MemSAC: Memory Augmented Sample Consistency for Large Scale Domain Adaptation

Tarun Kalluri, Astuti Sharma, and Manmohan Chandraker

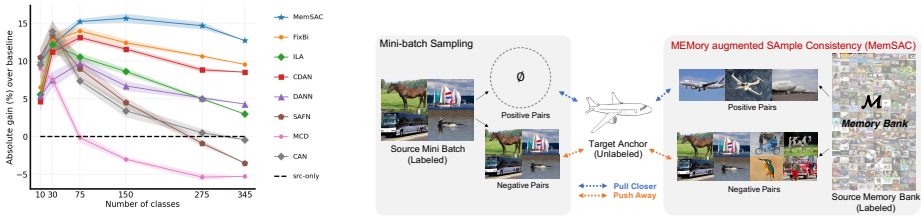
University of California San Diego, La Jolla CA 92093, USA
{sskallur,asharma,mkchandraker}@eng.ucsd.edu

Abstract. Practical real world datasets with plentiful categories introduce new challenges for unsupervised domain adaptation like small inter-class discriminability, that existing approaches relying on domain invariance alone cannot handle sufficiently well. In this work 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. We validate the advantages of MemSAC with significant improvements over previous state-of-the-art on multiple challenging transfer tasks designed for large-scale adaptation, such as DomainNet with 345 classes and fine-grained adaptation on Caltech-UCSD birds dataset with 200 classes. We also provide in-depth analysis and insights into the effectiveness of MemSAC. Code is available on the project webpage https://tarun005.github.io/MemSAC.

1 Introduction

It is well known that deep neural networks often do not generalize well when the distribution of test samples differ significantly 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 [3–6, 13, 18, 20, 24, 38–41, 43, 54, 60–62, 72]. Despite impressive gains, most approaches have been largely demonstrated on datasets with a limited number of categories [48, 50],

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 1a, which plots the absolute gain over a source-only model obtained by well-known adaptation methods (including DANN [20], MCD [55], SAFN [72], CAN [32], FixBi [43]) with respect to number of classes sampled from the DomainNet dataset [47]. While all methods provide similar



(a) (b)

Fig. 1: (a) Accuracy(%) of various methods proposed for unsupervised domain adaptation with respect to the number of training classes from DomainNet [47](**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. In contrast, proposed MemSAC obtains significant gains ($\sim 15\%$) even on large scale datasets with many classes [47]. (b) **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 ($\emptyset$) leading to noisy consistency estimates.

benefits over a source-only model in smaller-scale settings with 10-30 classes, the gains reduce significantly 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 [20, 32, 55, 71, 72] 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 [18, 32, 43, 43, 46, 54] suffer from noisy psuedo-labels leading to poor transfer. We observe that the contrastive loss is shown to be highly successful in learning better transferable features [11, 12, 22, 23, 27, 29, 42, 70] 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 [68], 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 [27, 67], 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. 1b). Our architecture scales remarkably well with the number of categories, including the case of fine-grained adaptation [66] where all classes belong to a single subordinate category [7, 77]. 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.

To summarize, in contrast to prior works, MemSAC achieves scalability in domain adaptation with a large number of classes. Our main contributions are:

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. 3.1).
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. 3.2).
3. Theoretical justification of the proposed losses in terms of the input-consistency regularization proposed in [68] for domain adaptation (Sec. 3.3).
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 DomainNet dataset with 345 categories and CUB-Drawings with 200 categories, respectively (Sec. 4).

2 Related Work

Unsupervised Domain Adaptation A suite of tools have been proposed recently under the umbrella of unsupervised domain adaptation (UDA) that enable training on a labeled source domain and deploy models on a different target

domain with few or no labels. A large body of these works aim to minimize some notion of divergence [3, 4, 50] between the source and target using an adversarial objective, resulting in domain invariant features [10, 20, 31, 39, 55, 60–62, 71]. Since domain invariance alone does not guarantee discriminative features in target [35], recent approaches propose class aware adaptation to align class conditional distributions across source and target [16, 18, 32, 43, 46, 53, 57, 69, 71]. ATT [53] assigns pseudo-labels based on predictions from classifiers, MADA [46] uses separate adversarial networks for each class, ILA [57] computes pairwise similarity between samples within a mini-batch for instance aware adaptation while SAFN [72] proposes re-normalizing features to achieve transferability. However, none of these works explicitly address the issue of scalability to adaptation with a large number of categories. Moreover, many clustering based methods [32, 45] and instance based methods [57, 65] proposed for UDA are not readily scalable to large datasets.

While partial adaptation [8, 9, 76], open set adaptation [44, 56] and universal adaptation [52] allow training on real world source datasets with many categories, they are only focused on adaptation across those categories that are shared between source and target which are generally few in number, and do not address the problem of discriminative transfer across *all* the categories which is a different practical problem, and focus of this work.

