Moment Matching for Multi-Source Domain Adaptation

Xingchao Peng
Boston University
xpeng@bu.edu

Qinxun Bai
Horizon Robotics
qinxun.bai@horizon.ai

Xide Xia
Boston University
xidexia@bu.edu

Zijun Huang
Columbia University
zijun.huang@columbia.edu

Kate Saenko
Boston University
saenko@bu.edu

Bo Wang
Vector Institute & Peter Munk Cardiac Center
bowang@vectorinstitute.ai

Abstract

Conventional unsupervised domain adaptation (UDA) assumes that training data are sampled from a single domain. This neglects the more practical scenario where training data are collected from multiple sources, requiring multi-source domain adaptation. We make three major contributions towards addressing this problem. First, we collect and annotate by far the largest UDA dataset, called DomainNet, which contains six domains and about 0.6 million images distributed among 345 categories, addressing the gap in data availability for multi-source UDA research. Second, we propose a new deep learning approach, Moment Matching for Multi-Source Domain Adaptation (M^3SDA), which aims to transfer knowledge learned from multiple labeled source domains to an unlabeled target domain by dynamically aligning moments of their feature distributions. Third, we provide new theoretical insights specifically for moment matching approaches in both single and multiple source domain adaptation. Extensive experiments are conducted to demonstrate the power of our new dataset in benchmarking state-of-the-art multi-source domain adaptation methods, as well as the advantage of our proposed model. Dataset and Code are available at http://ai.bu.edu/M3SDA/

1. Introduction

Generalizing models learned on one visual domain to novel domains has been a major obstacle in the quest for universal object recognition. The performance of the learned models degrades significantly when testing on novel domains due to the presence of *domain shift* [36].

Recently, transfer learning and domain adaptation methods have been proposed to mitigate the domain gap. For example, several UDA methods [27, 41, 25] incorporate Maximum Mean Discrepancy loss into a neural network to diminish the domain discrepancy; other models introduce different learning schema to align the source and target domains, including aligning second order correlation [39, 32],

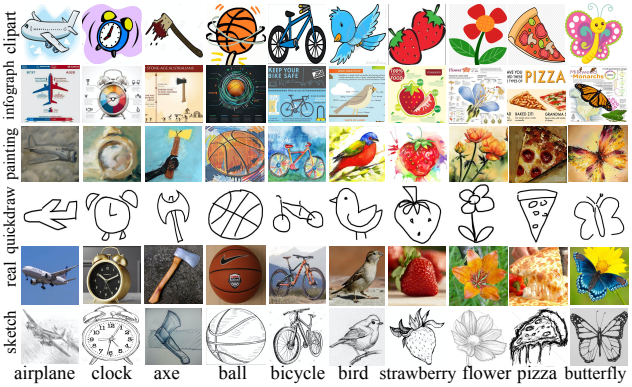

Figure 1. We address **Multi-Source Domain Adaptation** where source images come from multiple domains. We collect a large scale dataset, DomainNet, with six domains, 345 categories, and ∼0.6 million images and propose a model (M^3SDA) to transfer knowledge from multiple source domains to an unlabeled target domain.

moment matching [47], adversarial domain confusion [40, 8, 38] and GAN-based alignment [50, 15, 23]. However, most of current UDA methods assume that source samples are collected from a single domain. This assumption neglects the more practical scenarios where labeled images are typically collected from multiple domains. For example, the training images can be taken under different weather or lighting conditions, share different visual cues, and even have different modalities (as shown in Figure 1).

In this paper, we consider multi-source domain adaptation (MSDA), a more difficult but practical problem of knowledge transfer from multiple distinct domains to one unlabeled target domain. The main challenges in the research of MSDA are that: (1) the source data has multiple domains, which hampers the effectiveness of mainstream single UDA method; (2) source domains also possess domain shift with each other; (3) the lack of large-scale multi-domain dataset hinders the development of MSDA models.

