UNIVERSITY OF CALIFORNIA SAN DIEGO

Domain Adaptation for Fair and Robust Visual Categorization

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Computer Science

by

Subrahmanya Sai Tarun Kalluri

Committee in charge:

Professor Manmohan Chandraker, Chair
Professor Taylor Berg-Kirkpatrick
Professor Sanjoy Dasgupta
Professor Ravi Ramamoorthi
Professor Nuno Vasconcelos

2024

The Dissertation of Subrahmanya Sai Tarun Kalluri is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2024

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to first acknowledge the incredible support and guidance provided by my advisor, Prof. Manmohan Chandraker. Life is all about making the right choices, and I am extremely fortunate to have made the choice to pursue my PhD advised by Prof. Manmohan. I will be eternally grateful to him for trusting me to become a competent independent researcher and giving me complete freedom to pursue my interests while providing constant guidance and support throughout my graduate studies.

I also want to thank Prof. Taylor Berg-Kirkpatrick, Prof. Sanjoy Dasgupta, Prof. Ravi Ramamoorthi and Prof. Nuno Vasconcelos, the other members in my dissertation committee, for providing me several useful suggestions and directions in preparing the dissertation.

I want to thank Dr. Du Tran for giving me the opportunity to do summer internship at Facebook (FAIR) twice, and playing a significant role in shaping most of my research outlook during my PhD. His positive outlook towards life always served as a great source of optimism and energy throughout my internships, and his emphasis on placing value on the process of learning rather than the result greatly influenced my subsequent approach to research. I have to thank Dr. Jeremiah Liu for taking me as an intern at Google Research, and providing great support in handling a challenging project during my internship. I also want to thank several collaborators and mentors during these internships, specifically Deepak Pathak, Lorenzo Torresani, Weiyao Wang, Heng Wang (Facebook) and Kihyuk Sohn, Jihyeon Lee, Joseph Xu (Google), for helping me reshape and refine the project ideas and contributing to the success of those projects. I also want to thank Proj. C.V. Jawahar (I.I.I.T. Hyderabad) for first introducing me to computer vision research, and trusting me with opportunities that significantly reshaped my academic journey. His guidance and belief in my potential have been instrumental in my career growth and development.

I always feel incredibly lucky to spend my PhD years in the beautiful city of San Diego, and the time here was made much more memorable by the several friends and colleagues in UCSD and beyond. I want to thank Ravi, Tejaswi, Vedavyas and Akhil for help and support in

the initial years of my PhD while I navigated challenges of moving to a foreign country, and for making pandemic lockdown less stressful with all the late-night conversations and cooking adventures. I also want to thank all my friends and colleagues at UCSD - Yu-Ying, Rui, Zhengqin, Ishit, Kunal, Alex, Nithin, Yash, Kai-En, Bing, Sina and Hesper, Hemanth, Murali for standing as an unwavering source of strength and positivity during my tenure.

Finally, I want to thank my family - my parents and sister - for their encouragement, support and sacrifice, without which I would not have had the opportunity to pursue my goals and interests. For that, I am forever grateful.

I am also grateful to my co-authors who kindly approved the following publications and material to be included in my dissertation:

Chapter 3 is a reprint of the material as it appears in "Memsac: Memory augmented sample consistency for large scale domain adaptation" by Tarun Kalluri, Astuti Sharma, and Manmohan Chandraker, which was published in Proceedings of the European Conference on Computer Vision, 2022. The dissertation author was the primary investigator and author of this paper.

Chapter 4 is a reprint of the material as it appears in "Geonet: Benchmarking unsupervised adaptation across geographies." by Tarun Kalluri, Wangdong Xu, Manmohan Chandraker, which was published in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. The dissertation author was the primary investigator and author of this paper.

Chapter 5 is a reprint of the material as it appears in "Tell, Don't Show!: Language Guidance Eases Transfer Across Domains in Images and Videos" by Tarun Kalluri, Bodhisattwa Prasad Majumder, and Manmohan Chandraker, which was published in Proceedings of the International Conference on Machine Learning, 2024. The dissertation author was the primary investigator and author of this paper.

Chapter 6 is a reprint of the material as it appears in "UDA-Bench: Revisiting Common Assumptions in Unsupervised Domain Adaptation Using a Standardized Framework" by Tarun Kalluri, Sreyas Ravichandran, and Manmohan Chandraker, which was published in Proceedings

of the European Conference on Computer Vision, 2024. The dissertation author was the primary investigator and author of this paper.

VITA

2012-16	Bachelor of Technology in Engineering, Indian Institute of Technology (I.I.T.) Guwahati, India.
2016-17	Software Engineer, Oracle India Pvt.Ltd. Bangalore, India.
2017-19	Research Assistant, International Institute of Information Technology (I.I.I.T.) Hyderabad, India.
2019-24	Doctor of Philosophy, University of California San Diego, La Jolla, U.S.A.

ABSTRACT OF THE DISSERTATION

Domain Adaptation for Fair and Robust Visual Categorization

by

Subrahmanya Sai Tarun Kalluri

Doctor of Philosophy in Computer Science

University of California San Diego, 2024

Professor Manmohan Chandraker, Chair

Recent advancements in visual categorization have led to significant improvements across various applications, but these models still struggle to generalize effectively to under-represented regions and demographics, directly impacting the fairness and inclusivity of computer vision systems. While domain adaptation has been proposed as a solution to bridge domain gaps using unlabeled data, its effectiveness in handling complex distribution shifts remains insufficiently explored.

This dissertation explores efforts to enhance robustness and transferability in computer vision through domain adaptation. First, we introduce an efficient mechanism to scale domain adaptation to categorization tasks with hundreds of classes. Next, we introduce GeoNet, a dataset

designed to benchmark and analyze geographical disparities in visual categorization tasks. Later, we present our work on using language as a powerful tool to guide the learning of transferable representations across different domains in images and videos, followed by UDABench, a new unified framework aimed at standardizing the training and evaluation of domain adaptation algorithms along with share key insights from this framework. Lastly, we identify significant open research questions that could further advance the concepts discussed in this dissertation.

The primary contributions of this dissertation are to highlight the limitations of current domain adaptation methods in addressing new geographical shifts and to develop novel techniques for domain transfer using language guidance and contrastive learning methods.

Chapter 1

Introduction

In recent years, the field of computer vision and deep learning has been propelled by the emergence of large-scale foundational models, which unlock remarkable capabilities in various tasks across scene understanding, robot navigation, text-to-image synthesis and nuanced multi-modal dialogue. However, their reliance on uncurated, web-sourced data presents new challenges, where the biases in training data can lead to unfair outcomes for under-represented subgroups and their lack of robustness outside their training domain limits their universal adoption. These limitations extend to several applications such as autonomous driving, disease monitoring, remote surveillance and personal-assistive technologies. For instance self-driving technologies would significantly improve mobility and road-safety in high traffic-density geographies like Asia and Africa, but most benchmark datasets are instead collected from US or Europe with little to no representation from other countries, with notable domain gaps preventing robust transfer across geographies. This dissertation focuses on improving the generalizability of vision models across under-represented domains in various real-world settings using improved techniques for unsupervised adaptation.

As we transition to real-world adaptation scenarios, practical datasets often include numerous categories, leading to challenges like reduced inter-class discriminability. To address this, we introduce MemSAC, a novel variant of contrastive loss enhanced by a feature memory bank, designed to improve discriminative transfer across large-scale datasets. Our approach

efficiently handles an arbitrary number of classes with minimal negative alignment, setting new state-of-the-art results on challenging datasets such as DomainNet and CUB-200.

In the subsequent chapters, we address the lack of suitable benchmarks for assessing the geographical sensitivity of existing methods that has been a major obstacle to progress in geographical fairness research. To tackle this issue, we introduce GeoNet, a large-scale dataset and evaluation benchmark focused on studying geographical disparities in standard vision tasks. GeoNet is the largest dataset of its kind for training and evaluation in geographical adaptation. With it, we analyze key aspects of geographic distribution shifts and reveal the limitations of several modern algorithms in overcoming these domain gaps. GeoNet not only allows researchers to evaluate the effectiveness of state-of-the-art algorithms for universal deployment but also encourages the development of robust AI models that can adapt to dynamic geographic changes while maintaining high performance.

Recently, natural language has proven effective in enhancing the robustness and open-vocabulary capabilities of vision models. However, its role in addressing domain shifts remains underexplored. Building on the insight that language, with its richer semantics, tends to experience fewer domain shifts than images while offering better discriminative power, we developed a novel framework that leverages easily accessible text descriptions to guide the transfer of discriminative knowledge from labeled source data to unlabeled target data, effectively bridging domain gaps.

Finally, to foster open-source efforts in the field of unsupervised domain adaptation, we propose a standardized framework for implementing and evaluating domain adaptation algorithms using a unified testbed. Through an extensive empirical study, we discover several interesting and non-trivial observations pertaining to the role of backbone architectures, amount of unlabeled target data and pre-training data in unsupervised adaptation.

1.1 Outline

The dissertation is organized as follows. In Chapter 2, we first provide some background on the relevant topics presented in this dissertation. In the following chapters, we introduce different innovations in handling new challenges in unsupervised adaptation for large-scale real-world data.

In Chapter 3, we introduce a new method that leverages the similarity between labeled data from the source domain and unlabeled data from the target to improve knowledge transfer across domains. The method includes a memory-augmented approach that efficiently identifies relationships between pairs of data points when operating with large number of classes and small batch sizes. We also introduce a new variation of the contrastive loss function, which encourages the model to maintain consistency within the same class across domains while ensuring clear separation between different classes. This approach helps preserve the model's ability to discriminate between categories as it adapts from the source domain to the target domain, making it more effective for large-scale datasets with potentially fine-grained classes.

In Chapter 4, we introduce a novel problem of geographical domain adaptation, studied through the lens of a new large-scale dataset called GeoNet. We first explain the process of collection, curation and filtering the dataset to represent geographical diversity, and study various kinds of domain shifts unique to geographical disparity between domains including context shift, design shift and label shift. We then use our benchmark to study the competence of current domain adaptation methods in addressing geographical transfer and analyze the role of large-scale pre-training in imparting geographical robustness to downstream models.

In Chapter 5, we introduce a new language-guided domain adaptation mechanism specifically designed to excel in those scenarios where both source and target domains have natural language supervision in the form of captions or alt-text. We design a framework for efficiently leveraging naturally available or easily generated text supervision to reduce domain shifts and improve cross-domain transfer in image and video classification tasks. Through a novel cross-modal

distillation framework which transfers predictions from text to visual space, we achieve superior transfer performance on challenging GeoNet dataset for images as well as newly introduced Ego2Exo dataset for videos.

In Chapter 6, we explore the various factors that affect the success of modern unsupervised domain adaptation (UDA) methods through a controlled empirical study. We created UDA-Bench, a new PyTorch framework that standardizes training and evaluation for domain adaptation, allowing for fair comparisons across different UDA methods. Our study using UDA-Bench shows that: (i) the advantages of adaptation methods decrease with more advanced backbone architectures, (ii) current methods don't fully make use of unlabeled data, and (iii) pre-training datasets have a significant impact on later adaptation performance in both supervised and self-supervised settings. These findings provide valuable insights into unsupervised adaptation, challenging previous assumptions based on intuition or empirical observations without a standardized framework.

Finally, in Chapter 7, we summarize and discuss potential future directions inspired by the ideas presented in this dissertation for improving robustness for future computer vision models.

Chapter 2

Background

The traditional ML pipeline evaluates and tests a model's performance on data that matches the distribution used in training. However, learning-based methods often experience a significant drop in performance and accuracy when faced with test data from a different distribution than the training data. 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) [97, 138, 142, 20, 21, 139, 69, 195, 196, 260, 98, 248, 104, 207, 110, 238, 107, 106, 15, 269] that allow training using only unlabeled data from the target domain of interest while leveraging supervision from a different source domain with abundant labels.

There exists several design choices pertaining to the alignment objective used in domain adaptation including MMD distance [138, 142, 99, 252, 11, 161, 140, 226], higher-order correlations [150, 212, 211, 104], optimal-transport [47, 179, 50] or generative methods [19, 198], but the paradigm which has seen the greatest success has been adversarial discriminative training [69, 226, 225, 247, 139, 224, 28, 196]. A key goal of adversarial methods is to align the feature representations of the source and target domain using GAN-based objective to make these domains indistinguishable to a domain discriminator trained to classify between the source and target domains. This is realized by using a discriminator based loss on unlabeled samples

from source and target along with the classifier loss on the labeled source images.

$$\mathscr{L}_{\text{total}} = \mathscr{L}_{\text{sup}} + \lambda \cdot \mathscr{L}_{\text{adversarial}}, \tag{2.1}$$

where λ is a hyper-parameter to select the strength of adaptation. While adversarial objective adopted in popular works like DANN [69] and ADDA [225] rely on global domain alignment, in Chapter 3, we use the adversarial objective in CDAN [139] as the backbone to design our approach since it is capable of class-specific alignment more suited to our eventual objective of many class adaptation. More details on the working of CDAN are presented in Sec. 3.3.1.

On the other hand, geographic robustness is a key barrier to ensure equitable and fair deployment of computer vision models. While geographical sensitivity in large-scale data is a well-known artefact [57], the role of domain adaptation in bridging the domain gap across geographies with idiosyncratic distribution shifts is relatively under-studied. While algorithms for bridging geographic domain gaps have been proposed in [39, 108, 235], they are restricted to road scenes with limited number of classes. A major hindrance has been the lack of suitable benchmark datasets for geographic adaptation, so several datasets have been recently proposed to address this issue [204, 58, 170, 182]. Different from these, we aim to study this problem using evaluation benchmarks which are much larger in scale thereby facilitating training of domain adaptation algorithms.

A crucial observation through Chapter 4 is that a majority of domain adaptation algorithms do not efficiently work for bridging geography-specific shifts, and one of the major reasons for this is their sole reliance on pixel-level reasoning in aligning distributions. However, additional information in the form of text supervision, which is ubiquitously available for many web-sourced images, is shown to be highly effective in grounding the representation of images [172]. We leverage this fact to design a new domain adaptation method using cross-modal distillation from text modality to visual modality. A fundamental building block in our framework is the BERT [56] model for sentence classification. BERT (Bidirectional Encoder

Representations from Transformers) is a deep learning model that leverages a bidirectional transformer architecture for natural language processing tasks. The core idea is to pre-train the model using a masked language model (MLM) objective, where tokens are randomly masked, and the model predicts them based on surrounding context. Formally, BERT optimizes the likelihood:

$$\mathscr{L}_{\text{MLM}} = -\Sigma_{i \in \text{masked}} \log P(x_i | x_{\text{context}}). \tag{2.2}$$

BERT also uses a next sentence prediction (NSP) objective to learn sentence relationships. This approach enables BERT to achieve state-of-the-art results on tasks like QA and text classification, significantly improving transfer learning in NLP. In Chapter 5, we adopt a variant of pre-trained BERT model from HuggingFace [197] called Distill-BERT and fine-tune it on the captions for the task of sentence classification.

Furthermore, we also develop a new adaptation task between ego and exo views of a video, by sourcing data from the popular Ego4d [80, 81] dataset. Ego4D is a massive dataset of first-person (egocentric) videos collected from diverse individuals worldwide. Egocentric video recognition aims to understand actions, objects, and interactions from a wearer's perspective. Unlike traditional (exocentric) videos filmed from a fixed viewpoint, egocentric videos offer a unique first-person perspective, enabling development of AI systems that can mimic human perception and interaction with the environment. Owing to the natural dominance of exocentric videos in the internet, there exists a domain gap between ego and exocentric videos, which we seek to bridge.

Finally, following the observation that most domain adaptation methods are trained and evaluated using individual training frameworks and evaluation protocols, we try to unify them using standardized implementation for fairer comparisons. While prior works in the literature highlighted this issue in other related tasks in computer vision such as semi-supervised learning [159], metric learning [152, 185], transfer learning [146], domain generalization [84], optimization algorithms [41], contrastive learning [46], GANs [143] and self-supervised learn-

ing [78, 157, 75], we extend these efforts to the case of unsupervised domain adaptation.

More detailed related work for Chapters 4, 5, 3 and 6 will be provided in Chapters 4.2, 5.2, 3.2 and 6.2 respectively.

Chapter 3

Memory-augmented Sample Consistency for Large-Scale Domain Adaptation

While the field of domain adaptation decisively pushed the needle towards more equitable and robust models, their evaluation and utility is mostly restricted to small-scale datasets, preventing their use in real-world problems with plentiful categories. Carrying adaptation across larger-scale supervised datasets with many categories introduce additional challenges for unsupervised domain adaptation like small inter-class discriminability, that existing approaches relying only on domain invariance cannot handle sufficiently well. In this chapter, we propose MemSAC, which exploits sample level similarity across source and target domains to achieve discriminative transfer, along with architectures that scale to a large number of categories. For this purpose, we first introduce a memory augmented approach to efficiently extract pairwise similarity relations between labeled source and unlabeled target domain instances, suited to handle an arbitrary number of classes. Next, we propose and theoretically justify a novel variant of the contrastive loss to promote local consistency among within-class cross domain samples while enforcing separation between classes, thus preserving discriminative transfer from source to target. Overall, our algorithm proposed in this work served as state-of-the-art for long time on the highly challenging DomainNet dataset before the introduction of better vision-language models like CLIP [173] and SigLIP [255].

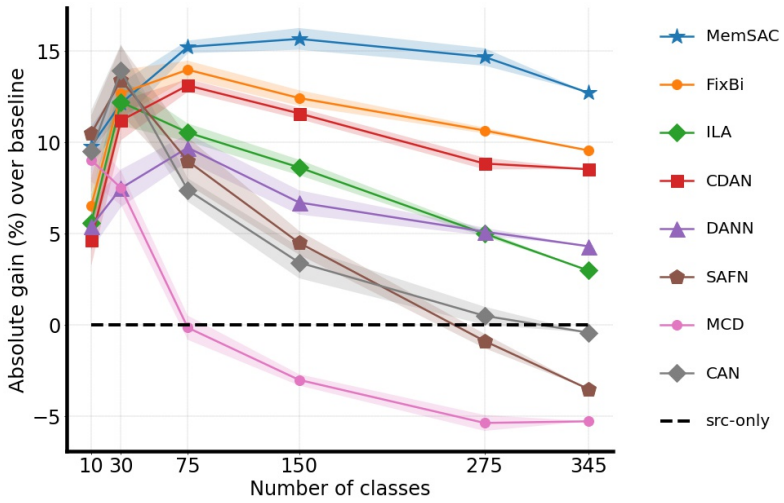


Figure 3.1. Accuracy(%) of various methods proposed for unsupervised domain adaptation with respect to the number of training classes from DomainNet[165](**R**→**C**). While most methods perform equally well for smaller number of categories (10-30), the benefits diminish with increasing number of classes in the dataset, to the extent that the performance drops *even below the source-only baseline* for few methods, while MemSAC obtains significant gains ($\sim 15\%$) even on large scale datasets with many classes [165].

3.1 Introduction

It is well known that deep neural networks often do not generalize well when the distribution of test samples significantly differs from those in training. Unsupervised domain adaptation seeks to improve transferability in the presence of such *domain shift*, for which a variety of approaches have been proposed [12, 13, 69, 139, 140, 138, 141, 21, 20, 195, 226, 224, 225, 248, 38, 83, 155, 62]. Despite impressive gains, most approaches have been largely demonstrated on datasets with a limited number of categories [187, 166].

We first ask the question of whether existing domain adaptation methods scale to a large number of categories. Surprisingly, the answer is usually no. To illustrate this, consider Figure 3.1, which plots the absolute gain over a source-only model obtained by well-known adaptation methods (including DANN [69], MCD [196], SAFN [248], CAN [110], FixBi [155]) with respect to number of classes sampled from the DomainNet dataset [165]. While all methods provide similar benefits over a source-only model in smaller-scale settings with 10-30 classes,

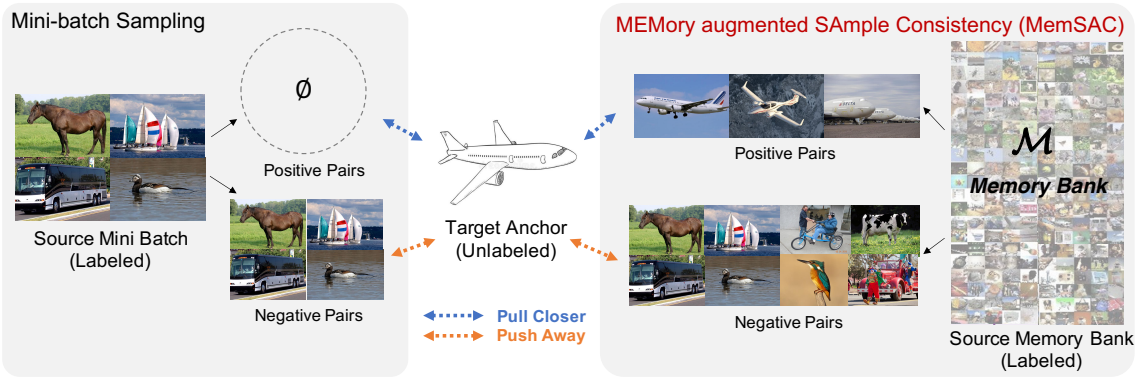


Figure 3.2. **(MEMory augmented SAmple Consistency (MemSAC)** The proposed method uses a memory bank and a sample consistency loss to identify source samples across a large number of categories that likely belong to the same class as an unlabeled target example, then pulls them together in feature space while pushing away samples from all other classes. Notice that without the proposed feature aggregation, a target anchor sample might not find any positive pairs (∅) leading to noisy consistency estimates.

the gains reduce when faced with a few hundred classes, where accuracies may even become *worse than a source-only model*.

We postulate that the above limitations with a larger number of categories arise due to lower inter-class separation and a greater possibility of negative transfer. Our key design choices stem from simple yet effective mechanisms developed in other areas such as self-supervised learning that can significantly benefit many-class domain adaptation. The resulting method, MemSAC (MEMory augmented SAmple Consistency), achieves impressive performance gains to establish new state-of-the-art on datasets such as DomainNet (345 classes) and CUB (200 classes). In the same illustration above, MemSAC obtains large improvements of 14.6% over a source-only baseline for 275 classes and 12.7% for 345 classes.

Our first insight for many-class domain adaptation pertains to class confusion, where several classes possibly look similar to each other. Classical adversarial approaches [69, 196, 247, 248, 110] which rely on domain alignment alone do not acknowledge this, giving rise to negative transfer as two seemingly close classes might align with each other. This problem is exacerbated in the extreme case of fine-grained datasets, where all the classes look similar to each

other. On the other hand, class specific alignment strategies [155, 62, 195, 164, 110, 155] suffer from noisy pseudo-labels leading to poor transfer. We observe that the contrastive loss is shown to be highly successful in learning better transferable features [89, 35, 32, 245, 92, 149, 76, 82] and seek to extend those benefits to many-class domain adaptation. We achieve this with a novel *cross-domain sample consistency* loss which tries to align each sample in source domain with related samples in target domain, achieving tighter clusters and improved adaptation in the process. We provide theoretical justification for the effectiveness of our proposed loss by showing that it is akin to minimizing an upper bound to the input-consistency regularization recently proposed in [237], thereby ensuring that locally consistent prediction provides accuracy guarantees on unlabeled target data for unsupervised domain adaptation.

Our second insight pertains to architectural choices for training with a large number of categories. While having access to plentiful positive and negative pairwise relations per training iteration is desirable to infer local structure, the number of possible pairs are inherently restricted by the batch-size which is in turn limited by the GPU memory. We efficiently tackle this challenge in MemSAC by augmenting the adaptation framework with a lightweight, non-parametric memory module. Distinct from prior works [89, 234], the memory module in our setting aggregates the *labeled* source domain features from multiple recent mini-batches, thus providing *unlabeled* target domain anchors meaningful interactions from sizeable positive and negative pairs even with reasonably small batch sizes that do not incur explosive growth in memory (Fig. 3.2). Our architecture scales remarkably well with the number of categories, including the case of fine-grained adaptation [233] where all classes belong to a single subordinate category [257, 22]. Moreover, MemSAC incurs negligible overhead in terms of speed and GPU memory during training and testing, making it an attractive choice for real-world usage of large-scale adaptation.

We highlight our key contributions as follows:

1. A novel cross-domain sample consistency loss to enforce closer clustering of same category

samples across source and target domains by exploiting pairwise relationships, thus achieving improved domain transfer even with many categories (Sec. 3.3.2).

2. A memory-based mechanism to handle limited batch-sizes by storing past features and effectively extracting similarity relations over a larger context for large scale datasets (Sec. 3.3.4).

3. Theoretical justification of the proposed losses in terms of the input-consistency regularization proposed in [237] for domain adaptation (Sec. 3.4).

4. A new state-of-the-art that outperforms all prior approaches by a significant margin on datasets with a large number of categories, such as 4.02% and 4.65% improvements in accuracy over the baseline which does not use our loss on the challenging DomainNetdataset with 345 categories and CUB-Drawings with 200 categories, respectively (Sec. 3.5).

3.2 Relation to Prior Literature

3.2.1 Fine Grained Domain Adaptation

Fine grained visual categorization deals with classifying images that belong to a single subordinate category, such as birds, trees or animal species [240, 229]. While fine grained classification on within domain samples has received much attention [258, 264, 214, 257, 22, 129, 265], the problem of unsupervised domain adaptation across fine-grained categories is relatively less studied [71, 49, 250, 233]. All prior works often demand additional annotations in the form of attributes [71], weak supervision [49], part annotations [250] or hierarchical relationships [233] in one of the domains which might not be universally available. In contrast, we propose a method that performs fine-grained adaptation requiring no such additional knowledge.

3.2.2 Contrastive Learning

The success of contrastive learning [87, 86, 6, 237] in extracting visual representations from unlabeled data has attracted wide interest [89, 35, 32, 245, 92, 149, 76, 82]. A unifying

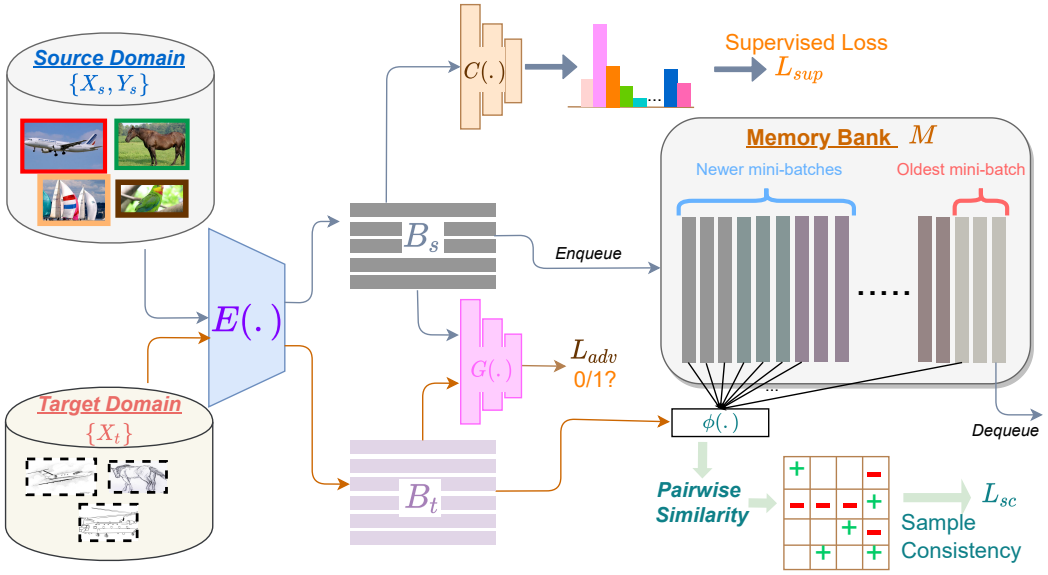


Figure 3.3. An overview of MemSAC for domain adaptation During each iteration, the 256-dim source feature embeddings computed using $\mathscr{E}$, along with their labels, are added to a memory bank $\mathscr{M}$ and the oldest set of features are removed. Pairwise similarities between each target feature in mini-batch and all source features in memory bank are used to extract possible within-class and other-class source samples from the memory bank. Using the proposed consistency loss ($\mathscr{L}_{sc}$) on these similar and dissimilar pairs, along with adversarial loss ($\mathscr{L}_{adv}$), we achieve both local alignment and global adaptation.

idea in those works is to encourage positive pairs, which are often augmented versions of the same image, to have similar representations in the feature space while pushing negative pairs far away. However, all those prior works assume that all positive and negative pairs in the contrastive loss come from the same domain. In contrast, we propose a variant of contrastive loss to handle multi-class discriminative transfer by enforcing sample consistency across similar samples extracted from different domains.

3.3 Unsupervised Adaptation using MemSAC

In unsupervised domain adaptation, we have labeled samples $\mathscr{X}^s$ from a source domain with a corresponding source probability distribution P_s, labeled according to a true labeling function f^*, and $\mathscr{Y}^s = f^*(\mathscr{X}^s)$. We are also given unlabeled data points $\mathscr{X}^t$ sampled according

to the target distribution P_t. We follow a *covariate shift assumption* [12], where we assume that the marginal source and target distributions P_s and P_t are different, while the true labeling function f^* is same across the domains. The labels belong to a fixed category set $\mathscr{Y} = \{1, 2, \ldots, C\}$ with C different categories. Provided with this information, the goal of any learner is to output a predictor that achieves good accuracy on the target data $\mathscr{X}_t$. A key novelty in our instantiation of this framework lies in proposing an adaptation approach that works well even with a large number of classes C, by efficiently handling class confusion and discriminative transfer. The overview of the proposed architecture is shown in Fig. 3.3. $\mathscr{E}$ and $\mathscr{C}$ are the feature extractor and the classifier respectively. The objective function for MemSAC is given by

$$\min_{\theta} \mathscr{L}_{sup}(\mathscr{X}^s, \mathscr{Y}^s; \theta) + \lambda_{adv}\mathscr{L}_{adv}(\mathscr{X}^s, \mathscr{X}^t; \theta) + \lambda_{sc}\mathscr{L}_{sc}(\mathscr{X}^s, \mathscr{Y}^s, \mathscr{X}^t; \theta), \tag{3.1}$$

where $\mathscr{L}_{sup}$ is the supervised loss on source data, or the cross-entropy loss between the predicted class probability distributions and ground truths computed on source data. $\mathscr{L}_{adv}$ is the domain adversarial loss which we implement using a class conditional discriminator (Eq. (3.2)) and $\mathscr{L}_{sc}$ is our novel cross-domain sample-consistency loss which is used to enforce the local similarity (or dissimilarity) between samples from source and target domains (Eq. (3.4)). λ_{adv} and λ_{sc} are the corresponding loss coefficients. We use $\mathscr{B}_s(\in \mathscr{X}^s)$ and $\mathscr{B}_t(\in \mathscr{X}^t)$ to denote labeled source and unlabeled target mini-batches respectively, which are chosen randomly at each iteration from the dataset.