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 [63, 64]. While fine grained classification on within domain samples has received much attention [7, 37, 58, 77–80], the problem of unsupervised domain adaptation across fine-grained categories is relatively less studied [17, 21, 66, 73]. All prior works often demand additional annotations in the form of attributes [21], weak supervision [17], part annotations [73] or hierarchical relationships [66] 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.

Contrastive Learning The success of contrastive learning [1, 25, 26, 68] in extracting visual representations from data has attracted wide interest in self-supervised [11, 12, 15, 22, 23, 27, 29, 42, 70], semi-supervised [2] and supervised learning [34]. A unifying 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 Unsupervised Adaptation using MemSAC

Problem Description In unsupervised domain adaptation, we have labeled samples $\mathcal{X}^s$ from a source domain with a corresponding source probability

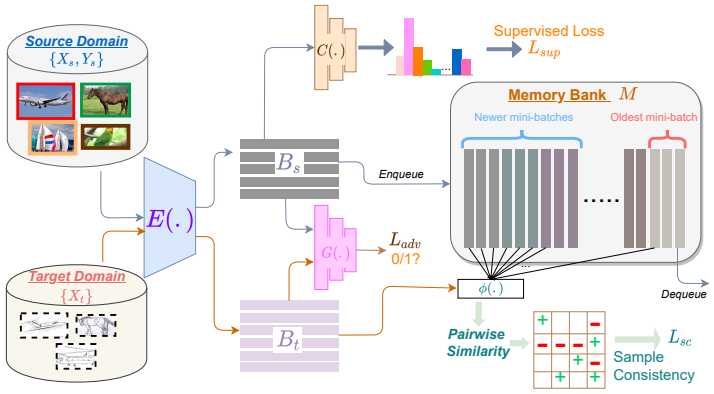


Fig. 2: **An overview of MemSAC for domain adaptation** During each iteration, the 256-dim source feature embeddings computed using $\mathcal{E}$, along with their labels, are added to a memory bank $\mathcal{M}$ and the oldest set of features are removed. Pairwise similarities between each target feature 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 ($\mathcal{L}_{sc}$) on these similar and dissimilar pairs, along with adversarial loss ($\mathcal{L}_{adv}$), we achieve both local alignment and global adaptation.

distribution P_s, labeled according to a true labeling function f^*, and $\mathcal{Y}^s = f^*(\mathcal{X}^s)$. We are also given unlabeled data points $\mathcal{X}^t$ sampled according to the target distribution P_t. We follow a *covariate shift assumption* [3], 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 $\mathcal{Y} = \{1, 2, \ldots, C\}$ 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 $\mathcal{X}_t$. 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. 2. $\mathcal{E}$ and $\mathcal{C}$ are the feature extractor and the classifier respectively. The objective function for MemSAC is given by

$$\min_{\theta} \mathcal{L}_{sup}(\mathcal{X}^s, \mathcal{Y}^s; \theta) + \lambda_{adv}\mathcal{L}_{adv}(\mathcal{X}^s, \mathcal{X}^t; \theta) + \lambda_{sc}\mathcal{L}_{sc}(\mathcal{X}^s, \mathcal{Y}^s, \mathcal{X}^t; \theta), \qquad (1)$$

where $\mathcal{L}_{sup}$ 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. $\mathcal{L}_{adv}$ is the domain adversarial loss which we implement using a class conditional discriminator (Eq. 2) and $\mathcal{L}_{sc}$ 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. 4). λ_{adv} and λ_{sc} are the corresponding loss coefficients. We use $\mathcal{B}_s (\in \mathcal{X}^s)$ and $\mathcal{B}_t (\in \mathcal{X}^t)$ to denote labeled

source and unlabeled target mini-batches respectively, which are chosen randomly at each iteration from the dataset.

Class conditional adversarial loss We adopt the widely used adversarial strategy to learn domain-invariant feature representations using a domain discriminator $\mathcal{G}(.,\omega)$ 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 [39] to fuse information from the deep features as well as the classifier predictions. Denoting $f = \mathcal{E}(x)$ and $g = \mathcal{C}(\mathcal{E}(x))$, the input $h(x)$ to the discriminator $\mathcal{G}$ 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

$$\mathcal{L}_d = \frac{1}{|\mathcal{B}_s|} \sum_{i \in \mathcal{B}_s} - \log(\mathcal{G}(h_i;\omega)) + \frac{1}{|\mathcal{B}_t|} \sum_{i \in \mathcal{B}_t} - \log(1 - \mathcal{G}(h_i;\omega)) \qquad \mathcal{L}_{adv} = -\mathcal{L}_d. \quad (2)$$

We note that our contributions are complementary to the type of alignment objective used. In Tab. 3a, we show significant gains starting from another adversarial objective (DANN [20]) and MMD objectives (CAN [33]) as well.