In the context of MSDA, some theoretical analysis [1, 28, 4, 49, 14] has been proposed for multi-source domain

Dataset	Year	Images	Classes	Domains	*Description*
Digit-Five	-	∼100,000	10	5	digit
Office [37]	2010	4,110	31	3	office
Office-Caltech [11]	2012	2,533	10	4	office
CAD-Pascal [33]	2015	12,000	20	6	animal,vehicle
Office-Home [43]	2017	15,500	65	4	office, home
PACS [21]	2017	9,991	7	4	animal, stuff
Open MIC [17]	2018	16,156	-	-	museum
Syn2Real [35]	2018	280,157	12	3	animal,vehicle
DomainNet (Ours)	-	**569,010**	**345**	**6**	see *Appendix*

Table 1. A collection of most notable datasets to evaluate domain adaptation methods. Specifically, "Digit-Five" dataset indicates five most popular digit datasets (*MNIST* [19], *MNIST-M* [8], Synthetic Digits [8], *SVHN*, and *USPS*) which are widely used to evaluate domain adaptation models. Our dataset is challenging as it contains more images, categories, and domains than other datasets. (see Table 10, Table 11, and Table 12 in *Appendix* for detailed categories.)

adaptation (MSDA). Ben-David et al [1] pioneer this direction by introducing an $\mathcal{H}\Delta\mathcal{H}$-divergence between the weighted combination of source domains and target domain. More applied works [6, 45] use an adversarial discriminator to align the multi-source domains with the target domain. However, these works focus only on aligning the source domains with the target, neglecting the domain shift between the source domains. Moreover, $\mathcal{H}\Delta\mathcal{H}$-divergence based analysis does not directly correspond to moment matching approaches.

In terms of data, research has been hampered due to the lack of large-scale domain adaptation datasets, as state-of-the-art datasets contain only a few images or have a limited number of classes. Many domain adaptation models exhibit saturation when evaluated on these datasets. For example, many methods achieve ∼90 accuracy on the popular Office [37] dataset; Self-Ensembling [7] reports ∼99% accuracy on the "Digit-Five" dataset and ∼92% accuracy on Syn2Real [35] dataset.

In this paper, we first collect and label a new multi-domain dataset called **DomainNet**, aiming to overcome benchmark saturation. Our dataset consists of six distinct domains, 345 categories and ∼0.6 million images. A comparison of DomainNet and several existing datasets is shown in Table 1, and example images are illustrated in Figure 1. We evaluate several state-of-the-art single domain adaptation methods on our dataset, leading to surprising findings (see Section 5). We also extensively evaluate our model on existing datasets and on DomainNet and show that it outperforms the existing single- and multi-source approaches.

Secondly, we propose a novel approach called M^{3}SDA to tackle MSDA task by aligning the source domains with the target domain, and aligning the source domains with each other simultaneously. We dispose multiple complex adversarial training procedures presented in [45], but directly align the moments of their deep feature distributions, leading to a more robust and effective MSDA model. To our best knowledge, we are the first to empirically demonstrate that aligning the source domains is beneficial for MSDA tasks.

Finally, we extend existing theoretical analysis [1, 14, 49] to the case of moment-based divergence between source and target domains, which provides new theoretical insight specifically for moment matching approaches in domain adaptation, including our approach and many others.

2. Related Work

Domain Adaptation Datasets Several notable datasets that can be utilized to evaluate domain adaptation approaches are summarized in Table 1. The Office dataset [37] is a popular benchmark for office environment objects. It contains 31 categories captured in three domains: office environment images taken with a high quality camera (DSLR), office environment images taken with a low quality camera (Webcam), and images from an online merchandising website (Amazon). The office dataset and its extension, Office-Caltech10 [11], have been used in numerous domain adaptation papers [25, 40, 27, 39, 45], and the adaptation performance has reached ∼90% accuracy. More recent benchmarks [43, 17, 34] are proposed to evaluate the effectiveness of domain adaptation models. However, these datasets are small-scale and limited by their specific environments, such as *office*, *home*, and *museum*. Our dataset contains about 600k images, distributed in 345 categories and 6 distinct domains. We capture various object divisions, ranging from *furniture*, *cloth*, *electronic* to *mammal, building*, etc.

Single-source UDA Over the past decades, various single-source UDA methods have been proposed. These methods can be taxonomically divided into three categories. The first category is the discrepancy-based DA approach, which utilizes different metric learning schemas to diminish the domain shift between source and target domains. Inspired by the kernel two-sample test [12], Maximum Mean Discrepancy (MMD) is applied to reduce distribution shift in various methods [27, 41, 9, 44]. Other commonly used methods include correlation alignment [39, 32], Kullback-Leibler (KL) divergence [51], and $\mathcal{H}$ divergence [1]. The second category is the adversarial-based approach [24, 40]. A domain discriminator is leveraged to encourage the domain confusion by an adversarial objective. Among these approaches, generative adversarial networks are widely used to learn domain-invariant features as well to generate fake source or target data. Other frameworks utilize only adversarial loss to bridge two domains. The third category is reconstruction-based, which assumes the data reconstruction helps the DA models to learn domain-invariant features. The reconstruction is obtained via an encoder-

