MemSAC: Memory Augmented Sample Consistency for Large-Scale Domain Adaptation

Tarun Kalluri
UC San Diego

Astuti Sharma
UC San Diego

Manmohan Chandraker
UC San Diego

Abstract

Practical real world datasets with plentiful categories introduce additional challenges for unsupervised domain adaptation like small inter-class discriminability, that existing approaches relying only on domain invariance cannot handle. In this work we propose MemSAC, a novel adaptation approach to address the challenge of large-scale domain adaptation. The main idea behind our approach is to exploit sample level similarity across source and target domains to achieve discriminative transfer 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 arbitrary number of classes. Next, we propose and justify a novel variant of the popular 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 on multiple challenging transfer tasks designed for large scale adaptation on DomainNet and fine-grained adaptation on Caltech-UCSD birds datasets. We show significant improvements over prior approaches on each of these tasks, and provide in-depth analysis and insights to understand the effectiveness of our approach. Source code and trained models implemented using MemSAC are provided with the supplementary material and will be publicly released.

1 Introduction

It is well known that deep learning models do not *generalize* or *transfer* well in the presence of domain shift, when the distribution of test samples is significantly different from those used in training [52]. The pioneering work of [2, 3] quantifies this notion of transferability across domains in terms of $\mathcal{H}\Delta\mathcal{H}$-divergence for unsupervised domain adaptation (UDA). Numerous works have since been proposed aimed at minimizing various notions of divergence to improve transferability [4, 5, 11, 17, 21, 33–35, 37, 45, 53–55, 64]. While prior works demonstrate the utility of adversarial adaptation to combat domain shift, they are usually evaluated on datasets with a limited number of categories. It turns out that many adaptation techniques that achieve state-of-the-art accuracies on these standard benchmarks like Office-31 [43] and visDA [41] do not deliver similar benefits on datasets with large number of categories like DomainNet, sometimes performing worse than just using a model trained on source data alone [42]. This motivates us to extend the current literature of unsupervised domain adaptation towards designing approaches that scale well to large datasets with a wide variety of categories.

To this end, we propose MemSAC (MEMory augmented SAmple Consistency), which is designed as an efficient approach for large-scale adaptation capable of handling multitude of classes. The main idea behind MemSAC is to enforce local consistency among related cross-domain samples to achieve global domain alignment. We work with the intuition that extracting pairwise sample-level similarity relations is more robust to discriminative transfer across domains when presented with large-scale datasets [13]. Therefore, we extract positive and negative pairs across source and target

Preprint. Under review.

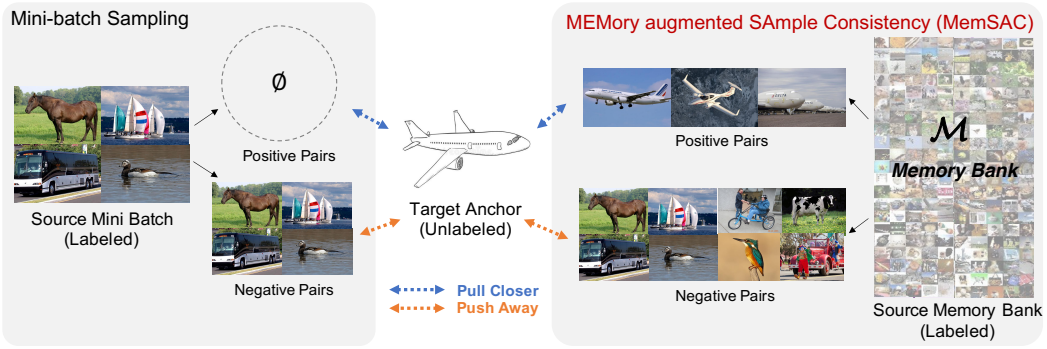

Figure 1: **MEMory augmented SAmple Consistency (MemSAC)** The proposed method uses a memory bank and a sample consistency loss to identify source samples across a large number of categories that likely belong to the same class as an unlabeled target example, then pulls them together in feature space while pushing away samples from all other classes.

datasets and propose a novel *cross-domain sample consistency* loss to pull similar samples from across domains closer to each other while pushing dissimilar samples farther apart in the feature space. We provide intuition for the effectiveness of our proposed loss by connecting it to the input-consistency regularization recently proposed in [61], which ensures that enforcing locally consistent prediction provides accuracy guarantees on unlabeled target data for unsupervised domain adaptation.