3.1 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 $\mathcal{B}_t$ as the anchor, we construct a *similar set* $\mathcal{B}_{s^+}^{x_t} = \{x \in \mathcal{B}_s | f^*(x) = f^*(x_t)\}$ and dissimilar set $\mathcal{B}_{s^-}^{x_t} = \mathcal{B}_s \backslash \mathcal{B}_{s^+}^{x_t}$ consisting of source samples and use this knowledge of sample-level similarity in the following *sample consistency loss*

$$\mathcal{L}_{sc,\mathcal{B}} = \frac{1}{|\mathcal{B}_t|} \sum_{j \in \mathcal{B}_t} - \log \left\{ \sum_{i \in \mathcal{B}_{s^+}^j} \frac{\exp(\phi_{ij}/\tau)}{\sum_{i \in \mathcal{B}_s} \exp(\phi_{ij}/\tau)} \right\} \quad (3)$$

where ϕ_{ij} measures the cosine similarity metric between two feature vectors i and j, $(\phi_{ij}{=}\phi(f_i, f_j){=}\frac{f_i \cdot f_j}{||f_i|| ||f_j||})$ and τ is the temperature parameter used to scale the contributions of positive and negative pairs [11,30]. $\mathcal{L}_{sc,\mathcal{B}}$ denotes the sample consistency loss computed using the mini-batch. Distinct from standard constrative loss [11, 42] that typically derives positive pairs from augmented versions of the same image, our loss in Eq. (3) is well-suited to handle multiple positive and negative pairs for each anchor, similar to SupCon loss [34]. However, in contrast to SupCon, our modified consistency loss allows us to scale domain adaptation to many-class settings.

kNN-based pseudo-labeling There are two challenges in directly using the sample consistency loss in (3). Firstly, unlike prior approaches [11,29,42] 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 ($f^*(x_t)$ is unknown). To address this issue, we use a k-NN based psuedo-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 $\mathcal{B}_t$, we find k nearest neighbors from the source training mini-batch $\mathcal{B}_s$, 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 psuedo-labeling is independent of, thus less sensitive to, noisy classifier boundaries helping us extract reliable target psuedo-labels during training. Once $\mathcal{B}_t$ is psuedo-labeled, it is straightforward to compute $\mathcal{B}_{s^+}^{x_t}$ in (3). The second challenge is lack of representation for all classes in a mini-batch, which we address next.

3.2 Memory augmented similarity extraction

From Eq. (3), we can observe that if the source and target mini-batches $\mathcal{B}_s$ and $\mathcal{B}_t$ 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 $\mathcal{B}_s$ and $\mathcal{B}_t$ usually contain 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 $\mathcal{M}$ 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 $|\mathcal{M}|$ is sufficiently large, then source samples from all the classes would be adequately present in $\mathcal{M}$, 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 $\mathcal{M}$, thus decoupling the similarity computation with mini-batch size or dataset size. Different from prior approaches that augment training with memory module [27,67,70], our approach aggregates features from multiple source batches, thus helping target samples to extract meaningful pairwise relationships from different classes.

Initializing and updating memory bank To initialize the memory bank, we first bootstrap the feature extractor for few hundred iterations by training only using $\mathcal{L}_{sup}$ and $\mathcal{L}_{adv}$ losses, and then introduce our consistency loss $\mathcal{L}_{sc}$ and start populating $\mathcal{M}$. After this, we follow a queue based approach for updating the memory bank similar to XBM [67]. 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 $\{\mathcal{E}(x)|x \in \mathcal{B}_s\}$) along with the corresponding

source labels. Alternative strategies for updating $\mathcal{M}$, such as a momentum encoder [27], yield similar results (details in the supplementary).

Sample consistency using memory bank We can now use $\mathcal{M}$ as a proxy for $\mathcal{B}_s$ (and similar set $\mathcal{M}_+^{x_t}$ as a proxy for $\mathcal{B}_{s+}^{x_t}$) in assigning the target psuedo labels and in the sample consistency loss in (3). $|\mathcal{M}|$ is often much higher than $|\mathcal{B}_s|$, so access to larger number of source samples from $\mathcal{M}$ provides k-NN pseudo labels that are more reliable, with richer variety of positive and negative pairwise relations (more details in the supplementary). The final sample consistency loss used in MemSAC is

$$\mathcal{L}_{sc} = \frac{1}{|\mathcal{B}_t|} \sum_{j \in \mathcal{B}_t} -\log \left\{ \sum_{i \in \mathcal{M}_+^j} \frac{\exp(\phi_{ij}/\tau)}{\sum_{i \in \mathcal{M}} \exp(\phi_{ij}/\tau)} \right\}. \tag{4}$$

3.3 Theoretical Insight

Recently, Wei et al. [68] 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 $R(\theta) \equiv \min_\theta \mathbb{E}_x[\max_{x' \in \mathcal{N}(x)} \mathbf{1}(C(x;\theta) \neq C(x';\theta))]$, where $\mathcal{N}(x)$ is the neighborhood of a sample x (Eq 1.2 in [68]). We now show the connections that can be drawn between our loss and the theory proposed in [68]. For this purpose, we work with the following approximations. Firstly, we approximate the neighborhood $\mathcal{N}(x)$ of a sample x with the *similar set* defined in Sec. 3.1, that is $\mathcal{N}(x) = \mathcal{B}_+^x$. 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