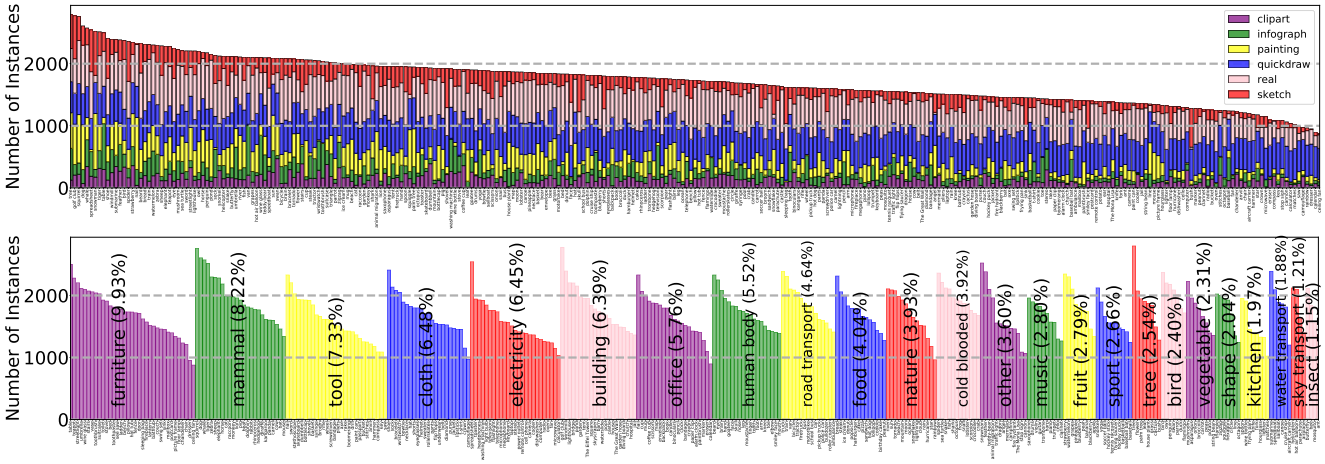

Figure 2. **Statistics for our DomainNet dataset.** The two plots show object classes sorted by the total number of instances. The top figure shows the percentages each domain takes in the dataset. The bottom figure shows the number of instances grouped by 24 different divisions. Detailed numbers are shown in Table 10, Table 11 and Table 12 in *Appendix*. (Zoom in to see the exact class names!)

decoder [3, 10] or a GAN discriminator, such as dualGAN [46], cycle-GAN [50], disco-GAN [16], and CyCADA [15]. Though these methods make progress on UDA, few of them consider the practical scenario where training data are collected from multiple sources. Our paper proposes a model to tackle multi-source domain adaptation, which is a more general and challenging scenario.

Multi-Source Domain Adaptation Compared with single source UDA, multi-source domain adaptation assumes that training data from multiple sources are available. Originated from the early theoretical analysis [1, 28, 4], MSDA has many practical applications [45, 6]. Ben-David et al [1] introduce an $\mathcal{H}\Delta\mathcal{H}$-divergence between the weighted combination of source domains and target domain. Crammer et al [4] establish a general bound on the expected loss of the model by minimizing the empirical loss on the nearest k sources. Mansour et al [28] claim that the target hypothesis can be represented by a weighted combination of source hypotheses. In the more applied works, Deep Cocktail Network (DCTN) [45] proposes a k-way domain discriminator and category classifier for digit classification and real-world object recognition. Hoffman et al [14] propose normalized solutions with theoretical guarantees for cross-entropy loss, aiming to provide a solution for the MSDA problem with very practical benefits. Duan et al [6] propose *Domain Selection Machine* for event recognition in consumer videos by leveraging a large number of loosely labeled web images from different sources. Different from these methods, our model directly matches all the distributions by matching the moments. Moreover, we provide a concrete proof of why matching the moments of multiple distributions works for multi-source domain adaptation.