3.3.1 Class Conditional Adversarial Loss

We adopt the widely used adversarial strategy to learn domain-invariant feature representations using a domain discriminator $\mathscr{G}(., \omega)$ parametrized by ω. To address the novel challenges presented by the current setting with large number of classes, we adopt the multilinear conditioning proposed in CDAN [139] to fuse information from the deep features as well as the classifier predictions. Denoting $f = \mathscr{E}(x)$ and $g = \mathscr{C}(\mathscr{E}(x))$, the input $h(x)$ to the discriminator

$\mathscr{G}$ is given by $h(x) = T_{\otimes}(g,f)(x) = f(x) \otimes g(x)$, where $\otimes$ refers to the multilinear product (or flattened outer product) between the feature embedding and the softmax output of the classifier. The discriminator and adversarial losses are then computed as

$$\mathscr{L}_d = \frac{1}{|\mathscr{B}_s|} \sum_{i \in \mathscr{B}_s} -\log(\mathscr{G}(h_i;\omega)) + \frac{1}{|\mathscr{B}_t|} \sum_{i \in \mathscr{B}_t} -\log(1-\mathscr{G}(h_i;\omega)) \qquad \mathscr{L}_{adv} = -\mathscr{L}_d. \tag{3.2}$$

We note that our contributions are complementary to the type of alignment objective used. In Tab. 3.3a, we show significant gains starting from another adversarial objective (DANN [69]) and MMD objectives (CAN [111]) as well.

3.3.2 Cross Domain Sample Consistency

To achieve category specific transfer from source to target, we propose using much finer sample-level information to enforce consistency between similar samples, while also separating dissimilar samples across domains. Since our final goal is to transfer the class discriminative capability from source to target, we define the notions of similarity and dissimilarity as follows. For each target sample x_t from a target mini-batch $\mathscr{B}_t$ as the anchor, we construct a *similar set* $\mathscr{B}_{s^+}^{x_t} = \{x \in \mathscr{B}_s | f^*(x) = f^*(x_t)\}$ and dissimilar set $\mathscr{B}_{s^-}^{x_t} = \mathscr{B}_s \backslash \mathscr{B}_{s^+}^{x_t}$ consisting of source samples and use this knowledge of sample-level similarity in the following *sample consistency loss*

$$\mathscr{L}_{sc,\mathscr{B}} = \frac{1}{|\mathscr{B}_t|} \sum_{j \in \mathscr{B}_t} -\log \left\{ \sum_{i \in \mathscr{B}_{s^+}^j} \frac{\exp(\phi_{ij}/\tau)}{\sum_{i \in \mathscr{B}_s} \exp(\phi_{ij}/\tau)} \right\} \tag{3.3}$$

where ϕ_{ij} measures the cosine similarity metric between two feature vectors i and j, given by the equation ($\phi_{ij}=\phi(f_i,f_j)=\frac{f_i \cdot f_j}{||f_i||||f_j||}$) and τ is the temperature parameter used to scale the contributions of positive and negative pairs [32, 96]. $\mathscr{L}_{sc,\mathscr{B}}$ denotes the sample consistency loss computed using the mini-batch. Distinct from standard constrative loss [149, 32] that typically derives positive pairs from augmented versions of the same image, our loss in Eq. (3.3) is well-suited to handle multiple positive and negative pairs for each anchor, similar to supervised

contrastive loss [113].

3.3.3 kNN-based Pseudo-Labeling

There are two challenges in directly using the sample consistency loss in Eq. (3.3). Firstly, unlike prior approaches [92, 32, 149] that use random transformations of same image to construct positives and negatives, the target data in unsupervised domain adaptation is completely unlabeled, so we do not have the similarity information readily. To address this issue, we use a k-NN based pseudo-labeling trick for all the target samples in a mini-batch. In every iteration of the training, for each target sample x_t from the target training mini-batch $\mathscr{B}_t$, we find k nearest neighbors from the source training mini-batch $\mathscr{B}_s$, which are computed using the feature similarity scores ϕ_{i,x_t}. x_t is then assigned the label corresponding to the majority class occurring among its neighbors. We use a value of k=5. Such an approach for pseudo-labeling is independent of, thus less sensitive to, noisy classifier boundaries helping us extract reliable target pseudo-labels during training. Once $\mathscr{B}_t$ is pseudo-labeled, it is straightforward to compute $\mathscr{B}_{s^+}^{x_t}$ in Eq. (3.3). The second challenge is lack of representation for all classes in a mini-batch, which we address next.

3.3.4 Memory Augmented Similarity Extraction

From Eq. (3.3), we can observe that if the source and target mini-batches $\mathscr{B}_s$ and $\mathscr{B}_t$ contain completely non-intersecting classes, then the pseudo labeling of targets and the subsequent sample consistency loss would be noisy and lead to negative impact. This problem is exacerbated in our setting with a large number of classes, as randomly sampled $\mathscr{B}_s$ and $\mathscr{B}_t$ usually contain many images with mutually non-intersecting categories. While one solution is to increase the size of mini-batch, it comes with significant growth in memory which is not scalable.

Therefore, we propose using a non-parametric memory bank $\mathscr{M}$ that aggregates the computation-free features, along with the corresponding labels, across multiple past mini-batches

from the source dataset. We note that if the size of the memory bank $|\mathscr{M}|$ is sufficiently large, then source samples from all the classes would be adequately present in $\mathscr{M}$, providing us with authentic positive and negative samples for use in the sample consistency loss. Furthermore, since the memory overhead of storing the features in the memory bank itself is negligible (we only store the computation-free features), proposed adaptation approach can be scaled to handle arbitrarily large number of classes, as datasets with larger classes only requires us to correspondingly increase the size of $\mathscr{M}$, thus decoupling the similarity computation with mini-batch size or dataset size. Different from prior approaches that augment training with memory module [245, 89, 234], our approach aggregates features from multiple source batches, thus helping target samples to extract meaningful pairwise relationships from different classes.

Initializing and updating the memory bank

To initialize the memory bank, we first bootstrap the feature extractor for few hundred iterations by training only using $\mathscr{L}_{sup}$ and $\mathscr{L}_{adv}$ losses before introducing our consistency loss $\mathscr{L}_{sc}$ and start populating $\mathscr{M}$. After this, we follow a queue based approach for updating the memory bank similar to XBM [234]. In each iteration, We remove (*dequeue*) the oldest batch of features from the queue and insert (*enqueue*) the fresh mini-batch of source features (computed as $\{\mathscr{E}(x)|x \in \mathscr{B}_s\}$) along with the corresponding source labels. Alternative strategies for updating $\mathscr{M}$, such as a momentum encoder [89], yielded similar results.

Sample consistency using memory bank

We can now use $\mathscr{M}$ as a proxy for $\mathscr{B}_s$ (and similar set $\mathscr{M}_+^{x_t}$ as a proxy for $\mathscr{B}_{s^+}^{x_t}$) in assigning the target pseudo labels and in the sample consistency loss in Eq. (3.3). $|\mathscr{M}|$ is often much higher than $|\mathscr{B}_s|$, so access to larger number of source samples from $\mathscr{M}$ provides k-NN pseudo labels that are more reliable, with richer variety of positive and negative pairwise relations. The final sample consistency loss used in MemSAC is

$$\mathscr{L}_{sc} = \frac{1}{|\mathscr{B}_t|} \sum_{j \in \mathscr{B}_t} -\log \left\{ \sum_{i \in \mathscr{M}_+^j} \frac{\exp(\phi_{ij}/\tau)}{\sum_{i \in \mathscr{M}} \exp(\phi_{ij}/\tau)} \right\}. \tag{3.4}$$

3.4 Theoretical Insight

Recently, Wei et al. [237] provide theoretical validation for contrastive learning. Specifically, under an *expansion* assumption which states that class conditional distribution of data is locally continuous, they bound the target error of a classifier C parametrized by θ by encouraging consistent predictions on neighboring examples. The regularization objective $R(\theta)$ is given by

$$R(\theta) \equiv \min_{\theta} \mathbb{E}_x[\max_{x' \in \mathscr{N}(x)} \mathbf{1}(C(x;\theta) \neq C(x';\theta))],$$

where $\mathscr{N}(x)$ is the neighborhood of a sample x (Eq 1.2 in [237]). We now show the connections that can be drawn between our loss and the theory proposed in [237]. For this purpose, we work with the following approximations. Firstly, we approximate the neighborhood $\mathscr{N}(x)$ of a sample x with the *similar set* defined in Sec. 3.3.2, that is $\mathscr{N}(x) = \mathscr{B}_+^x$. Next, we approximate the hard condition that the classifier outputs of two images be equal $\mathbf{1}(C(x;\theta) \neq C(x';\theta))$, with the soft probability $\mathbf{Pr}(C(x;\theta) \neq C(x';\theta))$. Starting with the above objective, we have

$$\begin{aligned}
&\max_{x' \in \mathscr{N}(x)} \mathbf{1}(C(x;\theta) \neq C(x';\theta)) \\
&\leq \sum_{x' \in \mathscr{N}(x)} \mathbf{Pr}(C(x;\theta) \neq C(x';\theta)) \\
&\approx |\mathscr{B}_+^x| - \sum_{x' \in \mathscr{B}_+^x} \mathbf{Pr}(C(x;\theta) = C(x';\theta)) \\
&\leq |\mathscr{B}_+^x| - \sum_{x' \in \mathscr{B}_+^x} \frac{\exp(\phi_{x,x'})}{\sum_{x' \in \mathscr{B}} \exp(\phi(x,x'))} \\
\implies R(\theta) &\equiv \max_{\theta} \mathbb{E}_x \left[\sum_{x' \in \mathscr{B}_+^x} \frac{\exp(\phi_{x,x'})}{\sum_{x' \in \mathscr{B}} \exp(\phi(x,x'))} \right]
\end{aligned}$$

where we used the softmax similarity between samples x, x' in the feature space as a proxy for the equality of their classifier outputs and changed max to sum with the bound. Under these specific assumptions, we can now see that the input-regularization objective $R(\theta)$ is strongly reminiscent of our sample consistency loss. Using Eq. (3.4), we minimize the negative log-likelihood of the similarity probability, which is equivalent to maximizing the similarity probability of like samples. Therefore, our sample consistency objective is akin to minimizing an upper bound on the input consistency regularization proposed in [237]. Furthermore, optimizing such an objective is shown to achieve bounded target error for unsupervised domain adaptation. Specifically, under the assumption that the pseudo label accuracy on target data is above a certain threshold, [237] showed that bounded error on target data is achievable using the consistency regularization (Theorem 4.3). In MemSAC, we realize this assumption by first training the feature extractor only using supervised ($\mathscr{L}_{sup}$) and adversarial ($\mathscr{L}_{adv}$) losses as explained in Sec. 3.3.4 before introducing our proposed sample consistency loss. To the best of our knowledge, we are the first to instantiate the regularization proposed in [237] for large scale domain adaptation, and showcase its effectiveness in achieving significant empirical gains.

3.5 Experimental Analysis for MemSAC

3.5.1 Datasets

Consistent with the key motivations that distinguish MemSAC from prior literature in domain adaptation, we focus on large-scale datasets with many categories to underline its benefits. First, DomainNet [165] is a large-scale dataset for UDA covering 6 domains and a total of 500k images from 345 different categories. It is an order of magnitude larger compared to prior benchmarks and serves as a useful testbed for evaluating many-class adaptation models. We follow the protocol established in prior works [190, 169, 216] to use data from 4 domains, namely real (**R**), clipart (**C**), sketch (**S**) and painting (**P**), showing results on all 12 transfer tasks across these 4 domains.

For evaluating our method on fine-grained dataset, we use CUB (Caltech-UCSD birds) which is a challenging dataset originally proposed for fine-grained classification of 200 categories of birds, while *CUB-Drawings* [233] consists of paintings corresponding to the 200 categories of birds in CUB. We use this dataset pair, consisting of 14k images in total, for evaluation of adaptation on images with fine-grained categories. This setting can be challenging as appearance variations across species can be subtle, while pose variations within a class can be high. Thus, discriminative transfer requires precisely mapping category-specific information from source to target to avoid negative transfer.

3.5.2 Training Details

We use a Resnet-50 [91] backbone pretrained on Imagenet, followed by a projection layer as the encoder $\mathscr{E}$ to obtain 256 dimensional feature embeddings. The discriminator $\mathscr{G}$ is implemented using an MLP with two hidden layers of 1024 dimensions. We use a standard batch size of 32 for both source and target in all experiments and for all methods. The reported accuracies are computed on the complete unlabeled target data for CUB-200 dataset following established protocol for UDA [139, 248, 233, 196], and the provided testset for DomainNet. The crucial hyper-parameters in our method are λ_{sc}, temperature τ and memory bank size $|\mathscr{M}|$. For all datasets, we choose $\lambda_{sc} = 0.1$ and $\tau = 0.07$ based on the adaptation performance on the $C \rightarrow D$ setting on the CUB-200 dataset. We use a memory bank size of 48k on DomainNet dataset and 24k on CUB-200 dataset owing to its smaller size. For experiments on MemSAC, we report mean and standard deviation over 3 random seeds. We compare MemSAC against traditional adversarial approaches (DANN [69], CDAN [139], MCD [196]) as well as the current state-of-the art (SAFN [248], BSP [38], RSDA [83], CAN [110], ILADA [207], FixBi [155], HDAN [48] and ToAlign [238]). We re-implement baselines using code and hyper-parameters provided online by respective authors.

Table 3.1. Accuracy scores on DomainNet-345 using Resnet-50 backbone. Best values are in **bold** and the next best are underlined. MemSAC performs better than all other methods on most of the tasks. †Uses hierarchical label annotation. ‡prediction uses ensemble classifiers. §Uses class-balanced sampling.

Source	**Real→**			**Clipart→**			**Painting→**			**Sketches→**			
Target	C	P	S	R	P	S	R	C	S	R	C	P	Avg.
ResNet-50	41.61	42.79	29.66	42.41	27.24	32.15	49.52	32.55	26.73	38.75	40.89	27.5	35.98
MSTN [247]	27.25	32.98	24.35	28.17	21.14	24.15	30.74	19.85	22.5	24.31	26.22	23.56	25.44
RSDA [83]	27.28	35.83	24.35	36.98	24.94	31.12	41.32	26.1	24.71	29.46	26.22	27.79	29.68
BSP [38]	34.51	39.14	27.57	40.56	26.71	30.72	40.83	24.56	26.85	36.54	32.37	28.08	32.37
MCD [196]‡	36.34	36.58	24.95	40.32	25.83	32.12	43.65	29.66	25.7	34.16	39.11	26.89	32.94‡
ILADA [207]§	46.45	39.01	35.4	47.94	26.68	36.33	43.00	26.62	27.3	48.85	47.68	32.23	38.12§
SAFN [248]	38.11	45.96	29.20	45.96	30.00	34.65	54.44	34.74	30.64	45.29	47.43	38.01	39.54
DANN [69]	45.93	44.51	35.47	46.85	30.52	36.77	48.02	34.76	32.15	47.1	46.45	38.47	40.58
CAN [110]§	40.71	37.77	33.7	**54.93**	31.41	37.37	51.05	33.64	30.95	52.13	42.19	32.04	39.82§
PAN [233]†	49.25	48.18	36.46	49.66	33.27	38.78	51.89	36.01	32.94	49.12	50.94	39.89	43.03 †
CDAN [139]	50.15	48.35	39.01	50.02	33.39	39.3	52.21	36.44	33.68	48.46	49.27	38.65	43.24
HDAN [48]	46.30	47.52	34.39	49.91	33.98	37.98	55.26	40.82	32.77	49.04	49.77	40.04	43.15
FixBi [155]‡	51.18	49.19	39.65	50.02	34.59	41.17	52.21	36.44	33.68	50.84	53.51	41.67	44.51‡
ToAlign [238]	50.82	50.72	35.17	49.52	33.88	41.41	**57.92**	**43.51**	36.29	47.96	**55.46**	41.61	45.45
MemSAC [Ours]	**54.34**$^{\pm.5}$	**52.27**$^{\pm.3}$	**41.74**$^{\pm.3}$	54.4$^{\pm.3}$	**36.87**$^{\pm.4}$	**42.45**$^{\pm.0}$	53.24$^{\pm.2}$	41.39$^{\pm.4}$	**37.22**$^{\pm.2}$	**53.33**$^{\pm.3}$	55.31$^{\pm.2}$	**44.56**$^{\pm.3}$	**47.26**
Tgt. Supervised	72.59	62.66	65.12	80.92	62.66	65.12	80.92	72.59	65.12	80.92	72.59	62.66	70.32

3.5.3 MemSAC Excels On Many-class Adaptation

The results for the 12 transfer tasks on DomainNet are provided in Tab. 3.1. Firstly, methods such as RSDA (29.68%) and SAFN (39.54%) that achieve best performance on smaller scale datasets (like Office-31 [187] and visDA-2017 [166]) provide only marginal or no benefits over the more traditional adversarial approaches such as DANN (40.58%) and CDAN (43.24%) on DomainNet with 345 classes, indicating that large-scale datasets need different techniques for adaptation. Next, we compare against PAN [233], which requires a label hierarchy as additional information for training. For this supervision, we use the one level of hierarchy proposed in DomainNet [165]. Even when provided with access to hierarchical grouping labels in source, PAN (43.03%) achieves no improvement over CDAN (43.24%). In contrast, our method MemSAC that combines global adaptation using a conditional adversarial approach and local alignment using sample consistency to alleviate negative achieves an average accuracy of 47.26%, with a significantly better performance than all the prior approaches across most of the tasks.

Table 3.2. Results on fine-grained adaptation on 200 categories from CUB-Drawings dataset. Bold and underline indicate the best and second best methods respectively. †Uses hierarchical label annotation.

	Resnet-50	MCD [196]	SAFN [248]	CAN [110]	RSDA [83]	DANN [69]	HDAN [48]	FixBi [155]	CDAN [139]	ToAlign [238]	PAN [233]	MemSAC
C → D	60.88	50.18	60.29	52.18	61.04	62.09	60.25	68.20	68.12	64.43	<u>70.53</u>	**71.78**
D → C	42.07	38.56	41.34	50.05	44.20	47.73	52.40	49.47	53.83	50.54	<u>55.38</u>	**59.48**
Avg.	51.47	44.37	50.82	51.11	52.62	54.91	56.33	58.84	60.98	57.48	<u>62.96</u>	**65.63**

3.5.4 MemSAC Achieves new State-of-the-art in Fine-grained Adaptation

We also illustrate the benefit of using MemSAC for adaptation on fine-grained categories in Tab. 3.2 on the CUB-Drawings dataset. Although fine-grained visual recognition is a well-studied area [258, 257, 22, 40, 64], domain adaptation for fine grained categories is a relevant but less-addressed problem. Notably, methods like MCD, SAFN and RSDA perform worse or only marginally better than a source only baseline. PAN [233] uses supervised hierarchical label relations in source across 3 levels and obtains an average accuracy of 62.96%, while MemSAC obtains a state-of-the art accuracy of 65.63% using only single level source labels, thus outperforming all prior approaches on this challenging setting with minimal assumptions. This underlines the benefit of enforcing sample consistency using MemSAC for adaptation even in the presence of fine-grained categories in order to effectively counter negative alignment issues.

3.5.5 MemSAC Complements Multiple Adaptation Methods

The proposed memory-augmented consistency loss is generic enough to improve many adaptation backbones. As shown in Tab. 3.3a for the case of R→C and C→R transfer tasks from DomainNet, MemSAC can be used with most adversarial as well as MMD based approaches. MemSAC improves adversarial approaches DANN and CDAN by 3.35% and 4.29% respectively, and MMD-based approach CAN by 1.75% indicating that our proposed framework is competitive yet complementary to many existing adaptation approaches.

Table 3.3. Ablations on DomainNet-345 dataset. In Tab. 3.3a, we show the complementary nature of our method, which works suitably well with other domain adaptation backbones, in addition to CDAN used in Tab. 3.1. In Tab. 3.3b, we show that MemSAC also is very efficient when used with other backbone architectures such as Resnet-101. For a more detailed analysis on the effect of architectures, please refer to Sec. 6.4.1 in Chapter 6.

(a) Accuracy values of MemSAC using DANN and CAN adaptation backbones on DomainNet-345 classes. Note improved accuracy using MemSAC on top of both the backbones.

Source	**Real**→			**Clipart**→			**Painting**→			**Sketches**→			
Target	C	P	S	R	P	S	R	C	S	R	C	P	Avg.
DANN [69]	45.93	44.51	35.47	46.85	30.52	36.77	48.02	34.76	32.15	47.1	46.45	38.47	40.58
DANN + MemSAC	49.67	48.61	39.14	49.81	35.1	40.59	50.04	38.51	36.61	50.31	50.8	42.73	**44.32**
CAN [110]	40.71	37.77	33.7	54.93	31.41	37.37	51.05	33.64	30.95	52.13	42.19	32.04	39.82
CAN + MemSAC	43.79	38.99	36.71	55.36	32.41	39.46	52.48	35.21	32.89	54.15	44.60	33.02	**41.59**

(b) Results on DomainNet-345 dataset with Resnet-101 backbone and batch size of 24.

Source	**Real**→			**Clipart**→			**Painting**→			**Sketches**→			
Target	C	P	S	R	P	S	R	C	S	R	C	P	Avg.
Resnet-101	45.62	44.24	33.12	41.96	27.07	33.07	48.54	34.92	29.84	35.87	42.64	28.01	37.07
DANN [69]	47.71	44.1	35.99	48.33	32.00	38.54	48.13	34.57	34.23	48.19	48.56	39.67	41.67
MCD [196]	41.11	39.01	26.1	40.77	28.26	33.02	45.49	33.03	29.1	38.29	42.3	29.51	35.49
CDAN [139]	52.47	48.0	40.42	46.63	32.42	39.18	48.81	37.92	35.39	45.69	48.92	37.31	42.76
SAFN [248]	44.93	46.52	28.2	37.2	31.11	36.3	53.32	36.95	32.48	44.12	53.46	40.05	40.38
ToAlign [238]	50.10	48.27	35.98	50.24	31.41	41.10	54.60	43.67	36.82	50.15	54.32	42.06	44.89
MemSAC	**56.25**	**52.96**	**42.22**	**53.52**	**37.46**	**43.46**	**53.38**	**42.69**	**39.65**	**53.17**	**55.29**	**44.29**	**47.86**

3.5.6 MemSAC Improves Adaptation Even With Larger Backbones

We employ Resnet-101 as a backbone in Tab. 3.3b and compare against other adaptation approaches with the same backbone. We note that the benefits obtained by MemSAC over prior adaptation approaches also hold for larger backbones, as shown for R→C and C→R of DomainNet dataset. For a more detailed analysis on the effect of architectures, including recent advances such as transformer backbones, please refer to Sec. 6.4.1 in Chapter 6.

3.5.7 Analysis and Discussion

Ablation studies

We show the influence of various design choices of our method in Tab. 3.4 on the CUB-200 dataset. First, we show in Tab. 3.4a that both the global domain adversarial method, which

Table 3.4. Ablation results. Effect of (a) Loss coefficients, (b) temperature scaling, and (c) choice of similarity functions on accuracy of MemSAC on the CUB-Drawing adaptation.

(a) Effect of various components of loss function in (3.1).

Method	$\mathscr{L}_{adv}$	$\mathscr{L}_{sc}$	C→D	D→C	Avg. Acc
Source	No	No	60.88	42.07	51.47
CDAN	yes	No	68.12	53.83	60.98
$\mathscr{L}_{sc}$ Only	No	yes	64.45	41.13	52.79
MemSAC	yes	yes	**71.78**	**59.48**	**65.63**

(b) Effect of the temperature τ in (3.4).

τ	C→D	D→C	Avg. Acc
1.0	68.36	53.46	60.91
0.07	**71.78**	**59.48**	**65.63**
0.007	71.25	57.21	64.23

(c) Accuracy using various choices for ϕ_{ij}.

Similarity	ϕ_{ij}	C→D	D→C	Avg. Acc
Inv. Euc.	$(1+\|\|f_i-f_j\|\|^2)^{-1}$	71.00	57.21	64.23
Gaussian	$exp(-\|\|f_i-f_j\|\|^2)$	70.10	50.84	60.47
Cosine	$f_i \cdot f_j$	**71.78**	**59.48**	**65.63**

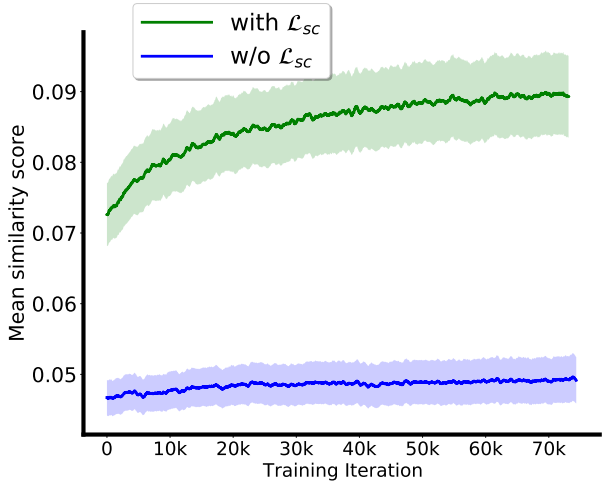


Figure 3.4. Mean similarity score for *within-class* samples vs. training iteration shown for **D→C** on CUB-Drawings.

we implement using CDAN, as well as local sample level consistency loss are important to achieve best accuracy, as evident from the drop in accuracy without either of those components. Next, we investigate the effect of the temperature parameter τ in Tab. 3.4b which we use to suitably scale the contributions of positive and negative pairs in $\mathscr{L}_{sc}$ loss function (Eq. (3.4)). We find that $\tau = 0.07$ gives the best performance on the cosine similarity metric. Finally, in Tab. 3.4c, we note that the performance using other choices of the similarity function $\phi(.)$, namely *Euclidean* similarity and *Gaussian* similarity is inferior to using *Cosine* similarity. We also observed that *cosine* similarity is more stable to train under severe domain shifts.

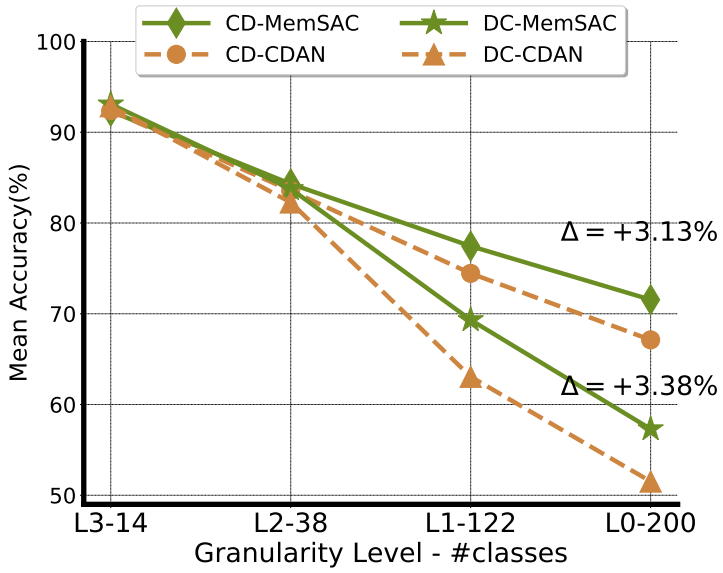


Figure 3.5. Comparison of accuracy vs. granularity of labels on CUB-Drawings dataset for 4 levels of label hierarchy.

Why does MemSAC help with large number of classes?

We propose our sample consistency loss in Eq. (3.4) to encourage tighter clustering of samples within each class, which is important in many-class datasets where class confusion is a significant problem. The main motivation of the proposed sample consistency loss is to bring within-class samples (that is, samples from the same class across source and target domains) closer to each other, so that a source classifier can be transferred to the target. To understand this further, in Fig. 3.4, we plot the *mean similarity score* during the training process. We define the *mean similarity score* as $\sum_{i \in \mathscr{M}_+^j} \phi_{ij}$, averaged over all the target samples $j \in \mathscr{B}_t$ in a mini-batch, which indicates the affinity score between same-class samples across domains. We observe that using the proposed loss, the similarity score is much higher and improves with training compared to the baseline without the consistency loss, which reflects in the overall accuracy (Tab. 3.1, Tab. 3.2).

MemSAC achieves larger gains with finer-grained classes

We show the appreciating benefits provided by MemSAC as the close granularity of the dataset becomes more pronounced. For this purpose, we chose the 4 levels of label hierarchy provided by PAN [233] on the CUB-Drawings dataset. The levels L3, L2, L1 and L0 contain different granularity of bird species, grouped into 14, 38, 122 and 200 classes, respectively. The L0 level contains the finest separation of classes, while the level L3 with 14 classes contains the coarsest separation. We observe from Fig. 3.5 that with coarser granularity, MemSAC performs as good as the baseline method CDAN, whereas with finer separation of the categories (L3 $\rightarrow$ L0), use of sample consistency loss provides much higher benefit ($>$ 3% improvement on both tasks). This confirms our intuition that sample level consistency benefits accuracies in fine-grained domain adaptation.

MemSAC alleviates class confusion for similar classes

In Fig. 3.6 we use the DomainNet dataset to show the accuracies on every *coarse* category, along with the number of finer classes in each coarse category. We find that MemSAC provides consistent improvement over CDAN (marked by $\uparrow$) on most categories and any drops in accuracy (marked by $\downarrow$) are negligible. Our improvements are especially greater on categories with fine-grained classes like *trees (+13.3%)*, *vegetables (+6.7%)* and *birds (+5.6%)*, underlining the advantage of MemSAC to overcome class confusion within dense categories.

Larger memory banks improve accuracy

A key design choice that we need to make in MemSAC is the size of the memory bank $\mathscr{M}$. Intuitively, small memory banks would not provide sufficient negative pairs in the sample consistency loss and lead to noisy gradients. We show in Fig. 3.7 for the two tasks in CUB-Drawings that accuracy indeed increases with larger sizes of memory banks (a memory size of 32, which is same as batch-size, indicates no memory at all and performs worse). We also find that the optimum capacity of the memory bank may even be much higher than the size of the

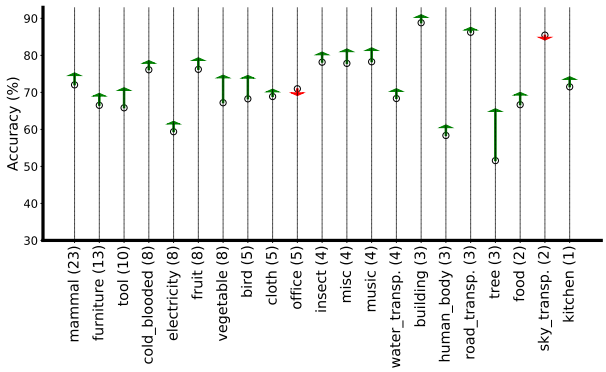


Figure 3.6. Category wise gain/drop in accuracy on **R**→**C** on DomainNet, compared to CDAN [139].

