomainNet, using Swin backbone with plain source-only training alone yields 55.5% accuracy, which is already higher than SOTA UDA methods that use ResNet-50 (54.5%) [41], indicating that *using an improved backbone may have the same effect as using a complex adaptation algorithm* on the target accuracy. While the general competence of ViT-backbones is well known, our study confirms that these improvements also extend to the case of out-of-domain robustness. We also observe that the relative ranking of different

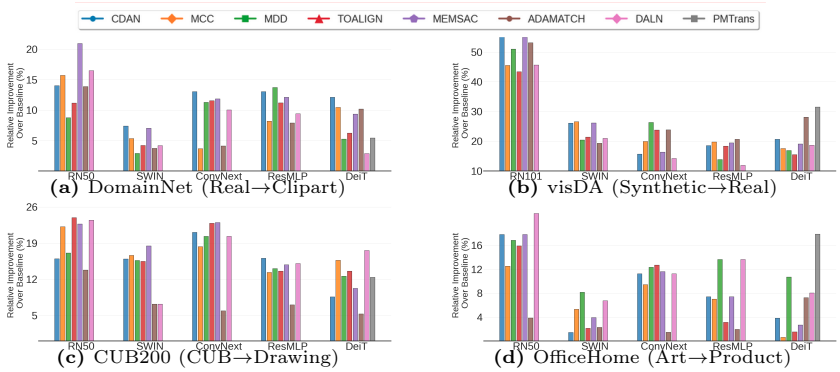


Fig. 3: Better backbones diminish gains from UDA. For each UDA method, we show the gain in accuracy relative to a baseline trained only using source-data. Across datasets, we observe that benefits offered by UDA approaches over the baseline diminish with backbones that have improved domain-robustness properties.

architectures widely varies across datasets, highlighting that the type of domain transfer influences domain robustness.

UDA Gains Diminish With Newer Architectures We next ask the question if these benefits are complementary to the UDA method itself, and explore the viability of incorporating these advanced architectures into existing UDA methods. From Fig. 3, we observe that most methods do yield complimentary benefits over a source-only trained baseline even with newer architectures, but the *relative improvement offered by UDA methods over this baseline tends to diminish when using better backbones*. Looking at the relative gain in accuracy over a source-only baseline, on Real→Clipart in Fig. 3a, the best adaptation method provides 20% relative gain over the baseline using ResNet-50, which falls to just 7% with Swin and 10% with DeiT backbone. Similarly, the relative gains offered by best UDA methods fall from 18% with ResNet-50 to 8% using Swin on Art→Product in Fig. 3d. These observation also holds for visDA Fig. 3b and CUB200 Fig. 3c datasets. The trends using the absolute accuracy drop also remain the same, while the relative drop further accounts for the strong source domain accuracy using advanced backbones. These results seem to suggest that the impact of many UDA methods is not really independent of the backbone used, and often tends to diminish in presence of better backbones which have better domain robustness properties. Furthermore, the *relative ranking of the best adaptation method and backbone changes across datasets*, and is not consistent. For example, an older and simpler method like CDAN gives best accuracies in Fig. 3a with Swin, ConvNext and DeiT, while MCC outperforms other methods with a ResMLP backbone. We also show the more results on DomainNet and OfficeHome in supplementary, and the results follow similar trends, where one of the more recent architectures significantly diminishes the returns yielded by all UDA methods.

Difference From Prior Works While prior works like [44] only show this trend for classical UDA methods [54, 83] without using a standardized

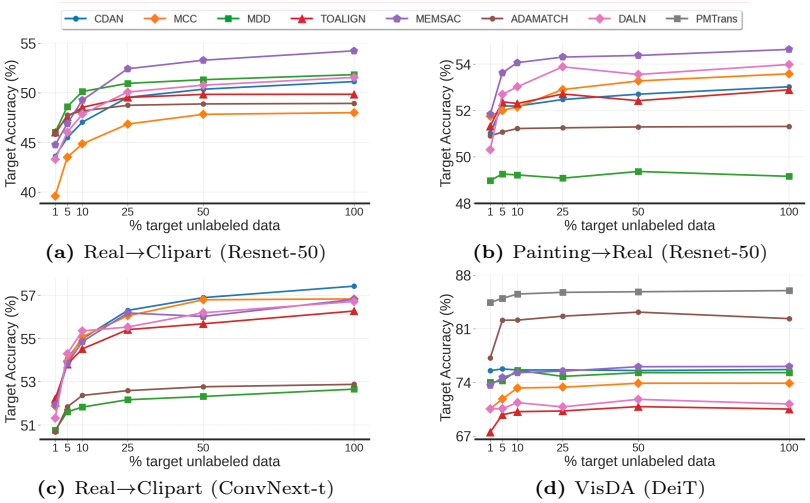
framework, we additionally show that this issue extends to more recent state-of-the-art UDA algorithms [14, 41, 103] as well, including methods using vision transformer backbones [115] using the proposed UDA-Bench, yielding several novel observations. For instance, we show in Fig. 3 that the current SOTA method PMTrans [115] performs worse than DALN on CUB200 and CDAN on DomainNet when all of them use the same DeiT backbone, highlighting the key need to standardize backbones and architectures before comparing different methods.