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

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. (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 [68]. Furthermore, optimizing such an objective is shown to achieve bounded target error for unsupervised domain adaptation. Specifically, under the assumption that the psuedo label accuracy on target data is above a certain threshold, [68] 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 ($\mathcal{L}_{sup}$) and adversarial ($\mathcal{L}_{adv}$) losses as explained in Sec. 3.2 before introducing our proposed sample consistency loss. To the best of our knowledge, we are the first to instantiate the regularization proposed in [68] for large scale domain adaptation, and showcase its effectiveness in achieving significant empirical gains.

4 Experiments and Analysis

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.

DomainNet [47] 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 [49, 51, 59] 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. In the supplementary material, we also provide results using a 126-class subset of DomainNet which contains much lesser label noise [36, 51, 75]. Nevertheless, our benefits persist on both these splits.

CUB (Caltech-UCSD birds) [64] is a challenging dataset originally proposed for fine-grained classification of 200 categories of birds, while *CUB-Drawings* [66] 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. Results on other fine-grained datasets like Birds-123 and CompCars [74] are present in the supplementary material.

Training Details We use a Resnet-50 [28] backbone pretrained on Imagenet, followed by a projection layer as the encoder $\mathcal{E}$ to obtain 256 dimensional feature embeddings. The discriminator $\mathcal{G}$ 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

Table 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 [71]	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 [24]	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 [13]	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 [55]‡	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 [57]§	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 [72]	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 [20]	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 [32]§	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 [66]†	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 [39]	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 [16]	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 [43]‡	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 [69]	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

dataset following established protocol for UDA [39, 55, 66, 72], and the provided testset for DomainNet. The crucial hyper-parameters in our method are λ_{sc}, temperature τ and memory bank size $|\mathcal{M}|$. 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, but 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 [20], CDAN [39], MCD [55]) as well as the current state-of-the art (SAFN [72], BSP [13], RSDA [24], CAN [32], ILADA [57], FixBi [43], HDAN [16] and ToAlign [69]).

MemSAC significantly outperforms others on many-class adaptation The results for the 12 transfer tasks on DomainNet are provided in Tab. 1. Firstly, methods such as RSDA (29.68%) and SAFN (39.54%) that achieve best performance on smaller scale datasets (like Office-31 [50] and visDA-2017 [48]) 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 [66], which requires a label hierarchy as additional information for training. For this supervision, we use the one level of hierarchy proposed in DomainNet [47]. 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. These trends and benefits also hold for the 126-class version of the DomainNet dataset, and results are presented in the supplementary material.

Table 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. §Uses class-balanced sampling.

	Resnet-50	MCD [55]	SAFN [72]	CAN§ [32]	RSDA [24]	DANN [20]	HDAN [16]	FixBi [43]	CDAN [39]	ToAlign [69]	PAN† [66]	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>	**73.97**
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>	**61.94**
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>	**67.95**

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. 2 on the CUB-Drawings dataset. Although fine-grained visual recognition is a well-studied area [7, 14, 19, 77, 78], 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 [66] 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 67.95% 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.

MemSAC complements multiple adaptation methods The proposed memory-augmented consistency loss is generic enough to improve many adaptation backbones. As shown in Tab. 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. Complete table for all the 12 transfer tasks is provided in supplementary material.

MemSAC improves adaptation even with larger backbones We employ Resnet-101 as a backbone in Tab. 3a 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, and complete table containing results on all transfer tasks is presented in the supplementary material.

4.1 Analysis and Discussion

Ablation studies We show the influence of various design choices of our method in Tab. 4 on the CUB-200 dataset. First, we show in Tab. 4a that both the global domain adversarial method, which 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,

Table 3: MemSAC is also complementary to most adversarial and adaptation methods, as shown in (a). We show the results using a larger backbone (Resnet-101) for training in (b). MemSAC adds negligible memory and time overhead to the training even with large queues, and zero overhead during inference, as shown in (c)

(a) MemSAC complements existing UDA methods.

	R→C	C→R	Avg.
DANN [20]	45.93	46.85	46.39
DANN+MemSAC	49.67	49.81	**49.74**(+3.35%)
CAN [32]	40.71	54.93	47.82
CAN+MemSAC	43.79	55.36	**49.57**(+1.75%)
CDAN [39]	50.15	50.02	50.08
CDAN+MemSAC	54.34	54.40	**54.37**(+4.29%)

(b) Results using Resnet-101 backbone

	R→C	C→R	Avg.
Resnet-101	45.62	41.96	43.79
DANN [20]	47.71	48.33	48.02
MCD [55]	41.11	40.77	40.94
CDAN [39]	52.47	46.63	49.55
SAFN [72]	44.93	37.20	41.06
ToAlign [69]	50.09	50.23	50.16
MemSAC	**56.25**	**53.52**	**54.88**

(c) Training times of various methods.