While having access to plentiful number of 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 [25, 60], 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. 1). Our architecture scales remarkably well with the number of categories, including the case of fine-grained adaptation [59] where all classes belong to a single subordinate category [6, 69].

In a significant advance for evaluation of unsupervised domain adaptation, we demonstrate the effectiveness of MemSAC on datasets with a large number of categories. On challenging transfer tasks on DomainNet designed for large scale adaptation, we achieve an average accuracy gain of $> 3\%$ compared to all current approaches. On fine-grained adaptation on CUB-Drawings, we achieve an accuracy of 64.23%, setting new state-of-the art that outperforms all prior approaches by a significant margin, indicating the utility of our approach on real world adaptation problems. In summary, MemSAC overcomes limitations of prior works in domain adaptation with respect to scalability, while being very effective on real-world datasets.

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 [2, 3, 43] between the source and target using an adversarial objective, resulting in domain invariant features [9, 17, 37, 46, 53–55, 63]. Since domain invariance alone does not guarantee discriminative features in target [30], recent approaches propose class aware adaptation to align class conditional distributions across source and target [28, 40, 44, 50, 63]. ATT [44] assigns pseudo-labels based on predictions from classifiers, MADA [40] uses separate adversarial networks for each class, while ILA [50] computes pairwise similarity between samples within a mini-batch for class aware adaptation. SAFN [64] and BSP [11] propose re-normalizing features to achieve transferability of source classifiers. However, none of these works explicitly address the issue of scalability to adaptation with a large number of categories.

While partial adaptation [7, 8, 67], open set adaptation [39, 47] and universal adaptation [49] allow training on real world source datasets with many categories, they, however, 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 more practical requirement, 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 [56, 58]. While fine grained classification on within domain samples has received much attention [6, 32, 51, 68–71], the problem of unsupervised domain adaptation across fine-grained categories is relatively less studied [15, 18, 59, 65]. Prior works often demand additional annotations in the form of attributes [18], target supervision [15], part annotations [65] or hierarchical relationships [59] 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, 22, 23, 61] in extracting visual representations from unlabeled data has attracted a lot of interest of late [10, 12, 19, 20, 25, 26, 38, 62]. A unifying idea in these 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. In our work, we propose a variant of contrastive loss to handle discriminative transfer, while also enforcing sample consistency across similar samples extracted from different domains.

3 Unsupervised domain 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 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 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, known as the *covariate shift assumption* [3]. 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. The objective function for MemSAC is given by

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

where $\mathcal{L}_{sup}$ is the supervised loss on source data, or the cross-entropy loss between the predicted and ground truth class probability distributions 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 (and dissimilarity) relations between samples from source and target domains (Eq. 4). λ_{adv} and λ_{sc} are the corresponding loss coefficients. The design and role of these losses is explained next. 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 [37] 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

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

We note that our contributions are complementary to the type of adversarial objective used and in supplementary, we show gains starting from a DANN [17] objective too.

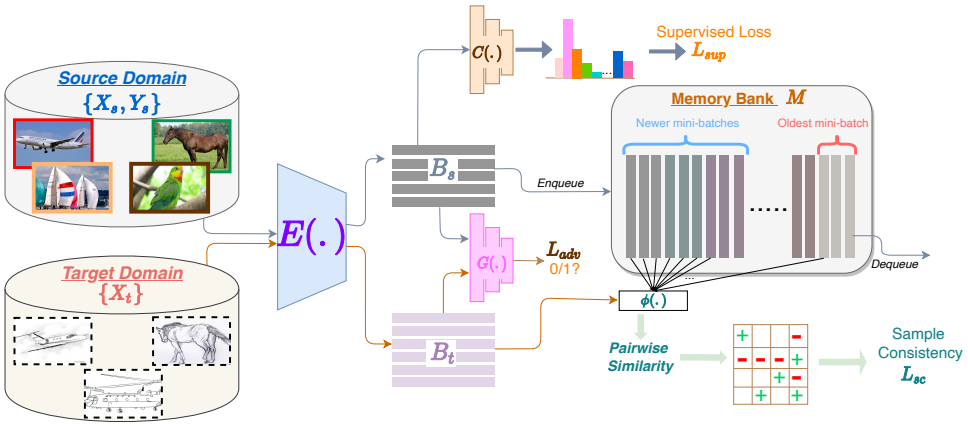

Figure 2: **An overview of MemSAC for domain adaptation** During each iteration, the 256-dim source feature embeddings computed using $\mathcal{E}$, along with their labels, are added to a memory bank $\mathcal{M}$ and the oldest set of features are removed. Pairwise similarities between each target 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 perform both local alignment and global adaptation.