4.2 How much unlabeled data can UDA methods use?

Motivation Although UDA holds great potential in leveraging unlabeled data from a target domain to enhance performance, an insight into their scalability properties in relation to the quantity of unlabeled data is lacking. These scaling properties are important to inform us which method has the greatest potential to improve performance when more unlabeled data becomes accessible, motivating us to study how much unlabeled data do UDA methods actually consume.

Experimental Setup To study the effects of data volume, we sample {1, 5, 10, 25, 50, 100}% of the data from the target domain and run the adaptation algorithm using each of these subsets as the unlabeled data. We repeat the experiment with three different seeds in each case and report the mean accuracy to eliminate sampling bias. To avoid tail effects, we perform stratified sampling so that the label distribution is constant across all the subsets. Specifically, we sample x% of data from each category individually which helps to preserve the tail properties of the resulting sub-sampled dataset. We also make sure that all categories have at least 1 image in the sub-sampled dataset. Note that the label information in the target is used only during sampling, but not during training. We note the possibility of hyper-parameter sensitivity to the amount of target unlabeled data, but do not preform any additional tuning to keep the number of experiments manageable. We restrict to using DomainNet and VisDA in our analysis as those are the largest available datasets for domain adaptation, and show results using another recent large-scale adaptation benchmark GeoNet in the supplementary. The already tiny data volume in OfficeHome and CUB200 prevents their use in a scalability study like this.

UDA Accuracy Does Not Increase With More Unlabeled Data. Remarkably the trends from Fig. 4 indicate that on all the settings *the accuracy achieved by the unsupervised adaptation saturates rather quickly with respect to the unlabeled data*. This trend holds for almost all of the studied adaptation methods, including adversarial [54], non-adversarial [7], consistency based [41] and pseudo-label based [115] methods. The gains remain less than 2% in most cases even when scaling unlabeled data four-fold (from 25% to 100%). For example, on R→C (Fig. 4a), the accuracy achieved at using just 25% of the unlabeled data is within 1% of the accuracy obtained at 100% of the data using any adaptation method. In P→R, (Fig. 4b) the accuracy plateaus much earlier, at around $10 - 15$% of the unlabeled data. Similar results are observed using a different backbone like DeiT with a purely transformer-based method PMTrans [115] (Fig. 4d), where the performance saturates after using only 10% of the unlabeled



(a) Real→Clipart (Resnet-50) **(b)** Painting→Real (Resnet-50)

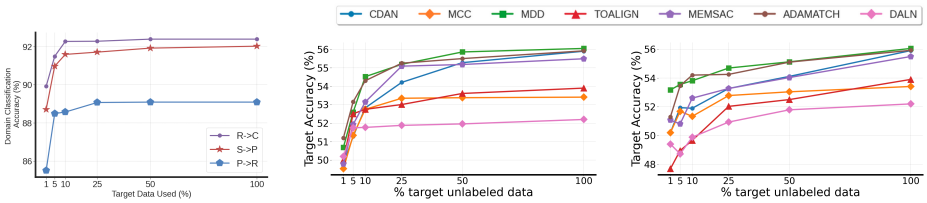
(c) Real→Clipart (ConvNext-t) **(d)** VisDA (DeiT)

Fig. 4: How much unlabeled data can UDA methods use? Across different adaptation datasets and backbones (Resnet50 in a, b, ConvNext in c and DeiT in d), we find that the performance of several UDA methods saturates quickly with respect to amount of target data, showing their limited efficiency in utilizing the unlabeled samples. In most cases, using only 25% of the data results in $< 1\%$ drop in accuracy.

data. These results suggest that even in cases where abundant unlabeled data becomes available, current UDA methods cannot leverage the potential benefits of this data to enhance performance. We also show more results on DomainNet in supplementary, where the observations follow similar trends.

Furthermore, we juxtapose this observation with a similar ablation using source labeled data in the supplementary, and identify that source supervision has a more pronounced effect on the target accuracy than target unlabeled data. Specifically, increasing source labeled data from 50% to 100% results in upto 10% gain in target accuracy (as opposed to $< 1\%$ observed using similar scale increase in target unlabeled data).

Investigating Poor Data Efficiency of UDA Methods We hypothesize that the main reason behind poor unlabeled sample efficiency is the underlying adaptation objective employed, which fails to effectively utilize growing amounts of unlabeled data. As an example, we take the objective of domain classification, which forms the backbone of several adversarial UDA methods [25, 54], and examine its data efficiency. We plot the accuracy of the domain discrimination objective itself against the quantity of unlabeled samples in Fig. 5a for different settings from DomainNet. We notice that the domain classification accuracy reaches a plateau after using approximately 25% of the data, potentially explaining the saturation of the adaptation accuracy in methods that rely on this objective for bridging the domain gap. While this explains adversarial alignment based methods, we posit that similar limitations impact other types of adaptation approaches including self-training, pseudo-label or consistency-based methods.



(a) Accuracy Saturation **(b)** Random Sampling **(c)** Class-split Sampling

Fig. 5: (a)**Saturation of the domain classification accuracy** is observed even with small amount of unlabeled data, potentially explaining the poor sample efficiency of UDA methods employing adversarial domain alignment. (b,c) **Role of the sampling technique adopted** We study the behavior of UDA methods with respect to target unlabeled data using two additional sampling techniques: random sampling in (b) and split-class sampling in (c). Our observation that UDA methods under-utilize unlabeled data holds for both of these cases as well.