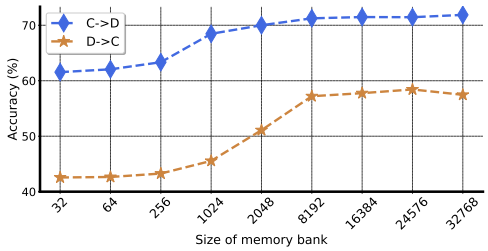
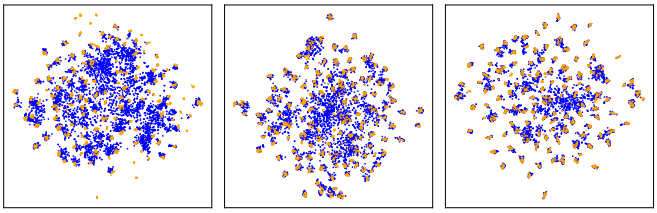


Figure 3.7. Effect of memory bank size on CUB-Drawings dataset.



(a) Source only **(b)** CDAN **(c)** MemSAC

Figure 3.8. tSNE for **R**→**C** on DomainNet. The two colors are source and target features. Notice improved alignment and feature separation with MemSAC.

dataset. For example, the "drawing" domain has around 4k examples, but from Fig. 3.7, **D**→**C** achieves best accuracy at memory size of 25k, indicating that it would help to have multiple copies of the same instance in the memory bank. This is in contrast to prior works using memory based contrastive learning [89, 245] since those works use a *single* positive sample from the memory and treat all other samples as negatives. But in our case, we can allow *multiple* positives and negatives into the sample consistency loss (Eq. (3.4)), so having multiple copies of the same instance is beneficial.

Feature alignment using MemSAC

In Fig. 3.8, we compare the feature alignment between a plain source-only baseline without any adaptation (Fig. 3.8a), as well as tSNE obtained after adaptation using CDAN Fig. 3.8b

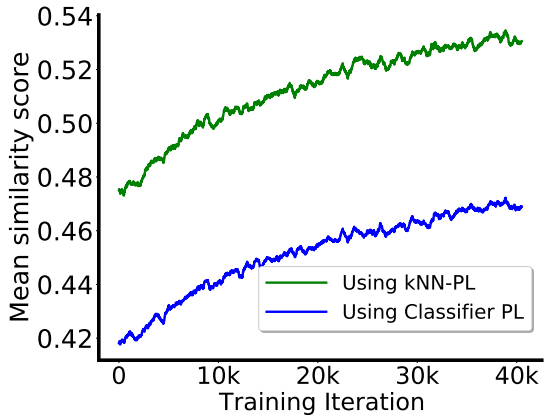


Figure 3.9. Similarity score of kNN based pseudo-labeling compared with classifier based pseudo-labeling.

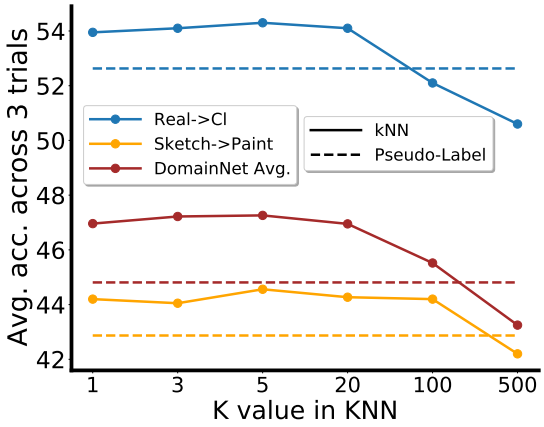


Figure 3.10. Effect of K in choosing the nearest neighbors on the target accuracy. As shown, a value of K in the range of 1-20 works best.

and MemSAC Fig. 3.8c. As shown, the source and target features are more perfectly aligned for the case of MemSAC compared to other approaches, showing evidence for stronger cross-domain alignment.

3.5.8 Ablations on KNN-based Pseudo-labeling

A crucial choice made in the design of MemSAC is the use of kNN-based pseudo-labeling instead of directly using the classifier predictions on unlabeled target samples as pseudo-labels for all the target samples. This follows from the observation that the kNN based pseudo-labeling is generally robust to noisy classifier boundaries, especially amidst domain shifts. Moreover, with the help of the memory bank, the neighborhood from which the nearest neighbors are computed is much larger than the size of the mini-batch. We verify this intuition in Fig. 3.9, where the mean similarity score between the samples from the same class is much higher when trained using the proposed kNN based pseudo-labeling technique as compared to the classifier based pseudo-labeling technique. Furthermore, we analyze the effect of the choice of the parameter K in Fig. 3.10. Our accuracy is robust to most values of K in the range of 1-20. At large values of K, however, the accuracy falls steeply due to large amounts of noise in the pseudo-labels.

Table 3.5. Role of memory module and kNN pseudo labeling. As shown, the best target accuracy is achieved when using memory bank in combination with kNN based pseudo-labeling technique, further validationg our design hypothesis.

	W/ kNN	Classifier PL
w/ Mem.	**47.26**	44.81
w/o Mem.	43.32	43.24

Table 3.6. MemSAC with different values of momentum parameter μ. The best accuracy is observed using no momentum updates, or at $\mu = 0$.

μ	C→D	D→C	Avg. Acc.
0	**73.97**	**61.94**	**67.95**
0.5	68.61	55.24	61.92
0.9	68.89	55.24	62.06
0.999	71.43	58.81	65.12

Additionally, in Tab. 3.5, we show that both the memory bank and kNN based pseudo-labeling are crucial to achieve performance gains using the consistency loss, as removing one of them (or both of them) results in significant drop in performance.

3.5.9 Queue Updates using Momentum Encoder

We now discuss possible alternative strategies to update the memory bank. For this purpose, we generalize the update rule using a *momentum encoder*, proposed in [89]. After the initial bootstrapping phase where we train the encoder on source data for few iterations, we initialize the momentum encoder $\mathscr{F}$ using the state of the encoder $\mathscr{E}$. After that, at every iteration, the parameters of the momentum encoder $\theta_{\mathscr{F}}$ are updated as follows.

$$\theta_{\mathscr{F}} = (1-\mu) * \theta_{\mathscr{E}} + (\mu) * \theta_{\mathscr{F}} \tag{3.5}$$

Here, μ is called the momentum parameter, and controls the speed of updates. The source features encoded in the memory bank $\mathscr{M}$ are obtained by a forward pass on $\mathscr{F}$, while the source features used to compute the supervised loss as well as all the target features are computed using

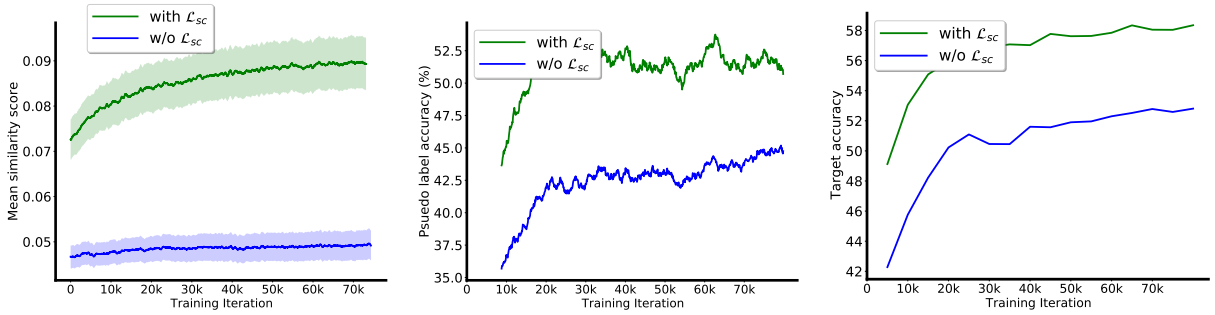


Figure 3.11. Training Curves for $D \rightarrow C$ (Fig. 3.11a) Mean similarity score of within class samples vs. Training iterations. (Fig. 3.11b) Pseudo-label accuracy vs. Training iterations. (Fig. 3.11c) Final target accuracy vs. Training iterations

a forward pass on $\mathscr{E}$. We note that the original update rule discussed in Sec. 3.3.4 is just a special case of Eq. (3.5), which is obtained by putting $\mu = 0$.

The intuition behind using such a momentum based encoder is that it gives features with a slow drift through the training, and hence can support larger queues. We use such a momentum update on MemSAC and show results for CUB-Drawings dataset in Tab. 3.6 We found no benefit using such a momentum encoder in our method. This might be because we already bootstrap the encoder until the features stabilize and achieve a slow-drift phenomenon, and using momentum based updates on top of that might not improve accuracy. In light of these results, designing better memory bank update schedules is left as a potential direction for future work.

3.5.10 Training Curves

In Fig. 3.11, we show the trends for the mean similarity score, pseudo label accuracy as well as the final target accuracy during training. We compare between MemSAC which uses a consistency based loss, with an approach which does not contain such a consistency constraint. We observe that using our sample consistency loss gives a higher value of mean similarity score, pseudo-label accuracy as well as final target accuracy during training, and each of them improve with training indicating the effectiveness of our proposed loss.

3.6 Summary

In this chapter, we presented MemSAC, a simple and effective approach for unsupervised domain adaptation designed to handle a large number of categories. We propose a sample consistency loss that pulls samples from similar classes across domains closer together, while pushing dissimilar samples further apart. Since minibatch sizes are limited, we devise a novel memory-based mechanism to effectively extract similarity relations for a large number of categories. We provide both theoretical intuition and empirical insights into the effectiveness of MemSAC for large-scale domain alignment and discriminative transfer. Through extensive experiments, we showcase the strong improvements achieved by MemSAC over prior works, setting new state-of-the-arts across challenging many-class adaptation on DomainNet and fine-grained adaptation on CUB-Drawings.

This chapter is a reprint of the material as it appears in "Memsac: Memory augmented sample consistency for large scale domain adaptation" by Tarun Kalluri, Astuti Sharma, and Manmohan Chandraker, which was published in Proceedings of the European Conference on Computer Vision, 2022. The dissertation author was the primary investigator and author of this paper.

Chapter 4

Benchmarking Unsupervised Adaptation Across Geographies

With the ever-increasing trend of collecting and using web-scale data for training large-scale visual categorization models, a pertinent questions needs to be raised regarding the geographic composition of such training data, and the resulting biases that will invariably trickle down into the downstream models when deployed in real world. For a rough estimation, an approximate geographic composition of several standard computer vision datasets is studied in [57], and presented in Fig. 4.1. As shown, it is apparent that there exists mismatch between the datasets fueling the progress in modern computer vision and the demographies potentially consuming this progress. While unsupervised adaptation is generally used to bridge such domain shifts, their effectiveness is not studied when faced with geographical variations. This chapter summarizes several contributions we made in studying this problem, which includes a new dataset, followed by extensive analysis of domain shifts between geographies and benchmarking robustness of several models against these shifts.

4.1 Introduction

In recent years, domain adaptation has emerged as an effective technique to alleviate dataset bias [220] during training and improve transferability of vision models to sparsely labeled target domains [138, 142, 139, 69, 195, 196, 98, 248, 110, 238, 107]. While being greatly

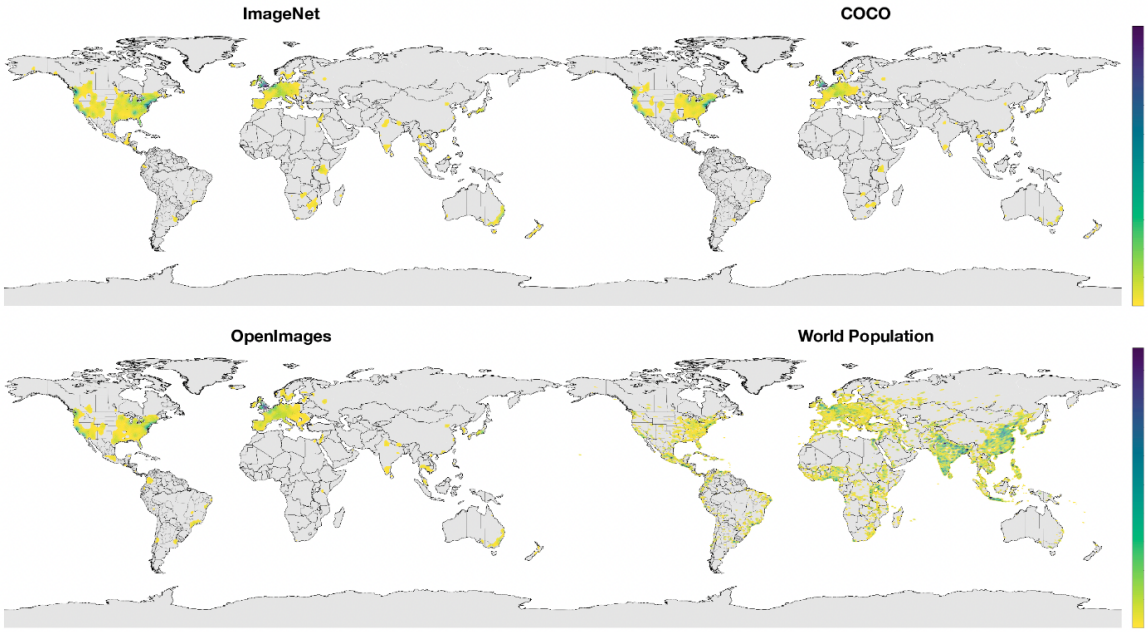


Figure 4.1. Geographical Distribution of Datasets (from [57]) The plot shows the geographical distribution of several standard training datasets used in computer vision literature such as ImageNet [186], MSCoCo [128] and OpenImages [120]. When compared to the population distribution of the world, which in turn serves as a rough proxy for the demographies consuming this technology, we see a clear bias and mismatch between the distributions highlighting a potential limitation of web-sourced datasets.

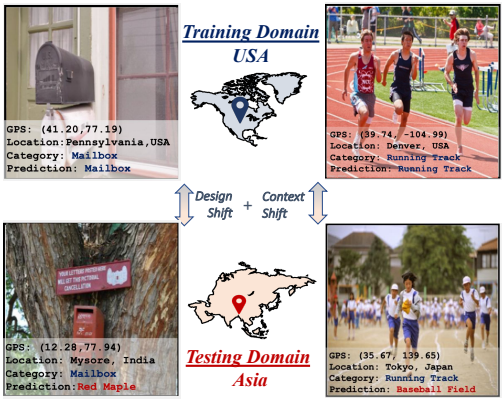
instrumental in driving research forward, methods and benchmark datasets developed for domain adaptation [188, 230, 166, 165] have been restricted to a narrow set of divergences between domains. However, the geographic origin of data remains a significant source of bias, attributable to several factors of variation between train and test data. Training on geographically biased datasets may cause a model to learn the idiosyncrasies of their geographies, preventing generalization to novel domains with significantly different geographic and demographic composition. Besides robustness, this also negatively impact the fairness and inclusivity of computer vision models, as most modern benchmark datasets like ImageNet [186] and COCO [128] suffer from a significant US or UK-centric bias in data [205, 57], with poor representation of images from various other geographies like Asia and Africa.

In this chapter, we study the problem of geographic adaptation by introducing a new

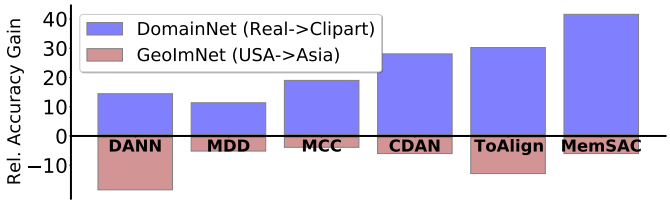
large-scale dataset called GeoNet, which constitutes three benchmarks – GeoPlaces for scene classification, GeoImNet for object recognition and GeoUniDA for universal domain adaptation. These benchmarks contain images from USA and Asia, which are two distinct geographical domains separated by various cultural, economic, demographic and climatic factors. We additionally provide rich metadata associated with each image, such as GPS location, captions and hashtags, to facilitate algorithms that leverage multimodal supervision.

GeoNet captures the multitude of novel challenges posed by varying image and label distributions across geographies. We analyze GeoNet through new sources of domain shift caused by geographic disparity, namely (i) *context shift*, where the appearance and composition of the background in images changes significantly across geographies, (ii) *design shift*, where the design and make of various objects changes across geographies, and (iii) *prior shift*, caused by different per-category distributions of images in both domains. We illustrate examples of performance drop caused by these factors in Fig. 4.2a, where models trained on images from USA fail to classify common categories such as *running track* and *mailbox* due to context and design shifts, respectively.

GeoNet is an order of magnitude larger than previous datasets for geographic adaptation [170, 182], allowing the training of modern deep domain adaptation methods. Importantly, it allows comparative analysis of new challenges posed by geographic shifts for algorithms developed on other popular adaptation benchmarks [188, 166, 230, 165]. Specifically, we evaluate the performance of several state-of-the-art unsupervised domain adaptation algorithms on GeoNet, and show their limitations in bridging domain gaps caused by geographic disparities. As illustrated in Fig. 4.2b for the case of DomainNet [165] vs. GeoNet, state-of-the-art models on DomainNet often lead to accuracies even worse than a source only baseline on GeoNet, resulting in negative *relative gain* in accuracy (defined as the gain obtained by an adaptation method over a source-only model as a percentage of gap between a source-only model and the target-supervised upper bound). Furthermore, we also conduct a study of modern architectures like vision transformers and various pre-training strategies, to conclude that larger models with



(a) Geographic bias manifested in proposed GeoNet dataset



(b) Unsupervised domain adaptation does not suffice on GeoNet

Figure 4.2. Summary of our contributions. (a): Training computer vision models on geographically biased datasets suffers from poor generalization to new geographies. We propose a new dataset called GeoNet to study this problem and take a closer look at the various types of domain shifts induced by geographic variations. (b) Prior unsupervised adaptation methods that efficiently handle other variations do not suffice for improving geographic transfer.

supervised and self-supervised pre-training offer improvements in accuracy, which however are not sufficient to address the domain gap. This highlights that the new challenges introduced by geographic bias such as context and design shift are relatively under-explored, where our dataset may motivate further research towards this important problem.

To summarize before we move into details, our contribution towards geographic domain adaptation is four-fold:

- A new large-scale dataset, GeoNet, with benchmarks for diverse tasks like scene classification and object recognition, with labeled images collected from geographically distant locations across hundreds of categories (Sec. 4.3).
- Analysis of domain shifts in geographic adaptation, which may be more complex and subtle than style or appearance variations (Sec. 4.3.5).
- Extensive benchmarking of unsupervised adaptation algorithms, highlighting their limitations in addressing geographic shifts (Sec. 4.5.2).

- Demonstration that large-scale pretraining and recent advances like vision transformers do not alleviate these geographic disparities (Sec. 4.5.3).

4.2 Relation to Prior Literature

4.2.1 Unsupervised Domain Adaptation

Unsupervised domain adaptation enables training models on a labeled source domain along with unlabeled samples from a different target domain to improve the target domain accuracy. A large body of prior works aim to minimize some notion of divergence [13, 12] between the source and target distributions based on MMD [216, 138, 142, 212] adversarial [69, 139, 21, 225, 195, 259, 28, 224], generative [199, 20, 98], class-level [164, 196, 144, 247, 116, 83] or instance-level alignment [238, 207, 232] techniques. Clustering [52, 105, 110, 162, 108] and memory-augmentation approaches [107] have also been shown to be effective. However, most of these works are shown to improve performance using standard datasets such as Office-31 [188], visDA [166], OfficeHome [230] or DomainNet [165], where the distribution shifts typically arise from unimodal variations in style or appearance between source and target. While prior works also study semantic shift [14] and sub-population shift [23], we aim to address a more practical problem of geographic domain adaptation with more complex variations not covered by prior works.

4.2.2 Geographic Robustness

Many prior works study biases of CNNs towards 3D poses [3, 262], textures [72], styles [93], natural variations [178, 16, 218] and adversarial inputs [93], but robustness of computer vision towards shift induced by geography is relatively under-explored. While algorithms for bridging geographic domain gaps have been proposed in [39, 108, 235], they are restricted to road scenes with limited number of classes. A major hindrance has been the lack of suitable benchmark datasets for geographic adaptation, so several datasets have been recently proposed to address this issue [204, 58, 170, 182]. Datasets based on dollar street images [182] highlight

Table 4.1. Summary Statistics of GeoNet Number of images in train and test splits in each of our benchmarks. While GeoPlaces and GeoImNet are developed for unsupervised adaptation, GeoUniDA is developed for universal domain adaptation across geographies.

	Split	GeoPlaces	GeoImNet	GeoUniDA
USA	Train	178110	154908	100136
	Test	17234	16784	25034
Asia	Train	187426	68722	33912
	Test	26923	9636	8478
classes-shared		205	600	62
classes-private		-	-	138

the geographic differences induced by income disparities between various countries, Ego4D [80] contains egocentric videos with actions from various geographies, while researchers in [170] design an adaptation dataset with images from YFCC-100M [66] to analyze geographic shift, authors in [175] use crowdsourcing to collect geographically diverse evaluation datasets. Adding to these efforts, we propose a much larger-scale dataset for geographic adaptation consisting of more diverse categories for place and object classification, across factors of variation beyond income disparities.

4.3 Dataset Creation and Analysis

We present an overview of our data collections pipeline in Fig. 4.3 and the overall summary of collected datasets in our benchmark in Tab. 4.1, including the number of images and categories from each of our settings. For creating our dataset, we broadly consider US and Asia as the two domains, as these two geographies have considerable separation in terms of underlying cultural, environmental and economical factors, while also providing the appropriate level of abstraction and leaving enough data from each domain to perform meaningful analysis. Although Asia is less homogeneous than USA with greater within-domain variance, our adopted geographical granularity follows from the amount of data we could retrieve from different countries using concepts in GeoNet, where we observed general paucity in images from many low-resource countries on Flickr.

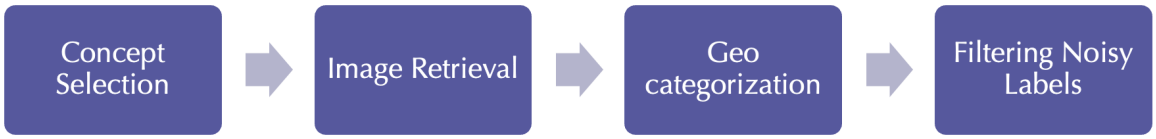


Figure 4.3. Pipeline for Collecting GeoNet Dataset We show a summary pipeline used to collect and curate images in the GeoNet dataset. We first select the concept names from existing datasets, followed by retrieving images from web pertaining to these concepts. Next, we use the geotags in the metadata of the collected images, whenever available, to geo-categorize the images into distinct regions. Finally, to curate the dataset we filter noisy labels and images before creating the train and test splits from each geography.

4.3.1 GeoPlaces

We propose GeoPlaces to study geographic adaptation in scene classification, which involves predicting the semantic category of the place or location present in the image [266]. In contrast to object classification, it is necessary to accurately identify and understand various interactions and relationships between the objects and people in the scene to predict the appropriate scene category. In spite of rapid progress in datasets [246, 266] and methods [29] for this task, robustness of scene classification networks to unseen domains in general, and across geographies in particular, has received little attention, for which our benchmark would set a new course.

Selecting concepts and images

We use the 205 scene categories from Places-205 [266] to build GeoPlaces, as these semantic categories cover a wide range of real world scenes commonly encountered in most geographies. We build our GeoPlaces benchmark from the labeled Places-205 dataset [267]. We first collect the unique Flickr identifier (`Flickr-id`) associated with each image in the Places-205 dataset, and then use the publicly available Flickr API[1] to extract the GPS location of the image. Since only a fraction of images belong to Flickr and a further smaller fraction contain valid geotags, we end up with around 400k images from 205 classes with associated geographical information. Of these, 190k images are from the US domain, and we use 178k

[1] Flickr.com/services/api/explore/Flickr.photos.geo.getLocation

of them for training and 17k for testing. In Asia domain however, we obtain only 27k images. To match the scale of images from both domains, we perform an additional step and manually collect more images as explained next.

Additional data

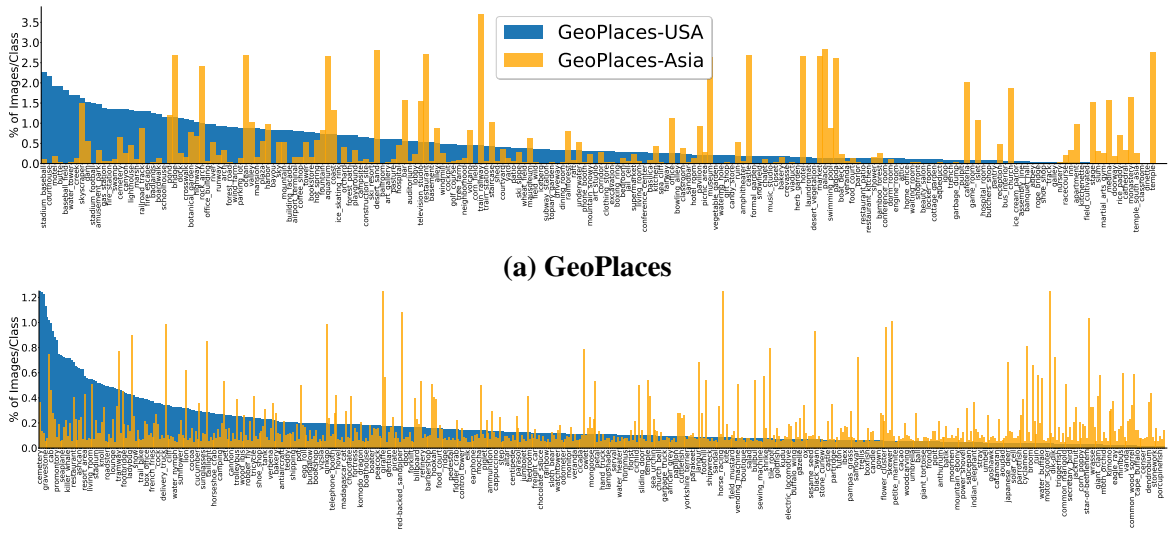
Due to the inherent US-centric bias of photo-sharing websites like Flickr, a major portion of images are US-based. In order to collect more images from the Asia domain, we directly scrape images from Flickr using the 205 category names from Places-205 as the *seed concepts*. As many Asian users often post descriptions and tags for pictures in languages other than English, we use translations of these seed concepts in English to Asian languages, namely {Hindi, Korean, Japanese, Chinese, Russian, Hebrew}, and use these along with the original concepts, as the augmented or *expanded concepts*. Then, we search Flickr for images which match the criterion that (i) they are geotagged in Asia, and (ii) the tags associated with the image match with exactly one of the categories in the expanded concept list (which we assign as the label). We collect around 190k images this way, and use this as the training set. Since images collected from web tend to be nosier than human labeled ones, we use the manually labeled 27k images from Places-205 as the test set for Asia domain to ensure robust benchmarking.

4.3.2 GeoImnet

We propose the GeoImNet benchmark to investigate the domain shift due to geographical disparities on object classification. Different from existing object-level datasets for domain adaptation [166, 165, 188, 230], GeoImNet provides domain shifts induced by geographic disparities.

Dataset curation

We collect images in the GeoImNet benchmark from the WebVision dataset [124], which itself is scraped from Flickr using queries generated from 5000 concepts in the Imagenet-5k



(a) GeoPlaces

(b) GeoImNet

Figure 4.4. Class distribution in GeoNet Percentage of images per class from USA and Asia domains shown for the GeoPlaces benchmark in a and GeoImNet benchmark in b. The label distributions are long-tailed in both, and the dominant and tail classes are widely different across geographies in each setting indicating a strong prior shift. (Best viewed in color, zoom in to see the class names).

dataset [51]. We then follow the same pipeline as explained above for GeoPlaces benchmark, and identify the GPS coordinates of each images using its `Flickr-id`.

Concept selection

Although the original dataset contains 5000 classes, many of these classes are indigenous to a particular geography. For example, *Bengal Tiger*s are found in Indian subcontinent, and *Bald Eagle* is a North-American bird. Since unsupervised domain adaptation typically demands matching label spaces across source and target, we select 600 categories out of the original 5000 with at least 20 images in each domain from each category. We then assign roughly 15% of images from each domain into the test set and use the remaining as the training images.

Dataset filtering

WebVision is *webly supervised* [36], which does not guarantee object-centric images or clean labels. Therefore, we remove all the images from the dataset which have more than one

(a) GeoPlaces **(b) GeoImNet**

Figure 4.5. Context Shift in GeoNet A few examples showing the nature of context shifts across categories from GeoPlaces benchmark in (a), and GeoImNet benchmark in (b), arising due to a variety of differences between geographical disparity. For example, outdoor scenes (shopfront, marketplace) reflect the demographies across geographies, indoor-scenes (living rooms, cafeteria) reflect cultural and economic variations and wildlife images reflect the habitat and climatic variations.

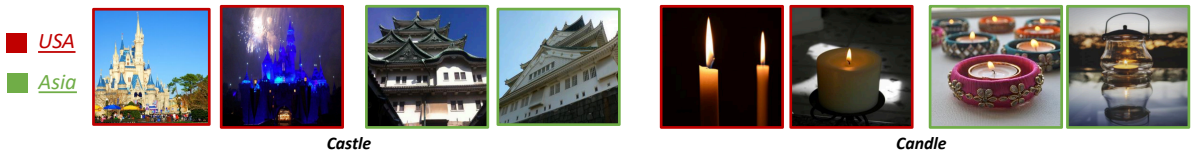
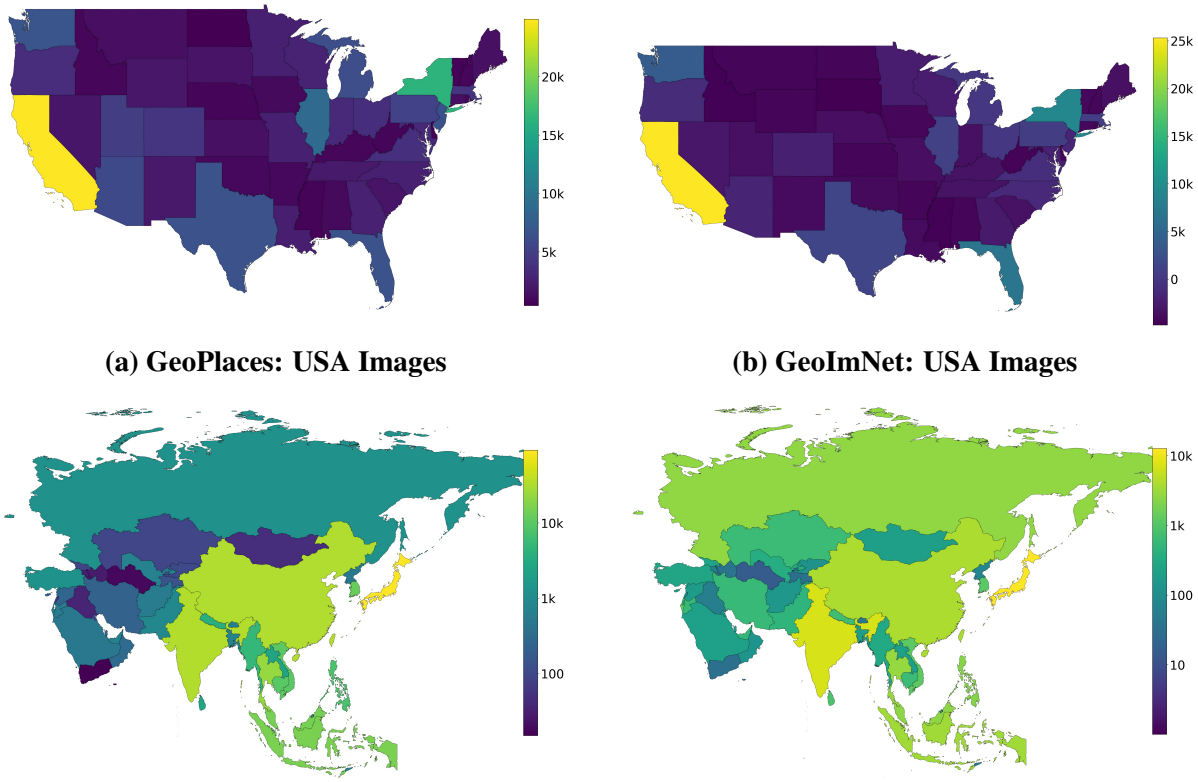


Figure 4.6. Design Shift in GeoNet We show examples illustrating the design shifts for the cases of *castle* from GeoPlaces and *candle* from GeoImNet. Note that differences in designs of castles as well as the variety of objects like candles found across geographies lead to design shifts between the domains.

tag that match our selected concepts (the 600 chosen categories) to handle multi-labeled images. Furthermore, we manually quality-check all the test images and remove all the images with noisy labels. Finally, we perform de-duplication to remove images from the training set which are very similar to those in the test set. The final label distribution for both US and Asia domains in both our benchmarks is shown in Fig. 4.4.