Method	Peak GPU Mem.	Training Time	Avg. Acc
DANN	7.9GB	11.7 Hrs	40.58%
CDAN	8.2GB	12 Hrs	43.24%
PAN	8.9GB	16.2 Hrs	43.03%
ToAlign	9.22GB	24.21 Hrs	45.45%
MemSAC	8.5GB	12.63 Hrs	47.26%

we investigate the effect of the temperature parameter τ in Tab. 4b which we use to suitably scale the contributions of positive and negative pairs in $\mathcal{L}_{sc}$ loss function (Eq. (4)). We find that $\tau = 0.07$ gives the best performance on the cosine similarity metric. Finally, in Tab. 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.

Why does MemSAC help with large number of classes? We propose our sample consistency loss in (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, we plot the *mean similarity score* during the training process. We define the *mean similarity score* as $\sum_{i \in \mathcal{M}_+^j} \phi_{ij}$, averaged over all the target samples $j \in \mathcal{B}_t$ 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. 1, Tab. 2).

MemSAC achieves larger gains with finer-grained classes We show the appreciating benefits provided by MemSAC as the fine-grainedness of the dataset becomes more pronounced. For this purpose, we chose the 4 levels of label hierarchy provided by PAN [66] 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. 4 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.

Table 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 (1).

Method	$\mathcal{L}_{adv}$	$\mathcal{L}_{sc}$	C→D	D→C	Avg. Acc
Source	✗	✗	60.88	42.07	51.47
CDAN	✓	✗	68.12	53.83	60.98
$\mathcal{L}_{sc}$ Only	✗	✓	64.45	41.13	52.79
MemSAC	✓	✓	**73.97**	**61.94**	**67.95**

(b) Effect of the temperature τ in (4).

τ	C→D	D→C	Avg. Acc
1.0	68.36	53.46	60.91
0.07	**73.97**	**61.94**	**67.95**
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	$f_i \cdot f_j$	**73.97**	**61.94**	**67.95**

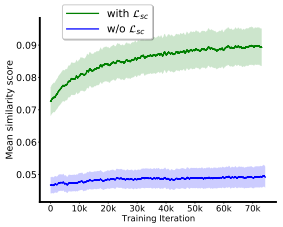


Fig. 3: Mean similarity score for *within-class* samples vs. training iteration shown for **D→C** on CUB-Drawings.

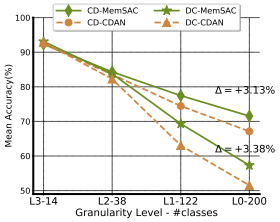


Fig. 4: Comparison of accuracy vs. granularity of labels on CUB-Drawings dataset for 4 levels of label hierarchy.

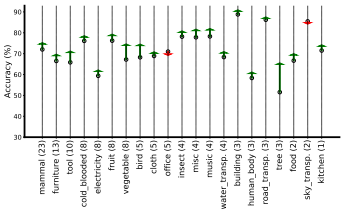


Fig. 5: Category wise gain/drop in accuracy on **R→C** on DomainNet, compared to CDAN [39].

MemSAC alleviates class confusion for similar classes In Fig. 5 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 ↑) on most categories and any drops in accuracy (marked by ↓) 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. Similar plots for other tasks in DomainNet are provided with the supplementary material.

Larger memory banks improve accuracy A key design choice that we need to make in MemSAC is the size of the memory bank $\mathcal{M}$. Intuitively, small memory banks would not provide sufficient negative pairs in the sample consistency loss and lead to noisy gradients. We show in Fig. 6 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 dataset. For example, the "drawing" domain has around 4k examples, but from Fig. 6, **D→C** achieves best accuracy at memory size of 25k. Since the feature encoder is simultaneously trained while updating memory bank, two copies of the same instance need not necessarily be exact duplicates of each other, but instead provide complementary "views" of the same

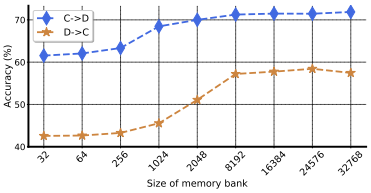
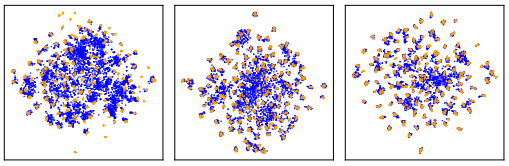


Fig. 6: Effect of memory bank size on CUB-Drawings dataset.



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

Fig. 7: tSNE for **R→C** on DomainNet. The two colors are source and target features. Notice improved alignment and feature separation with MemSAC.

sample. Thus, large queues help in enriching the positive and negative sample set, improving the accuracy.