Moment Matching The moments of distributions have been studied by the machine learning community for a long time. In order to diminish the domain discrepancy between two domains, different moment matching schemes have been proposed. For example, MMD matches the first moments of two distributions. Sun et al [39] propose an approach that matches the second moments. Zhang et al [48] propose to align infinte-dimensional covariance matrices in RKHS. Zellinger et al [47] introduce a moment matching regularizer to match high moments. As the generative adversarial network (GAN) becomes popular, many GAN-based moment matching approaches have been proposed. *McGAN* [29] utilizes a GAN to match the mean and covariance of feature distributions. GMMN [22] and MMD GAN [20] are proposed for aligning distribution moments with generative neural networks. Compared to these methods, our work focuses on matching distribution moments for multiple domains and more importantly, we demonstrate that this is crucial for multi-source domain adaptation.

3. The DomainNet dataset

It is well-known that deep models require massive amounts of training data. Unfortunately, existing datasets for visual domain adaptation are either small-scale or limited in the number of categories. We collect by far the largest domain adaptation dataset to date, **DomainNet** . The DomainNet contains six domains, with each domain containing 345 categories of common objects, as listed in Table 10, Table 11, and Table 12 (see *Appendix*). The domains include **Clipart** (*clp*, see *Appendix*, Figure 9): collection of clipart images; **Infograph** (*inf*, see Figure 10): infographic images with specific object; **Painting** (*pnt*, see Figure 11): artistic depictions of objects in the form of paintings; **Quickdraw** (*qdr*, see Figure 12): drawings of the worldwide players of game "Quick Draw!"[1]; **Real** (*rel*, see Figure 13): photos and real world images; and **Sketch**

[1]https://quickdraw.withgoogle.com/data

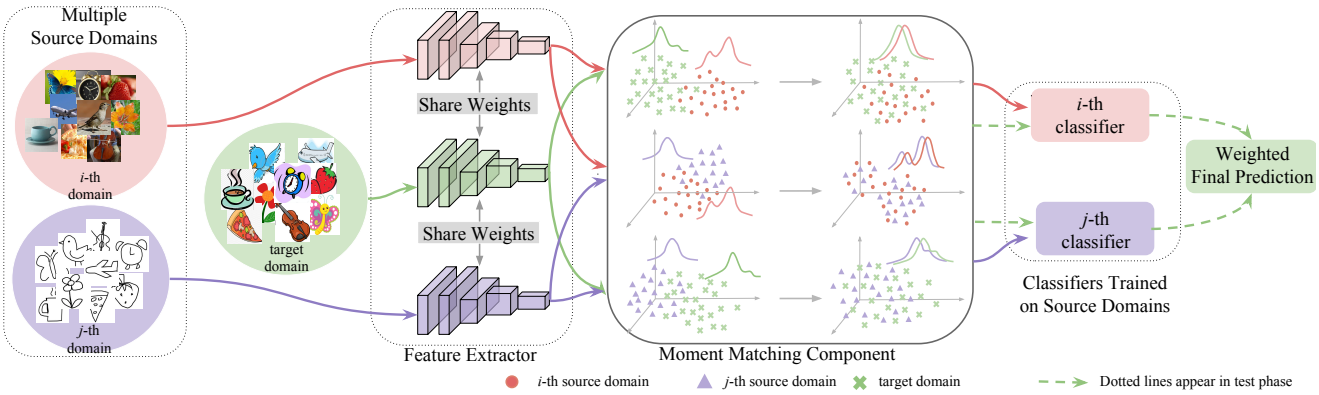

Figure 3. The framework of **Moment Matching for Multi-source Domain Adaptation** (M^{3}SDA). Our model consists of three components: i) feature extractor, ii) moment matching component, and iii) classifiers. Our model takes multi-source annotated training data as input and transfers the learned knowledge to classify the unlabeled target samples. Without loss of generality, we show the *i*-th domain and *j*-th domain as an example. The feature extractor maps the source domains into a common feature space. The moment matching component attempts to match the *i*-th and *j*-th domains with the target domain, as well as matching the *i*-th domain with the *j*-th domain. The final predictions of target samples are based on the weighted outputs of the *i*-th and *j*-th classifiers. (Best viewed in color!)

(*skt*, see Figure 14): sketches of specific objects.

The images from *clipart*, *infograph*, *painting*, *real*, and *sketch* domains are collected by searching a category name combined with a domain name (*e.g.* "aeroplane painting") in different image search engines. One of the main challenges is that the downloaded data contain a large portion of outliers. To clean the dataset, we hire 20 annotators to manually filter out the outliers. This process took around 2,500 hours (more than 2 weeks) in total. To control the annotation quality, we assign two annotators to each image, and only take the images agreed by both annotators. After the filtering process, we keep 423.5k images from the 1.2 million images crawled from the web. The dataset has an average of around 150 images per category for *clipart* and *infograph* domain, around 220 per category for *painting* and *sketch* domain, and around 510 for *real* domain. A statistical overview of the dataset is shown in Figure 2.