(a) GeoPlaces: USA Images **(b) GeoImNet: USA Images**

(c) GeoPlaces: Asia Images **(d) GeoImNet: Asia Images**

Figure 4.7. Geographical Distribution of images from USA and Asia domains. We show the images per geographical sub-region in both domains on GeoNet. As shown, in Asia, a majority of images are from Japan, India, Korea, China and Taiwan while in USA, a majority of images are from populous regions like California and New York. Note that the color-bar scale is linear for USA and log-scale for Asia.

4.3.3 GeoUniDA

Universal Domain Adaptation (UniDA) [254] facilitates domain adaptation between source and target domains that have few private classes, in addition to shared classes which are common to both. While this is a realistic problem, prior works [254, 193, 191, 119] use benchmarks created from existing UDA datasets for evaluation. However, our proposed geographical adaptation setting gives us an unique opportunity to design benchmarks for UniDA such that the private categories from the source and the target are a natural reflection of the presence or absence of these categories in the respective geographical domains. In order to select

the shared and private categories for our Geo-UniDA benchmark, we first start with the 1000 categories in the original Imagenet-1k dataset [186], and select top 200 categories each in the USA and Asia domains that have the most number of images from the WebVision dataset. Out of these, we use the 62 common classes as the shared categories, and the remaining 138 as the private classes in each domain.

4.3.4 Geographic Distribution of Images

While we broadly categorize Asia and USA to be the two major geographical domains, not all sub-regions in these geographies have equal representation. We show the geographic distribution over respective geographies in Fig. 4.7, by leveraging the per-image GPS metadata provided in GeoNet. For images from Asia from Fig. 4.7c for GeoPlaces and Fig. 4.7d for GeoImNet, we observe a large fraction of images from Japan, India, Korea, China and Taiwan, while some countries are more sparsely represented. Likewise, in USA in Fig. 4.7a and Fig. 4.7b, we observe a significant share of images from California, New York and Florida than other regions. These distributions reflect the larger user demographic biases in photo-sharing websites like Flickr from where all our images have been taken from.

4.3.5 Analysis of Distribution Shifts

We denote the source dataset using D_s=$\{X_s, Y_s\}$, and assume that $X_s \sim P_s(x)$, and the joint distribution $(X_s, Y_s) \sim P_s(x,y)$, where $P_s(x)$ and $P_s(x,y)$ are the image marginal and image-label joint distribution respectively. Target dataset $D_t = \{X_s, Y_s\}$ and target distributions $P_t(x)$ and $P_t(x,y)$ are defined similarly, and the domain discrepency assumption states that $P_s(x,y) \neq P_t(x,y)$. In order to formulate domain shift across geographies, we define f_x as the part of image referring to the foreground objects (corresponds to the salient objects in a scene) and b_x to be the rest of the image corresponding to the background regions (corresponding to the surrounding regions or context). For example, for the task of classifying *living room* in Fig. 4.5a from GeoPlaces, common objects like sofa and table are foreground, while floor, roof and walls are backgrounds.

We make a simplifying assumption that an image is completely explainable using its foreground and background and replace the class-conditional distribution of the images $P(x|y)$ with the joint class-conditional $P(b_x, f_x|y)$. Further, we also assume that given a class label, the background is conditionally independent of the foreground. Then,

$$\begin{aligned} P(x,y) &= P(x|y) \cdot P(y) \\ &= P(b_x, f_x|y) \cdot P(y) \\ &= P(b_x|y) \cdot P(f_x|b_x, y) \cdot P(y) \\ \implies P(x,y) &= \underbrace{P(b_x|y)}_{\text{context}} \cdot \underbrace{P(f_x|y)}_{\text{design}} \cdot \underbrace{P(y)}_{\text{prior}} \end{aligned} \tag{4.1}$$

We define the class-conditional background distribution $P(b_x|y)$ as context, while the class-conditional object distribution $P(f_x|y)$ as design and the label distribution $P(y)$ as prior. Note that standard covariate shift assumption [12] assumes uniform domain discrepency across all the images ($P_s(x) \neq P_t(x)$), which does not hold for geographic adaptation due to the diverse source of variations. We analyze each of these from a geographic adaptation perspective next.

Context shift

We define context shift to be the changes in the context around an object or scene given by $P_s(b_x|y) \neq P_t(b_x|y)$. Deep learning models are generally sensitive to object contexts and backgrounds, and learn spurious correlations that impede their ability to recognize objects and scenes in novel contexts [44, 42, 209, 184]. In geographic adaptation, context shift can be caused by differences in cultural or economic factors across geographies, and few examples illustrating context shift from GeoPlaces and GeoImNet are shown in Fig. 4.5. While prior works already introduce context shift for domain adaptation [170], a key difference lies in their modeling assumption that the context is irrelevant while training, while in our case context might play a key role in improving scene classification on GeoPlaces.

Design shift

We define "design" shift as the change in object structure, shape and appearance, where the foreground objects belonging to the same semantic category look different across geographies, given by $P_s(f_x|y) \neq P_t(f_x|y)$. Few examples are shown in Fig. 4.6, where categories like *castle* from GeoPlaces and *candle* from GeoImNet datasets look widely different due to high intra-class variance, although they belong to the same semantic category. It is important to note that context and design shifts might also occur within a domain or within a geography. However, it is easier to account for intra-domain variations on labeled source datasets than ensuring robustness to new and unlabeled geographies.

Prior shift

The label distributions across the domains in our benchmarks widely differ due to natural prominence or rarity of the classes according to the geography, as shown in Fig. 4.4, where the head classes of one domain might be tail classes in another. This leads to a prior shift where $P_s(y) \neq P_t(y)$. For example, categories like *railway station, outdoor markets, monasteries* are common in Asia while *baseball stadiums* are more common in USA. Prior works examining prior shift or label shift across domains [8, 70, 256, 130, 4] generally assume that the class conditionals remain the same, which is not true in the case of geographic adaptation due to context and design shifts as illustrated above.

4.4 Visualizing Sample Images

We show few sample images from selected classes across both USA and Asia domains in GeoPlaces benchmark in Fig. 4.8, Fig. 4.9 and GeoImNet benchmark in Fig. 4.10, Fig. 4.11.

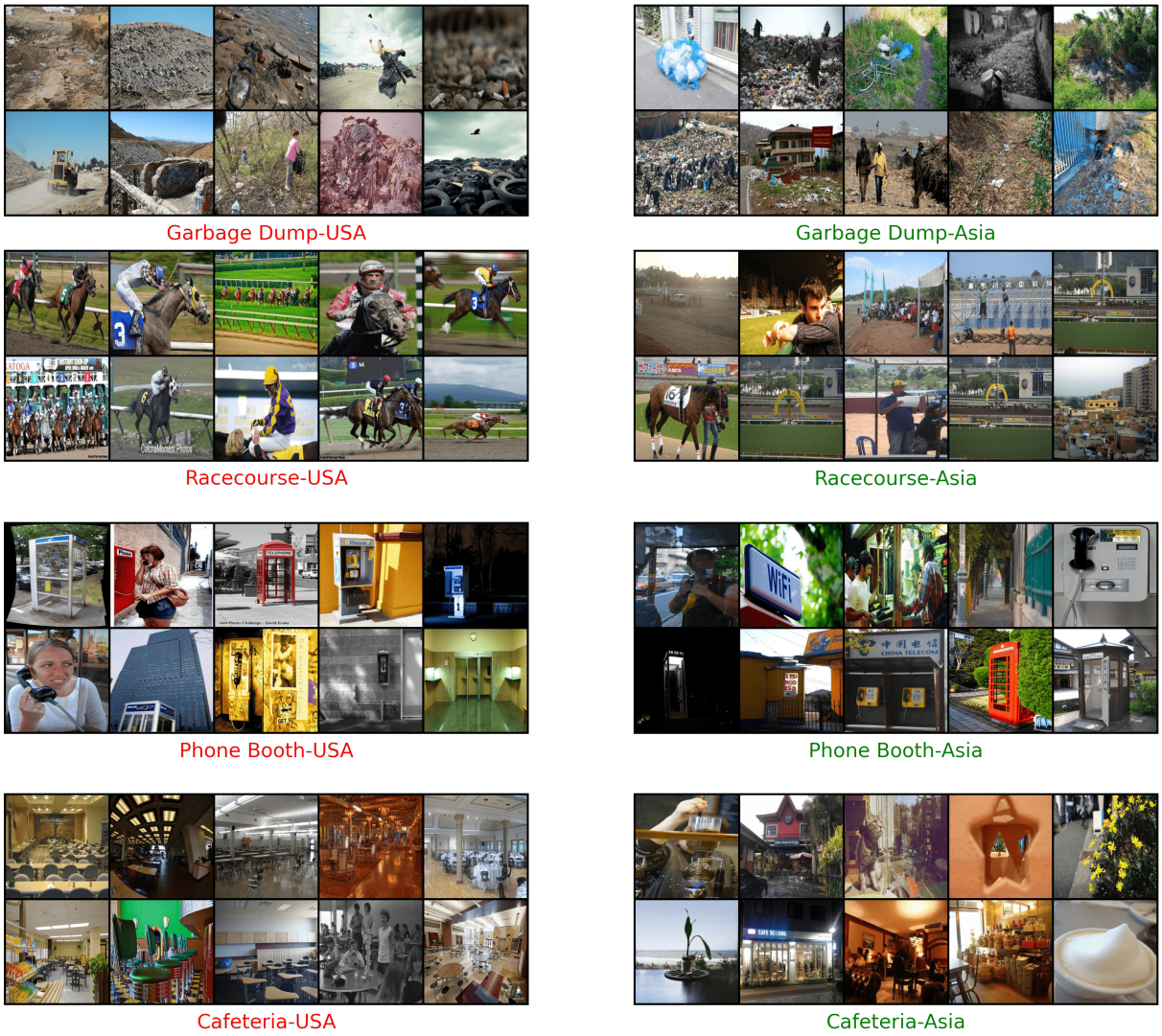


Figure 4.8. Sample images showing the domain gap between USA (left) and Asia (right) domains for classes `garbage dump`, `race course`, `phone booth` and `cafetaria` from GeoPlaces.

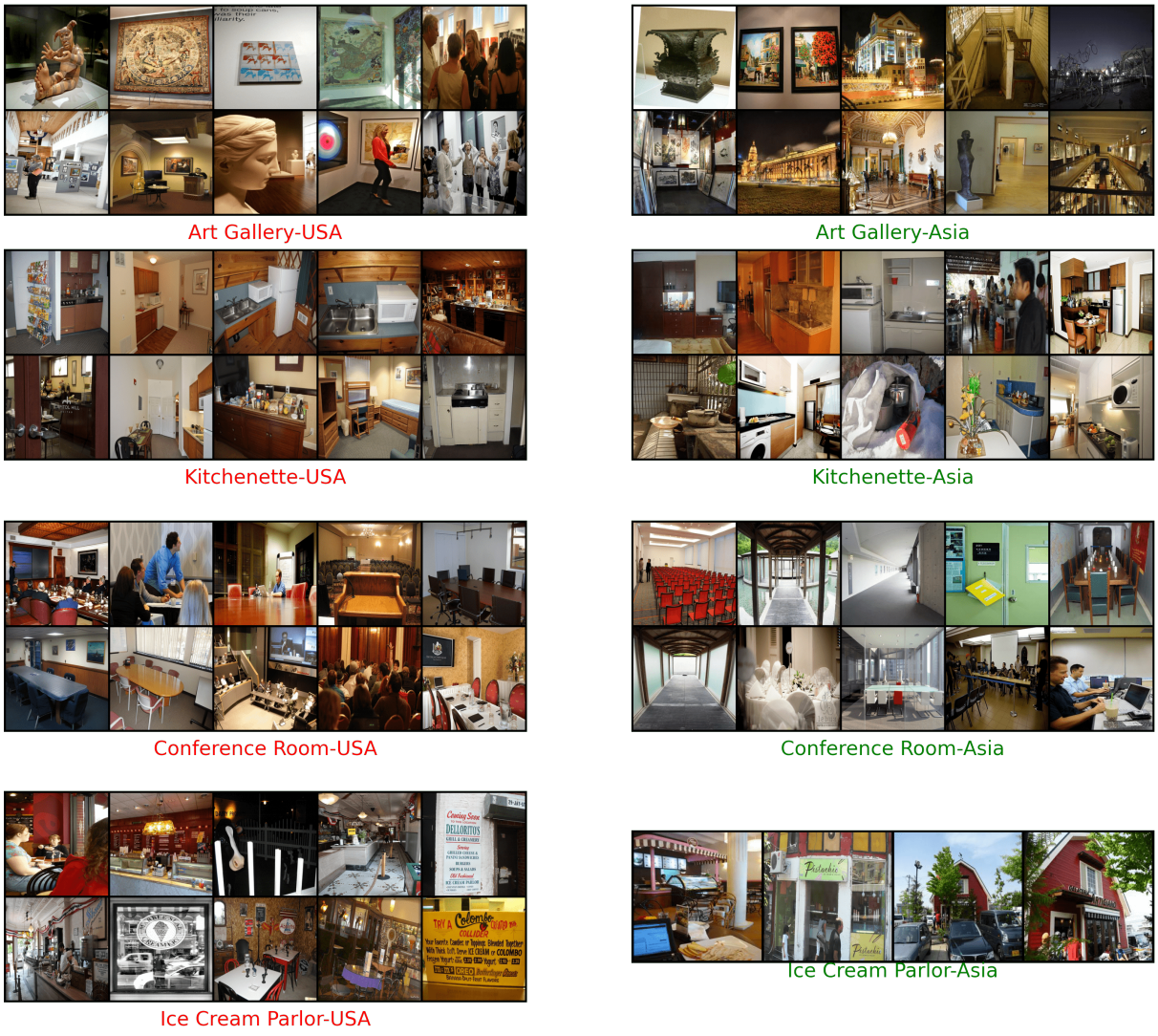


Figure 4.9. Sample images showing the domain gap between USA (left) and Asia (right) domains for classes `art gallery`, `kitchenette`, `conference room` and `ice-cream parlor` from GeoPlaces.

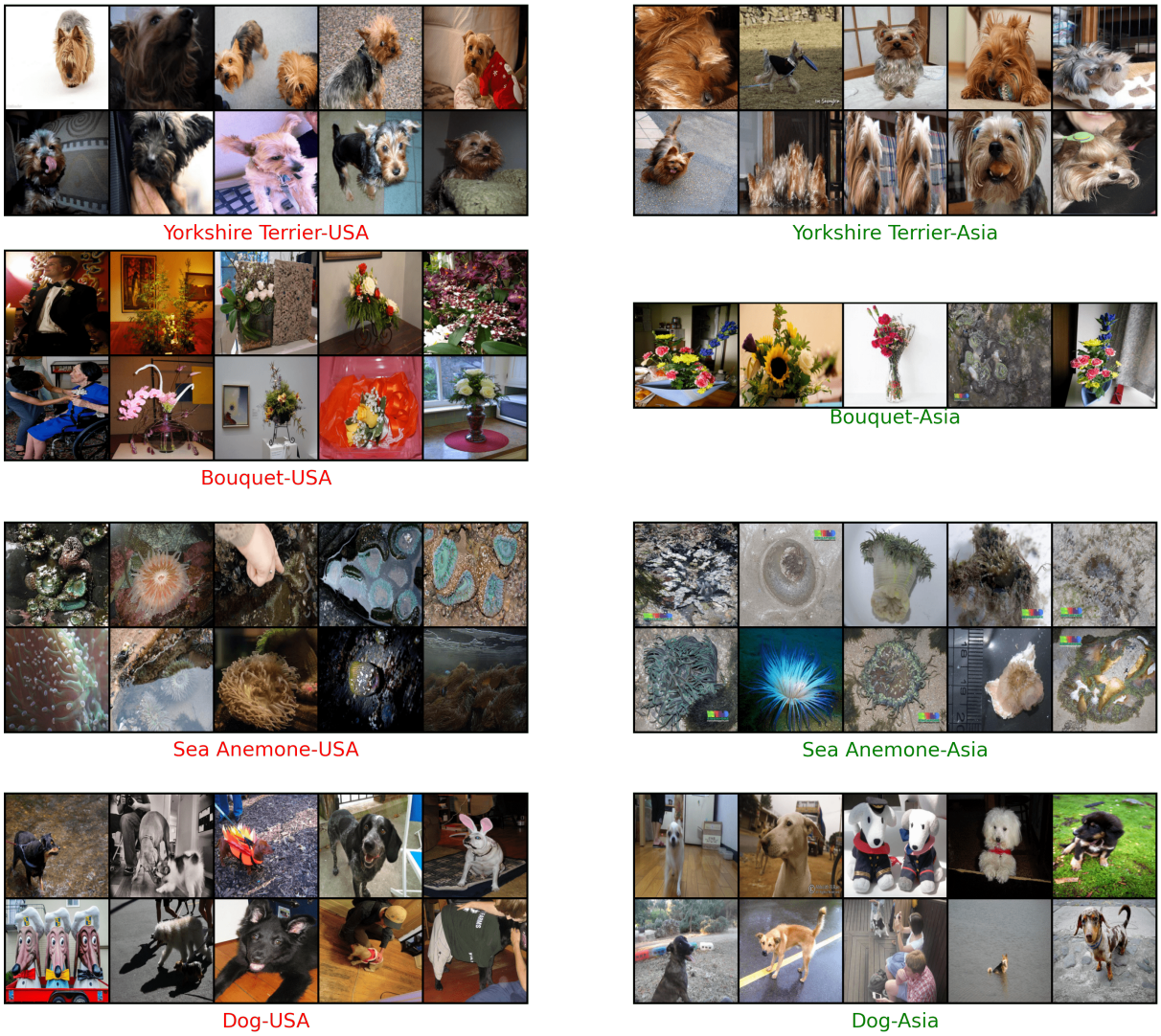


Figure 4.10. Sample images showing the domain gap between USA (left) and Asia (right) domains for classes `Yorkshire Terrier`, `bouquet`, `sea anemone` and `dog` from GeoImNet.

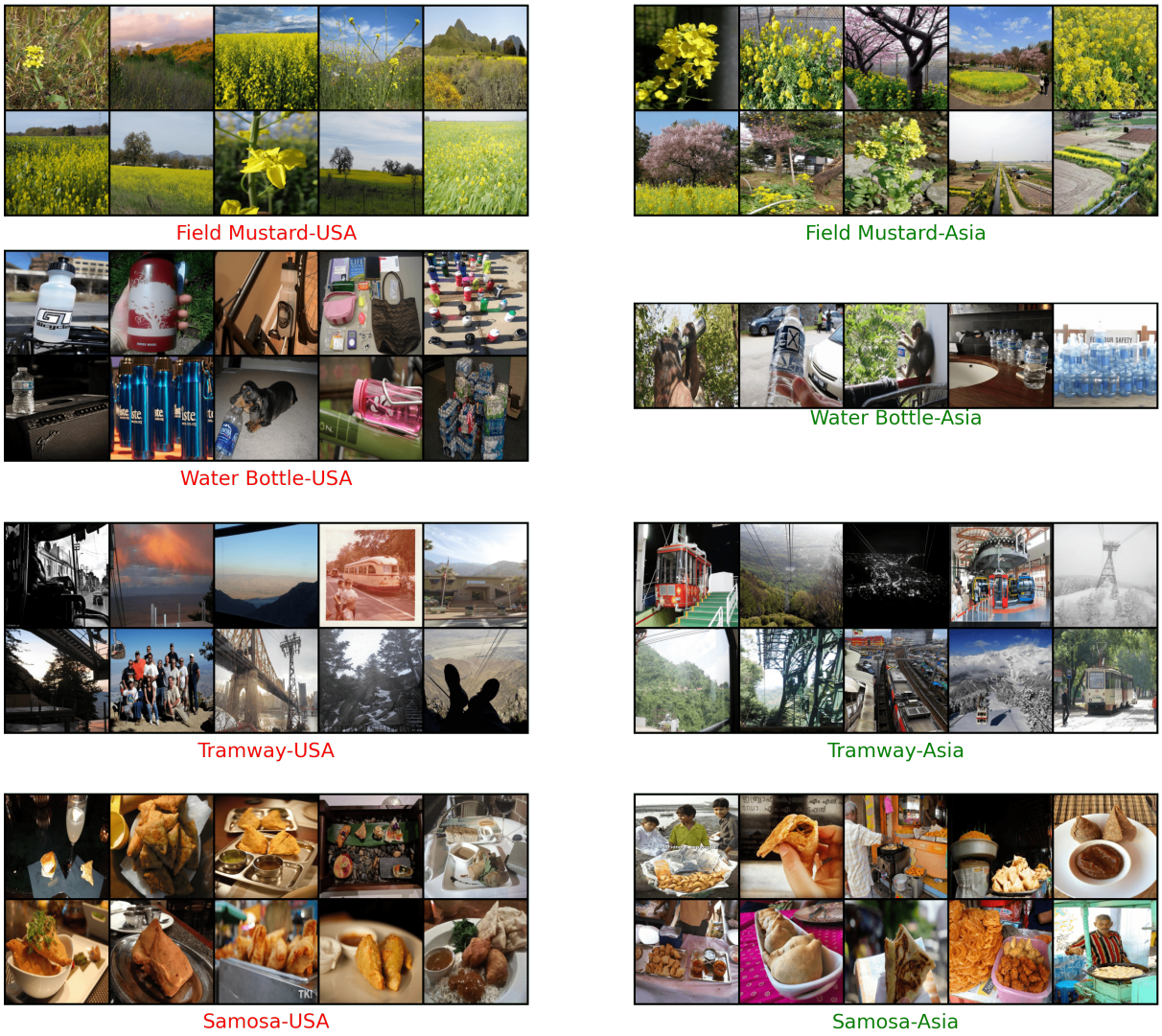


Figure 4.11. Sample images showing the domain gap between USA (left) and Asia (right) domains for classes `Field Mustard`, `Water Bottle`, `Tramway` and `Samosa` from GeoImNet.

4.5 Experimental Results

4.5.1 Domain Shifts in Proposed Datasets

We illustrate the severity of domain differences across geographies using the drop in accuracy caused by cross-geography transfer in Tab. 4.2. Specifically, we train a Resnet-50 [91] model using images only from one domain, and compute the accuracies on both within-domain

Table 4.2. Top-1/Top-5 accuracies of Resnet-50 models across geographically different train and test domains. Note the significant drop in accuracies caused by the geographical domain shifts in each setting.

	GeoPlaces		
Train ↓ / Test →	USA	Asia	Drop (%)
USA	56.35/85.15	36.27/63.27	-20.08/-21.88
Asia	21.03/44.81	49.63/78.45	-28.60/-33.64
	GeoImNet		
Train ↓ / Test →	USA	Asia	Drop (%)
USA	56.35/77.95	36.98/63.42	-19.37/-14.53
Asia	40.43/64.60	60.37/80.22	-19.94/-15.62

Table 4.3. USA → Asia comparison between GeoNet and its label-balanced version. Non-trivial gaps between the geographies still exist even after accounting for prior shift between the domains.

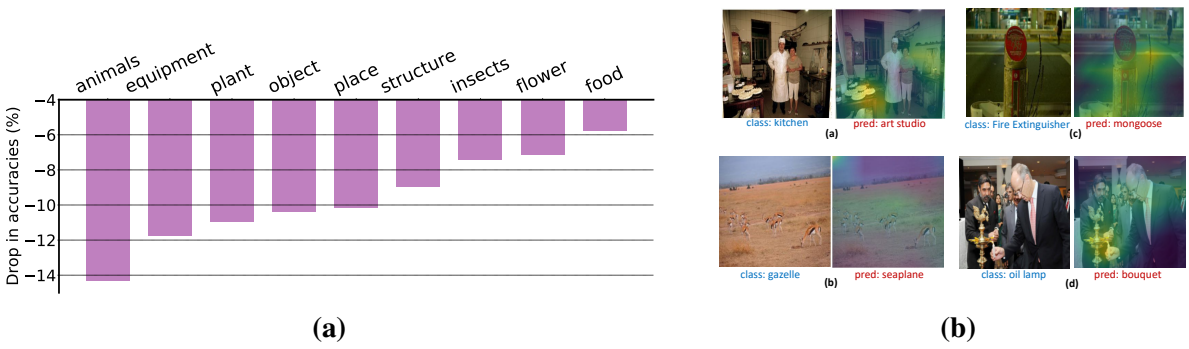
	Original			Balanced		
	USA	Asia	Δ	USA	Asia	Δ
GeoPlaces	56.35	36.27	20.08%	55.52	42.6	12.92%
GeoImNet	56.35	36.98	19.37%	52.72	37.3	15.42%

and cross-domain test sets. Since a lot of categories in GeoNet are close (example, *train station* vs. *subway station*), we use both top-1 and top-5 accuracies to report the performance. We observe a significant drop in accuracy caused by direct transfer of models across domains which can be attributed to the geographic bias in the training data. For example, a model trained on GeoPlaces benchmark on US images gives 56.35% Top-1 accuracy on US images, but only 36.27% on images from Asia with a notable drop of 20%.

On the GeoImNet benchmark, within-domain testing on images collected from USA gives 56.35% top-1 accuracy while cross-domain testing on Asia images gives only 36.98% with a drop of 19.37%. The 36.98% accuracy is also much inferior to the supervised accuracy on the Asia domain (60.37%) which can be considered as the target upper bound.

Meta-category wise error analysis for GeoImNet

We relate the drop in performances across geographies to the proposed notions of domain discrepency in geographic adaptation like context and domain shifts in Fig. 4.12a. Specifically, since the concepts in GeoImNet are sourced from ILSVRC, we leverage the wordnet hierarchy



(a) **(b)**

Figure 4.12. (a) Drop in accuracies for each meta-category in GeoImNet. Groups that showcase context and design shifts suffer a larger drop in accuracy. (b) GradCAM visualization of predictions of a USA-trained model on Asia images show that prominent context and design shifts across geography hurts accuracy. (a) is from GeoPlaces, (b,c,d) are from GeoImNet.

to group our 600 classes into 9 meta-labels. We then average the accuracy within each meta-class from USA→Asia domain transfer, and plot the difference in accuracy across domains per meta-label in Fig. 4.12a. We note that categories in the meta-label "animals" have minimum design-shift across domains, but suffer from context shift due to shifts in weather and habitats across geographies leading to significant drop in accuracy. On the other hand, many categories in "equipment" and "object"(like *candle, broom, sewing machine*) have prominent design shifts (Fig. 4.6) leading to notable performance drop. Finally, categories in "food" (like *bottled water, ice-cream*) have minimum change in both design and context and hence suffer the least fall in accuracy across domains.

GradCAM visualization of the failure cases

We present few examples in Fig. 4.12b of predictions made on Asia test images by a model trained on USA, along with their GradCAM visualizations. As shown, when the model focuses on the context and background, it fails to generalize to new scenes from target geographies with notable shifts in context (*kitchen* classified as *art studio*). Even in cases when the model accurately focuses on the foreground object, it sometimes leads to incorrect predictions due to design shifts between geographies, where *oil lamp* is accurately localized, but predicted as *bouquet*.

Table 4.4. UDA on GeoNet Top-1 and Top-5 accuracies of various unsupervised adaptation methods on GeoNet. Most of the methods fail to sufficiently handle cross-geography transfer on both GeoPlaces and GeoImNet benchmarks and often give lower accuracies even compared to a baseline model trained only using source data calling attention to the need for novel methods that can handle domain shifts beyond style and appearance.

Method	GeoPlaces				GeoImNet			
	USA → Asia		Asia → USA		USA → Asia		Asia → USA	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Source Only	**36.27**	**63.27**	**21.03**	**44.81**	**36.98**	**63.43**	**40.43**	**64.6**
DANN [69]	29.58	55.23	16.59	35.32	32.88	57.77	38.42	62.90
CDAN [139]	30.48	55.94	17.01	36.26	35.94	60.21	39.88	63.74
MCC [104]	30.09	55.85	17.17	36.85	35.71	60.48	39.86	64.00
SAFN [248]	32.50	57.93	14.34	35.68	32.40	58.43	36.26	61.58
MDD [260]	34.18	59.10	17.81	36.44	36.26	62.13	40.15	63.91
MCD [196]	33.49	59.41	16.57	34.74	25.60	48.45	36.69	60.68
ToAlign [238]	29.86	56.16	16.32	33.58	32.13	58.64	37.98	63.17
MemSAC [107]	34.68	60.52	15.75	32.83	36.71	63.16	40.34	64.40
Tgt. Supervised	49.63	78.45	56.35	85.15	60.37	80.22	56.35	77.95

Separating the prior shift

To further delineate prior shift from context and design shifts, we curate a balanced subset out of GeoNet such that each category has about 200-300 images, and drop categories which have fewer images (about $3/4^{th}$ of the categories remain). From Tab. 4.3, the drop in accuracy after addressing the prior shift is 12.9% on GeoPlaces and 15.4% on GeoImNet, compared to 20.08% and 19.37% on the original datasets, showing that non-trivial accuracy drops caused by context and design shifts still exist even after accounting for label imbalance between the domains.

4.5.2 Benchmarking Domain Adaptation

We study the effectiveness of prior unsupervised adaptation algorithms in bridging novel notions of domain gaps like context shift and design shift on GeoNet. We review various standard as well as current state-of-the-art domain adaptation methods to examine their geographical robustness.

Table 4.5. Universal domain adaptation methods on GeoUniDA.*closed-set* and *open-set* refer to the closed set and open set accuracies, and *H-Score* is the harmonic-mean of the two. Note the significant gap that still exists with target supervised accuracy on closed-set labels with the best adaptation method DANCE [192].