Computational cost and resources We show the training time and GPU memory consumed for MemSAC compared to other baseline approaches in Tab. 3c. In summary, MemSAC incurs negligible overhead in memory during training and no overhead during inference even for large queues.

Limitations, future work and impact Although we report outstanding performance using MemSAC, we assume that the list of categories present in the data is known beforehand. Therefore, an avenue of future work is to relax this assumptions and extend MemSAC to open world adaptation approaches. While domain adaptation may have the positive impact of equitable performance of machine learning across geographic or social factors, MemSAC shares with other deep domain adaptation approaches the limitation of lack of explainability, which may have a negative impact on applications where decisions based on domain adaptation have a bearing on safety. We further note that significant room for improvement remains in achieving unsupervised domain adaptation that approach fully supervised performances.

5 Conclusion

We proposed 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. In extensive experiments and analysis across the main paper and supplementary, we showcase the strong improvements achieved by MemSAC over prior works, setting new state-of-the-arts across challenging many-class adaptation on DomainNet (126 and 345 classes) and fine-grained adaptation on CUB-Drawings (200 classes).

Acknowledgements We thank NSF CAREER 1751365, NSF Chase-CI 1730158, Google Award for Inclusion Research and IPE PhD Fellowship.

References

1. Arora, S., Khandeparkar, H., Khodak, M., Plevrakis, O., Saunshi, N.: A theoretical analysis of contrastive unsupervised representation learning. arXiv preprint arXiv:1902.09229 (2019) 4
2. Assran, M., Caron, M., Misra, I., Bojanowski, P., Joulin, A., Ballas, N., Rabbat, M.: Semi-supervised learning of visual features by non-parametrically predicting view assignments with support samples. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8443–8452 (2021) 4
3. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W.: A theory of learning from different domains. Machine learning **79**(1-2), 151–175 (2010) 1, 4, 5
4. Ben-David, S., Blitzer, J., Crammer, K., Pereira, F.: Analysis of representations for domain adaptation. Advances in neural information processing systems **19**, 137–144 (2006) 1, 4
5. Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., Krishnan, D.: Unsupervised pixel-level domain adaptation with generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3722–3731 (2017) 1
6. Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D., Erhan, D.: Domain separation networks. In: Advances in neural information processing systems. pp. 343–351 (2016) 1
7. Branson, S., Van Horn, G., Belongie, S., Perona, P.: Bird species categorization using pose normalized deep convolutional nets. arXiv preprint arXiv:1406.2952 (2014) 3, 4, 11
8. Cao, Z., Long, M., Wang, J., Jordan, M.I.: Partial transfer learning with selective adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2724–2732 (2018) 4
9. Cao, Z., Ma, L., Long, M., Wang, J.: Partial adversarial domain adaptation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 135–150 (2018) 4
10. Chen, C., Xie, W., Huang, W., Rong, Y., Ding, X., Huang, Y., Xu, T., Huang, J.: Progressive feature alignment for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 627–636 (2019) 4
11. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. arXiv preprint arXiv:2002.05709 (2020) 2, 4, 6, 7
12. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020) 2, 4
13. Chen, X., Wang, S., Long, M., Wang, J.: Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In: International conference on machine learning. pp. 1081–1090. PMLR (2019) 1, 10
14. Chen, Y., Bai, Y., Zhang, W., Mei, T.: Destruction and construction learning for fine-grained image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5157–5166 (2019) 11
15. Cole, E., Yang, X., Wilber, K., Mac Aodha, O., Belongie, S.: When does contrastive visual representation learning work? arXiv preprint arXiv:2105.05837 (2021) 4
16. Cui, S., Jin, X., Wang, S., He, Y., Huang, Q.: Heuristic domain adaptation. In: Advances in Neural Information Processing Systems. vol. 33, pp. 7571–7583. Curran Associates, Inc. (2020) 4, 10, 11