UDA Empirical Data Efficiency Does Not Match Theory. The above observation stands in stark contrast to the theoretical framework of domain adaptation established by Ben-David et al. [5], which underpins several UDA methods. Their theoretical analysis suggests an inverse relationship between target sample size and target error (Theorem 2 from [5]), further highlighting the importance of empirical study like ours using a unified framework like UDA-Bench to understand the bridge between theory and practice. Our observation from UDA is also different from prior scalability studies in supervised [89], weakly-supervised [86] and self-supervised learning [29] literature, where increasing labeled or unlabeled data significantly enhances performance.

Similar Results Hold For Other Sampling Techniques In addition to the class-balanced sampling procedure in Fig. 4, we also show results using two other sampling techniques, random sampling and split-class sampling in Fig. 5b and Fig. 5c respectively. In Fig. 5b, we randomly select x% of images from the whole dataset without any class-aware sampling, and show the general observation that UDA methods reach a performance plateau after utilizing a limited amount of unlabeled data holds, where using only 50% of the unlabeled data resulted no drop in performance for most of the methods. In Fig. 5c, we adopt a *split-class* sampling technique, where we first randomly select half the classes, and remove $2x$% of data from these classes while keeping images from the rest of the classes the same. This sampling technique would reveal insights into scenarios where the tail properties of the category distribution exhibit significant skewness, and adding unlabeled data translates to correcting the skewed tail property of the dataset. However, the gains yielded from adding more unlabeled data is still limited. Even when the overall trends look positive with non-saturated performance, the absolute gain *is still less than 2%* while doubling the amount of unlabeled data from 50% to 100%, matching the observations made with other sampling techniques.

Table 2: In-task Supervised pre-training helps domain adaptation. We analyze the relationship between data used for supervised pre-training and downstream adaptation for source-only transfer as well as several UDA methods including MemSAC [41], ToAlign [103], MDD [111] and DALN [14]. We show that *in-task* supervised pre-training significantly helps adaptation. All models use ResNet-50 backbone. IN:ImageNet, PL:Places-205, NAT:iNaturalist.

	Plain Transfer (no adapt)			ToAlign [103]			MemSAC [41]			MDD [111]			DALN [14]		
Pre-training	DNet	GeoP	CUB	DNet	GeoP	CUB	DNet	GeoP	CUB	DNet	GeoP	CUB	DNet	GeoP	CUB
IN-1M	**41.46**	34.55	50.20	**49.29**	30.42	62.78	**50.75**	32.98	62.92	**42.40**	30.84	59.84	**47.59**	26.85	61.45
PL-1M	35.14	**41.95**	40.83	38.55	**34.9**	55.29	41.93	**40.16**	54.22	34.94	**37.90**	51.14	39.21	**36.23**	50.74
NAT-1M	33.77	31.53	**58.77**	37.65	26.81	**67.47**	38.67	29.99	**67.34**	32.29	26.79	**63.72**	37.30	24.69	**66.80**

4.3 Does pre-training data matter in UDA?

Motivation Following recent works that reveal the importance of pre-training data in influencing downstream accuracy [19], we revisit a standard practice in UDA to adopt ImageNet pre-trained backbone irrespective of the downstream adaptation task. While Kim et al. [44] share similar motivations as ours, a notable distinction lies in their focus on *scaling* pre-training data and architectures, while we offer complementary insights by exploring the relationship between the *type* of pre-training and downstream adaptation maintaining a *constant* datasize.

Experimental Setup We use ImageNet [78], Places-205 [113] and iNaturalist-2021 [99] as datasets during pre-training. While ImageNet contains images from diverse natural and object categories, Places-205 is designed for scene classification and iNaturalist contains images of bird species. We select 1M images each from ImageNet, Places-205 and iNaturalist datasets (indicated as IN-1M, PL-1M and NAT-1M respectively) to keep the size of the pre-training datasets constant, allowing us to decouple the impact of nature of data from the volume of the dataset. In terms of pre-training methods, we use supervised pre-training using labeled data, along with recent state-of-the-art self-supervised methods SwAV [12], MoCo-V3 [16] and MAE [32], which broadly cover the three families of clustering, contrastive and masked auto-encoding based methods for self-supervised learning. We train SwAV on ResNet-50, MoCo on ViT-S/16 and MAE on ViT-B/16 architectures, along with supervised pre-training on ResNet-50, thereby extending our inferences to a diverse pool of pretraining data and architectures. For the downstream adaptation tasks, we use Real→Clipart on DomainNet, CUB→Drawing on CUB200 and USA→Asia on GeoPlaces covering three distinct application scenarios for adaptation on objects, birds and scenes respectively. To prevent overlap between pre-training and adaptation data, we remove images from Places-205 that are also present in GeoPlaces and remove images from iNaturalist that belong to the same class as those in CUB200.

Supervised Pre-training Using In-Task Data Helps UDA In our analysis, we loosely consider pre-training on ImageNet, iNaturalist and Places205 to be *in-task pre-training* for downstream adaptation on DomainNet, CUB200 and GeoPlaces respectively due to the matching style of images. We show our results using supervised pre-training on Resnet-50 in Tab. 2 for plain source-only transfer (no adaptation), as well as adaptation using ToAlign, MemSAC, MDD and DALN.