The *quickdraw* domain is downloaded directly from https://quickdraw.withgoogle.com/. The raw data are presented as a series of discrete points with temporal information. We use the B-spline [5] algorithm to connect all the points in each strike to get a complete drawing. We choose 500 images for each category to form the *quickdraw* domain, which contains 172.5k images in total.

4. Moment Matching for Multi-Source DA

Given $\mathcal{D}_S = \{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_N\}$ the collection of labeled source domains and $\mathcal{D}_T$ the unlabeled target domain, where all domains are defined by bounded rational measures on input space $\mathcal{X}$, the multi-source domain adaptation problem aims to find a hypothesis in the given hypothesis space $\mathcal{H}$, which minimizes the testing target error on $\mathcal{D}_T$.

Definition 1. *Assume* $\mathbf{X}_1$, $\mathbf{X}_2$, ..., $\mathbf{X}_N$, $\mathbf{X}_T$ *are collections of i.i.d. samples from* $\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_N, \mathcal{D}_T$ *respectively, then the Moment Distance between* $\mathcal{D}_S$ *and* $\mathcal{D}_T$ *is defined as*

$$MD^2(\mathcal{D}_S, \mathcal{D}_T) = \sum_{k=1}^{2} \Big(\frac{1}{N} \sum_{i=1}^{N} \|\mathbb{E}(\mathbf{X}_i^k) - \mathbb{E}(\mathbf{X}_T^k)\|_2 + \binom{N}{2}^{-1} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \|\mathbb{E}(\mathbf{X}_i^k) - \mathbb{E}(\mathbf{X}_j^k)\|_2 \Big). \quad (1)$$

M^{3}SDA We propose a moment-matching model for MSDA based on deep neural networks. As shown in Figure 3, our model comprises of a feature extractor G, a moment-matching component, and a set of N classifiers $\mathcal{C} = \{C_1, C_2, ..., C_N\}$. The feature extractor G maps $\mathcal{D}_S$, $\mathcal{D}_T$ to a common latent feature space. The moment matching component minimizes the moment-related distance defined in Equation 1. The N classifiers are trained on the annotated source domains with cross-entropy loss. The overall objective function is:

$$\min_{G,\mathcal{C}} \sum_{i=1}^{N} \mathcal{L}_{\mathcal{D}_i} + \lambda \min_{G} MD^2(\mathcal{D}_S, \mathcal{D}_T), \quad (2)$$

where $\mathcal{L}_{\mathcal{D}_i}$ is the softmax cross entropy loss for the classifier C_i on domain $\mathcal{D}_i$, and λ is the trade-off parameter.

M^{3}SDA assumes that $p(y|x)$ will be aligned automatically when aligning $p(x)$, which might not hold in practice. To mitigate this limitation, we further propose M^{3}SDA-β.

M^{3}SDA-β In order to align $p(y|x)$ and $p(x)$ at the same time, we follow the training paradigm proposed by [38]. In particular, we leverage two classifiers per domain to form N pairs of classifiers $\mathcal{C}' = \{(C_1, C_1'), (C_2, C_2'), ..., (C_N, C_N')\}$. The training procedure includes three steps. **i).** We train G and $\mathcal{C}'$ to classify the multi-source samples correctly. The objective is similar

to Equation 2. **ii).** We then train the classifier pairs for a fixed G. The goal is to make the discrepancy of each pair of classifiers as large as possible on the target domain. For example, the outputs of C_1 and $C_1{}'$ should possess a large discrepancy. Following [38], we define the discrepancy of two classifiers as the L1-distance between the outputs of the two classifiers. The objective is:

$$\min_{C'} \sum_{i=1}^{N} \mathcal{L}_{\mathcal{D}_i} - \sum_{i}^{N} |P_{C_i}(D_T) - P_{C_i{}'}(D_T)|, \quad (3)$$

where $P_{C_i}(D_T)$, $P_{C_i{}'}(D_T)$ denote the outputs of C_i, $C_i{}'$ respectively on the target domain. **iii).** Finally, we fix $\mathcal{C}'$ and train G to minimize the discrepancy of each classifier pair on the target domain. The objective function is as follows:

$$\min_{G} \sum_{i}^{N} |P_{C_i}(D_T) - P_{C_i{}'}(D_T)|, \quad (4)$$

These three training steps are performed periodically until the whole network converges.