17. Cui, Y., Song, Y., Sun, C., Howard, A., Belongie, S.: Large scale fine-grained categorization and domain-specific transfer learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4109–4118 (2018) 4
18. Du, Z., Li, J., Su, H., Zhu, L., Lu, K.: Cross-domain gradient discrepancy minimization for unsupervised domain adaptation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2021) 1, 2, 4
19. Dubey, A., Gupta, O., Raskar, R., Naik, N.: Maximum-entropy fine-grained classification. arXiv preprint arXiv:1809.05934 (2018) 11
20. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: International conference on machine learning. pp. 1180–1189. PMLR (2015) 1, 2, 4, 6, 10, 11, 12
21. Gebru, T., Hoffman, J., Fei-Fei, L.: Fine-grained recognition in the wild: A multi-task domain adaptation approach. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1349–1358 (2017) 4
22. Gordon, D., Ehsani, K., Fox, D., Farhadi, A.: Watching the world go by: Representation learning from unlabeled videos. arXiv preprint arXiv:2003.07990 (2020) 2, 4
23. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., et al.: Bootstrap your own latent: A new approach to self-supervised learning. arXiv preprint arXiv:2006.07733 (2020) 2, 4
24. Gu, X., Sun, J., Xu, Z.: Spherical space domain adaptation with robust pseudo-label loss. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9101–9110 (2020) 1, 10, 11
25. Gutmann, M., Hyvärinen, A.: Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. pp. 297–304 (2010) 4
26. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). vol. 2, pp. 1735–1742. IEEE (2006) 4
27. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738 (2020) 2, 3, 4, 7, 8
28. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 9
29. Hénaff, O.J., Srinivas, A., De Fauw, J., Razavi, A., Doersch, C., Eslami, S., Oord, A.v.d.: Data-efficient image recognition with contrastive predictive coding. arXiv preprint arXiv:1905.09272 (2019) 2, 4, 7
30. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015) 6
31. Kalluri, T., Varma, G., Chandraker, M., Jawahar, C.: Universal semi-supervised semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5259–5270 (2019) 4
32. Kang, G., Jiang, L., Yang, Y., Hauptmann, A.G.: Contrastive adaptation network for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4893–4902 (2019) 1, 2, 4, 10, 11, 12
33. Kang, G., Zheng, L., Yan, Y., Yang, Y.: Deep adversarial attention alignment for unsupervised domain adaptation: the benefit of target expectation maximization.

In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 401–416 (2018) 6
34. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. arXiv preprint arXiv:2004.11362 (2020) 4, 6
35. Kumar, A., Sattigeri, P., Wadhawan, K., Karlinsky, L., Feris, R., Freeman, B., Wornell, G.: Co-regularized alignment for unsupervised domain adaptation. In: Advances in Neural Information Processing Systems. pp. 9345–9356 (2018) 4
36. Liang, J., Hu, D., Feng, J.: Combating domain shift with self-taught labeling. arXiv preprint arXiv:2007.04171 (2020) 9
37. Lin, T.Y., RoyChowdhury, A., Maji, S.: Bilinear cnn models for fine-grained visual recognition. In: Proceedings of the IEEE international conference on computer vision. pp. 1449–1457 (2015) 4
38. Long, M., Cao, Y., Wang, J., Jordan, M.: Learning transferable features with deep adaptation networks. In: International conference on machine learning. pp. 97–105. PMLR (2015) 1
39. Long, M., Cao, Z., Wang, J., Jordan, M.I.: Conditional adversarial domain adaptation. In: Advances in Neural Information Processing Systems. pp. 1640–1650 (2018) 1, 4, 6, 10, 11, 12, 13
40. Long, M., Wang, J., Ding, G., Sun, J., Yu, P.S.: Transfer feature learning with joint distribution adaptation. In: Proceedings of the IEEE international conference on computer vision. pp. 2200–2207 (2013) 1
41. Long, M., Zhu, H., Wang, J., Jordan, M.I.: Unsupervised domain adaptation with residual transfer networks. In: Advances in neural information processing systems. pp. 136–144 (2016) 1
42. Misra, I., Maaten, L.v.d.: Self-supervised learning of pretext-invariant representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6707–6717 (2020) 2, 4, 6, 7
43. Na, J., Jung, H., Chang, H.J., Hwang, W.: Fixbi: Bridging domain spaces for unsupervised domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1094–1103 (2021) 1, 2, 4, 10, 11
44. Panareda Busto, P., Gall, J.: Open set domain adaptation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 754–763 (2017) 4
45. Park, C., Lee, J., Yoo, J., Hur, M., Yoon, S.: Joint contrastive learning for unsupervised domain adaptation. arXiv preprint arXiv:2006.10297 (2020) 4
46. Pei, Z., Cao, Z., Long, M., Wang, J.: Multi-adversarial domain adaptation. arXiv preprint arXiv:1809.02176 (2018) 2, 4
47. Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., Wang, B.: Moment matching for multi-source domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1406–1415 (2019) 1, 2, 9, 10
48. Peng, X., Usman, B., Kaushik, N., Hoffman, J., Wang, D., Saenko, K.: Visda: The visual domain adaptation challenge. arXiv preprint arXiv:1710.06924 (2017) 1, 10
49. Prabhu, V., Khare, S., Kartik, D., Hoffman, J.: Sentry: Selective entropy optimization via committee consistency for unsupervised domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8558–8567 (2021) 9
50. Saenko, K., Kulis, B., Fritz, M., Darrell, T.: Adapting visual category models to new domains. In: European conference on computer vision. pp. 213–226. Springer (2010) 1, 4, 10