Method	closed-set	open-set	H-Score	Target Sup.
UniDA [254]	27.64	43.93	33.93	
DANCE [192]	38.54	78.73	51.75	70.70%
OVANet [193]	36.54	66.89	47.26	

Architecture and training details

We follow the standard protocol established in prior works [139, 196, 107] and use an ImageNet pre-trained Resnet-50 [91] as the feature extractor backbone and a randomly intialized classifier layer. We use a batch size of 32 and SGD with a learning rate of 0.01 for the classifier head and 0.001 for the already pretrained backbone. We report the top-1 and top-5 accuracy numbers using the test splits from each benchmarks. We perform comparisons between traditional adversarial methods (DANN [69], CDAN [139]), class-aware adaptation methods (MCC [104], MDD [260]), non-adversarial methods (SAFN [248], MCD [196]) as well as recent state-of-the-art (ToAlign [238], MemSAC [107]). We train prior works using their publicly available code and adopt all hyper-parameters as recommended in the respective papers.

Existing UDA methods do not suffice on GeoNet

We show the Top-1 and Top-5 accuracies of all the transfer settings from GeoNet in Tab. 4.4. A key observation is that most of the domain adaptation approaches are no better, or sometimes even worse, than the baseline model trained only using source domain data, indicating their limitations for geographic domain adaptation. For example, on GeoPlaces, training using data from USA achieves a top-1 accuracy of 36.27% on test data from Asia test images, while the best adaptation method (MemSAC) obtains lesser accuracy of 34.7%, indicating negative transfer. Likewise, on GeoImNet, a USA-trained source model achieves 36.98% on test images from Asia which is comparable to the best adaptation accuracy of 36.71%. To further illustrate this, we define relative accuracy gain as the improvement in accuracy obtained by a method over a source-

only model as a percentage of gap between a source-only model and the target-supervised upper bound (which is 100% if the method achieves the target supervised upper bound). From Fig. 4.2b, it is notable that the same adaptation methods that yield significantly high relative accuracy gains on DomainNet [165] yield negative relative accuracy gains on GeoNet, highlighting the unique the nature of distribution shifts in real-world settings like geographic adaptation that challenge existing methods. These observations also suggest that future research should focus on context-aware and object-centric representations in addition to domain invariant features to improve cross-domain transfer amidst context and design shifts.

Universal domain adaptation on Geo-UniDA

We run SOTA universal domain adaptation methods (You et.al. [254], DANCE [192] and OvaNET [193]) on the Geo-UniDA benchmark of GeoNet. Following prior works [193], we adopt the H-score metric which is a harmonic mean of closed-set and open-set accuracies giving equal importance to closed set transfer as well as open set accuracy. In Tab. 4.5, we show that DANCE [192] outperforms both You et.al. [254] and OVANet [193] on the Geo-UniDA benchmark. We also show that a significant gap still exists between target supervised accuracy when trained using supervision (70.7%) and best adaptation accuracy (38.5%) on our benchmark, highlighting the limitations of existing methods to efficiently address universal adaptation in a geographic context.

4.5.3 Large-scale Pre-training and Architectures

It is common to use large scale self-supervised [89, 25, 32, 37, 26, 88] and weakly-supervised [103, 210, 145] pre-trained models as starting points in various downstream applications. While recent works explored role of pre-training on domain robustness [114], we are interested in the extent to which large scale pre-training effectively preserved robustness when fine-tuned on geographically under-represented datasets. We investigate the performance of a variety of methods on GeoNet in terms of backbone architectures, pre-training strategies and

supervision.

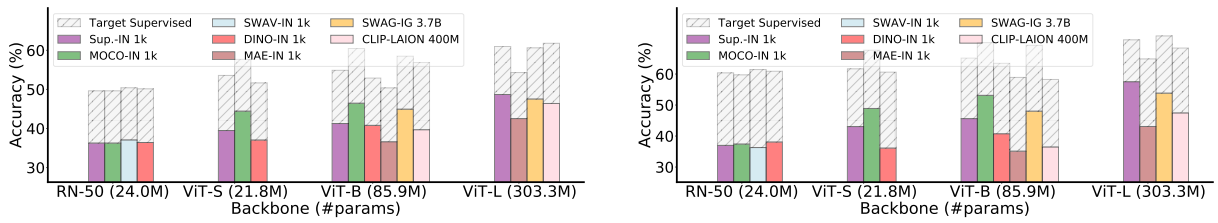
Experimental setup

Our backbone architectures include Resnet50 [91] as well as the small (ViT-S), base (ViT-B) and large (ViT-L) vision transformers [61]. In terms of supervision, in addition to the standard supervised pre-training on ImageNet-1k, we also consider self-supervised methods MoCo-V3 [37], SwAV [25], DINO [26], MAE [88] trained on ImageNet-1k, the weakly supervised SWAG [210] trained on 3.6B uncurated instagram images and CLIP [172] trained on 400M image-language pairs [202]. We denote {Backbone-Supervision-Data} for different model choices (for example, Resnet50-sup-IN1k indicates a Resnet50 pre-trained on supervised data from ImageNet-1k).

For evaluating geographic robustness of these models, we first take the pre-trained model and fine-tune it on training data from a "source" geography, then evaluate the performance on test data from the "target" geography. We show the results using USA as the source and Asia as the target from the GeoPlaces and GeoImNet benchmarks in Fig. 4.13. For reference, we also report accuracy after fine-tuning on labeled data from the target geography for each {Backbone-Supervision-Data} pair (denoted as target-supervised), which serves as an upper bound for the transfer performance.

Large-scale pretraining is not geographically robust

From Fig. 4.13, we make a few observations. Firstly, comparison between Resnet50 and ViT-S which have roughly the same number of parameters suggests the superiority of the vision transformer architectures over CNNs. For example, ViT-S-sup-IN1k is better than Resnet50-sup-IN1k, and ViT-S-moco-IN1k is better than Resnet50-moco-IN1k, indicating that global reasoning using self-attention layers in vision transformers benefits context-dependent tasks like GeoPlaces. Next, comparing different pre-training strategies, we observe that MoCo gives best accuracy on ViT-S and ViT-B, while supervised pre-training outperforms other approaches



(a) GeoPlaces: USA → Asia.

(b) GeoImnet: USA → Asia.

Figure 4.13. Effect of large-scale pre-training on geographical robustness We show that most architectures and pre-training strategies exhibit significant cross-domain drops when fine-tuned on geographically biased datasets. Shown for USA→Asia on GeoPlaces and GeoImNet.

on large models like ViT-L. However, the gap between target supervised accuracy and the best adaptation accuracy achieved using either Resnet50 or any of the vision transformers is still high, highlighting the need for better transfer strategies.

In terms of data, weakly-supervised pre-training using billion-scale dataset IG3.6B (ViT-B-swag-3B) shows significant improvements over self-supervised training methods like MAE (ViT-B-mae-IN1k) and DINO (ViT-B-dino-IN1k). But despite training on massive-scale data, ViT-L-swag-3B and ViT-L-clip-400M are still inferior to the target supervised accuracies, revealing the limitations of current pre-training strategies towards robust cross-geography transfer after fine-tuning. While the success of large-scale pre-training strategies are well-documented on popular datasets like ImageNet, our results indicate that similar benefits might not be observed when application domains significantly differ from pre-training or fine-tuning datasets [45].

4.5.4 Zeroshot Classification Using Vision-Language Models

While the benchmarking analysis so far has focused on the robustness of models when fine-tuned on geographically biased data (such as USA data) and tested on unseen geographies (such as Asia data), an emerging application prospect lies in directly using strong vision-language frontier models such as CLIP [172] for prompt-based zero-shot classification. It is important to examine the geopgraphical robustness of these models because of their reliance on web-scale data during training, and wide adoption either as an open-vocabulary recognition model [121, 73, 60] or strong pre-training models for downstream fine-tuning [243].

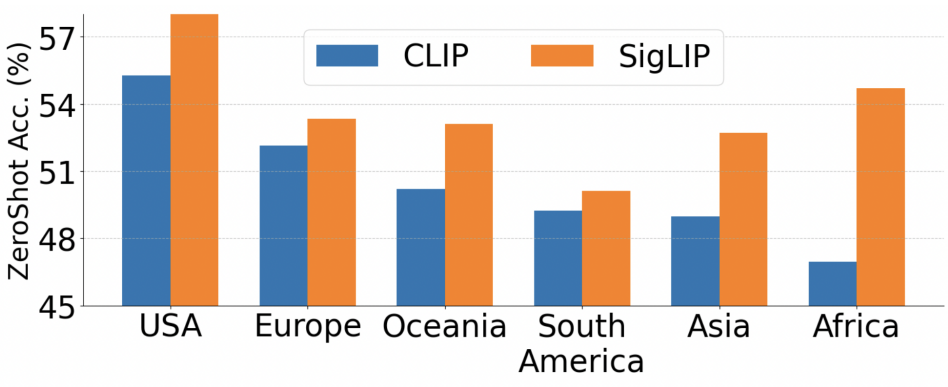


Figure 4.14. Zeroshot Accuracy of VLM models on GeoNet We show the zeroshot accuracy of large-scale VLM models CLIP [173] and SigLIP [255] on the GeoNet dataset across various geographies. Despite being trained on billion-scale image-text pairs, there still exists wide disparity in recognizing concepts from dominant geographies like USA and under-represented geographies like Africa and South America.

We adopted two strong zeroshot models, CLIP [173] and SigLIP [255] to verify their robustness properties on GeoNet dataset. Since no training is required, we use more diverse data from all over the globe, and compute accuracy for each continent separately. We use standard text-prompts such as *An image of a classname* for probing the models, and show our results in Fig. 4.14. As shown, there still exists significant disparity between the accuracy obtained by these models on dominant geographies like USA and under-represented geographies like Africa and South America. For instance, CLIP achieves an accuracy of 55.2% on USA images but only 46.9% on Africa and 49.2% on images tagged from South America. While SigLIP is better then CLIP on most domains owing to its larger pool of training data [34], it still suffers from accuracy drops of upto 5% between geographies, highlighting the need for developing geographically robust vision-language models.

4.6 Summary

Through this chapter, we introduce a new dataset called GeoNet for the problem of geographic adaptation with benchmarks covering the tasks of scene and object classification. In contrast to existing datasets for domain adaptation [166, 165, 188, 230], our dataset with

images collected from different locations contains domain shifts captured by natural variations due to geographies, cultures and weather conditions from across the world, which is a novel and understudied direction in domain adaptation. Through GeoNet, we analyze the sources of domain shift caused by changes in geographies such as context and design shift. We conduct extensive benchmarking on GeoNet and highlight the limitations of current domain adaptation methods as well as large-scale pretraining methods towards geographical robustness. In Chapter 5, we will introduce a language-guided solution for addressing the challenging problem of geographical transfer.

This chapter is a reprint of the material as it appears in "Geonet: Benchmarking unsupervised adaptation across geographies." by Tarun Kalluri, Wangdong Xu, Manmohan Chandraker, which was published in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. The dissertation author was the primary investigator and author of this paper.

Chapter 5

Improving Domain Transfer in Images and Videos using Language Guidance

In this chapter, we investigate the limitations of classical image-alignment based adaptation methods in bridging geographic domain shifts, and propose text-based adaptation as an effective alternative. We introduce a new solution for domain transfer using language as an effective guiding mechanism. By making simple assumptions about the availability of weakly-correlated textual metadata during training from source and target domains, we design a simple knowledge transfer mechanism to handle domain shifts. This is particularly important in current scenarios where there are growing applications with access to textual metadata which can be used for robust fine-tuning across geographies. Further, we extend the boundaries of unsupervised adaptation into videos by first introducing a new benchmark called Ego2Exo and showing the effectiveness of our method on this unique setting.

5.1 Introduction

Despite great strides in the performance in several applications of computer vision recent years, achieving robustness to distribution shifts at test-time still remains a challenge. In particular, a fundamental need to improve generalization to domains without manual supervision arises due to the cost and scarcity of acquiring labeled images. A dominant paradigm to address this limitation has been unsupervised domain adaptation (UDA), which uses labels

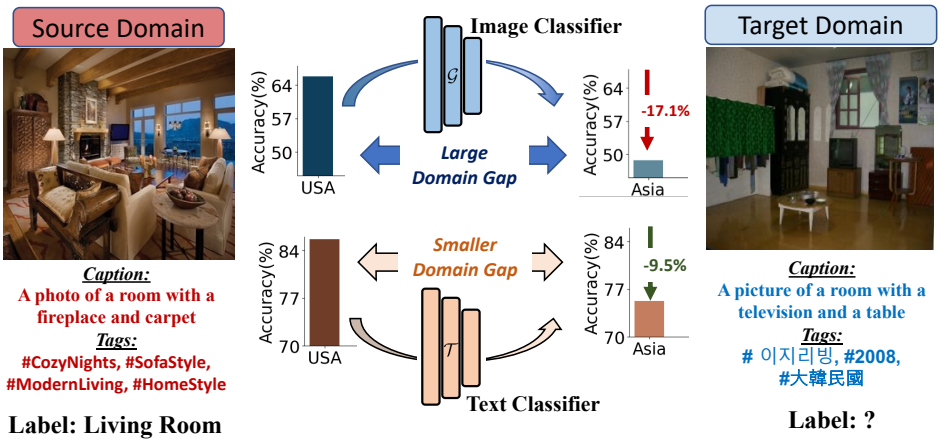


Figure 5.1. A summary of our insights for LaGTran: In a domain transfer setting with labeled source and unlabeled target domain data, we observe significantly more drop incurred while transferring an image-classifier trained on source images to target (17.1%), compared to a text-classifier trained on corresponding text descriptions of source images (9.5%). We use this insight to build a simple framework called LaGTran that leverages these text descriptions easily available in both domains to improve transfer in images and videos.

from a related source domain along with distribution alignment techniques to bridge the domain gap [69, 139, 196, 248, 207, 238, 30, 270]. Despite their noted success, their limitations in addressing challenging transfer beyond regular domain shifts [187, 230, 166] is recently highlighted [170, 109]. We posit that a part of this limitation potentially stems from their dependence on pixel-level data alone to bridge domain gaps, as accurately characterizing shifts and devising bridging strategies solely based on images becomes challenging beyond standard domain shift scenarios.

In contrast, we propose an alternative approach to ease transfer across such challenging shifts by instead leveraging ubiquitously available language guidance during training. Our framework, called **LaGTran** for **La**nguage **G**uided **Tr**ansfer **A**cross Domai**n**s, is surprisingly simple to implement, yet shows extreme effectiveness and competence in handling transfer across challenging domain shifts in images and videos compared to any image-based adaptation method. Our key insight lies in observing that text guidance, which is readily available in the form of metadata for internet-sourced datasets or easily generated with emerging image captioning models, requires no human annotation while offering a more suitable avenue in transferring

discriminative knowledge even across challenging domain shifts.

We further illustrate this property in Fig. 5.1, where we examine the transferability of image and text classifiers trained using image or text supervision respectively between USA and Asia domains from the GeoNet dataset [109]. We observe significantly less drop (9.5%) when applying a text classifier trained on the source text to target text, compared to 17.1% drop incurred when transferring an image classifier to classify target images. As text operates in a significantly lower-dimensional space, language modality naturally tends to have lesser domain gaps as opposed to images or videos. Furthermore, text descriptions often contain valuable attributes and identifiers that enhance the ability to accurately recognize images in a standard classification setting, suggesting more favorable domain robustness and discriminative properties of language descriptions compared to images.

We incorporate these observations to improve transfer in a scenario where the source domain has text descriptions accessible along with the labels, but the target domain only has text descriptions corresponding to the images. Accordingly, we first train a text classifier using the source domain language descriptions and labels and transfer this classifier to assign pseudo-labels to the target text descriptions, which, from Fig. 5.1, would yield more robust pseudo-labels compared to the common image-based transfer [134, 215, 117]. We, therefore, directly use these pseudo-labels as supervision for the unlabeled target images to train an image classifier jointly with source labels. This simple technique, free of any complicated adaptation mechanisms, shows remarkably strong performance surpassing competitive baselines and prior UDA methods.

To further demonstrate the broad usefulness of LaGTran beyond images, we introduce and study a novel benchmark for transfer learning in videos called Ego2Exo, which focuses on the previously under-explored challenge of transferring action recognition between ego (first-person) and exo (third-person) perspectives in videos [125, 158, 171, 251] from a transfer learning standpoint. We curate Ego2Exo benchmark using cooking videos from the Ego-Exo4D dataset utilizing key step annotations to assign action labels and atomic action descriptions for textual guidance. Our language-aided transfer shows remarkable utility in this challenging

setting, significantly outperforming prior Video-UDA methods [31, 239].

To summarize before we delve into details, our contributions in addressing domain transfer using language is three-fold.

- A novel framework LaGTran highlighting the feasibility of incorporating various forms of readily available text supervision in enhancing transfer across domain shifts (Sec. 5.3.1).
- A new dataset Ego2Exo to study the problem of cross-view transfer in videos with fine-grained labels covering a diverse pool of actions and free-form text descriptions providing language guidance (Sec. 5.4.5).
- Demonstration of the competence of LaGTran across a variety of domain shifts, with non-trivial gains over UDA methods on challenging datasets like GeoNet (+10%), DomainNet (+3%) and the proposed Ego2Exo (+4%) datasets (Sec. 5.4).

5.2 Relation to Prior Literature

In this section, we review the existing literature which are closely related to our problem.

5.2.1 Language Supervision in Computer Vision.

The recent proliferation of internet-sourced datasets highlights the ready availability of natural language supervision without the need for any labeling or annotation efforts in images [219, 27, 201, 145, 54] and videos [147, 10, 80, 81]. This availability of language supervision has been effectively utilized to learn scalable weakly supervised models [145, 210], robust vision-language representations [172, 103, 167, 53, 200, 127, 263, 77], text-conditioned generative models [183, 176, 189] and improving sampling techniques for self-supervised learning [68]. Even in the absence of associated language supervision, recent innovations showed the potential of generating correlated descriptions for images using image-to-text or image captioning models [122, 133, 1]. Despite this ubiquity and proven effectiveness of language supervision for vision tasks, little attention has been directed at leveraging their utility

in improving transfer learning across domains. In this work, we use language guidance to develop a straightforward mechanism to improve image and video classification on domains without manual supervision.

5.2.2 Domain Robustness Using Language Supervision.

Recent emergence of large-scale pre-trained vision-language foundational models such as CLIP [172] enabled strong zero-shot generalization across diverse domains and tasks [55]. However, the zeroshot inference using frozen pre-trained models still fall short of supervised fine-tuning [172, 167, 5], which in-turn suffers from poor generalization to distributions outside the fine-tuning data [118, 242]. Prior works explored robust fine-tuning of zero-shot models, but do not leverage target domain data [227] or language supervision [242, 79] during fine-tuning. While recent works incorporate language guidance for domain generalization [67, 236, 131, 101, 148], they mostly rely on domain or class descriptors and do not leverage semantically richer free form text supervision from target images during transfer. In contrast to these efforts, we show that incorporating language aided transfer yields a remarkably effective framework for improving domain robustness.

5.3 Method Details

Problem description and background

We consider the setting of unsupervised cross-domain transfer, with access to labeled data from a source domain $\mathscr{D}_s : \{X_s^i, y_s^i\}_{i=1}^{N_s}$ along with unlabeled data from a target domain $\mathscr{D}_t : \{X_t^i\}_{i=1}^{N_t}$, where $X_s \sim P_s$, $X_t \sim P_t$, N_s and N_t are the number of samples in source and target domains, and the covariate shift assumption means marginal distributions $P_s \neq P_t$ [13, 12]. However, different from prior works, we additionally assume access to natural language descriptions, denoted by c_i, corresponding to each image or video input in both source and target domains during training. Consequently, we denote the labeled source domain with $\mathscr{D}_s : \{X_s^i, y_s^i, c_s^i\}_{i=1}^{N_s}$ and the unlabeled target domain with $\mathscr{D}_t : \{X_t^i, c_t^i\}_{i=1}^{N_t}$. These text descriptions are readily available

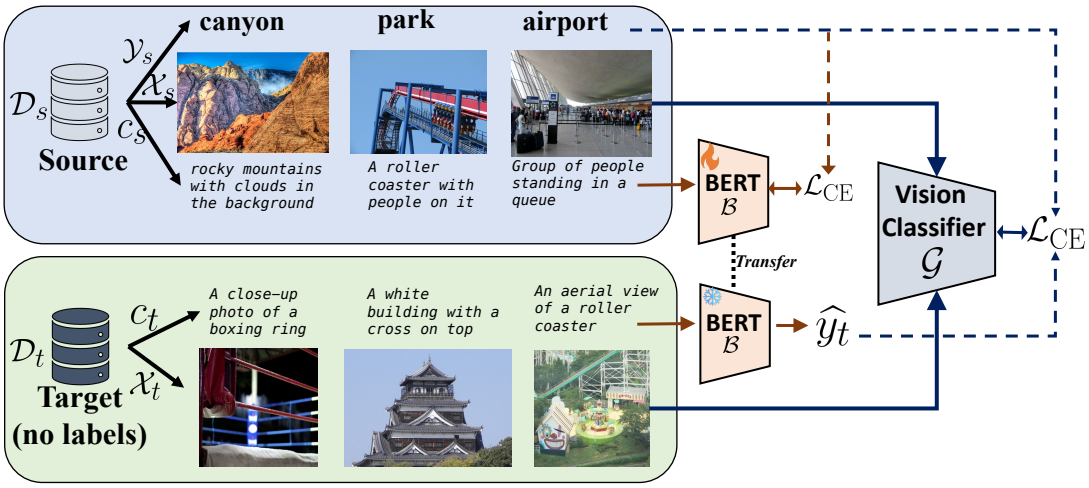


Figure 5.2. An overview of training using LaGTran: We operate in a setting where the labeled source domain and unlabeled target domain data possess cheaply available or easily generated language descriptions for each image. LaGTran proceeds by first training a BERT-classifier $\mathscr{B}$ using source captions and labels (Eq. (5.1)), and using the trained model to generate pseudo-labels $\hat{y}_t$ for the target captions and corresponding images (Eq. (5.2)). We then use this generated supervision along with source domain data in jointly training a **Vision classifier** $\mathscr{G}$ for image or video classification (Eq. (5.3)).

through associated metadata in web-collected images [145], or can be effortlessly generated with state-of-the-art image-to-text models [122]. In Sec. 5.4, we show robust performance using text descriptions derived from a variety of sources, including: image metadata (e.g., alt-text, hashtags) for web-sourced images, state-of-the-art image captioners for manually curated datasets, as well as action descriptions or narrations in videos. Note that our setting requires language descriptions c_i only during training and not during inference or deployment, and therefore incurs no speed or memory overhead at test-time when compared with prior approaches.

5.3.1 LaGTran for Cross-Domain Transfer

The training pipeline used in LaGTran for cross-domain transfer is summarized in Fig. 5.2, where we first train a BERT sentence classifier using the (text, label) pairs from the source domain dataset, and utilize this trained classifier to infer predictions on all the descriptions from the target domain. We then use these predictions as pseudo-labels for the target images, and train a joint vision classifier along with the labeled source domain images.

Training the text classifier

We use the supervised text-label pairs from the source domain (c_i^s, y_i^s) and train a BERT [56] sentence classifier $\mathscr{B}$ to predict the category label from an input text description, using the training objective

$$\phi^* = \arg\min_{\phi} \mathbb{E}_{(c_i, y_i)\sim\mathscr{D}_s} \mathscr{L}_{\text{CE}}(\mathscr{B}(c_i;\phi), y_i), \tag{5.1}$$

where ϕ denotes the parameters of the BERT classifier and $\mathscr{L}_{\text{CE}}$ is the supervised cross-entropy loss. We adopt a pre-trained Distill-BERT [197] model from HuggingFace as the sentence classifier $\mathscr{B}(;\phi)$, and fine-tune it on the source domain data. We observed sub-optimal performance using other pre-trained backbones such as T5 [174] or GPT-2 [173] (Tab. 5.5). Across all datasets and experiment settings used in this work, we feed the raw text descriptions directly into the sentence classifier network without any preprocessing or manual curation. We observed remarkable robustness of the trained classifier in handling several challenges posed by unfiltered text, including their variable lengths across images, language barriers prevalent in geographically diverse data, unrelated tags and descriptions commonly found in web-sourced images or potentially imperfect captions from state-of-the-art captioning models.

To further illustrate our motivation to use text classifier for label transfer, we show the tSNE visualizations of the feature embeddings derived from a source-trained sentence classifier, and compare them to the features derived from a source-trained image classifier in Fig. 5.3. Evidently, the features computed using the text classifier (Figs. 5.3c and 5.3d) are well-separated (more intra-class separation) and well-aligned (less inter-domain separation) compared to image classifier (Figs. 5.3a and 5.3b) further validating our hypothesis that the text descriptions of same-class images from both within and across domains lie close to each other.

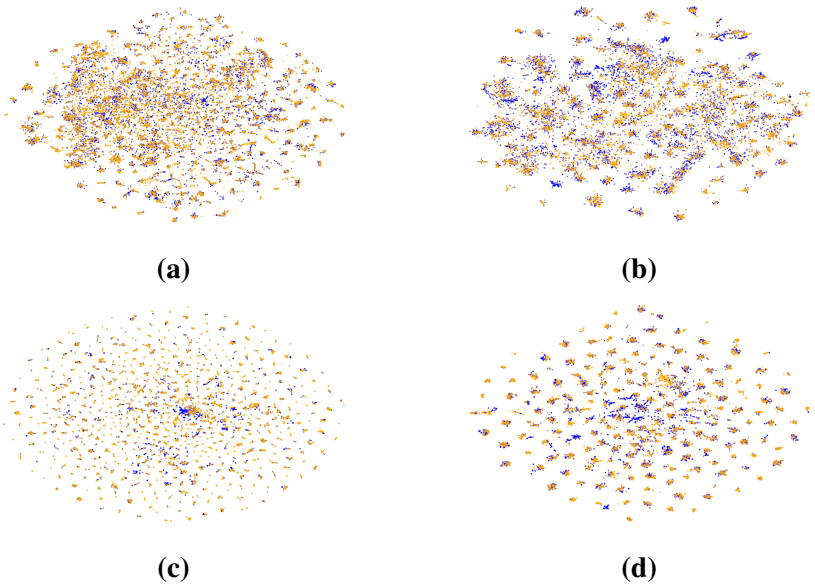


Figure 5.3. tSNE visualization of cross-domain features on GeoNet. We show improved domain-alignment with better class-separation in source and target when extracting features from a text-classifier (Figs. 5.3c to 5.3d) compared to features from image-classifier (Figs. 5.3a to 5.3b) highlighting better transferability through text modality. (Source in orange and target in blue).

Cross-modal supervision transfer

We distill the powerful discriminative knowledge learned from text into images through cross-modal (text to image) supervision transfer in the target domain. Specifically, we first freeze the weights of the source-trained BERT classifier $\mathscr{B}$ and compute pseudo-labels on all the target images using their corresponding text descriptions. For an image x_i^t with caption c_i^t,

$$\hat{y}_i^t = \arg\max_{\mathscr{C}} \mathscr{B}(c_i^t; \phi^*), \tag{5.2}$$

where $\mathscr{C}$ is the set of categories in the classification task. Using these predictions, we construct a pseudo-labeled target dataset, given by $\widehat{\mathscr{D}_t} = \{x_i^t, \hat{y}_i^t\}_{i=1}^{N_t}$. Finally, we combine this pseudo-labeled target images along with manually labeled source domain images to train an image classifier

backbone $\mathscr{G}$.

$$\arg\min_{\theta} \mathop{\mathbb{E}}_{(x_i,y_i)\sim\mathscr{D}_s} \mathscr{L}_{\text{CE}}(\mathscr{G}(x_i;\theta),y_i) + \mathop{\mathbb{E}}_{(x_i,\hat{y}_i)\sim\widehat{\mathscr{D}}_t} \mathscr{L}_{\text{CE}}(\mathscr{G}(x_i;\theta),y_i) \tag{5.3}$$

Note that the inference is performed exclusively using the trained image-based classifier $\mathscr{G}(;\theta^*)$ on image inputs, and neither the text inputs nor the sentence-classifier $\mathscr{B}$ is needed or used at test-time.

5.3.2 Extending LaGTran to Handle Outliers

Owing to the simplicity in the design, LaGTran can easily be extended to the case where the target domain potentially contains outlier samples from outside the category set, also called open-world or universal adaptation (UniDA) [254, 191]. While classical transfer necessitates complete matching between source and target category spaces, open-world transfer relaxes this requirement, allowing the possibility of encountering images from previously unseen and outlier categories during test-time in the target domain [254, 191]. The task is then to accurately classify a test-image into one of $\mathscr{C}_s$ categories shared between source and target domains while simultaneously detecting outlier images from target private classes. To suit LaGTran for UniDA, we modify Eq. (5.2) to additionally label predictions made by the text-classifier network $\mathscr{B}$ with an *outlier* class using maximum softmax probability threshold [95] after training.

$$\hat{y}_i^t = \begin{cases} \arg\max_{\mathscr{C}_s} \mathscr{B}(c_i^t;\phi^*) & \text{if } \max_{\mathscr{C}_s} \mathscr{B}(c_i^t;\phi^*) > \tau \\ |\mathscr{C}_s|+1 & \text{otherwise,} \end{cases} \tag{5.4}$$

where τ is a threshold used to detect outlier samples during inference. We then proceed to train a downstream classifier on $|\mathscr{C}|_s{+}1$ classes using data from supervised source and pseudo-labeled target data from shared classes as well as the outlier class. During inference, we assign

a test-image to one of the $\mathscr{C}_s$ classes or the special outlier class based on the prediction. We heuristically choose $\tau = 0.75$ and do not ablate on this. We show in Sec. 5.4.4 that this simple extension yields highest accuracy on the challenging GeoUniDA dataset [109], highlighting the versatility of LaGTran to handle diverse styles of domain transfer.

5.4 Experimental Results

We first study the effectiveness of LaGTran on standard image datasets (Sec. 5.4.1) and extensions to open world transfer (Sec. 5.4.4). We then show our results on a new dataset for transfer between ego-exo views in videos (Sec. 5.4.5) followed by extensive ablations and insights into our framework (Sec. 5.4.6).

5.4.1 LaGTran for Image Classification

Datasets

We adopt GeoNet [109] and DomainNet [165] datasets which together cover a range of domain shifts across varying difficulty levels. GeoNet is the largest dataset for domain adaptation with more than 750k images, proposed to study a practical real-world problem of geographic disparities in images for two tasks - GeoImnet for image classification from 600 classes and GeoPlaces for scene recognition from 205 classes. DomainNet is a challenging dataset proposed for adaptation with 400,000 images from 345 classes. Following prior work [238, 107], we show our results on all 12 transfer settings from the 4 most studied domains *real, clipart, sketch* and *painting*. We use a ViT-base [61] backbone as the image encoder on the GeoNet, and follow prior work [270] and use Swin-base backbone [136] for experiments on DomainNet.

Training details

We use a ViT-base [61] backbone as the image encoder on the GeoNet dataset, and follow prior work [270] and use Swin-base backbone [136] for experiments on the DomainNet data. Both the backbones are pre-trained on ImageNet-1k, and we add a 2-layer MLP on top of the

computed features as the classifier head. Across all transfer settings, we train these backbones for 90,000 iterations using the objective function specified in Eq. (5.3), employing SGD with a learning rate of 3e-4 and batch size of 64 from each domain, along with a cosine decay schedule.