Table 3: Self-supervised pre-training and domain adaptation. We find that self-supervised pre-training on object-centric images (on ImageNet) help downstream accuracy on object-centric adaptation (on DomainNet and CUB200), while scene-centric pre-training (on Places205) benefit adaptation on scene-centric GeoPlaces task. IN:ImageNet, PL:Places-205, NAT:iNaturalist

	SwAV (ResNet50) [12]			MoCo-V3 (ViT-s/16) [16]			MAE (ViT-b/16) [32]		
Pretraining	DNet	GeoP	CUB	DNet	GeoP	CUB	DNet	GeoP	CUB
IN-1M	**36.51**	35.76	**31.59**	**30.48**	31.13	**40.7**	**38.58**	35.85	**52.34**
PL-1M	30.86	**42.26**	27.44	27.45	**35.89**	39.49	34.76	**38.1**	45.25
NAT-1M	28.01	29.01	30.12	25.66	27.82	40.03	33.78	31.68	49.4

(a) Plain Transfer (No Adaptation)

	SwAV (ResNet50) [12]			MoCo-V3 (ViT-s/16) [16]			MAE (ViT-b/16) [32]		
Pretraining	DNet	GeoP	CUB	DNet	GeoP	CUB	DNet	GeoP	CUB
IN-1M	**44.6**	36.33	**51.81**	**34.33**	30.35	**52.61**	**44.91**	34.07	**64.26**
PL-1M	36.48	**41.14**	39.49	30.83	**35.51**	46.99	39.56	**37.00**	53.68
NAT-1M	31.6	28.75	45.65	28.24	26.01	48.46	38.48	28.74	59.7

(b) Using MemSAC Adaptation

Across the board, we observe that *in-task pre-training always yields better results on downstream adaptation* even when using the same amount of data. Focusing on plain transfer from Tab. 2, the de-facto choice of ImageNet pre-training gives 50.2% on CUB→Drawing transfer task, while just switching the pre-training dataset to iNaturalist2021 yields 58.7% accuracy with an absolute improvement of 8.5%. Likewise, we observe a non-trivial improvement of 7.4% absolute accuracy for GeoPlaces (34.5% to 41.9%) using Places205 for pre-training even without any adaptation, challenging the common assumption of using an ImageNet-pretrained model irrespective of the downstream task. We hypothesize that supervised pre-training on in-task data creates strong priors with more relevant features, thereby enhancing generalization on similar downstream tasks. Consequently, we conclude that selecting in-task pre-trained models is a viable approach to improve accuracy, particularly when target unlabeled data is unavailable. While similar observations have been made before in continual pre-training [75] or language models [31], our difference lies in highlighting this behavior for the specific case of UDA through a unified framework and controlled empirical study.

In-Task Pre-training is complementary to UDA method We also observe that these benefits obtained from in-task supervised pre-training complement the advantages potentially obtained using UDA methods, resulting in additional improvements in accuracy. From Tab. 2, on CUB200, we observe 17.1% and 17.3% improvement using MemSAC and ToAlign respectively together with in-task pre-training, over standard practice of ImageNet-pretraining and fine-tuning on source data (12% from changing the backbone and further 5% from the adaptation), *setting a new state-of-the-art on CUB200 dataset* using in-task pre-training. On the other hand, a significant mismatch between the pre-training dataset and the downstream domain adaptation dataset (such as Places and Birds datasets), noticeably reduces the accuracy by >10% in most cases, underlining the dependence of model's generalization ability to the pre-training data. While these findings may seem intuitive, it is important to note that all UDA methods consistently utilize ImageNet pre-training as the default, irrespective of the adaptation dataset. This may lead to practitioners assuming ImageNet pre-training as the optimal choice, potentially overlooking performance gains achievable by employing alternative pre-trained models tailored to the target task, as demonstrated by our empirical study.

Nature of Pre-training Images matter for Self-supervised Learning We show results for self-supervised setting in Tab. 3. We first note that supervised pre-training (Tab. 2) achieves much higher accuracies after downstream adaptation compared to self-supervised pre-training. This is expected, as supervised pre-training captures richer object semantics through labels inherently benefiting any downstream task, while self-supervised learning relies on pretext tasks that may not impart equivalent semantic understanding. In terms of pre-training data, we observe that both CUB200 and DomainNet benefit from self-supervised pre-training on ImageNet, while GeoPlaces still benefits from pre-training on Places205. This observation holds for both source-only transfer (Tab. 3a) as well as adaptation using MemSAC (Tab. 3b). We posit that in a self-supervised setting, *the nature of images in the datasets (whether object-centric or scene-centric) plays a crucial role in downstream transfer*. Specifically, unsupervised pre-training on object-centric images from ImageNet leads to improved image classification accuracies on DomainNet and CUB200. Conversely, unsupervised pre-training on scene-centric Places205 showcase better transfer performance in place recognition tasks on the GeoPlaces dataset. Among the two object-centric datasets, we find that the diversity of images in ImageNet is better for effective transfer compared to specific domain-based datasets like iNaturalist, as also highlighted in prior works for self-supervised learning [19]. Furthermore, this property is consistent across different kinds of self-supervised pretext tasks like SwAV, MoCo and MAE.