51. Saito, K., Kim, D., Sclaroff, S., Darrell, T., Saenko, K.: Semi-supervised domain adaptation via minimax entropy. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8050–8058 (2019) 9
52. Saito, K., Kim, D., Sclaroff, S., Saenko, K.: Universal domain adaptation through self supervision. arXiv preprint arXiv:2002.07953 (2020) 4
53. Saito, K., Ushiku, Y., Harada, T.: Asymmetric tri-training for unsupervised domain adaptation. arXiv preprint arXiv:1702.08400 (2017) 4
54. Saito, K., Ushiku, Y., Harada, T., Saenko, K.: Adversarial dropout regularization. arXiv preprint arXiv:1711.01575 (2017) 1, 2
55. Saito, K., Watanabe, K., Ushiku, Y., Harada, T.: Maximum classifier discrepancy for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3723–3732 (2018) 1, 2, 4, 10, 11, 12
56. Saito, K., Yamamoto, S., Ushiku, Y., Harada, T.: Open set domain adaptation by backpropagation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 153–168 (2018) 4
57. Sharma, A., Kalluri, T., Chandraker, M.: Instance level affinity-based transfer for unsupervised domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5361–5371 (2021) 4, 10
58. Sun, M., Yuan, Y., Zhou, F., Ding, E.: Multi-attention multi-class constraint for fine-grained image recognition. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 805–821 (2018) 4
59. Tan, S., Peng, X., Saenko, K.: Class-imbalanced domain adaptation: an empirical odyssey. In: European Conference on Computer Vision. pp. 585–602. Springer (2020) 9
60. Tzeng, E., Hoffman, J., Darrell, T., Saenko, K.: Simultaneous deep transfer across domains and tasks. In: Proceedings of the IEEE international conference on computer vision. pp. 4068–4076 (2015) 1, 4
61. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7167–7176 (2017) 1, 4
62. Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., Darrell, T.: Deep domain confusion: Maximizing for domain invariance. arXiv preprint arXiv:1412.3474 (2014) 1, 4
63. Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The inaturalist species classification and detection dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8769–8778 (2018) 4
64. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 Dataset. Tech. Rep. CNS-TR-2011-001, California Institute of Technology (2011) 4, 9
65. Wang, R., Wu, Z., Weng, Z., Chen, J., Qi, G.J., Jiang, Y.G.: Cross-domain contrastive learning for unsupervised domain adaptation. arXiv preprint arXiv:2106.05528 (2021) 4
66. Wang, S., Chen, X., Wang, Y., Long, M., Wang, J.: Progressive adversarial networks for fine-grained domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9213–9222 (2020) 3, 4, 9, 10, 11, 12
67. Wang, X., Zhang, H., Huang, W., Scott, M.R.: Cross-batch memory for embedding learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6388–6397 (2020) 3, 7

68. Wei, C., Shen, K., Chen, Y., Ma, T.: 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), https://openreview.net/forum?id=rC8sJ4i6kaH 3, 4, 8, 9
69. Wei, G., Lan, C., Zeng, W., Zhang, Z., Chen, Z.: Toalign: Task-oriented alignment for unsupervised domain adaptation. In: NeurIPS (2021) 4, 10, 11, 12
70. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3733–3742 (2018) 2, 4, 7
71. Xie, S., Zheng, Z., Chen, L., Chen, C.: Learning semantic representations for unsupervised domain adaptation. In: International Conference on Machine Learning. pp. 5423–5432 (2018) 2, 4, 10
72. Xu, R., Li, G., Yang, J., Lin, L.: Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1426–1435 (2019) 1, 2, 4, 10, 11, 12
73. Xu, Z., Huang, S., Zhang, Y., Tao, D.: Webly-supervised fine-grained visual categorization via deep domain adaptation. IEEE transactions on pattern analysis and machine intelligence **40**(5), 1100–1113 (2016) 4
74. Yang, L., Luo, P., Change Loy, C., Tang, X.: A large-scale car dataset for fine-grained categorization and verification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3973–3981 (2015) 9
75. Yang, L., Wang, Y., Gao, M., Shrivastava, A., Weinberger, K.Q., Chao, W.L., Lim, S.N.: Mico: Mixup co-training for semi-supervised domain adaptation. arXiv preprint arXiv:2007.12684 (2020) 9
76. Zhang, J., Ding, Z., Li, W., Ogunbona, P.: Importance weighted adversarial nets for partial domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8156–8164 (2018) 4
77. Zhang, N., Donahue, J., Girshick, R., Darrell, T.: Part-based r-cnns for fine-grained category detection. In: European conference on computer vision. pp. 834–849. Springer (2014) 3, 4, 11
78. Zhang, N., Farrell, R., Darrell, T.: Pose pooling kernels for sub-category recognition. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 3665–3672. IEEE (2012) 4, 11
79. Zheng, H., Fu, J., Mei, T., Luo, J.: Learning multi-attention convolutional neural network for fine-grained image recognition. In: Proceedings of the IEEE international conference on computer vision. pp. 5209–5217 (2017) 4
80. Zheng, H., Fu, J., Zha, Z.J., Luo, J.: 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. pp. 5012–5021 (2019) 4