For the text classifier, We use a pre-trained Distill-BERT [197] model from HuggingFace as the sentence classification model $\mathscr{B}(;\phi)$, and fine-tune it for five epochs over the source domain data using AdamW optimizer with a learning rate of 5e-5 and cosine decay over the training schedule. We observed sub-optimal performance using other pre-trained backbones such as T5 [174], GPT-2 [173] or text encoder in CLIP [172] (Sec. 5.4.6).

Source of text supervision

For GeoNet, we use text supervision from the metadata publicly released along with the dataset, and concatenate the tags, alt-text and free-form captions provided for each image to create the text descriptions. For the DomainNet dataset, since no associated text descriptions are provided, we use a BLIP-2 [122] model to generate short captions for each image from all the domains. Note that our method only requires text during training, and inference is done solely based on images.

Baselines used for comparison

A possible argument for the effectiveness of text supervision might be the direct presence of label information in the text description, eliminating the need for any manual supervision at all. To study this in greater detail, we devise two strong baselines to derive pseudo-labels directly using the text descriptions in the target without using any source domain data as follows. We first use a pre-trained Sentence-BERT [181] encoder, and compute the label embeddings of all the category names as $\mathbf{L} \in \mathbb{R}^{|\mathscr{C}| \times d}$, where d is the embedding dimension of the sentence encoder, followed by zero-shot inference using: (i) **TextMatch,** where we compute the embedding of each text description $e_i^t \in \mathbb{R}^{1 \times d}$ from the target domain, and assign pseudo-label to the label with the highest similarity score with the text embeddings: $\hat{y}_i = \arg\max_{|\mathscr{C}|}(e_i^t \cdot \mathbf{L}^T)$, and

Table 5.1. LaGTran outperforms all prior methods by >10% on average with the challenging GeoImnet benchmark with 600 classes and GeoPlaces with 205 classes designed for geographical transfer (Sec. 5.4.2). All methods use a ViT-B backbone. †denotes domain aware-prompting. Best values in **bold**, second best underlined. U:USA, A:Asia.

	GeoImnet		GeoPlaces		Average
	U→A	A→U	U→A	A→U	
Unsupervised Adaptation					
Source Only	52.46	51.91	44.90	36.85	46.53
CDAN [139]	54.48	53.87	42.88	36.21	46.86
MemSAC [107]	53.02	54.37	42.05	38.33	46.94
ToAlign [238]	55.67	55.92	42.32	38.40	48.08
MDD [260]	51.57	50.73	42.54	39.23	46.02
DALN [30]	55.36	55.77	41.06	40.41	48.15
PMTrans [270]	56.76	57.60	46.18	40.33	50.22
Zeroshot Classification					
CLIP† [172]	49.84	53.83	43.41	54.34	50.36
TextMatch	49.68	54.82	53.06	50.11	51.92
nGramMatch	49.53	51.02	51.70	49.87	50.93
LaGTran	**63.67**	**64.16**	**56.14**	**57.02**	**60.24**

(ii) **nGramMatch,** where we additionally compute the set of all n-grams $\{w\}$ for each text description c_i for $n = \{1,2,3,4\}$ and find the embeddings for each of these ngrams separately: $\mathbf{W} \in \mathbb{R}^{|w| \times d}$. The pseudo-label is then assigned to the label with the highest similarity score with the best matching ngram: $\hat{y}_i = \arg\max_{|\mathscr{C}|} \max_w (\mathbf{W} \cdot \mathbf{L}^T)$. Once the pseudo-labels are generated, we proceed with training a joint model using Eq. (5.3) as before. In addition to these, we also compare the zero-shot classification obtained using CLIP [172] with ViT-base backbone. We adopt domain-aware prompting following prior work [67, 131], where we incorporate the domain information into the prompt-text (eg: `A sketch of a <class>` instead of `A photo of a <class>` to classify sketch images).

5.4.2 LaGTran Outperforms Prior Works on GeoNet

We present results for GeoPlaces and GeoImnet benchmarks in Tab. 5.1. As noted in [109], previous UDA methods often fall short of bridging geographic disparities, highlighting the

Table 5.2. LaGTran sets new state-of-the-art on DomainNet-345 dataset, outperforming prior methods and baselines in most tasks. All models use Swin-B backbone, and UDA numbers are taken from [270]. †denotes domain aware-prompting. Best values in **bold**, second best underlined. R:Real, C:Clipart, S:Sketch, P:Painting.

Source	**Real**→			**Clipart**→			**Sketch**→			**Painting**→			
Target	C	S	P	R	S	P	R	C	P	R	C	S	Avg.
Unsupervised Adaptation													
Source Only	63.02	49.47	60.48	70.52	56.09	52.53	70.42	65.91	54.47	73.34	60.09	48.25	60.38
MCD [196]	39.40	25.20	41.20	44.60	31.20	25.50	34.50	37.30	27.20	48.10	31.10	22.80	34.01
MDD [260]	52.80	41.20	47.80	52.50	42.10	40.70	54.20	54.30	43.10	51.20	43.70	41.70	47.11
CGDM [63]	49.40	38.20	47.20	53.50	36.90	35.30	55.60	50.10	43.70	59.40	37.70	33.50	45.04
SCDA [123]	54.00	42.50	51.90	55.00	44.10	39.30	53.20	55.60	44.70	56.20	44.10	42.00	48.55
SSRT-B [215]	69.90	58.90	66.00	75.80	59.80	60.20	73.20	70.60	62.20	71.40	61.70	55.20	65.41
MemSAC [107]	63.49	42.14	60.32	72.33	54.92	46.14	73.46	68.04	52.75	74.42	57.79	43.57	59.11
CDTrans [249]	66.20	52.90	61.50	72.60	58.10	57.20	72.50	69.00	59.00	72.10	62.90	53.90	63.16
PMTrans [270]	74.10	61.10	**70.00**	79.30	63.70	62.70	77.50	73.80	62.60	79.80	69.70	61.20	69.63
Zero-shot Classification													
CLIP† [172]	72.39	60.90	66.81	**81.37**	60.90	66.81	**81.37**	72.39	66.81	**81.37**	72.39	60.90	70.38
TextMatch	71.36	64.30	65.32	81.25	65.65	64.85	81.09	72.65	63.94	81.08	70.84	**64.17**	70.14
nGramMatch	68.92	59.82	63.15	76.35	61.72	62.87	76.35	69.28	62.51	76.04	68.52	60.52	67.17
LaGTran	**77.30**	**68.25**	67.35	81.31	**67.03**	**66.81**	80.78	**75.62**	**68.08**	79.23	**73.80**	63.44	**72.41**

challenge of geographical transfer with image data alone. Notably, LaGTran achieves 60.24% average Top-1 accuracy on all transfer tasks, beating all previous UDA methods and strong baselines by significant margins, providing solid validation to our transfer approach using language guidance. Specifically, LaGTran outperforms the source-only baseline by $\sim$14% and best adaptation approach PMTrans [270] by $\sim$10% on the average accuracy, highlighting the natural benefit conferred by training while leveraging text supervision in source and target domains. LaGTran even surpasses zeroshot accuracy using domain-aware prompting on CLIP [172] by $\sim$10%, while being trained on order of magnitude fewer data compared to CLIP's hundreds of millions of image-text pairs. Remarkably, we also outperform the strongest baseline *TextMatch* by $\sim$8%, underlining the fact that in cases when the text descriptions might not always have embedded label information directly, using labels from a source domain still has significant advantage.

Table 5.3. Results on open-world transfer on GeoUniDA shows strong performance of LaGTran even with target outlier classes, achieving the highest H-score. Baseline numbers takes from [109].

Method	Closed Set Acc.	Open Set Acc.	H-score
Source Only w/MSP	38.00	73.90	50.20
UniDA [254]	27.64	43.93	33.93
DANCE [191]	38.54	78.73	<u>51.75</u>
OVANet [193]	36.54	66.89	47.26
LaGTran	52.98	72.35	**61.16**

5.4.3 LaGTran is Highly Effective on DomainNet

We summarize the results on DomainNet in Tab. 5.2, where LaGTran yields large gains over several prior UDA methods and all the competitive baselines, setting new state-of-the-art on this challenging dataset. Notably, many prior methods return lesser numbers than directly training on a source model [196, 260, 63, 123], indicating their poor scalability to natural domain shifts in large-scale data. While more recent innovations in UDA such as self-training [215] and patch-based mixing [270], as well as zeroshot inference using CLIP offer improved performance, LaGTran still outperforms these methods on most tasks. Finally, our superior accuracy compared to both baselines *TextMatch* and *nGramMatch*, that employ target-only pseudo-labeling, underscores the significance of having access to supervised text data and labels from a source domain for enhanced target accuracy.

5.4.4 LaGTran Improves Transfer with Outliers

We show our results on open-world transfer setting using the GeoUniDA dataset [109], which examines unsupervised transfer across geographies in the presence of geographically unique classes in both source and target along with common classes. Specifically, GeoUniDA contains 62 shared classes between source and target, along with 138 private categories in each domain. We follow OVANet [193] to adopt the H-score evaluation metric, which gives equal importance to closed-set and open-set accuracies by measuring the harmonic mean of both. In

addition to standard works that address outlier detection through universal adaptation [254, 193, 191], we also train a baseline model using only the source domain data performing test-time outlier detection using MSP thresholding [95]. As shown in Tab. 5.3, LaGTran achieves a H-score of 61.16%, significantly surpassing the baseline source-only accuracy as well as all prior universal adaptation approaches by $>$10%, indicating that language guidance naturally provides a strong signal to detect target samples while handling outliers in open-set target domain data.

5.4.5 LaGTran for Video Domain Adaptation

So far, we showed the effectiveness of LaGTran for domain transfer for image datasets. However, videos also provide new challenges for cross-domain transfer, where language guidance can aid in helping bridge the domain shifts. We next extend these ideas to a new video adaptation dataset.

Ego2Exo dataset

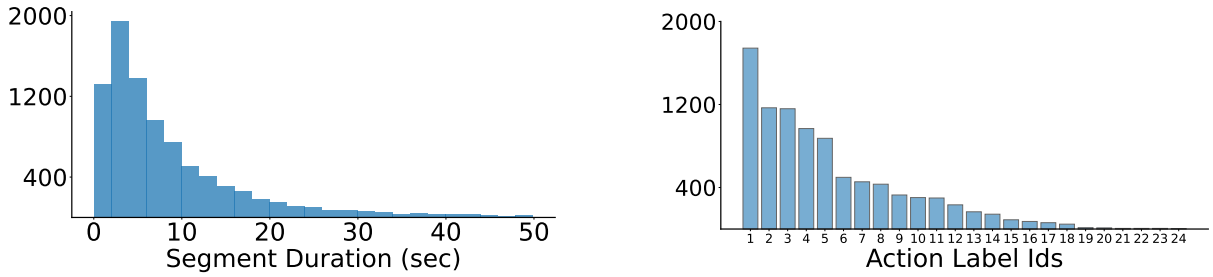
Despite rapid advances in methods [31, 151, 43, 239] and benchmarks [151, 168] for video domain adaptation, little insight is available into their ability to address challenging settings such as transfer between ego (first-person) and exo (third-person) perspectives in videos. While prior efforts studying ego-exo transfer require paired videos from both views [171, 208] or do not leverage unlabeled data in the target [125, 158, 251], limited works looked into the aspect of unsupervised domain transfer from ego to exo views due to the lack of a suitable benchmark.

Therefore, we introduce a new benchmark called Ego2Exo to study transfer between the ego and exo views in videos. We curate our dataset using the recently proposed Ego-Exo4D [81], utilizing their keystep annotations for action labels, and atomic descriptions as text supervision. We manually remap the labels to a coarser hierarchy to ease the difficult task of predicting very fine-grained action classes from short segments (eg: `add coffee beans` vs. `add coffee grounds`).

Our proposed Ego2Exo consists of video segments labeled with actions from one of the

24 keysteps, and we split these video segments into two equal groups classwise, and collect ego-videos from one group and exo-videos from the other to create our adaptation benchmark. We finally obtain 4100 ego-videos and 4986 exo-videos capturing mutually exclusive actions and scenes. The atomic action descriptions from all the timestamps within each segment, whenever available, form the text supervision for that segment. The same procedure applied to the validation videos yields 3147 segments with both ego and exo views. The distribution of the duration of segments in the benchmark, along with the label distribution for ego and exo domains is presented in Fig. 5.4. The final category list in Ego2Exo is as follows:

1. Cook
2. Serve
3. Clean up
4. Add water
5. Make dough
6. Make pasta
7. Make salad
8. Make chai tea
9. Make milk tea
10. Get Ingredients
11. Prepare dressing
12. Prepare a skillet
13. Add spring onions
14. Turn off the stove
15. Check paper recipe
16. Prepare ingredients
17. Prepare milk (boiled)
18. Construct undressed salad
19. Cook noodles in a skillet
20. Get kitchenware & utensils
21. Brew coffee (instant coffee)
22. Boil noodles in boiling water
23. Brew coffee (manual pour-over)
24. Mix noodles with sauce in a bowl



(a) Segment Duration

(b) Label Distribution

Figure 5.4. Dataset Statistics for Ego2Exo: Fig. 5.4a Shows the distribution of segment durations of action videos from Ego2Exo which range from 0.4sec-1min. Fig. 5.4b shows the long-tail of category distribution in Ego2Exo indicating the challenge in robust classification and transfer.

Curating the dataset

To provide text supervision to our algorithm, we use the *atomic action descriptions* provided in Ego-Exo4D dataset. These descriptions provide a narrative of the events in the video, presented in free-form text from the perspective of a third-party observer. Unlike keystep labels, which are defined between specific start and end times within a video, these text descriptions are associated with distinct timestamps, or a single point in time within the video. To create correspondence mapping between the keystep segments and text descriptions, we adopt the method outlined in EgoVLP [127] as follows: to generate a text description for a segment, we compile all text descriptions that fall within the timestamps defined by the start and end times of that segment. If multiple timestamps exist, we concatenate the corresponding texts; if no timestamps are available, we include no associated text with the segment. Furthermore, we concatenate the annotations provided by multiple annotators in creating the text description.

Our proposed Ego2Exo consists of video segments labeled with actions from one of the 24 keysteps, with corresponding text descriptions for each segment. We split these video segments into two equal groups classwise, and collect ego-videos from one group and exo-videos from the other to create our adaptation benchmark. The same procedure applied to the validation videos yields 3147 validation segments with both ego and exo views. An illustration of the

Figure 5.5. Illustration of transfer setting in Ego2Exo Our Ego2Exo benchmark studies domain transfer between the ego and exo modalities in videos. We provide textual descriptions of the actions to serve as the language supervision for each snippet of video.

transfer setting in Ego2Exo is presented in Fig. 5.5.

Training details for videos

We use the pre-computed Omnivore-base [74] features which are provided along with the EgoExo4D dataset for training and evaluation. Since different keysteps may be represented by largely different timespans (Fig. 5.4a), we collect all features that fall within the start and end times of a segment, and pool these features together to form a 1536-dimensional feature representation of that segment. We then train a 2-layer MLP classifier on top of these features, using the labeled source feature as well as pseudo-labeled target features following Eq. (5.3). Note that this training strategy is equivalent to training an MLP classifier on top of frozen Omnivore backbone. For fair comparison, we follow the same strategy for training all the other baselines as well as prior adaptation methods. For methods that require a temporal sequence of features [239, 31], we sample 8 equally spaced features from the complete set of segment features, and use this feature sequence as input. We follow similar strategy for evaluation, and use features pre-extracted from the validation videos for testing. We use the top-1 accuracy on the validation

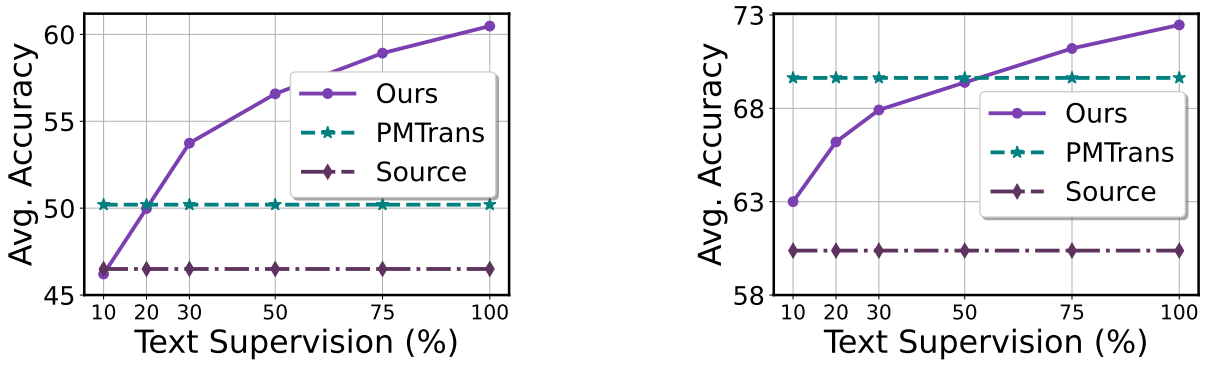
Table 5.4. Results on Ego2Exo benchmark LaGTran achieves the highest accuracy compared to prior video UDA methods as well as zeroshot video-text pre-trained models. Best values in **bold**, second best underlined. All methods use pre-extracted omnivore-base features, EgoVLP and LaVILA use Timesformer-base backbone.

	Ego→Exo	Exo→Ego	Avg.
Unsupervised Adaptation			
Source Only	8.39	15.66	12.03
TA3N [31]	6.92	27.95	17.44
TransVAE [239]	12.06	23.34	17.70
Zero-shot Video Recognition			
EgoVLP [127]	5.89	19.35	12.62
LaVILA [263]	5.86	23.16	14.51
TextMatch	10.36	13.57	11.97
nGramMatch	11.50	15.46	13.98
LaGTran	**12.34**	**30.76**	**21.55**
Target Sup.	17.91	33.19	25.55

set for evaluation. We compare LaGTran for video with prior UDA approaches [28, 239] as well as Video-CLIP based methods with domain-aware prompting [127, 263].

LaGTran efficiently handles cross-view transfer in videos

Firstly, we highlight the importance of studying robustness across ego and exo views in Tab. 5.4 by examining the ego-test accuracy of a model trained directly on ego videos, which achieves 33.19%, compared to a model transferred from exo-videos, which only achieves 15.66%. Similarly, a model trained on ego videos achieve only 8.4% for recognition in exo view, compared to a potential 17.91% achievable by training directly on exo videos, indicating a significant domain shift. Current state-of-the-art video adaptation methods [239] yield limited gains to bridge these gaps, highlighting the need for novel approaches to address this challenge. Moreover, zeroshot video classification accuracy using EgoVLP [127] and LaVILA [263] also show limited gains. Notably, LaGTran which efficiently leverages action descriptions available alongside the videos, achieves an accuracy of 21.55% on average significantly outperforming the source-only



(a) Accuracy on GeoNet

(b) Accuracy on DomainNet

Figure 5.6. Impact of the amount of text supervision on the target accuracy. LaGTran outperforms strong UDA methods while requiring text supervision from only 20% of samples in GeoNet and 50% in DomainNet, with potential for further enhancement with increased text data.

baseline by 9% and prior adaptation methods by >4%. LaGTran also outperforms pseudo-labeling using *nGramMatch* or *TextMatch*, as the text descriptions, independently developed from keystep labels, often lack utility for deciphering the action category labels on their own. We also note the substantial scope for further improvement in future, both in terms of the low within-domain accuracy as well as the remaining gap to supervised target accuracy.

5.4.6 Analysis and Ablations

How much text supervision is needed for LaGTran?

Since natural language supervision is fundamental to LaGTran, we analyze the impact of the amount of supervision available on the eventual target accuracy. We retrain LaGTran by assuming text supervision from only μ% of images in both source and target domains, where $\mu = \{10, 20, 30, 50, 75, 100\}$%, and simply discard the target images that do not have corresponding textual supervision. As shown in Fig. 5.6, LaGTran outperforms image-only method PMTrans [270] even with just 20% text supervision in GeoNet(Fig. 5.6a) and 50% in DomainNet(Fig. 5.6b), indicating its high data efficiency. Notably, the graph remains unsaturated, suggesting the potential for further improvement through the collection of more cheaply available text supervision in the target domain.

Table 5.5. Comparison of text-classifier backbones using text-classification accuracy on GeoNet and DomainNet datasets. BERT backbone outperforms other text-pretrained backbones and vision-language pre-trained CLIP-T.

Model	params (M)	GeoImnet	GeoPlaces	DomainNet
T5-small [174]	60.87	73.93	63.61	68.57
CLIP-T [172]	63.16	79.87	66.45	71.15
GPT-2 [173]	124	77.88	66.65	69.60
DistilBERT [56]	67.1	**83.53**	**69.31**	**71.43**

Effect of text classifier backbone.

We compare different choices of text classifiers such as DistilBERT [197], T5-Small [174], GPT2 [173] as well as text branch of CLIP [172] (CLIP-T) using text-classification accuracy on our datasets. We refer readers to the respective papers for details on their architectures and pre-training datasets. From Tab. 5.5, DistilBERT yields best text-classification accuracy on all our three benchmarks, outperforming text-only models like T5 and GPT2. Despite large-scale vision-language pre-training, CLIP-T did not yield substantial benefits.

Nearest neighbors using image and text features.

We show the top-2 nearest neighbor retrievals using text-features computed from source-trained text-classifier as opposed to image-features in Fig. 5.7. We observe more robust retrievals based on text-features corresponding to the captions of the images, rather than the images directly signifying the reduced domain gap in the text space. We also note a failure case in the last row of Fig. 5.7, where neither the text features nor the image features could retrieve the image from the correct class *church*.

Importance of source domain images.

While the majority of our accuracy gains stem from the text guidance, the source domain images providing noise-free supervision are also important in learning strong models on the target domain. We observed that joint training using source and pseudo-labeled target yields

Figure 5.7. Visualization of nearest neighbors of the leftmost source image, using text-trained and image-trained features, along with *ground truth* labels for each image from GeoNet. We observe better "same-class" retrievals using text-captions due to reduced domain gap, as opposed to images.

improvements of 1.57% for DomainNet and 0.8% on Ego2Exo benchmarks compared to target-only training. More importantly, training jointly on source and target allows deploying a single joint model across both domains as opposed to domain specific models, greatly optimizing inference costs.

5.5 Summary

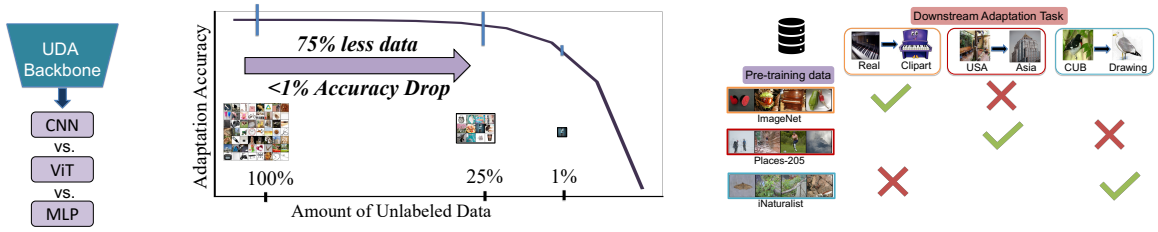
We introduce a novel framework called LaGTran to use readily available text supervision and enhance target performance in unsupervised domain transfer scenarios. We first start with the observation that traditional domain alignment approaches yield limited benefits beyond well-understood domain shifts, followed by insights that language provides a semantically richer medium of transfer with reduced domain gaps. This leads to a language-guided transfer mechanism where we train a text classifier on language descriptions from a source domain and then use its predictions on descriptions from a different target domain as supervision for the corresponding images. Despite being conceptually simple and straightforward, we show the remarkable ability of our method to outperform competitive prior approaches on challenging benchmarks like GeoNet and DomainNet for images and proposed Ego2Exo for videos. Through an emphasis on cost-effective or easily producible text supervision, we open new possibilities for advancing domain transfer in scenarios with limited manual supervision. Although LaGTran achieves state-of-the-art performance across all studied datasets, it relies on external vision-language models for textual guidance in the absence of metadata, potentially constraining its applicability in scenarios where the pre-trained VLM models fail to offer reliable and discriminative text supervision. Additionally, while exhibiting fewer domain discrepancies, there remain non-trivial gaps even within the text modality that may reduce the accuracy of pseudo-labels in the target domain, which can be potentially addressed by incorporating text-adaptation mechanisms into our framework[102].

This chapter is a reprint of the material as it appears in "Tell, Don't Show!: Language Guidance Eases Transfer Across Domains in Images and Videos" by Tarun Kalluri, Bodhisattwa Prasad Majumder, and Manmohan Chandraker, which was published in Proceedings of the International Conference on Machine Learning, 2024. The dissertation author was the primary investigator and author of this paper.

Chapter 6

Revisiting Common Assumptions in Unsupervised Domain Adaptation Using a Standardized Framework

In order to truly understand the utility of various domain adaptation algorithms, it is imperative to conduct fair training and standardized evaluation among all the current methods. In this chapter, we highlight our efforts towards this objective, where we 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, we conduct comprehensive empirical study into the impact of backbone architectures, unlabeled data quantity, and pre-training datasets revealing 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.



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

Figure 6.1. A summary of our contributions though UDABench. We examine the effectiveness of SOTA UDA approaches using our proposed framework UDA-Bench by revisiting the role of backbone architectures (Fig. 6.1a, Sec. 6.4.1), unlabeled data (Fig. 6.1b, Sec. 6.4.2) and pre-training data (Fig. 6.1c, Sec. 6.4.3) with several useful observations.

6.1 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) [97, 138, 142, 20, 21, 139, 69, 195, 196, 260, 98, 248, 104, 207, 110, 238, 107, 106, 15, 269] 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 [187, 166, 230, 24, 165]. 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 work, we address this through a large-scale empirical study of three major factors that potentially influence performance the most, namely,

1. **Choice of backbone architecture:** With recent advances in architecture designs such as vision transformers [61, 222, 136] and improved CNNs [137] 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 [114], 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. 6.1, and can be summarized as follows:

1. Recent advancements in vision transformers such as Swin [135] and DeiT [223] exhibit superior robustness against diverse domain shifts when compared to the conventional choice of ResNet-50 (see Tab. 6.1). 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. 6.3 and Sec. 6.4.1).

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. 6.4), suggesting that *that current UDA methods saturate quickly, and are not well-equipped to exploit the increasing availability of inexpensive unlabeled data* (see Sec. 6.4.2). This observation also contradicts the prevailing theory underpinning modern UDA research proposed in Ben-David et. al. [12], 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. 6.2).

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. 6.3). This trend holds across different types of pre-text tasks in self-supervised pre-training (see Sec. 6.4.3).

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.

We build our codebase using PyTorch following several open-source deep-learning libraries like Detectron [244] and PyTorch3D [177]. The overarching motivation in designing UDA-Bench is to standardize evaluation and training of existing unsupervised adaptation methods to facilitate fair comparative studies like ours, while also enabling quick prototyping and design

of new adaptation methods in the future. UDA-Bench is designed to be flexible to incorporate newer architecture backbones, classifier modules, optimizers, loss functions, dataloaders and training methods with minimal effort and design overhead, allowing researchers to build upon existing adaptation methods to develop new innovations in unsupervised adaptation.

Our framework is publicly available to continue improving our understanding of UDA methods.

6.2 Relation to Prior Literature

While the primary focus of most works in unsupervised adaptation is on algorithmic innovations to improve adaptation, our emphasis throughout this chapter 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.

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 [159], SLAM [156], metric learning [152, 185], transfer learning [146], domain generalization [84], optimization algorithms [41], few-shot learning [33], contrastive learning [46], GANs [143], fairness [77] and self-supervised learning [78, 157, 75]. Prior works also established standardized benchmarks to facilitate fair comparisons and quick prototyping [152, 217, 156]. 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 [114], which carries UDA study but without a unified training framework, [154, 153], which study UDA methods through fair validation methods and [112] 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.

6.3 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 [12] between the domains, which arises when $P_s \neq P_t$, although other forms of shift have also been studied in literature [216, 7, 70, 2]. 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, we instead aim 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.

6.3.1 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. 6.2. 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

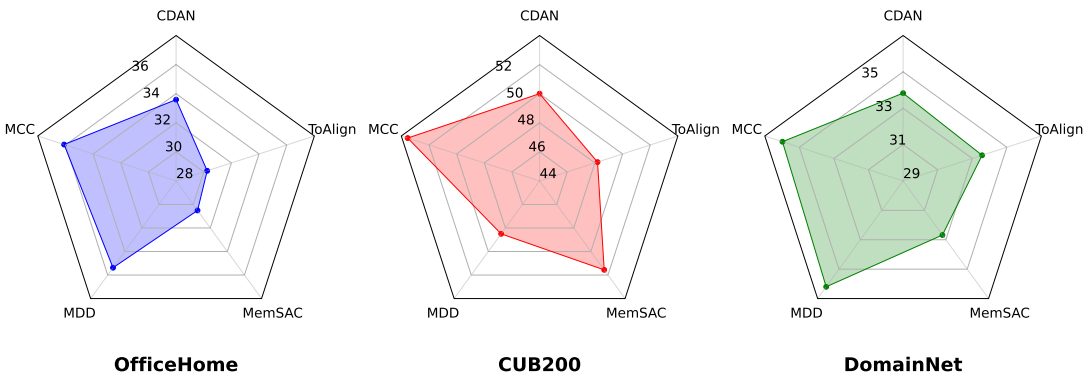


Figure 6.2. Need for UDA-Bench. We illustrate the disparity between various codebases proposed for prior UDA methods by highlighting the different accuracy numbers obtained for a plain source only model. Computed without any adaptation, it should ideally match across implementations which is clearly 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.

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, MDD [260] uses a deeper MLP as a classifier, MCC [104] uses batchnorm layers in the bottleneck layer, CDAN [139] uses 10-crop evaluation and AdaMatch [15] uses stronger augmentation on source data.

To alleviate this issue, we create a new framework in PyTorch [163] 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 the current chapter 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.

6.3.2 Axes of Variation

We choose backbone architecture (Sec. 6.4.1), amount of unlabeled data in the target (Sec. 6.4.2) and the nature of data/algorithm used in pre-training the backbone (Sec. 6.4.3) 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.

6.3.3 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 [69], CDAN [139]), *non-adversarial* (MDD [260], MCC [104], DALN [30]), *consistency-based* (MemSAC [107], AdaMatch [15]), *alignment-based* (ToAlign [238]) and *pseudo-label based* [269] methods.

6.3.4 Adaptation Datasets

Following popular choices in UDA literature, we use visDA [166], OfficeHome [230], DomainNet [165] and CUB200 [233] 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.

6.3.5 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.

6.3.6 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. 6.4.2) 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 $\sim$14000 GPU hours. Likewise, the experiments in Sec. 6.4.1 cost 18640 GPU hours and Sec. 6.4.3 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.