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 following [50] as follows. For each sample x_t from a target mini-batch $\mathcal{B}_t$ as the anchor, we consider the *similar set* $\mathcal{B}_{s+}^{x_t} = \{x \in \mathcal{B}_s | f^*(x) = f^*(x_t)\}$ and dissimilar set $\mathcal{B}_{s-}^{x_t} = \mathcal{B}_s \backslash \mathcal{B}_{s+}^{x_t}$. We use this knowledge of sample-level similarity in the following *sample consistency loss*

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

where ϕ_{ij} measures the inverse euclidean similarity metric between two feature vectors i and j, ($\phi_{ij}{=}\phi(f_i, f_j){=}(1 + ||f_i - f_j||^2)^{-1}$) and τ is the temperature parameter used to scale the contributions of positive and negative pairs [10, 27]. $\mathcal{L}_{sc,\mathcal{B}}$ denotes the sample consistency loss computed using the mini-batch. Distinct from standard constrative loss [10, 38] that typically derives positive pairs from augmented versions of the same image, our loss in Eq. (3) is well-suited to handling multiple positive and negative pairs for each anchor.

3.2 kNN-based pseudo-labeling

There are two challenges to directly use the sample consistency loss in (3). Firstly, unlike prior approaches [10, 26, 38] 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 psuedo-labeling trick for all the target samples in a mini-batch. For each target sample $x_t \in \mathcal{B}_t$, we find k nearest neighbors from source domain, 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 $k{=}5$. We believe that 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).

3.3 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 ineffective 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 $\mathcal{B}_t$ would often contain many images with mutually non-intersecting categories with a high probability. While one solution is to increase the size of mini-batch, it comes with significant growth in memory and hence is not feasible.

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 using in the sample consistency loss. Furthermore, since the memory overhead of storing the features in the memory bank itself is negligible, 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 [25, 60, 62], 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 on source data after which the features stabilize and follow the *slow-drift* phenomenon in subsequent iterations [60], meaning that the distribution shift of features across mini-batches would be small. For updating the memory bank during training, we follow a queue based approach similar to XBM [60]. Specifically, after each iteration, we remove (*dequeue*) the oldest batch of features from the queue and insert (*enqueue*) the fresh mini-batch of features (computed as $\{\mathcal{E}(x)|x \in \mathcal{B}_s\}$) along with the corresponding labels. Alternative strategies for updating $\mathcal{M}$, such as using a momentum encoder [25], yielded similar results (discussed 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 in Sec. 3.2, as well as 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 $\mathcal{M}$ means that the k-NN pseudo labels are more reliable, with richer variety of positive and negative pairwise relations. The final sample consistency loss used in MemSAC is

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

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 [42] is designed as a large-scale dataset for domain adaptation, with 6 domains and a total of 500k images from 345 different categories consisting of an order of magnitude larger number of images and categories compared to prior benchmarks. Since some domains and classes have noisy labels and image duplication, we follow the protocol established in prior works [31, 48, 66] to use the 126 class version of DomainNet from 4 domains, namely real (**R**), clipart(**C**), sketch(**S**) and painting(**P**), showing results on all 12 transfer tasks across these 4 domains. The benefits and improvements using MemSAC persist even on complete DomainNet with 345 classes, with results presented in the supplementary material.

CUB (Caltech-UCSD birds) [58] is a challenging dataset originally proposed for fine-grained classification of 200 categories of birds, while *CUB-Drawings* [59] 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.

OfficeHome [57] is a popular benchmark for domain adaptation with 65 classes from four domains, namely Real World (**Rw**), Clipart(**Cl**), Product(**Pr**) and Art(**Ar**) with a total of 15.5k images. We show results on all 12 transfer tasks.