5 Conclusion

In this work, we provide a holistic analysis of factors that impact the effectiveness UDA methods developed for image-classification, most of which are not apparent from standard training and evaluation practices. Through our innovation called UDA-bench that facilitates fair comparisons across UDA methods, we perform a controlled empirical study revealing key insights regarding the sensitivity of these methods to the backbone architecture, their limited efficiency in utilizing unlabeled data, and the potential for enhancing performance through in-task pre-training - where existing UDA theory proves highly inadequate for explaining several of our novel empirical observations. In terms of limitations of the study, we only consider UDA designed for classification in this work, and our findings might or might not hold for other problem areas such as domain adaptive semantic segmentation. We also acknowledge the potential existence of other unexplored factors that may impact the performance of UDA methods beyond those studied here, and offer UDA-Bench as a suitable avenue for future research in this direction. Further, we mainly focus on the standard setting in unsupervised adaptation, but believe that a deeper understanding of algorithms in such conventional settings forms the backbone for future studies in other variants including source-free [48], semi-supervised [80] and universal [109] DA methods. Several other avenues like adaptation of vision-language models [73, 110] and emerging generative models [49, 76] are also left to a future work.

Acknowledgements

We acknowledge support from NSF and a Google Award for Inclusion Research.

References

1. Alabdulmohsin, I., Chiou, N., D'Amour, A., Gretton, A., Koyejo, S., Kusner, M.J., Pfohl, S.R., Salaudeen, O., Schrouff, J., Tsai, K.: Adapting to latent subgroup shifts via concepts and proxies. In: International Conference on Artificial Intelligence and Statistics. pp. 9637–9661. PMLR (2023)
2. Azizzadenesheli, K.: Importance weight estimation and generalization in domain adaptation under label shift. IEEE Trans. Pattern Anal. Mach. Intell. **44**(10), 6578–6584 (2022). https://doi.org/10.1109/TPAMI.2021.3086060, https://doi.org/10.1109/TPAMI.2021.3086060
3. Bai, Y., Mei, J., Yuille, A.L., Xie, C.: Are transformers more robust than cnns? Advances in Neural Information Processing Systems **34**, 26831–26843 (2021)
4. Baktashmotlagh, M., Harandi, M., Salzmann, M.: Distribution-matching embedding for visual domain adaptation. Journal of Machine Learning Research **17**, Article–number (2016)
5. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W.: A theory of learning from different domains. Machine learning **79**, 151–175 (2010)
6. Ben-David, S., Blitzer, J., Crammer, K., Pereira, F.: Analysis of representations for domain adaptation. Advances in neural information processing systems **19**, 137–144 (2006)
7. Berthelot, D., Roelofs, R., Sohn, K., Carlini, N., Kurakin, A.: Adamatch: A unified approach to semi-supervised learning and domain adaptation. arXiv preprint arXiv:2106.04732 (2021)
8. Bhojanapalli, S., Chakrabarti, A., Glasner, D., Li, D., Unterthiner, T., Veit, A.: Understanding robustness of transformers for image classification. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10231–10241 (2021)
9. Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., Krishnan, D.: Unsupervised pixel-level domain adaptation with generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3722–3731 (2017)
10. Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D., Erhan, D.: Domain separation networks. In: Advances in neural information processing systems. pp. 343–351 (2016)
11. Caputo, B., Müller, H., Martinez-Gomez, J., Villegas, M., Acar, B., Patricia, N., Marvasti, N., Üsküdarlı, S., Paredes, R., Cazorla, M., et al.: Imageclef 2014: Overview and analysis of the results. In: Information Access Evaluation. Multilinguality, Multimodality, and Interaction: 5th International Conference of the CLEF Initiative, CLEF 2014, Sheffield, UK, September 15-18, 2014. Proceedings 5. pp. 192–211. Springer (2014)
12. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. Advances in neural information processing systems **33**, 9912–9924 (2020)

13. Chen, C., Xie, W., Huang, W., Rong, Y., Ding, X., Huang, Y., Xu, T., Huang, J.: Progressive feature alignment for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 627–636 (2019)
14. Chen, L., Chen, H., Wei, Z., Jin, X., Tan, X., Jin, Y., Chen, E.: Reusing the task-specific classifier as a discriminator: Discriminator-free adversarial domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7181–7190 (June 2022)
15. Chen, W.Y., Liu, Y.C., Kira, Z., Wang, Y.C.F., Huang, J.B.: A closer look at few-shot classification. arXiv preprint arXiv:1904.04232 (2019)
16. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9640–9649 (2021)
17. Chen, X., Wang, S., Long, M., Wang, J.: Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In: International conference on machine learning. pp. 1081–1090. PMLR (2019)
18. Choi, D., Shallue, C.J., Nado, Z., Lee, J., Maddison, C.J., Dahl, G.E.: On empirical comparisons of optimizers for deep learning. arXiv preprint arXiv:1910.05446 (2019)
19. Cole, E., Yang, X., Wilber, K., Mac Aodha, O., Belongie, S.: When does contrastive visual representation learning work? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14755–14764 (2022)
20. Courty, N., Flamary, R., Tuia, D., Rakotomamonjy, A.: Optimal transport for domain adaptation. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(9), 1853–1865 (2017). https://doi.org/10.1109/TPAMI.2016.2615921
21. Cui, S., Jin, X., Wang, S., He, Y., Huang, Q.: Heuristic domain adaptation. In: Advances in Neural Information Processing Systems. vol. 33, pp. 7571–7583. Curran Associates, Inc. (2020)
22. Damodaran, B.B., Kellenberger, B., Flamary, R., Tuia, D., Courty, N.: Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In: Proceedings of the European conference on computer vision (ECCV). pp. 447–463 (2018)
23. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
24. Du, Z., Li, J., Su, H., Zhu, L., Lu, K.: Cross-domain gradient discrepancy minimization for unsupervised domain adaptation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3937–3946 (2021)
25. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: International conference on machine learning. pp. 1180–1189. PMLR (2015)
26. Garg, S., Wu, Y., Balakrishnan, S., Lipton, Z.: A unified view of label shift estimation. Advances in Neural Information Processing Systems **33**, 3290–3300 (2020)
27. Goldblum, M., Souri, H., Ni, R., Shu, M., Prabhu, V., Somepalli, G., Chattopadhyay, P., Ibrahim, M., Bardes, A., Hoffman, J., Chellappa, R., Wilson, A.G., Goldstein, T.: Battle of the backbones: A large-scale comparison of pretrained models across computer vision tasks. arXiv preprint arXiv: 2310.19909 (2023)
28. Goyal, P., Duval, Q., Seessel, I., Caron, M., Misra, I., Sagun, L., Joulin, A., Bojanowski, P.: Vision models are more robust and fair when pretrained on uncurated images without supervision. arXiv preprint arXiv:2202.08360 (2022)