Ensemble Schema In the testing phase, testing data from the target domain are forwarded through the feature generator and the N classifiers. We propose two schemas to combine the outputs of the classifiers:

- average the outputs of the classifiers, marked as M^{3}SDA*
- Derive a weight vector $\mathcal{W} = (w_1, \ldots, w_{N-1})$ ($\sum_{i=1}^{N-1} w_i = 1$, assuming N-th domain is the target). The final prediction is the weighted average of the outputs.

To this end, how to derive the weight vector becomes a critical problem. The main philosophy of the weight vector is to make it represent the intrinsic closeness between the target domain and source domains. In our setting, the weighted vector is derived by the source-only accuracy between the i-th domain and the N-th domain, *i.e.* $w_i = acc_i / \sum_{j=1}^{N-1} acc_j$.

4.1. Theoretical Insight

Following [1], we introduce a rigorous model of multi-source domain adaptation for binary classification. A domain $\mathcal{D} = (\mu, f)$ is defined by a probability measure (distribution) μ on the input space $\mathcal{X}$ and a labeling function $f : \mathcal{X} \to \{0, 1\}$. A hypothesis is a function $h : \mathcal{X} \to \{0, 1\}$. The probability that h disagrees with the domain labeling function f under the domain distribution μ is defined as:

$$\epsilon_{\mathcal{D}}(h) = \epsilon_{\mathcal{D}}(h, f) = \mathbb{E}_{\mu}[|h(\boldsymbol{x}) - f(\boldsymbol{x})|]. \quad (5)$$

For a source domain $\mathcal{D}_S$ and a target domain $\mathcal{D}_T$, we refer to the source error and the target error of a hypothesis h as $\epsilon_S(h) = \epsilon_{\mathcal{D}_S}(h)$ and $\epsilon_T(h) = \epsilon_{\mathcal{D}_T}(h)$ respectively. When the expectation in Equation 5 is computed with respect to an empirical distribution, we denote the corresponding empirical error by $\hat{\epsilon}_{\mathcal{D}}(h)$, such as $\hat{\epsilon}_S(h)$ and $\hat{\epsilon}_T(h)$. In particular, we examine algorithms that minimize convex combinations of source errors, i.e., given a weight vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)$ with $\sum_{j=1}^{N} \alpha_j = 1$, we define the $\boldsymbol{\alpha}$-weighted source error of a hypothesis h as $\epsilon_{\boldsymbol{\alpha}}(h) = \sum_{j=1}^{N} \alpha_j \epsilon_j(h)$, where $\epsilon_j(h)$ is the shorthand of $\epsilon_{\mathcal{D}_j}(h)$. The empirical $\boldsymbol{\alpha}$-weighted source error can be defined analogously and denoted by $\hat{\epsilon}_{\boldsymbol{\alpha}}(h)$.

Previous theoretical bounds [1, 14, 49] on the target error are based on the $\mathcal{H}\Delta\mathcal{H}$-divergence between the source and target domains. While providing theoretical insights for general multi-source domain adaptation, these $\mathcal{H}\Delta\mathcal{H}$-divergence based bounds do not directly motivate moment-based approaches. In order to provide a specific insight for moment-based approaches, we introduce the k-th order cross-moment divergence between domains, denoted by $d_{CM^k}(\cdot, \cdot)$, and extend the analysis in [1] to derive the following moment-based bound for multi-source domain adaptation. See *Appendix* for the definition of the cross-moment divergence and the proof of the theorem.

Theorem 1. *Let $\mathcal{H}$ be a hypothesis space of VC dimension d. Let m be the size of labeled samples from all sources $\{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_N\}$, S_j be the labeled sample set of size $\beta_j m$ ($\sum_j \beta_j = 1$) drawn from μ_j and labeled by the groundtruth labeling function f_j. If $\hat{h} \in \mathcal{H}$ is the empirical minimizer of $\hat{\epsilon}_{\boldsymbol{\alpha}}(h)$ for a fixed weight vector $\boldsymbol{\alpha}$ and $h_T^* = \min_{h \in \mathcal{H}} \epsilon_T(h)$ is the target error minimizer, then for any $\delta \in (0, 1)$ and any $\epsilon > 0$, there exist N integers $\{n_\epsilon^j\}_{j=1}^N$ and N constants $\{a_{n_\epsilon^j}\}_{j=1}^N$, such that with probability at least $1 - \delta$,*