6.4 Methodology and Evaluation

6.4.1 Which Backbone Architectures Suit UDA Best?

Motivation

Although ResNet-50 [90] backbone is a widely adopted standard in domain adaptation research [139, 196, 194, 15, 107, 238], several recent architectures [61, 222, 137] have emerged

Table 6.1. 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

as feasible alternatives with better performance. While a more recent method PMTrans [269] 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 [269] compare to classical methods while using a same backbone. While robustness properties of vision transformers to adversarial and out-of-context examples have been widely studied [9, 17, 206, 261, 268], 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 [222], Swin [136], ResMLP [221], and ConvNext [137]. We use newer versions of DeiT (DeiT-III [222]) and Swin (Swin-V2 [136]) 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 [241].

Across the architectures, we uniformly use a batch size of 32, SGD optimizer with an initial learning rate of 0.003 and cosine decay. For data augmentation, we first resize the images so that the shorter size is 256 and then choose a random 224×224 crop followed by random horizontal flip. However, we use a crop size of 256 instead of 224 for Swin transformer due to its input size. We train the networks for a total of 75k iterations on DomainNet and CUB200 with validation performed at every 5k steps, and for 30k iterations on the smaller OfficeHome dataset with validation at every 500 steps. We use early stopping on the test set to choose the best accuracy.

For the classifier, we use a 2-layer MLP with a hidden dimension of 256. The input dimension for the MLP, though, varies depending on the output dimension of the backbone architecture used. For Resnet-50, it is 2048, for Swin-t and ConvNext-t it is 768 and for Deit-s and ResMLP-s it is 384.

Newer architectures show better domain transfer

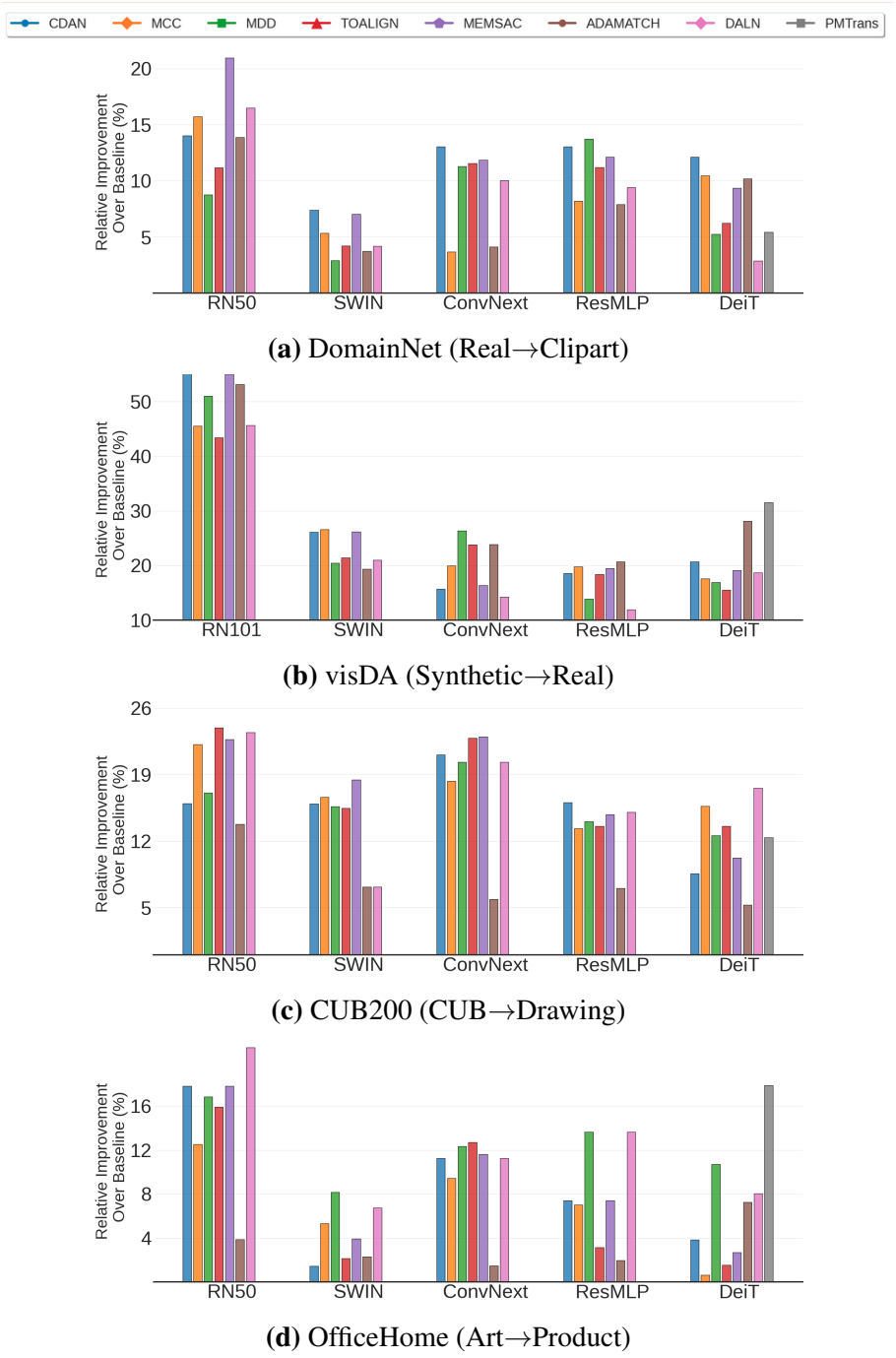
For a model trained only on source-domain data (no adaptation), we use the accuracy on the source test-set (λ_s) and the accuracy on the target test-set (λ_t), to define relative cross-domain accuracy drop $\sigma_{st} = \frac{\lambda_s - \lambda_t}{\lambda_s} * 100$. While this metric is sensitive to 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. 6.1, 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$\rightarrow$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%) [107], 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 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. 6.3, 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. 6.3a, 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. 6.3d. These observation also holds for visDA Fig. 6.3b and CUB200 Fig. 6.3c 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. 6.3a with Swin, ConvNext and DeiT, while MCC outperforms other methods with a ResMLP backbone.

While prior works like [114] only show this trend for classical UDA methods [139, 196] without using a standardized framework, we additionally show that this issue extends to more recent state-of-the-art UDA algorithms [238, 30, 107] as well, including methods using vision



(a) DomainNet (Real→Clipart)

(b) visDA (Synthetic→Real)

(c) CUB200 (CUB→Drawing)

(d) OfficeHome (Art→Product)

Figure 6.3. 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.

transformer backbones [269] using the proposed UDA-Bench, yielding several novel observations. For instance, we show in Fig. 6.3 that the current SOTA method PMTrans [269] 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.

6.4.2 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

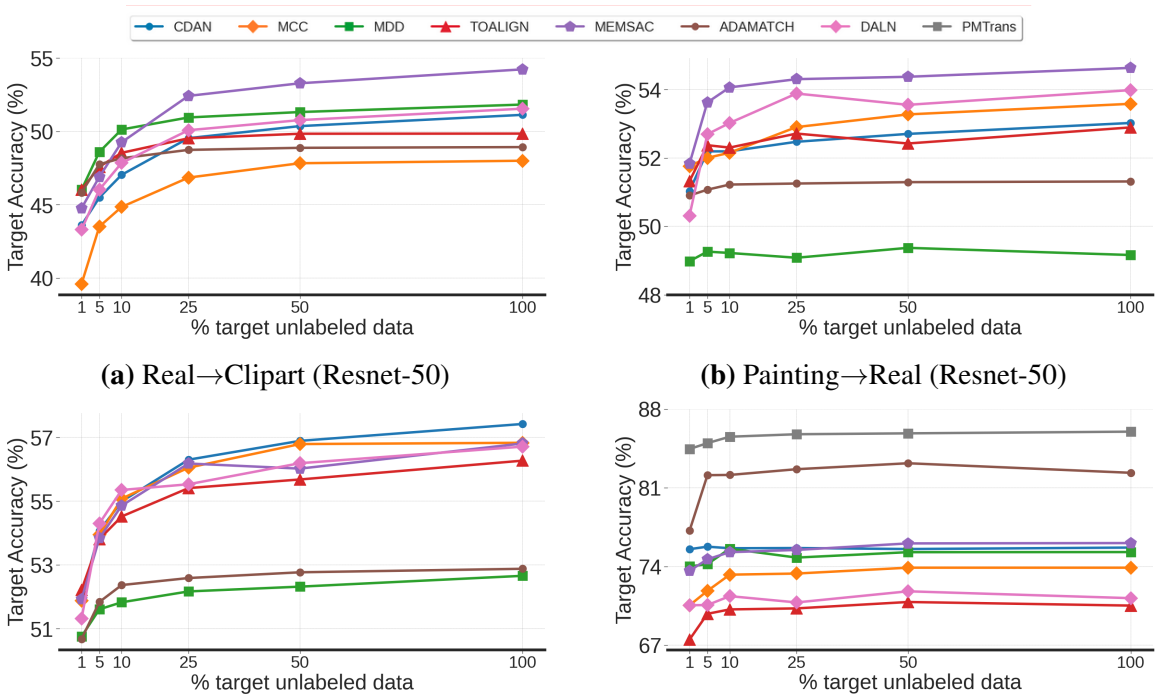
To study the effects of data volume, we sample $\{1, 5, 10, 25, 50, 100\}\%$ of the data from the target domain and run the adaptation algorithm using each of these subsets as the unlabeled data. We repeat the experiment with three different seeds in each case and report the mean accuracy to eliminate 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. 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. 6.4 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 [139], non-adversarial [15], consistency based [107] and pseudo-label based [269] 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. 6.4a), 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. 6.4b) 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 [269] (Fig. 6.4d), where the performance saturates after using only 10% of the unlabeled 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.

Furthermore, we juxtapose this observation with a similar ablation using source labeled data, and identify that source supervision has a more pronounced effect on the target accuracy than target unlabeled data. Specifically, we use $\{1,5,10,25,50,100\}\%$ of source labels and train the UDA methods on each subset. We run three random seeds and plot the mean accuracy in Fig. 6.5. We observe that the scaling trends of target accuracy with respect to source labeled data are much more favorable towards improving performance. For example, doubling the number of source labels from 50% to 100% improves target accuracy by $\sim 9\%$ on average across UDA methods. In contrast, the improvement in doubling the target unlabeled data from 50% to 100% is less than 0.5% on average. This confirms the fact that labels have a more pronounced impact on target accuracy even when they arise from a different domain, compared to unlabeled data from the same domain.



(a) Real→Clipart (Resnet-50)

(b) Painting→Real (Resnet-50)

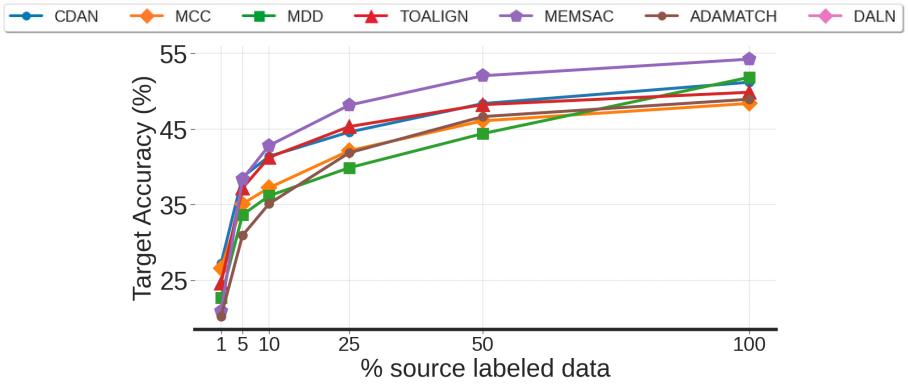
(c) Real→Clipart (ConvNext-t)

(d) VisDA (DeiT)

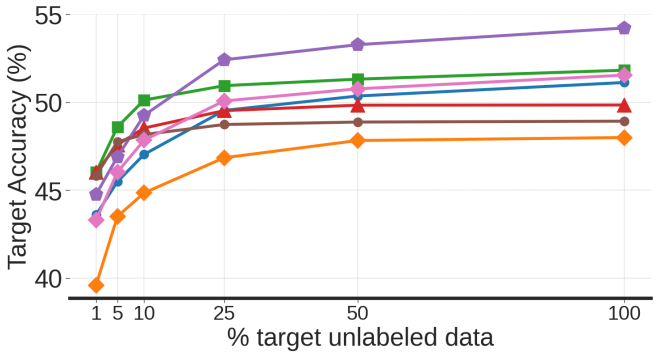
Figure 6.4. How much unlabeled data can UDA methods use? Across different adaptation datasets and backbones (Resnet50 in a, b, ConvNext in c and DeiT in d), we find that the performance of several UDA methods saturates quickly with respect to amount of target data, showing their limited efficiency in utilizing the unlabeled samples. In most cases, using only 25% of the data results in $< 1\%$ drop in accuracy.

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 [69, 139], and examine its data efficiency. We plot the accuracy of the domain discrimination objective itself against the quantity of unlabeled samples in Fig. 6.6a 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



(a) Source Labeled Data



(b) Target Unlabeled Data

Figure 6.5. Source labels vs. Target unsupervised data We show that collecting more labels from source dataset, even when it is from a different domain, has a more profound influence on the target accuracy (a) compared to collecting more unlabeled data from the target domain using current UDA methods (b). Results shown on Real→Clipart setting from DomainNet dataset.

that similar limitations impact other types of adaptation approaches including self-training, pseudo-label or consistency-based methods.

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.[12], which underpins several UDA methods. Their theoretical analysis suggests an inverse relationship between target sample size and target error (Theorem 2 from [12]), 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 [213], weakly-supervised [210] and self-supervised learning [78] literature, where increasing labeled or unlabeled data significantly enhances performance.

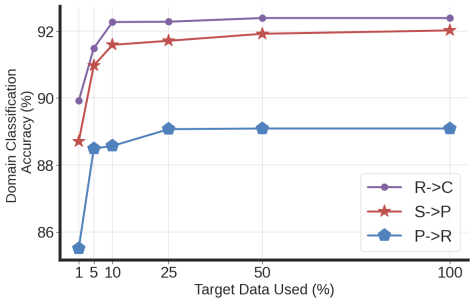
Similar results hold for other sampling techniques

In addition to the class-balanced sampling procedure in Fig. 6.4, we also show results using two other sampling techniques, random sampling and split-class sampling in Fig. 6.6b and Fig. 6.6c respectively. In Fig. 6.6b, 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. 6.6c, 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.

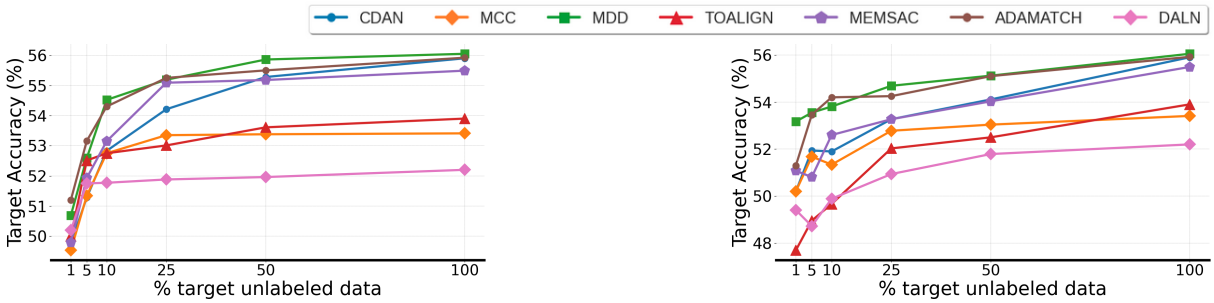
6.4.3 Does Pre-training Data Matter in UDA?

Motivation

Following recent works that reveal the importance of pre-training data in influencing downstream accuracy [46], we revisit a standard practice in UDA to adopt ImageNet pre-trained backbone irrespective of the downstream adaptation task. While Kim et al. [114] 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



(a) Accuracy Saturation



(b) Random Sampling

(c) Class-split Sampling

Figure 6.6. (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.

type of pre-training and downstream adaptation maintaining a *constant* datasize.

Experimental setup

We use ImageNet [186], Places-205 [266] and iNaturalist-2021 [228] 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-

Table 6.2. 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 [107], ToAlign [238], MDD [260] and DALN [30]. 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 [238]			MemSAC [107]			MDD [260]			DALN [30]		
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**

of-the-art self-supervised methods SwAV [25], MoCo-V3 [37] and MAE [88], 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. 6.2 for plain source-only transfer (no adaptation), as well as adaptation using ToAlign, MemSAC, MDD and DALN. 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. 6.2, the de-facto choice of ImageNet pre-training gives 50.2% on CUB→Drawing transfer task, while just switching the

Table 6.3. 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) [25]			MoCo-V3 (ViT-s/16) [37]			MAE (ViT-b/16) [88]		
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) [25]			MoCo-V3 (ViT-s/16) [37]			MAE (ViT-b/16) [88]		
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

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 [180] or language models [85], our difference lies in highlighting this behavior for the specific case of UDA through a unified framework and controlled empirical study.

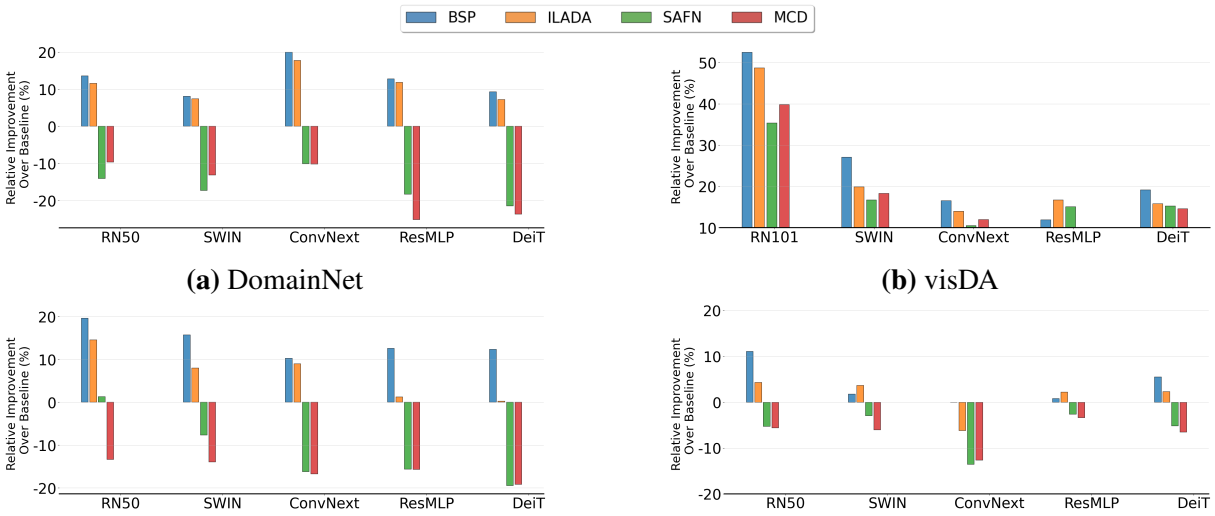
In-task pre-training is complementary to UDA method

We also observe that these benefits obtained from in-task supervised pre-training complement the advantages potentially obtained using UDA methods, resulting in additional improve-

ments in accuracy. From Tab. 6.2, 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. 6.3. We first note that supervised pre-training (Tab. 6.2) 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. 6.3a) as well as adaptation using MemSAC (Tab. 6.3b). 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



(a) DomainNet **(b)** visDA

(c) CUB200 **(d)** OfficeHome

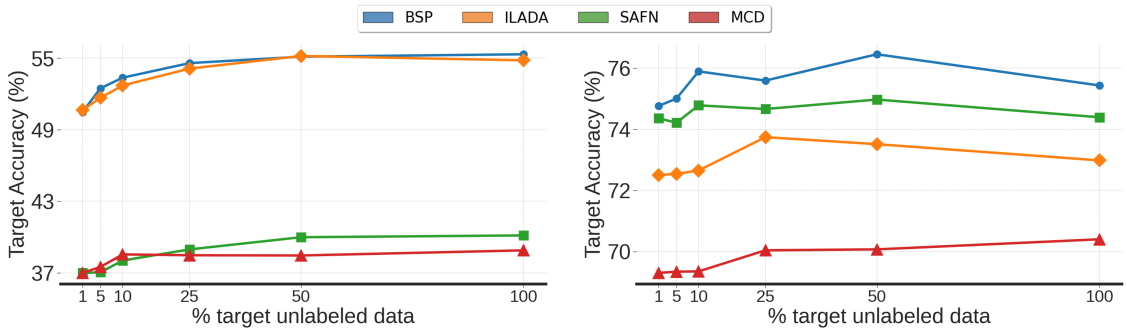
Figure 6.7. Newer backbones give limited returns or perform worse than baseline. For each of the UDA methods, we show the gain in accuracy relative to a baseline trained only using source-data. For methods like SAFN [248] and MCD [196], we observe that the relative improvement over a source-only baseline is negative in most cases. Further, the gains observed by other methods like BSP [38] and ILADA [207] are not same across architectures.

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 [46]. Furthermore, this property is consistent across different kinds of self-supervised pretext tasks like SwAV, MoCo and MAE.

6.5 Additional Results on Other UDA Methods

In addition to the wide variety of UDA methods already studied, we reinforce our observations using results from four additional adaptation methods: BSP [38], ILADA [207], SAFN [248] and MCD [196]. The observations for the effect of backbone architecture is presented in Fig. 6.7 and the study for the effect of unlabeled target domain data is presented in Fig. 6.8.

Resonating with the observations made in Sec. 6.4.1, we show in Fig. 6.7 that the gains

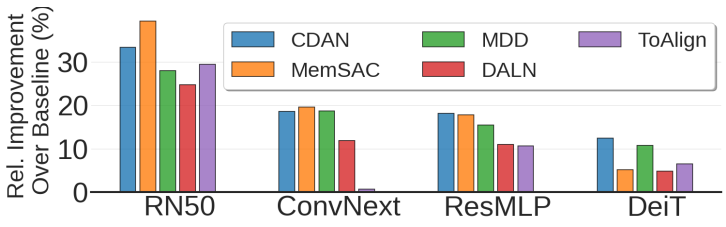


(a) DomainNet **(b)** VisDA

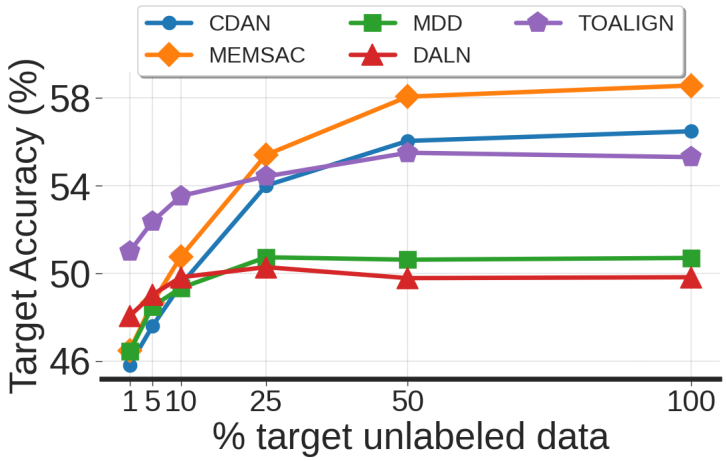
Figure 6.8. Unlabeled Data-efficiency of UDA algorithms Across both DomainNet and visDA datasets, the performance of UDA methods exhibits diminishing returns with increasing amounts of unlabeled data. In most cases, utilizing only 25% of the available unlabeled data results in a performance drop of less than 1%, suggesting that collecting additional unlabeled data is unlikely to yield significant improvements for these methods.

obtained by UDA method are not independent of the backbone. For instance, on CUB200 dataset, BSP [38] and ILADA [207] gives 20% and 15% relative gain respectively, but using DeiT diminishes these gains to 12% and 3% respectively. Similarly, on visDA, the improvements using ResNet is much higher than improvements offered on other backbones like ConvNext and DeiT. Moreover, as demonstrated in previous research [107], other unsupervised domain adaptation (UDA) algorithms, such as SAFN [248] and MCD [196], under-perform compared to a source-only baseline, and the disparity worsens when employing these algorithms with newer architectures.

Similarly, from Fig. 6.8, the performance of the additional adaptation methods studied also plateaus quickly, reaching near saturation after utilizing only 20% of the available unlabeled data. Further addition of unlabeled data yields negligible performance gains. This suggests that collecting additional unlabeled data is unlikely to yield significant improvements for these methods, corroborating the observations noted in Sec. 6.4.2 for several other UDA methods.



(a) Varying backbones



(b) Varying target unlabeled data

Figure 6.9. Results on TinyImageNet vs. TinyImageNet-C We show the similar observations regarding backbone architectures and data volume hold also for a non-standard adaptation dataset. We use images from TinyImageNet as the source and *snow-3* perturbations from TinyImageNet-C as the target.

6.6 Additional Results using TinyImageNet

To further examine the presented trends on non-standard adaptation datasets, we show results using images from the TinyImageNet dataset as the source domain and *snow* perturbations from TinyImageNet-C [94] as the target domain. We train models using the 200 classes in each dataset, and use report accuracy on the target domain. In Fig. 6.9, we show that the broad trends observed for other adaptation datasets also hold for this novel setting. Specifically, from a, adaptation gains are much lesser with recent architectures (like *ConvNext* and *DeiT*) and from b, performance saturates in-spite of adding unlabeled data, further corroborating the main inferences from our study.

6.7 Summary

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 [126], semi-supervised [190] and universal [254] DA methods. Several other avenues like adaptation of vision-language models [172, 255] and emerging generative models [183, 132] are also left to a future work.

This chapter is a reprint of the material as it appears in "UDA-Bench: Revisiting Common Assumptions in Unsupervised Domain Adaptation Using a Standardized Framework" by Tarun Kalluri, Sreyas Ravichandran, and Manmohan Chandraker, which was published in Proceedings of the European Conference on Computer Vision, 2024. The dissertation author was the primary investigator and author of this paper.

Chapter 7

Conclusion and Future Work

In this dissertation, we addressed several challenges in training and deploying robust computer vision models and introduced innovations to tackle these issues effectively. However, with the growing trend of using vast amounts of web-scale data in open-ended frontier model training [18], there is a pressing need to reconsider key formulations of domain adaptation to suit the evolving landscape of computer vision. In this chapter, we explore several potential extensions of the ideas presented in this dissertation, offering a range of avenues for future research and practical application. These suggestions aim to build on the foundations laid out in the previous chapters, providing researchers and practitioners with new directions to further advance the field of domain adaptation for fair and robust computer vision.

While visual categorization is the most fundamental task in computer vision, there is an increasing emphasis on moving towards more structured prediction tasks like semantic segmentation [203], instance segmentation [115], image captioning [122] and open-world question answering [231]. Therefore, it is important to understand the key aspects of geographical domain robustness for these tasks and derive effective solutions that suit large-model training. Specifically, a potential direction exists in extending ideas in language-guided adaptation for tasks like segmentation in autonomous driving scenarios where geographic transferability is a fundamental necessity, but pixel-level annotation are costly and cumbersome to gather. Likewise, evaluating geographic diversity of various personal assistants deployed in mobile phones for solving

tasks like open-ended dialogue, knowledge-based question answering or image generation is an important direction to be studied.

A major focus in this dissertation is on closed-world prediction tasks where the categories are known beforehand, but there have also been several parallel efforts to extend these ideas to more open-world scenarios where new categories might be encountered at test-time during deployment [254, 100, 253]. Therefore, investigating the solutions for geographic transferability amidst such open world setting yields models which can be deployed with better robustness guarantees. Furthermore, an increasing trend in machine learning of late is to download a large-scale pre-trained model trained on noisy web-scale data provided through an api [160] or open-weight access [65] and then fine-tune the model on custom data for various downstream tasks. In this setting, the access to the original source data is restricted or not available at all, so a potential future work is to extend the ideas of domain adaptation presented in this dissertation to suit the setting with no access to the source data or source weights, yet attain robustness to downstream data through efficient fine-tuning techniques.

A significant challenge also lies in addressing constantly evolving domains during a model deployment in the wild. This calls for innovations in the areas of continual adaptation of frontier models where an efficient feedback loop should be designed that can automatically identify potential biases in an open-world and continuously adapt to the changing needs of robustness with minimal human annotation efforts. Certifying guarantees to problems in open-set biases [59] is rarely studied in the literature, and it would be an important and exciting research direction to extend the ideas of domain adaptation to such a challenging scenario.

Bibliography

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Ibrahim Alabdulmohsin, Nicole Chiou, Alexander D'Amour, Arthur Gretton, Sanmi Koyejo, Matt J Kusner, Stephen R Pfohl, Olawale Salaudeen, Jessica Schrouff, and Katherine Tsai. Adapting to latent subgroup shifts via concepts and proxies. In *International Conference on Artificial Intelligence and Statistics*, pages 9637–9661. PMLR, 2023.

[3] Michael A Alcorn, Qi Li, Zhitao Gong, Chengfei Wang, Long Mai, Wei-Shinn Ku, and Anh Nguyen. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4845–4854, 2019.

[4] Amr Alexandari, Anshul Kundaje, and Avanti Shrikumar. Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation. In *International Conference on Machine Learning*, pages 222–232. PMLR, 2020.

[5] Anders Andreassen, Yasaman Bahri, Behnam Neyshabur, and Rebecca Roelofs. The evolution of out-of-distribution robustness throughout fine-tuning. *arXiv preprint arXiv:2106.15831*, 2021.

[6] Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019.

[7] Kamyar Azizzadenesheli. Importance weight estimation and generalization in domain adaptation under label shift. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(10):6578–6584, 2022.

[8] Kamyar Azizzadenesheli, Anqi Liu, Fanny Yang, and Animashree Anandkumar. Regularized learning for domain adaptation under label shifts. *arXiv preprint arXiv:1903.09734*, 2019.

[9] Yutong Bai, Jieru Mei, Alan L Yuille, and Cihang Xie. Are transformers more robust than cnns? *Advances in Neural Information Processing Systems*, 34:26831–26843, 2021.

[10] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021.

[11] Mahsa Baktashmotlagh, Mehrtash Harandi, and Mathieu Salzmann. Distribution-matching embedding for visual domain adaptation. *Journal of Machine Learning Research*, 17:Article–number, 2016.

[12] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79:151–175, 2010.

[13] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19:137–144, 2006.

[14] Ryan Y Benmalek, Sabhya Chhabria, Pedro O Pinheiro, Claire Cardie, and Serge Belongie. Learning to adapt to semantic shift. 2021.

[15] David Berthelot, Rebecca Roelofs, Kihyuk Sohn, Nicholas Carlini, and Alex Kurakin. Adamatch: A unified approach to semi-supervised learning and domain adaptation. *arXiv preprint arXiv:2106.04732*, 2021.

[16] Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020.

[17] Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10231–10241, 2021.

[18] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi

Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models. *arXiv preprint arXiv: 2108.07258*, 2021.

[19] Konstantinos Bousmalis, N. Silberman, David Dohan, D. Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. *Computer Vision and Pattern Recognition*, 2016.

[20] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3722–3731, 2017.