29. Goyal, P., Mahajan, D., Gupta, A., Misra, I.: Scaling and benchmarking self-supervised visual representation learning. In: Proceedings of the ieee/cvf International Conference on computer vision. pp. 6391–6400 (2019)
30. Gulrajani, I., Lopez-Paz, D.: In search of lost domain generalization. arXiv preprint arXiv:2007.01434 (2020)
31. Gururangan, S., Marasovic, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., Smith, N.A.: Don't stop pretraining: Adapt language models to domains and tasks. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J.R. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020. pp. 8342–8360. Association for Computational Linguistics (2020). https://doi.org/10.18653/V1/2020.ACL-MAIN.740, https://doi.org/10.18653/v1/2020.acl-main.740
32. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16000–16009 (2022)
33. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arxiv 2015. arXiv preprint arXiv:1512.03385 **14** (2015)
34. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. arXiv preprint arXiv:1903.12261 (2019)
35. Hoffman, J., Rodner, E., Donahue, J., Darrell, T., Saenko, K.: Efficient learning of domain-invariant image representations. arXiv preprint arXiv:1301.3224 (2013)
36. Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A., Darrell, T.: Cycada: Cycle-consistent adversarial domain adaptation. In: International conference on machine learning. pp. 1989–1998. Pmlr (2018)
37. Hsu, T.M.H., Chen, W.Y., Hou, C.A., Tsai, Y.H.H., Yeh, Y.R., Wang, Y.C.F.: Unsupervised domain adaptation with imbalanced cross-domain data. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4121–4129 (2015)
38. Jin, Y., Wang, X., Long, M., Wang, J.: Minimum class confusion for versatile domain adaptation. In: European Conference on Computer Vision. pp. 464–480. Springer (2020)
39. Kalluri, T., Chandraker, M.: Cluster-to-adapt: Few shot domain adaptation for semantic segmentation across disjoint labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4121–4131 (2022)
40. Kalluri, T., Majumder, B.P., Chandraker, M.: Tell, don't show!: Language guidance eases transfer across domains in images and videos. arXiv preprint arXiv:2403.05535 (2024)
41. Kalluri, T., Sharma, A., Chandraker, M.: Memsac: Memory augmented sample consistency for large scale domain adaptation. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXX. pp. 550–568. Springer (2022)
42. Kang, G., Jiang, L., Yang, Y., Hauptmann, A.G.: Contrastive adaptation network for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4893–4902 (2019)
43. Kareer, S., Vijaykumar, V., Maheshwari, H., Chattopadhyay, P., Hoffman, J., Prabhu, V.: We're not using videos effectively: An updated domain adaptive video segmentation baseline. arXiv preprint arXiv: 2402.00868 (2024)
44. Kim, D., Wang, K., Sclaroff, S., Saenko, K.: A broad study of pre-training for domain generalization and adaptation. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII. pp. 621–638. Springer (2022)