Source	Real→			Clipart→			Painting→			Sketches→			
Target	C	P	S	R	P	S	R	C	S	R	C	P	Avg.
Resnet-50	54.60	57.92	43.71	50.87	38.37	43.92	66.65	50.33	39.87	48.28	52.46	43.47	49.20
MCD [46]	52.94	57.29	40.38	55.71	43.69	47.57	67.80	51.88	44.95	56.83	56.32	50.83	52.18
RSDA [21]	54.60	61.54	50.94	56.56	45.50	48.63	60.41	45.74	48.64	58.62	56.09	54.00	53.44
DANN [17]	61.67	60.27	53.86	58.23	46.46	51.63	64.17	52.70	52.88	61.55	62.73	56.70	56.90
BSP [11]	55.16	60.80	48.60	58.73	45.66	55.47	65.18	48.59	48.58	61.40	56.78	55.79	55.06
SAFN [64]	55.81	64.82	48.50	58.68	49.96	52.42	**73.71**	56.25	53.54	64.32	60.65	59.53	58.18
CDAN [37]	70.41	66.87	57.73	61.61	50.90	54.72	68.47	59.43	55.49	64.27	64.22	59.14	61.11
PAN [59]	67.56	66.73	55.86	65.16	**58.87**	54.55	70.46	57.54	53.14	66.55	64.40	60.22	61.75
MemSAC	**73.23**$^{\pm 0.09}$	**70.46**$^{\pm 0.13}$	**61.51**$^{\pm 0.08}$	**66.51**$^{\pm 0.21}$	53.61$^{\pm 0.39}$	**58.79**$^{\pm 0.68}$	71.23$^{\pm 0.20}$	**63.17**$^{\pm 0.75}$	**58.11**$^{\pm 0.63}$	**67.60**$^{\pm 0.16}$	**68.77**$^{\pm 0.52}$	**64.09**$^{\pm 0.51}$	**64.76**

Table 1: Accuracy scores on 126 classes on DomainNet. **Bold** and underline indicate the best and next best methods respectively. All the baselines have been re-implemented by us.

Training Details We use a Resnet-50 [24] 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 the experiments and for all the methods. The reported accuracies are computed on the complete unlabeled target data in each case, following established protocol for UDA [37, 46, 59, 64]. The crucial hyper-parameters in our method are λ_{sc}, temperature τ and memory bank size $|\mathcal{M}|$. We choose $\lambda_{sc} = 0.1$ and $\tau = 0.007$ based on the adaptation performance on the $C \rightarrow D$ setting on the *CUB-Drawings* dataset. For fairness, we use these values across all the experiments on all the datasets. We reimplement all the baselines using the code and hyper-parameters provided online by the respective authors. Complete training details, along with the source code and trained models, are provided with the supplementary material.

4.1 Results

The results for the 12 transfer tasks on DomainNet are provided in Tab. 1. We compare MemSAC against traditional adversarial approaches (DANN [17], CDAN [37], MCD [46]) as well as the current state-of-the art (SAFN [64], BSP [11], RSDA [21]). We make the following observations.

Firstly, methods such as RSDA (53.44%) and SAFN (58.54%) that achieve best performance on smaller scale datasets (like Office-31 [43] and visDA-2017 [41]) provide only marginal or no benefits over the more traditional adversarial approaches such as DANN (56.90%) and CDAN (61.11%) on DomainNet with 126 classes, indicating that large-scale datasets needs different techniques for adaptation. Additionally, we also compare against PAN [59], which requires a label hierarchy as additional information for training, for which we use the DomainNet grouping proposed in [42] to construct one level of hierarchy. Even when provided with access to hierarchical labels in source, PAN (61.75%) achieves only a small improvement over CDAN. Finally, our method MemSAC, that combines global adaptation using a conditional adversarial approach and local alignment using sample consistency, achieves an average accuracy of 64.76%, with a significantly better performance than all the prior approaches across most of the tasks. These trends and benefits also hold for the much larger and noisier 345-class version of the DomainNet dataset, and results are included in the supplementary material.

Method	C → D	D → C	Avg.
Resnet-50* [24]	60.88	42.07	51.47
DANN* [17]	62.09	47.73	54.91
JAN [36]	62.42	40.37	51.40
ADDA [55]	60.12	40.65	50.36
MCD* [46]	54.38	39.79	47.08
SAFN* [64]	60.29	43.77	52.03
PAN [59]	67.40	50.92	59.16
CDAN* [37]	68.12	53.83	60.98
MemSAC	**71.25**$^{\pm 0.12}$	**57.21**$^{\pm 0.22}$	**64.23**

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. Baselines marked by (*) are reimplemented by us, rest taken from PAN [59].