$$\epsilon_T(\hat{h}) \leq \epsilon_T(h_T^*) + \eta_{\boldsymbol{\alpha},\boldsymbol{\beta},m,\delta} + \epsilon + \sum_{j=1}^{N} \alpha_j \Big(2\lambda_j + a_{n_\epsilon^j} \sum_{k=1}^{n_\epsilon^j} d_{CM^k}(\mathcal{D}_j, \mathcal{D}_T) \Big), \quad (6)$$

where $\eta_{\boldsymbol{\alpha},\boldsymbol{\beta},m,\delta} = 4\sqrt{(\sum_{j=1}^{N} \frac{\alpha_j^2}{\beta_j})(\frac{2d(\log(\frac{2m}{d})+1)+2\log(\frac{4}{\delta})}{m})}$ *and* $\lambda_j = \min_{h \in \mathcal{H}} \{\epsilon_T(h) + \epsilon_j(h)\}$.

Theorem 1 shows that the upper bound on the target error of the learned hypothesis depends on the pairwise moment divergence $d_{CM^k}(\mathcal{D}_S, \mathcal{D}_T)$ between the target domain and each source domain.[2] This provides a direct motivation for moment matching approaches beyond ours. In particular, it motivates our multi-source domain adaptation approach to align the moments between each target-source pair. Moreover, it is obvious that the last term of the bound, $\sum_k d_{CM^k}(\mathcal{D}_j, \mathcal{D}_T)$, is lower bounded by the pairwise divergences between source domains. To see this, consider

[2] Note that single source is just a special case when $N = 1$.

Standards	Models	*mt,up,sv,sy* → *mm*	*mm,up,sv,sy* → *mt*	*mm,mt,sv,sy* → *up*	*mm,mt,up,sy* → *sv*	*mm,mt,up,sv* → *sy*	Avg
Source Combine	Source Only	63.70±0.83	92.30±0.91	90.71±0.54	71.51±0.75	83.44±0.79	80.33±0.76
	DAN [25]	67.87±0.75	97.50± 0.62	93.49±0.85	67.80±0.84	86.93±0.93	82.72± 0.79
	DANN [8]	70.81±0.94	97.90±0.83	93.47±0.79	68.50±0.85	87.37±0.68	83.61±0.82
Multi-Source	Source Only	63.37±0.74	90.50±0.83	88.71±0.89	63.54±0.93	82.44±0.65	77.71±0.81
	DAN [25]	63.78±0.71	96.31±0.54	94.24±0.87	62.45±0.72	85.43±0.77	80.44±0.72
	CORAL [39]	62.53±0.69	97.21±0.83	93.45±0.82	64.40±0.72	82.77±0.69	80.07±0.75
	DANN [8]	71.30±0.56	97.60±0.75	92.33±0.85	63.48±0.79	85.34±0.84	82.01±0.76
	JAN [27]	65.88±0.68	97.21±0.73	95.42±0.77	75.27±0.71	86.55±0.64	84.07±0.71
	ADDA [40]	71.57± 0.52	97.89±0.84	92.83±0.74	75.48±0.48	86.45±0.62	84.84±0.64
	DCTN [45]	70.53±1.24	96.23±0.82	92.81±0.27	77.61±0.41	86.77±0.78	84.79±0.72
	MEDA [44]	71.31±0.75	96.47±0.78	97.01±0.82	78.45±0.77	84.62±0.79	85.60± 0.78
	MCD [38]	72.50±0.67	96.21±0.81	95.33±0.74	78.89±0.78	87.47±0.65	86.10±0.73
	M^{3}SDA (ours)	69.76±0.86	**98.58**±0.47	95.23±0.79	78.56±0.95	87.56±0.53	86.13±0.64
	M^{3}SDA-β (ours)	**72.82**±1.13	98.43±0.68	**96.14**±0.81	**81.32**±0.86	**89.58**±0.56	**87.65**± 0.75

Table 2. **Digits Classification Results. mt**, **up**, **sv**, **sy**, **mm** are abbreviations for *MNIST*, *USPS*, *SVHN*, *Synthetic Digits*, *MNIST-M*, respectively. Our model M^{3}SDA and M^{3}SDA-β achieve **86.13%** and **87.65%** accuracy, outperforming other baselines by a large margin.

the toy example consisting of two sources $\mathcal{D}_1, \mathcal{D}_2$, and a target $\mathcal{D}_T$, since $d_{CM^k}(\cdot,\cdot)$ is a metric, triangle inequality implies the following lower bound:

$$d_{CM^k}(\mathcal{D}_1, \mathcal{D}_T) + d_{CM^k}(\mathcal{D}_2, \mathcal{D}_T) \geq d_{CM^k}(\mathcal{D}_1, \mathcal{D}_2).$$

This motivates our algorithm to also align the moments between each pair of source domains. Intuitively, it is not possible to perfectly align the target domain with every source domain, if the source domains are not aligned themselves. Further discussions of Theorem 1 and its relationship with our algorithm are provided in the *Appendix*.