[21] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *Advances in neural information processing systems*, pages 343–351, 2016.

[22] Steve Branson, Grant Van Horn, Serge Belongie, and Pietro Perona. Bird species categorization using pose normalized deep convolutional nets. *arXiv preprint arXiv:1406.2952*, 2014.

[23] Tianle Cai, Ruiqi Gao, Jason Lee, and Qi Lei. A theory of label propagation for subpopulation shift. In *International Conference on Machine Learning*, pages 1170–1182. PMLR, 2021.

[24] Barbara Caputo, Henning Müller, Jesus Martinez-Gomez, Mauricio Villegas, Burak Acar, Novi Patricia, Neda Marvasti, Suzan Üsküdarlı, Roberto Paredes, Miguel Cazorla, 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*, pages 192–211. Springer, 2014.

[25] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

[26] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[27] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3558–3568, 2021.

[28] Chaoqi Chen, Weiping Xie, Wenbing Huang, Yu Rong, Xinghao Ding, Yue Huang, Tingyang Xu, and Junzhou Huang. Progressive feature alignment for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 627–636, 2019.

[29] Gongwei Chen, Xinhang Song, Bohan Wang, and Shuqiang Jiang. See more for scene: Pairwise consistency learning for scene classification. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 4987–4999. Curran Associates, Inc., 2021.

[30] Lin Chen, Huaian Chen, Zhixiang Wei, Xin Jin, Xiao Tan, Yi Jin, and Enhong Chen. 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)*, pages 7181–7190, June 2022.

[31] Min-Hung Chen, Zsolt Kira, Ghassan AlRegib, Jaekwon Yoo, Ruxin Chen, and Jian Zheng. Temporal attentive alignment for large-scale video domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6321–6330, 2019.

[32] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[33] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. *arXiv preprint arXiv:1904.04232*, 2019.

[34] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022.

[35] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[36] Xinlei Chen and Abhinav Gupta. Webly supervised learning of convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1431–1439, 2015.

[37] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021.

[38] Xinyang Chen, Sinan Wang, Mingsheng Long, and Jianmin Wang. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In *International conference on machine learning*, pages 1081–1090. PMLR, 2019.

[39] Yi-Hsin Chen, Wei-Yu Chen, Yu-Ting Chen, Bo-Cheng Tsai, Yu-Chiang Frank Wang, and Min Sun. No more discrimination: Cross city adaptation of road scene segmenters. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1992–2001, 2017.

[40] Yue Chen, Yalong Bai, Wei Zhang, and Tao Mei. Destruction and construction learning for fine-grained image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5157–5166, 2019.

[41] Dami Choi, Christopher J Shallue, Zachary Nado, Jaehoon Lee, Chris J Maddison, and George E Dahl. On empirical comparisons of optimizers for deep learning. *arXiv preprint arXiv:1910.05446*, 2019.

[42] Jinwoo Choi, Chen Gao, Joseph CE Messou, and Jia-Bin Huang. Why can't i dance in the mall? learning to mitigate scene bias in action recognition. *Advances in Neural Information Processing Systems*, 32, 2019.

[43] Jinwoo Choi, Gaurav Sharma, Manmohan Chandraker, and Jia-Bin Huang. Unsupervised and semi-supervised domain adaptation for action recognition from drones. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1717–1726, 2020.

[44] Myung Jin Choi, Antonio Torralba, and Alan S Willsky. Context models and out-of-context objects. *Pattern Recognition Letters*, 33(7):853–862, 2012.

[45] Elijah Cole, Xuan Yang, Kimberly Wilber, Oisin Mac Aodha, and Serge J. Belongie. When does contrastive visual representation learning work? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 1–10. IEEE, 2022.

[46] Elijah Cole, Xuan Yang, Kimberly Wilber, Oisin Mac Aodha, and Serge Belongie. When does contrastive visual representation learning work? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14755–14764, 2022.

[47] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865, 2017.

[48] Shuhao Cui, Xuan Jin, Shuhui Wang, Yuan He, and Qingming Huang. Heuristic domain adaptation. In *Advances in Neural Information Processing Systems*, volume 33, pages 7571–7583. Curran Associates, Inc., 2020.

[49] Yin Cui, Yang Song, Chen Sun, Andrew Howard, and Serge Belongie. Large scale fine-grained categorization and domain-specific transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4109–4118, 2018.

[50] Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia, and Nicolas Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 447–463, 2018.

[51] 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*, pages 248–255. Ieee, 2009.

[52] Zhijie Deng, Yucen Luo, and Jun Zhu. Cluster alignment with a teacher for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9944–9953, 2019.

[53] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11162–11173, 2021.

[54] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. Redcaps: Web-curated image-text data created by the people, for the people. *arXiv preprint arXiv:2111.11431*, 2021.

[55] Benjamin Devillers, Bhavin Choksi, Romain Bielawski, and Rufin VanRullen. Does language help generalization in vision models? *arXiv preprint arXiv:2104.08313*, 2021.

[56] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *North American Chapter of the Association for Computational Linguistics*, 2019.

[57] Terrance DeVries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. Does object recognition work for everyone? In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 52–59. Computer Vision Foundation / IEEE, 2019.

[58] Terrance DeVries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. Does object recognition work for everyone? *CoRR*, abs/1906.02659, 2019.

[59] Moreno D'Incà, Elia Peruzzo, Massimiliano Mancini, Dejia Xu, Vidit Goel, Xingqian Xu, Zhangyang Wang, Humphrey Shi, and Nicu Sebe. Openbias: Open-set bias detection in text-to-image generative models. *CVPR*, 2024.

[60] Zheng Ding, Jieke Wang, and Z. Tu. Open-vocabulary universal image segmentation with maskclip. *International Conference on Machine Learning*, 2022.

[61] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[62] Zhekai Du, Jingjing Li, Hongzu Su, Lei Zhu, and Ke Lu. Cross-domain gradient discrepancy minimization for unsupervised domain adaptation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021.

[63] Zhekai Du, Jingjing Li, Hongzu Su, Lei Zhu, and Ke Lu. Cross-domain gradient discrepancy minimization for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3937–3946, 2021.

[64] Abhimanyu Dubey, Otkrist Gupta, Ramesh Raskar, and Nikhil Naik. Maximum-entropy fine-grained classification. *arXiv preprint arXiv:1809.05934*, 2018.

[65] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman,

Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre

Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The llama 3 herd of models. *arXiv preprint arXiv: 2407.21783*, 2024.

[66] Abhimanyu Dubey, Vignesh Ramanathan, Alex Pentland, and Dhruv Mahajan. Adaptive methods for real-world domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14340–14349, 2021.

[67] Lisa Dunlap, Clara Mohri, Devin Guillory, Han Zhang, Trevor Darrell, Joseph E. Gonzalez, Aditi Raghunathan, and Anna Rohrbach. Using language to extend to unseen domains. In *The Eleventh International Conference on Learning Representations*, 2023.

[68] Mohamed El Banani, Karan Desai, and Justin Johnson. Learning visual representations via language-guided sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19208–19220, 2023.

[69] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

[70] Saurabh Garg, Yifan Wu, Sivaraman Balakrishnan, and Zachary Lipton. A unified view of label shift estimation. *Advances in Neural Information Processing Systems*, 33:3290–3300, 2020.

[71] Timnit Gebru, Judy Hoffman, and Li Fei-Fei. Fine-grained recognition in the wild: A multi-task domain adaptation approach. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1349–1358, 2017.

[72] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

[73] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXVI*, volume 13696 of *Lecture Notes in Computer Science*, pages 540–557. Springer, 2022.

[74] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16102–16112, 2022.

[75] Micah Goldblum, Hossein Souri, Renkun Ni, Manli Shu, Viraj Prabhu, Gowthami Somepalli, Prithvijit Chattopadhyay, Mark Ibrahim, Adrien Bardes, Judy Hoffman, Rama Chellappa, Andrew Gordon Wilson, and Tom Goldstein. Battle of the backbones: A large-scale comparison of pretrained models across computer vision tasks. *arXiv preprint arXiv: 2310.19909*, 2023.

[76] Daniel Gordon, Kiana Ehsani, Dieter Fox, and Ali Farhadi. Watching the world go by: Representation learning from unlabeled videos. *arXiv preprint arXiv:2003.07990*, 2020.

[77] Priya Goyal, Quentin Duval, Isaac Seessel, Mathilde Caron, Ishan Misra, Levent Sagun, Armand Joulin, and Piotr Bojanowski. Vision models are more robust and fair when pretrained on uncurated images without supervision. *arXiv preprint arXiv:2202.08360*, 2022.

[78] Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and benchmarking self-supervised visual representation learning. In *Proceedings of the ieee/cvf International Conference on computer vision*, pages 6391–6400, 2019.

[79] Sachin Goyal, Ananya Kumar, Sankalp Garg, Zico Kolter, and Aditi Raghunathan. Finetune like you pretrain: Improved finetuning of zero-shot vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19338–19347, 2023.

[80] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022.

[81] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. *arXiv preprint arXiv:2311.18259*, 2023.

[82] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

[83] Xiang Gu, Jian Sun, and Zongben Xu. Spherical space domain adaptation with robust pseudo-label loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9101–9110, 2020.

[84] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

[85] Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8342–8360. Association for Computational Linguistics, 2020.

[86] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304, 2010.

[87] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.

[88] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[89] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[90] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. arxiv 2015. *arXiv preprint arXiv:1512.03385*, 14, 2015.

[91] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[92] Olivier J Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, SM Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding. *arXiv preprint arXiv:1905.09272*, 2019.

[93] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021.

[94] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv: 1903.12261*, 2019.

[95] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

[96] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv: 1503.02531*, 2015.

[97] Judy Hoffman, Erik Rodner, Jeff Donahue, Trevor Darrell, and Kate Saenko. Efficient learning of domain-invariant image representations. *arXiv preprint arXiv:1301.3224*, 2013.

[98] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. Pmlr, 2018.

[99] Tzu Ming Harry Hsu, Wei Yu Chen, Cheng-An Hou, Yao-Hung Hubert Tsai, Yi-Ren Yeh, and Yu-Chiang Frank Wang. Unsupervised domain adaptation with imbalanced cross-domain data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4121–4129, 2015.

[100] Jiaxing Huang, Jingyi Zhang, Han Qiu, Sheng Jin, and Shijian Lu. Prompt ensemble self-training for open-vocabulary domain adaptation. *arXiv preprint arXiv: 2306.16658*, 2023.

[101] Zeyi Huang, Andy Zhou, Zijian Ling, Mu Cai, Haohan Wang, and Yong Jae Lee. A sentence speaks a thousand images: Domain generalization through distilling clip with language guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11685–11695, 2023.

[102] Chia-Chien Hung, Lukas Lange, and Jannik Strotgen. Tada: Efficient task-agnostic domain adaptation for transformers. *Annual Meeting of the Association for Computational Linguistics*, 2023.

[103] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[104] Ying Jin, Ximei Wang, Mingsheng Long, and Jianmin Wang. Minimum class confusion for versatile domain adaptation. In *European Conference on Computer Vision*, pages 464–480. Springer, 2020.

[105] Tarun Kalluri and Manmohan Chandraker. 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*, pages 4121–4131, 2022.

[106] Tarun Kalluri, Bodhisattwa Prasad Majumder, and Manmohan Chandraker. Tell, don't show!: Language guidance eases transfer across domains in images and videos. *arXiv preprint arXiv:2403.05535*, 2024.

[107] Tarun Kalluri, Astuti Sharma, and Manmohan Chandraker. 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*, pages 550–568. Springer, 2022.

[108] Tarun Kalluri, Girish Varma, Manmohan Chandraker, and CV Jawahar. Universal semi-supervised semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5259–5270, 2019.

[109] Tarun Kalluri, Wangdong Xu, and Manmohan Chandraker. Geonet: Benchmarking unsupervised adaptation across geographies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15368–15379, June 2023.

[110] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4893–4902, 2019.

[111] Guoliang Kang, Liang Zheng, Yan Yan, and Yi Yang. Deep adversarial attention alignment for unsupervised domain adaptation: the benefit of target expectation maximization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 401–416, 2018.

[112] Simar Kareer, Vivek Vijaykumar, Harsh Maheshwari, Prithvijit Chattopadhyay, Judy Hoffman, and Viraj Prabhu. We're not using videos effectively: An updated domain adaptive video segmentation baseline. *arXiv preprint arXiv: 2402.00868*, 2024.

[113] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020.

[114] Donghyun Kim, Kaihong Wang, Stan Sclaroff, and Kate Saenko. 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*, pages 621–638. Springer, 2022.

[115] A. Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, A. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. *IEEE International Conference on Computer Vision*, 2023.

[116] Abhishek Kumar, Prasanna Sattigeri, Kahini Wadhawan, Leonid Karlinsky, Rogerio Feris, Bill Freeman, and Gregory Wornell. Co-regularized alignment for unsupervised domain adaptation. In *Advances in Neural Information Processing Systems*, pages 9345–9356, 2018.

[117] Ananya Kumar, Tengyu Ma, and Percy Liang. Understanding self-training for gradual domain adaptation. In *International Conference on Machine Learning*, pages 5468–5479. PMLR, 2020.

[118] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*, 2022.

[119] Jogendra Nath Kundu, Suvaansh Bhambri, Akshay Kulkarni, Hiran Sarkar, Varun Jampani, and R Venkatesh Babu. Subsidiary prototype alignment for universal domain adaptation. *arXiv preprint arXiv:2210.15909*, 2022.

[120] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint arXiv: 1811.00982*, 2018.

[121] Boyi Li, Kilian Q. Weinberger, Serge J. Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[122] Junnan Li, Dongxu Li, S. Savarese, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *International Conference on Machine Learning*, 2023.

[123] Shuang Li, Mixue Xie, Fangrui Lv, Chi Harold Liu, Jian Liang, Chen Qin, and Wei Li. Semantic concentration for domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9102–9111, 2021.

[124] Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862*, 2017.

[125] Yanghao Li, Tushar Nagarajan, Bo Xiong, and Kristen Grauman. Ego-exo: Transferring visual representations from third-person to first-person videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6943–6953, 2021.

[126] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International conference on machine learning*, pages 6028–6039. PMLR, 2020.

[127] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z XU, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *Advances in Neural Information Processing Systems*, 35:7575–7586, 2022.

[128] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context. *arXiv preprint arXiv: 1405.0312*, 2014.

[129] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 1449–1457, 2015.

[130] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, pages 3122–3130. PMLR, 2018.

[131] Geng Liu and Yuxi Wang. Tdg: Text-guided domain generalization. *arXiv preprint arXiv:2308.09931*, 2023.

[132] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.

[133] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NEURIPS*, 2023.

[134] Hong Liu, Jianmin Wang, and Mingsheng Long. Cycle self-training for domain adaptation. *Advances in Neural Information Processing Systems*, 34:22968–22981, 2021.

[135] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022.

[136] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[137] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022.

[138] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015.

[139] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pages 1640–1650, 2018.

[140] Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S Yu. Transfer feature learning with joint distribution adaptation. In *Proceedings of the IEEE international conference on computer vision*, pages 2200–2207, 2013.

[141] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. In *Advances in neural information processing systems*, pages 136–144, 2016.

[142] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, pages 2208–2217. PMLR, 2017.

[143] Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are gans created equal? a large-scale study. *Advances in neural information processing systems*, 31, 2018.

[144] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2507–2516, 2019.

[145] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, pages 181–196, 2018.

[146] Thomas Mensink, Jasper Uijlings, Alina Kuznetsova, Michael Gygli, and Vittorio Ferrari. 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.

[147] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640, 2019.

[148] Seonwoo Min, Nokyung Park, Siwon Kim, Seunghyun Park, and Jinkyu Kim. Grounding visual representations with texts for domain generalization. In *European Conference on Computer Vision*, pages 37–53. Springer, 2022.

[149] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2020.

[150] Pietro Morerio, Jacopo Cavazza, and Vittorio Murino. Minimal-entropy correlation alignment for unsupervised deep domain adaptation. *arXiv preprint arXiv:1711.10288*, 2017.

[151] Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 122–132, 2020.

[152] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pages 681–699. Springer, 2020.

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

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

[155] Jaemin Na, Heechul Jung, Hyung Jin Chang, and Wonjun Hwang. Fixbi: Bridging domain spaces for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1094–1103, 2021.

[156] Luigi Nardi, Bruno Bodin, M. Z. Zia, John Mawer, A. Nisbet, Paul H. J. Kelly, Andrew J. Davison, M. Luján, Michael F. P. O'Boyle, G. Riley, N. Topham, and Steve Furber. Introducing slambench, a performance and accuracy benchmarking methodology for slam. *IEEE International Conference on Robotics and Automation*, 2014.

[157] Alejandro Newell and Jia Deng. How useful is self-supervised pretraining for visual tasks? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7345–7354, 2020.

[158] Takehiko Ohkawa, Takuma Yagi, Taichi Nishimura, Ryosuke Furuta, Atsushi Hashimoto, Yoshitaka Ushiku, and Yoichi Sato. Exo2egodvc: Dense video captioning of egocentric procedural activities using web instructional videos. *arXiv preprint arXiv:2311.16444*, 2023.

[159] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. *Advances in neural information processing systems*, 31, 2018.

[160] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke

Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report. *arXiv preprint arXiv: 2303.08774*, 2023.

[161] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE transactions on neural networks*, 22(2):199–210, 2010.

[162] Changhwa Park, Jonghyun Lee, Jaeyoon Yoo, Minhoe Hur, and Sungroh Yoon. Joint contrastive learning for unsupervised domain adaptation. *arXiv preprint arXiv:2006.10297*, 2020.

[163] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[164] Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. Multi-adversarial domain adaptation. *arXiv preprint arXiv:1809.02176*, 2018.

[165] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1406–1415, 2019.

[166] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017.

[167] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, et al. Combined scaling for zero-shot transfer learning. *Neurocomputing*, 555:126658, 2023.

[168] Chiara Plizzari, Toby Perrett, Barbara Caputo, and Dima Damen. What can a cook in italy teach a mechanic in india? action recognition generalisation over scenarios and locations. *arXiv preprint arXiv: 2306.08713*, 2023.

[169] Viraj Prabhu, Shivam Khare, Deeksha Kartik, and Judy Hoffman. Sentry: Selective entropy optimization via committee consistency for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8558–8567, 2021.

[170] Viraj Prabhu, Ramprasaath R Selvaraju, Judy Hoffman, and Nikhil Naik. Can domain adaptation make object recognition work for everyone? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3981–3988, 2022.

[171] Camillo Quattrocchi, Antonino Furnari, Daniele Di Mauro, Mario Valerio Giuffrida, and Giovanni Maria Farinella. Synchronization is all you need: Exocentric-to-egocentric transfer for temporal action segmentation with unlabeled synchronized video pairs. *arXiv preprint arXiv:2312.02638*, 2023.

[172] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning

transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[173] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[174] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[175] Vikram V Ramaswamy, Sing Yu Lin, Dora Zhao, Aaron Adcock, Laurens van der Maaten, Deepti Ghadiyaram, and Olga Russakovsky. Geode: a geographically diverse evaluation dataset for object recognition. *Advances in Neural Information Processing Systems*, 36, 2024.

[176] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[177] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020.

[178] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5389–5400. PMLR, 09–15 Jun 2019.

[179] Ievgen Redko, Nicolas Courty, Rémi Flamary, and Devis Tuia. Optimal transport for multi-source domain adaptation under target shift. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 849–858. PMLR, 2019.

[180] Colorado J Reed, Xiangyu Yue, Ani Nrusimha, Sayna Ebrahimi, Vivek Vijaykumar, Richard Mao, Bo Li, Shanghang Zhang, Devin Guillory, Sean Metzger, et al. Self-supervised pretraining improves self-supervised pretraining. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2584–2594, 2022.

[181] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[182] William A Gaviria Rojas, Sudnya Diamos, Keertan Ranjan Kini, David Kanter, Vijay Janapa Reddi, and Cody Coleman. The dollar street dataset: Images representing the geographic and socioeconomic diversity of the world. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[183] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[184] Amir Rosenfeld, Richard Zemel, and John K Tsotsos. The elephant in the room. *arXiv preprint arXiv:1808.03305*, 2018.

[185] Karsten Roth, Timo Milbich, Samarth Sinha, Prateek Gupta, Bjorn Ommer, and Joseph Paul Cohen. Revisiting training strategies and generalization performance in deep metric learning. In *International Conference on Machine Learning*, pages 8242–8252. PMLR, 2020.

[186] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[187] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. 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*, pages 213–226. Springer, 2010.

[188] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010.

[189] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

[190] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8050–8058, 2019.

[191] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, and Kate Saenko. Universal domain adaptation through self supervision. *Advances in neural information processing systems*, 33:16282–16292, 2020.

[192] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, and Kate Saenko. Universal domain adaptation through self-supervision. 2020.

[193] Kuniaki Saito and Kate Saenko. Ovanet: One-vs-all network for universal domain adaptation. *IEEE International Conference on Computer Vision*, 2021.

[194] Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. Asymmetric tri-training for unsupervised domain adaptation. *arXiv preprint arXiv:1702.08400*, 2017.

[195] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Adversarial dropout regularization. *arXiv preprint arXiv:1711.01575*, 2017.

[196] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3723–3732, 2018.

[197] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *NEURIPS*, 2019.

[198] Swami Sankaranarayanan, Yogesh Balaji, Carlos D. Castillo, and Rama Chellappa. Generate to adapt: Aligning domains using generative adversarial networks. *arXiv preprint arXiv: 1704.01705*, 2017.

[199] Swami Sankaranarayanan, Yogesh Balaji, Carlos D Castillo, and Rama Chellappa. Generate to adapt: Aligning domains using generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8503–8512, 2018.

[200] Mert Bulent Sariyildiz, Julien Perez, and Diane Larlus. Learning visual representations with caption annotations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pages 153–170. Springer, 2020.

[201] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[202] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[203] Manuel Schwonberg, J. Niemeijer, Jan-Aike Termöhlen, Jörg P. Schäfer, Nico M. Schmidt, H. Gottschalk, and T. Fingscheidt. Survey on unsupervised domain adaptation for semantic segmentation for visual perception in automated driving. *IEEE Access*, 2023.

[204] D. Sculley, Eric Breck, Igor Ivanov, James Atwood, Miha Skalic, Pallavi Baljekar, Pavel Ostyakov, Roman Solovyev, Weimin Wang, and Yoni Halpern. *The Inclusive Images Competition*. 2019.

[205] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv preprint arXiv:1711.08536*, 2017.

[206] Rulin Shao, Zhouxing Shi, Jinfeng Yi, Pin-Yu Chen, and Cho-Jui Hsieh. On the adversarial robustness of vision transformers. *ArXiv*, abs/2103.15670, 2021.

[207] Astuti Sharma, Tarun Kalluri, and Manmohan Chandraker. Instance level affinity-based transfer for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5361–5371, 2021.

[208] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Actor and observer: Joint modeling of first and third-person videos. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7396–7404, 2018.

[209] Krishna Kumar Singh, Dhruv Mahajan, Kristen Grauman, Yong Jae Lee, Matt Feiszli, and Deepti Ghadiyaram. Don't judge an object by its context: learning to overcome contextual bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11070–11078, 2020.

[210] Mannat Singh, Laura Gustafson, Aaron Adcock, Vinicius de Freitas Reis, Bugra Gedik, Raj Prateek Kosaraju, Dhruv Mahajan, Ross Girshick, Piotr Dollár, and Laurens Van Der Maaten. Revisiting weakly supervised pre-training of visual perception models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 804–814, 2022.

[211] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

[212] Baochen Sun and Kate Saenko. 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*, pages 443–450. Springer, 2016.

[213] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and A. Gupta. Revisiting unreasonable effectiveness of data in deep learning era. *IEEE International Conference on Computer Vision*, 2017.

[214] Ming Sun, Yuchen Yuan, Feng Zhou, and Errui Ding. Multi-attention multi-class constraint for fine-grained image recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 805–821, 2018.

[215] Tao Sun, Cheng Lu, Tianshuo Zhang, and Haibin Ling. Safe self-refinement for transformer-based domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7191–7200, 2022.

[216] Shuhan Tan, Xingchao Peng, and Kate Saenko. Class-imbalanced domain adaptation: an empirical odyssey. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 585–602. Springer, 2020.

[217] Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, J. Kerr, Terrance Wang, Alexander Kristoffersen, J. Austin, Kamyar Salahi, Abhik Ahuja, David McAllister, and Angjoo Kanazawa. Nerfstudio: A modular framework for neural radiance field development. *International Conference on Computer Graphics and Interactive Techniques*, 2023.

[218] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33, 2020.

[219] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[220] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

[221] Hugo Touvron, Piotr Bojanowski, Mathilde Caron, Matthieu Cord, Alaaeldin El-Nouby, Edouard Grave, Gautier Izacard, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, et al. Resmlp: Feedforward networks for image classification with data-efficient training. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[222] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

[223] Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, pages 516–533. Springer, 2022.

[224] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE international conference on computer vision*, pages 4068–4076, 2015.

[225] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017.

[226] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

[227] Vishaal Udandarao, Ankush Gupta, and Samuel Albanie. Sus-x: Training-free name-only transfer of vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2725–2736, 2023.

[228] Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking representation learning for natural world image collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12884–12893, 2021.

[229] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification

and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018.

[230] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5018–5027, 2017.

[231] Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Luowei Zhou, Yucheng Zhao, Yujia Xie, Ce Liu, Yu-Gang Jiang, and Lu Yuan. Omnivl:one foundation model for image-language and video-language tasks. *arXiv preprint arXiv: 2209.07526*, 2022.

[232] Rui Wang, Zuxuan Wu, Zejia Weng, Jingjing Chen, Guo-Jun Qi, and Yu-Gang Jiang. Cross-domain contrastive learning for unsupervised domain adaptation. *arXiv preprint arXiv:2106.05528*, 2021.

[233] Sinan Wang, Xinyang Chen, Yunbo Wang, Mingsheng Long, and Jianmin Wang. Progressive adversarial networks for fine-grained domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9213–9222, 2020.

[234] Xun Wang, Haozhi Zhang, Weilin Huang, and Matthew R Scott. Cross-batch memory for embedding learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6388–6397, 2020.

[235] Yan Wang, Xiangyu Chen, Yurong You, Li Erran Li, Bharath Hariharan, Mark Campbell, Kilian Q Weinberger, and Wei-Lun Chao. Train in germany, test in the usa: Making 3d object detectors generalize. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11713–11723, 2020.

[236] Zhenbin Wang, Lei Zhang, Lituan Wang, and Minjuan Zhu. Landa: Language-guided multi-source domain adaptation. *arXiv preprint arXiv:2401.14148*, 2024.

[237] Colin Wei, Kendrick Shen, Yining Chen, and Tengyu Ma. Theoretical analysis of self-training with deep networks on unlabeled data. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[238] Guoqiang Wei, Cuiling Lan, Wenjun Zeng, Zhizheng Zhang, and Zhibo Chen. Toalign: Task-oriented alignment for unsupervised domain adaptation. In *NeurIPS*, 2021.

[239] Pengfei Wei, Lingdong Kong, Xinghua Qu, Yi Ren, Zhiqiang Xu, Jing Jiang, and Xiang Yin. Unsupervised video domain adaptation for action recognition: A disentanglement perspective. In *Advances in Neural Information Processing Systems*, 2023.

[240] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010.

[241] Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

[242] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7959–7971, 2022.

[243] Mitchell Wortsman, Gabriel Ilharco, Mike Li, Jong Wook Kim, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. *Computer Vision and Pattern Recognition*, 2021.

[244] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

[245] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018.

[246] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010.

[247] Shaoan Xie, Zibin Zheng, Liang Chen, and Chuan Chen. Learning semantic representations for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 5423–5432, 2018.

[248] Ruijia Xu, Guanbin Li, Jihan Yang, and Liang Lin. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1426–1435, 2019.

[249] Tongkun Xu, Weihua Chen, Pichao Wang, Fan Wang, Hao Li, and Rong Jin. Cdtrans: Cross-domain transformer for unsupervised domain adaptation. *arXiv preprint arXiv:2109.06165*, 2021.

[250] Zhe Xu, Shaoli Huang, Ya Zhang, and Dacheng Tao. Webly-supervised fine-grained visual categorization via deep domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 40(5):1100–1113, 2016.

[251] Zihui Sherry Xue and Kristen Grauman. Learning fine-grained view-invariant representations from unpaired ego-exo videos via temporal alignment. *Advances in Neural Information Processing Systems*, 36, 2023.

[252] Hongliang Yan, Yukang Ding, Peihua Li, Qilong Wang, Yong Xu, and Wangmeng Zuo. 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*, pages 2272–2281, 2017.

[253] Gonca Yilmaz, Songyou Peng, Francis Engelmann, Marc Pollefeys, and Hermann Blum. Opendas: Domain adaptation for open-vocabulary segmentation. *arXiv preprint arXiv: 2405.20141*, 2024.

[254] Kaichao You, Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Universal domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2720–2729, 2019.

[255] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. *ICCV*, 2023.

[256] Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *International conference on machine learning*, pages 819–827. PMLR, 2013.

[257] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell. Part-based r-cnns for fine-grained category detection. In *European conference on computer vision*, pages 834–849. Springer, 2014.

[258] Ning Zhang, Ryan Farrell, and Trever Darrell. Pose pooling kernels for sub-category recognition. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3665–3672. IEEE, 2012.

[259] Weichen Zhang, Wanli Ouyang, Wen Li, and Dong Xu. Collaborative and adversarial network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3801–3809, 2018.

[260] Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael I Jordan. Bridging theory and algorithm for domain adaptation. *arXiv preprint arXiv:1904.05801*, 2019.

[261] Bingchen Zhao, Shaozuo Yu, Wufei Ma, Mingxin Yu, Shenxiao Mei, Angtian Wang, Ju He, Alan Yuille, and Adam Kortylewski. Robin: a benchmark for robustness to individual nuisances in real-world out-of-distribution shifts. *arXiv preprint arXiv:2111.14341*, 2021.

[262] Bingchen Zhao, Shaozuo Yu, Wufei Ma, Mingxin Yu, Shenxiao Mei, Angtian Wang, Ju He, Alan Yuille, and Adam Kortylewski. Ood-cv: A benchmark for robustness to individual nuisances in real-world out-of-distribution shifts. In *ICML 2022 Shift Happens Workshop*, 2022.

[263] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6586–6597, 2023.

[264] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 5209–5217, 2017.

[265] Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5012–5021, 2019.

[266] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

[267] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. *Advances in neural information processing systems*, 27, 2014.

[268] Daquan Zhou, Zhiding Yu, Enze Xie, Chaowei Xiao, Animashree Anandkumar, Jiashi Feng, and Jose M Alvarez. Understanding the robustness in vision transformers. In *International Conference on Machine Learning*, pages 27378–27394. PMLR, 2022.

[269] Jinjing Zhu, Haotian Bai, and Lin Wang. Patch-mix transformer for unsupervised domain adaptation: A game perspective. *Computer Vision and Pattern Recognition*, 2023.

[270] Jinjing Zhu, Haotian Bai, and Lin Wang. Patch-mix transformer for unsupervised domain adaptation: A game perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3561–3571, 2023.