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 [6, 14, 16, 68, 69], domain adaptation for fine grained categories is a relevant but less-addressed problem. PAN [59] uses label relations in source across 3 levels of hierarchy and obtains an average accuracy of 59.16%, while MemSAC obtains an average accuracy of 64.23%, thus outperforming all prior approaches on this challenging setting. This shows the benefit of enforcing sample consistency using MemSAC for adaptation in the presence of fine-grained categories. Furthermore, we also evaluate MemSAC on all 12 tasks from OfficeHome dataset with 65 classes in Tab. 3 to achieve an accuracy of 68.42% thus outperforming most established baselines, demonstrating the utility of our

Method	Ar → Cl	Ar → Pr	Ar → Rw	Cl → Ar	Cl → Pr	Cl → Rw	Pr → Ar	Pr → Cl	Pr → Rw	Rw → Ar	Rw → Cl	Rw → Pr	AVG
Resnet-50	34.9	50.0	58.0	37.4	41.9	46.2	38.5	31.2	60.4	53.9	41.2	59.9	46.1
DANN [17]	45.6	59.3	70.1	47.0	58.5	60.9	46.1	43.7	68.5	63.2	51.8	76.8	57.6
JAN [36]	45.9	61.2	68.9	50.4	59.7	61.0	45.8	43.4	70.3	63.9	52.4	76.8	58.3
CDAN [37]	50.7	70.6	76.0	57.6	70.0	70.0	57.4	50.9	77.3	70.9	56.7	81.6	65.8
BSP [11]	52.0	68.6	76.1	58.0	70.3	70.2	58.6	50.2	77.6	72.2	59.3	81.9	66.3
SAFN [64]	52.0	71.7	76.3	**64.2**	69.9	71.9	**63.7**	51.4	77.1	70.9	57.1	81.5	67.3
MemSAC	**53.10**	**73.7**	**77.8**	62.9	**71.22**	**72.32**	61.22	**51.93**	**79.22**	**75.0**	**59.39**	**83.35**	**68.42**

Table 3: Accuracy scores on 65 categories on OfficeHome [57] dataset. **Bold** and underline denote the best and next best performing methods respectively. Numbers for baselines taken from [11] and [64].

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	✓	✓	**71.25**	**57.21**	**64.23**

(a) Ablation on various components of loss function in Eq. (1).

τ	C→D	D→C	Avg. Acc
1.0	68.36	53.46	60.91
0.07	69.94	55.04	62.49
0.007	**71.25**	**57.21**	**64.23**

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

Similarity	ϕ_{ij}	C→D	D→C	Avg. Acc
Inv. Euc.	$(1+\|\|f_i - f_j\|\|^2)^{-1}$	**71.25**	**57.21**	**64.23**
Cosine	$f_i \cdot f_j$	69.94	57.00	63.47
Gaussian	$exp(-\|\|f_i - f_j\|\|^2)$	70.10	50.84	60.47

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

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.

approach even on medium-sized datasets. We next provide an in-depth analysis into the proposed method delineating its salient aspects.

4.2 Analysis and Discussion

Ablation studies We show the influence of various design choices of our method in Tab. 4. 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, 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 $\tau = 0.007$ gives the best performance on the inverse euclidean similarity metric. Finally, in Tab. 4c, we note that the performance using other choices of $\phi(.)$, namely *Cosine* similarity and *Gaussian* similarity is inferior to using *Euclidean* similarity, while also being more stable to train under severe domain shifts. A discussion in greater detail on these choices, along with more ablations, is presented in the supplementary material.

Category-wise accuracy for MemSAC We show the accuracies on every *coarse* category, along with the number of finer classes for each coarse category, in Fig. 3. 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 (+13.2%)*, *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

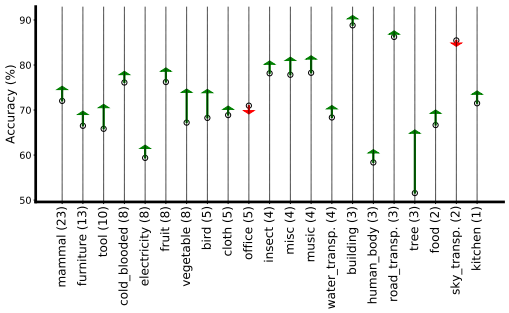

Figure 3: Category wise gain/drop in accuracy on **R→C** on DomainNet, compared to CDAN [37].

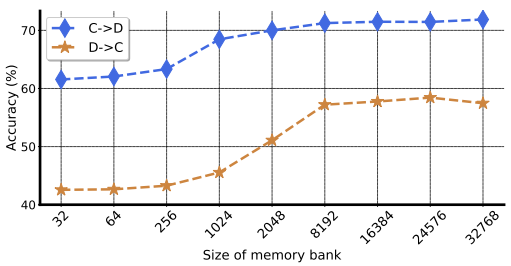

Figure 4: Effect of memory bank size on **C→D** and **D→C** on CUB-Drawings dataset.