45. Kumar, A., Sattigeri, P., Wadhawan, K., Karlinsky, L., Feris, R., Freeman, B., Wornell, G.: Co-regularized alignment for unsupervised domain adaptation. In: Advances in Neural Information Processing Systems. pp. 9345–9356 (2018)
46. Lai, X., Tian, Z., Xu, X., Chen, Y.C., Liu, S., Zhao, H., Wang, L., Jia, J.: Decouplenet: Decoupled network for domain adaptive semantic segmentation. In: European Conference on Computer Vision (2022)
47. Le, Y., Yang, X.: Tiny imagenet visual recognition challenge. CS 231N **7**(7), 3 (2015)
48. Liang, J., Hu, D., Feng, J.: Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In: International conference on machine learning. pp. 6028–6039. PMLR (2020)
49. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning (2023)
50. Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., et al.: Swin transformer v2: Scaling up capacity and resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12009–12019 (2022)
51. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
52. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11976–11986 (2022)
53. Long, M., Cao, Y., Wang, J., Jordan, M.: Learning transferable features with deep adaptation networks. In: International conference on machine learning. pp. 97–105. PMLR (2015)
54. Long, M., Cao, Z., Wang, J., Jordan, M.I.: Conditional adversarial domain adaptation. In: Advances in Neural Information Processing Systems. pp. 1640–1650 (2018)
55. Long, M., Wang, J., Ding, G., Sun, J., Yu, P.S.: Transfer feature learning with joint distribution adaptation. In: Proceedings of the IEEE international conference on computer vision. pp. 2200–2207 (2013)
56. Long, M., Zhu, H., Wang, J., Jordan, M.I.: Deep transfer learning with joint adaptation networks. In: International conference on machine learning. pp. 2208–2217. PMLR (2017)
57. Lucic, M., Kurach, K., Michalski, M., Gelly, S., Bousquet, O.: Are gans created equal? a large-scale study. Advances in neural information processing systems **31** (2018)
58. Mensink, T., Uijlings, J., Kuznetsova, A., Gygli, M., Ferrari, V.: Factors of influence for transfer learning across diverse appearance domains and task types. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(12), 9298–9314 (2021)
59. Morerio, P., Cavazza, J., Murino, V.: Minimal-entropy correlation alignment for unsupervised deep domain adaptation. arXiv preprint arXiv:1711.10288 (2017)
60. Musgrave, K., Belongie, S., Lim, S.N.: A metric learning reality check. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16. pp. 681–699. Springer (2020)
61. Musgrave, K., Belongie, S., Lim, S.N.: Unsupervised domain adaptation: A reality check. arXiv preprint arXiv:2111.15672 (2021)

62. Musgrave, K., Belongie, S.J., Lim, S.N.: Three new validators and a large-scale benchmark ranking for unsupervised domain adaptation. ArXiv, abs/2208.07360 (2022)
63. Na, J., Jung, H., Chang, H.J., Hwang, W.: Fixbi: Bridging domain spaces for unsupervised domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1094–1103 (2021)
64. Nardi, L., Bodin, B., Zia, M.Z., Mawer, J., Nisbet, A., Kelly, P.H.J., Davison, A.J., Luján, M., O'Boyle, M.F.P., Riley, G., Topham, N., Furber, S.: Introducing slambench, a performance and accuracy benchmarking methodology for slam. IEEE International Conference on Robotics and Automation (2014). https://doi.org/10.1109/ICRA.2015.7140009
65. Newell, A., Deng, J.: How useful is self-supervised pretraining for visual tasks? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7345–7354 (2020)
66. Oliver, A., Odena, A., Raffel, C.A., Cubuk, E.D., Goodfellow, I.: Realistic evaluation of deep semi-supervised learning algorithms. Advances in neural information processing systems **31** (2018)
67. Pan, S.J., Tsang, I.W., Kwok, J.T., Yang, Q.: Domain adaptation via transfer component analysis. IEEE transactions on neural networks **22**(2), 199–210 (2010)
68. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)
69. Pei, Z., Cao, Z., Long, M., Wang, J.: Multi-adversarial domain adaptation. arXiv preprint arXiv:1809.02176 (2018)
70. Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., Wang, B.: Moment matching for multi-source domain adaptation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1406–1415 (2019)
71. Peng, X., Usman, B., Kaushik, N., Hoffman, J., Wang, D., Saenko, K.: Visda: The visual domain adaptation challenge. arXiv preprint arXiv:1710.06924 (2017)
72. Prabhu, V., Khare, S., Kartik, D., Hoffman, J.: Sentry: Selective entropy optimization via committee consistency for unsupervised domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8558–8567 (2021)
73. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
74. Redko, I., Courty, N., Flamary, R., Tuia, D.: Optimal transport for multi-source domain adaptation under target shift. In: The 22nd International Conference on Artificial Intelligence and Statistics. pp. 849–858. PMLR (2019)
75. Reed, C.J., Yue, X., Nrusimha, A., Ebrahimi, S., Vijaykumar, V., Mao, R., Li, B., Zhang, S., Guillory, D., Metzger, S., et al.: Self-supervised pretraining improves self-supervised pretraining. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2584–2594 (2022)
76. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
77. Roth, K., Milbich, T., Sinha, S., Gupta, P., Ommer, B., Cohen, J.P.: Revisiting training strategies and generalization performance in deep metric learning. In: International Conference on Machine Learning. pp. 8242–8252. PMLR (2020)

78. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
79. Saenko, K., Kulis, B., Fritz, M., Darrell, T.: Adapting visual category models to new domains. In: Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11. pp. 213–226. Springer (2010)
80. Saito, K., Kim, D., Sclaroff, S., Darrell, T., Saenko, K.: Semi-supervised domain adaptation via minimax entropy. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8050–8058 (2019)
81. Saito, K., Ushiku, Y., Harada, T.: Asymmetric tri-training for unsupervised domain adaptation. arXiv preprint arXiv:1702.08400 (2017)
82. Saito, K., Ushiku, Y., Harada, T., Saenko, K.: Adversarial dropout regularization. arXiv preprint arXiv:1711.01575 (2017)
83. Saito, K., Watanabe, K., Ushiku, Y., Harada, T.: Maximum classifier discrepancy for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3723–3732 (2018)
84. Shao, R., Shi, Z., Yi, J., Chen, P.Y., Hsieh, C.J.: On the adversarial robustness of vision transformers. ArXiv **abs/2103.15670** (2021)
85. Sharma, A., Kalluri, T., Chandraker, M.: Instance level affinity-based transfer for unsupervised domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5361–5371 (2021)
86. Singh, M., Gustafson, L., Adcock, A., de Freitas Reis, V., Gedik, B., Kosaraju, R.P., Mahajan, D., Girshick, R., Dollár, P., Van Der Maaten, L.: Revisiting weakly supervised pre-training of visual perception models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 804–814 (2022)
87. Sun, B., Feng, J., Saenko, K.: Return of frustratingly easy domain adaptation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 30 (2016)
88. Sun, B., Saenko, K.: Deep coral: Correlation alignment for deep domain adaptation. In: Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14. pp. 443–450. Springer (2016)
89. Sun, C., Shrivastava, A., Singh, S., Gupta, A.: Revisiting unreasonable effectiveness of data in deep learning era. IEEE International Conference on Computer Vision (2017). https://doi.org/10.1109/ICCV.2017.97
90. Tan, S., Peng, X., Saenko, K.: Class-imbalanced domain adaptation: an empirical odyssey. In: Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. pp. 585–602. Springer (2020)
91. Tancik, M., Weber, E., Ng, E., Li, R., Yi, B., Kerr, J., Wang, T., Kristoffersen, A., Austin, J., Salahi, K., Ahuja, A., McAllister, D., Kanazawa, A.: Nerfstudio: A modular framework for neural radiance field development. International Conference on Computer Graphics and Interactive Techniques (2023). https://doi.org/10.1145/3588432.3591516
92. Touvron, H., Bojanowski, P., Caron, M., Cord, M., El-Nouby, A., Grave, E., Izacard, G., Joulin, A., Synnaeve, G., Verbeek, J., et al.: Resmlp: Feedforward networks for image classification with data-efficient training. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)

93. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International conference on machine learning. pp. 10347–10357. PMLR (2021)
94. Touvron, H., Cord, M., Jégou, H.: Deit iii: Revenge of the vit. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV. pp. 516–533. Springer (2022)
95. Tsai, Y.H., Hung, W.C., Schulter, S., Sohn, K., Yang, M.H., Chandraker, M.: Learning to adapt structured output space for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7472–7481 (2018)
96. Tzeng, E., Hoffman, J., Darrell, T., Saenko, K.: Simultaneous deep transfer across domains and tasks. In: Proceedings of the IEEE international conference on computer vision. pp. 4068–4076 (2015)
97. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7167–7176 (2017)
98. Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., Darrell, T.: Deep domain confusion: Maximizing for domain invariance. arXiv preprint arXiv:1412.3474 (2014)
99. Van Horn, G., Cole, E., Beery, S., Wilber, K., Belongie, S., Mac Aodha, O.: Benchmarking representation learning for natural world image collections. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12884–12893 (2021)
100. Venkateswara, H., Eusebio, J., Chakraborty, S., Panchanathan, S.: Deep hashing network for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5018–5027 (2017)
101. Vu, T.H., Jain, H., Bucher, M., Cord, M., Pérez, P.: Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2517–2526 (2019)
102. Wang, S., Chen, X., Wang, Y., Long, M., Wang, J.: Progressive adversarial networks for fine-grained domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9213–9222 (2020)
103. Wei, G., Lan, C., Zeng, W., Zhang, Z., Chen, Z.: Toalign: Task-oriented alignment for unsupervised domain adaptation. In: NeurIPS (2021)
104. Wightman, R.: Pytorch image models. `https://github.com/rwightman/pytorch-image-models` (2019)
105. Xie, S., Zheng, Z., Chen, L., Chen, C.: Learning semantic representations for unsupervised domain adaptation. In: International Conference on Machine Learning. pp. 5423–5432 (2018)
106. Xu, R., Li, G., Yang, J., Lin, L.: Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1426–1435 (2019)
107. Xu, T., Chen, W., Wang, P., Wang, F., Li, H., Jin, R.: Cdtrans: Cross-domain transformer for unsupervised domain adaptation. arXiv preprint arXiv:2109.06165 (2021)
108. Yan, H., Ding, Y., Li, P., Wang, Q., Xu, Y., Zuo, W.: Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2272–2281 (2017)

109. You, K., Long, M., Cao, Z., Wang, J., Jordan, M.I.: Universal domain adaptation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2720–2729 (2019)
110. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11975–11986 (2023)
111. Zhang, Y., Liu, T., Long, M., Jordan, M.I.: Bridging theory and algorithm for domain adaptation. arXiv preprint arXiv:1904.05801 (2019)
112. Zhao, B., Yu, S., Ma, W., Yu, M., Mei, S., Wang, A., He, J., Yuille, A., Kortylewski, A.: Robin: a benchmark for robustness to individual nuisances in real-world out-of-distribution shifts. arXiv preprint arXiv:2111.14341 (2021)
113. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE transactions on pattern analysis and machine intelligence **40**(6), 1452–1464 (2017)
114. Zhou, D., Yu, Z., Xie, E., Xiao, C., Anandkumar, A., Feng, J., Alvarez, J.M.: Understanding the robustness in vision transformers. In: International Conference on Machine Learning. pp. 27378–27394. PMLR (2022)
115. Zhu, J., Bai, H., Wang, L.: Patch-mix transformer for unsupervised domain adaptation: A game perspective. Computer Vision and Pattern Recognition (2023). https://doi.org/10.1109/CVPR52729.2023.00347