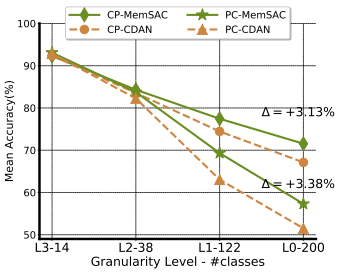

Figure 5: Comparison of accuracy vs. granularity of labels on **C**→**D** and **D**→**C** on CUB-Drawings dataset for 4 levels of label hierarchy.

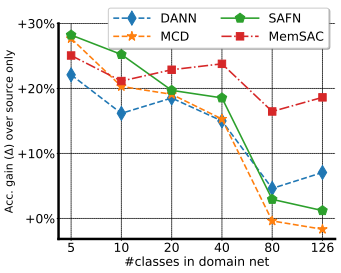

Figure 6: *Gain in accuracy* (Δ) over a model trained only on source data, using various adaptation approaches vs. the #classes from DomainNet for **R**→**C**.

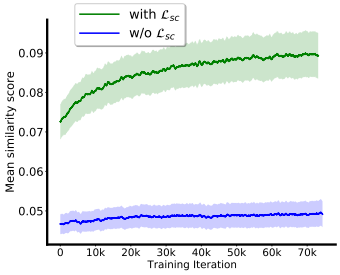

Figure 7: Mean similarity score for *within-class* samples vs. training iteration shown for **D**→**C** on CUB-Drawings.

negative pairs in the sample consistency loss and lead to noisy gradients. We show in Fig. 4 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). 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. 4, **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 [25, 62] since those works use a *single* positive sample from the memory and treat all other samples as negatives. But in our case, we can admit *multiple* positives and negatives into the sample consistency loss (Eq. (4)), so having multiple copies of the same instance is beneficial.

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 [59] 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. We observe from Fig. 5 that with coarser granularity, MemSAC performs as good as the baseline method CDAN, whereas with finer separation of the categories (L3 → L0), use of sample consistency loss provides greater benefit. This conforms our intuition that fine-grained domain adaptation requires enforcing sample level consistency.

MemSAC scales well with number of classes In Fig. 6, we randomly sample $\{5, 10, 20, 40, 80, 126\}$ classes from DomainNet, then plot the gain in accuracy achieved by each adaptation approach compared to a model trained only on the source data, for the **R**→**C** setting. We observe that for a smaller number of classes, prior methods like MCD [46] and AFN [64] perform the best. However, their benefits over a source-only model gradually decrease with increase in classes, while MemSAC (shown in red) still delivers non-trivial benefits even for a large number of classes.

MemSAC increases similarity of within-class samples 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. We test this hypothesis in Fig. 7, in which 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. We observe that using the proposed loss, the similarity score is much higher and improves with training compared to the baseline with no such constraint, which also reflects in the overall target accuracy (Tab. 2).

Computational cost and resources In general, MemSAC incurs negligible overhead in memory during training and none at all during inference. Training for each task is performed on a regular 12GB GPU. We used a total of 2400 GPU hours to carry all the experiments in this paper. For the experiments on MemSAC, we report mean and standard deviation over 3 random trials.

4.3 Discussion

Our sample consistency algorithm stems from recent success of contrastive learning [10, 12, 19, 20, 25, 26, 29, 38, 62] in computer vision. Recently, Wei et al. [61] 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 that encourages predictions of a classifier C parametrized by θ to be *consistent* on neighboring examples by minimizing a regularization objective $R(\theta)$ given by. $\min_\theta \mathbb{E}_x[\max_{x' \in \mathcal{N}(x)} \mathbf{1}(C(x;\theta) \neq C(x';\theta))]$, where $\mathcal{N}(x)$ is the neighborhood of a sample x (Eq 1.2 in [61]). We now show the connections that can be drawn between our loss and the theory proposed in [61]. For this purpose, we work with the following approximations. Firstly, we approximate the neighborhood $\mathcal{N}(x)$ of a data sample x in our case with the *similar set* defined in Sec. 3.1, that is $\mathcal{N}(x) = \mathcal{B}^x_+$. Furthermore, we approximate the hard condition that the classifier outputs of two images be equal ($\mathbf{1}(C(x;\theta) \neq C(x';\theta))$), with their soft probability $\mathbf{Pr}(C(x;\theta) \neq C(x';\theta))$. Starting with the above objective, we have the following.

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

where we used the softmax similarity between samples x, x' in the feature space as a proxy for the equality of their classifier outputs. 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 [61]. Furthermore, optimizing such an objective is shown to achieve bounded target error (Theorem 4.3 in [61]) for unsupervised domain adaptation under suitable assumptions. To the best of our knowledge, we are the first to instantiate the regularization proposed in [61] for large scale domain adaptation, and showcase its effectiveness in achieving significant empirical gains.

Limitations, future work and impact Although we report outstanding performance using MemSAC, we assume that we 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 approaches fully supervised performances. To further advance research in these areas, our code and models will be made publicly available to the community.

5 Conclusion

We have proposed MemSAC, a novel approach for unsupervised domain adaptation designed to handle a large number of categories. We propose a novel 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 intuition for the effectiveness of such memory-augmented sample consistency loss for achieving large-scale domain alignment and discriminative transfer. In extensive experiments across the main paper and supplementary material, 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).

References

[1] S. Arora, H. Khandeparkar, M. Khodak, O. Plevrakis, and N. Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019. 3

[2] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19:137–144, 2006. 1, 2

[3] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010. 1, 2, 3

[4] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan. Domain separation networks. In *Advances in neural information processing systems*, pages 343–351, 2016. 1

[5] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. 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. 1

[6] S. Branson, G. Van Horn, S. Belongie, and P. Perona. Bird species categorization using pose normalized deep convolutional nets. *arXiv preprint arXiv:1406.2952*, 2014. 2, 3, 6

[7] Z. Cao, M. Long, J. Wang, and M. I. Jordan. Partial transfer learning with selective adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2724–2732, 2018. 2

[8] Z. Cao, L. Ma, M. Long, and J. Wang. Partial adversarial domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 135–150, 2018. 2

[9] C. Chen, W. Xie, W. Huang, Y. Rong, X. Ding, Y. Huang, T. Xu, and J. Huang. Progressive feature alignment for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 627–636, 2019. 2

[10] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 3, 4, 8

[11] X. Chen, S. Wang, M. Long, and J. Wang. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In *International conference on machine learning*, pages 1081–1090. PMLR, 2019. 1, 2, 6, 7

[12] X. Chen, H. Fan, R. Girshick, and K. He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 3, 8

[13] Y. Chen, E. K. Garcia, M. R. Gupta, A. Rahimi, and L. Cazzanti. Similarity-based classification: Concepts and algorithms. *Journal of Machine Learning Research*, 10(3), 2009. 1

[14] Y. Chen, Y. Bai, W. Zhang, and T. 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. 6

[15] Y. Cui, Y. Song, C. Sun, A. Howard, and S. 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. 3

[16] A. Dubey, O. Gupta, R. Raskar, and N. Naik. Maximum-entropy fine-grained classification. *arXiv preprint arXiv:1809.05934*, 2018. 6

[17] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 1, 2, 3, 6, 7

[18] T. Gebru, J. Hoffman, and L. 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. 3

[19] D. Gordon, K. Ehsani, D. Fox, and A. Farhadi. Watching the world go by: Representation learning from unlabeled videos. *arXiv preprint arXiv:2003.07990*, 2020. 3, 8

[20] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020. 3, 8

[21] X. Gu, J. Sun, and Z. 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. 1, 6

[22] M. Gutmann and A. 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. 3

[23] R. Hadsell, S. Chopra, and Y. 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. 3

[24] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[25] K. He, H. Fan, Y. Wu, S. Xie, and R. 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. 2, 3, 5, 8

[26] O. J. Hénaff, A. Srinivas, J. De Fauw, A. Razavi, C. Doersch, S. Eslami, and A. v. d. Oord. Data-efficient image recognition with contrastive predictive coding. *arXiv preprint arXiv:1905.09272*, 2019. 3, 4, 8

[27] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 4

[28] G. Kang, L. Jiang, Y. Yang, and A. 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. 2

[29] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020. 8

[30] A. Kumar, P. Sattigeri, K. Wadhawan, L. Karlinsky, R. Feris, B. Freeman, and G. Wornell. Co-regularized alignment for unsupervised domain adaptation. In *Advances in Neural Information Processing Systems*, pages 9345–9356, 2018. 2

[31] J. Liang, D. Hu, and J. Feng. Combating domain shift with self-taught labeling. *arXiv preprint arXiv:2007.04171*, 2020. 5

[32] T.-Y. Lin, A. RoyChowdhury, and S. Maji. Bilinear cnn models for fine-grained visual recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 1449–1457, 2015. 3

[33] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu. Transfer feature learning with joint distribution adaptation. In *Proceedings of the IEEE international conference on computer vision*, pages 2200–2207, 2013. 1

[34] M. Long, Y. Cao, J. Wang, and M. Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015.

[35] M. Long, H. Zhu, J. Wang, and M. I. Jordan. Unsupervised domain adaptation with residual transfer networks. In *Advances in neural information processing systems*, pages 136–144, 2016. 1

[36] M. Long, H. Zhu, J. Wang, and M. I. Jordan. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, pages 2208–2217. PMLR, 2017. 6, 7

[37] M. Long, Z. Cao, J. Wang, and M. I. Jordan. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pages 1640–1650, 2018. 1, 2, 3, 6, 7

[38] I. Misra and L. v. d. 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. 3, 4, 8

[39] P. Panareda Busto and J. Gall. Open set domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 754–763, 2017. 2

[40] Z. Pei, Z. Cao, M. Long, and J. Wang. Multi-adversarial domain adaptation. *arXiv preprint arXiv:1809.02176*, 2018. 2

[41] X. Peng, B. Usman, N. Kaushik, J. Hoffman, D. Wang, and K. Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017. 1, 6

[42] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1406–1415, 2019. 1, 5, 6

[43] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010. 1, 2, 6

[44] K. Saito, Y. Ushiku, and T. Harada. Asymmetric tri-training for unsupervised domain adaptation. *arXiv preprint arXiv:1702.08400*, 2017. 2

[45] K. Saito, Y. Ushiku, T. Harada, and K. Saenko. Adversarial dropout regularization. *arXiv preprint arXiv:1711.01575*, 2017. 1

[46] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3723–3732, 2018. 2, 6, 8

[47] K. Saito, S. Yamamoto, Y. Ushiku, and T. Harada. Open set domain adaptation by backpropagation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 153–168, 2018. 2

[48] K. Saito, D. Kim, S. Sclaroff, T. Darrell, and K. Saenko. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8050–8058, 2019. 5

[49] K. Saito, D. Kim, S. Sclaroff, and K. Saenko. Universal domain adaptation through self supervision. *arXiv preprint arXiv:2002.07953*, 2020. 2

[50] A. Sharma, T. Kalluri, and M. Chandraker. Instance level affinity-based transfer for unsupervised domain adaptation. *arXiv preprint arXiv:2104.01286*, 2021. 2, 4

[51] M. Sun, Y. Yuan, F. Zhou, and E. 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. 3

[52] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011. 1

[53] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014. 1, 2

[54] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE international conference on computer vision*, pages 4068–4076, 2015.

[55] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017. 1, 2, 6

[56] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018. 3

[57] H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5018–5027, 2017. 5, 7

[58] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 3, 5

[59] S. Wang, X. Chen, Y. Wang, M. Long, and J. 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. 2, 3, 5, 6, 8

[60] X. Wang, H. Zhang, W. Huang, and M. 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. 2, 5

[61] C. Wei, K. Shen, Y. Chen, and T. Ma. Theoretical analysis of self-training with deep networks on unlabeled data. *arXiv preprint arXiv:2010.03622*, 2020. 2, 3, 8, 9

[62] Z. Wu, Y. Xiong, S. X. Yu, and D. 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. 3, 5, 8

[63] S. Xie, Z. Zheng, L. Chen, and C. Chen. Learning semantic representations for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 5423–5432, 2018. 2

[64] R. Xu, G. Li, J. Yang, and L. 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. 1, 2, 6, 7, 8

[65] Z. Xu, S. Huang, Y. Zhang, and D. Tao. Webly-supervised fine-grained visual categorization via deep domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 40(5):1100–1113, 2016. 3

[66] L. Yang, Y. Wang, M. Gao, A. Shrivastava, K. Q. Weinberger, W.-L. Chao, and S.-N. Lim. Mico: Mixup co-training for semi-supervised domain adaptation. *arXiv preprint arXiv:2007.12684*, 2020. 5

[67] J. Zhang, Z. Ding, W. Li, and P. Ogunbona. Importance weighted adversarial nets for partial domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8156–8164, 2018. 2

[68] N. Zhang, R. Farrell, and T. Darrell. Pose pooling kernels for sub-category recognition. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3665–3672. IEEE, 2012. 3, 6

[69] N. Zhang, J. Donahue, R. Girshick, and T. Darrell. Part-based r-cnns for fine-grained category detection. In *European conference on computer vision*, pages 834–849. Springer, 2014. 2, 6

[70] H. Zheng, J. Fu, T. Mei, and J. 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.

[71] H. Zheng, J. Fu, Z.-J. Zha, and J. 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. 3