5. Experiments

We perform an extensive evaluation on the following tasks: digit classification (*MNIST, SVHN, USPS, MNIST-M, Sythetic Digits*), and image recognition (*Office-Caltech10*, DomainNet dataset). In total, we conduct 714 experiments. The experiments are run on a GPU-cluster with 24 GPUs and the total running time is more than 21,440 GPU-hours. Due to space limitations, we only report major results; more implementation details are provided in the supplementary material. Throughout the experiments, we set the trade-off parameter λ in Equation 2 as 0.5. In terms of the parameter sensitivity, we have observed that the performance variation is not significant if λ is between 0.1$\sim$1. All of our experiments are implemented in the PyTorch[3] platform.

5.1. Experiments on Digit Recognition

Five digit datasets are sampled from five different sources, namely *MNIST* [19], *Synthetic Digits* [8], *MNIST-M* [8], *SVHN*, and *USPS*. Following *DCTN* [45], we sample 25000 images from training subset and 9000 from testing subset in *MNIST, MINST-M, SVHN*, and *Synthetic Digits*. *USPS* dataset contains only 9298 images in total, so we take the entire dataset as a domain. In all of our experiments, we take turns to set one domain as the target domain and the rest as the source domains.

We take four state-of-the-art discrepancy-based approaches: Deep Adaptation Network [25] (**DAN**), Joint Adaptation Network (**JAN**), Manifold Embedded Distribution Alignment (**MEDA**), and Correlation Alignment [39] (**CORAL**), and four adversarial-based approaches: Domain Adversarial Neural Network [8] (**DANN**), Adversarial Discriminative Domain Adaptation [40] (**ADDA**), Maximum Classifier Discrepancy (**MCD**) and Deep Cocktail Network [45] (**DCTN**) as our baselines. In the *source combine* setting, all the source domains are combined to a single domain, and the baseline experiments are conducted in a traditional single domain adaptation manner.

The results are shown in Table 2. Our model M^{3}SDA achieves an **86.13%** average accuracy, and M^{3}SDA-β boosts the performance to **87.65%**, outperforming other baselines by a large margin. One interesting observation is that the results on MNIST-M dataset is lower. This phenomenon is probably due to the presence of *negative transfer* [31]. For a fair comparison, all the experiments are based on the same network architecture. For each experiment, we run the same setting for five times and report the mean and standard deviation. (See *Appendix* for detailed experiment settings and analyses.)

5.2. Experiments on Office-Caltech10

The Office-Caltech10 [11] dataset is extended from the standard Office31 [37] dataset. It consists of the same 10 object categories from 4 different domains: *Amazon*, *Caltech*, *DSLR*, and *Webcam*.

The experimental results on Office-Caltech10 dataset are shown in Table 4. Our model M^{3}SDA gets a 96.1% average accuracy on this dataset, and M^{3}SDA-β further boosts the performance to **96.4%**. All the experiments are based on ResNet-101 pre-trained on ImageNet. As far as we

[3] http://pytorch.org

AlexNet	*clp*	*inf*	*pnt*	*qdr*	*rel*	*skt*	Avg.
clp	65.5	8.2	21.4	10.5	36.1	10.8	17.4
inf	32.9	27.7	23.8	2.2	26.4	13.7	19.8
pnt	28.1	7.5	57.6	2.6	41.6	20.8	20.1
qdr	13.4	1.2	2.5	68.0	5.5	7.1	5.9
rel	36.9	10.2	33.9	4.9	72.8	23.1	21.8
skt	35.5	7.1	21.9	11.8	30.8	56.3	21.4
Avg.	29.4	6.8	20.7	6.4	28.1	15.1	**17.8**

DAN	*clp*	*inf*	*pnt*	*qdr*	*rel*	*skt*	Avg.
clp	N/A	9.1	23.4	16.2	37.9	29.7	23.2
inf	17.2	N/A	15.6	4.4	24.8	13.5	15.1
pnt	29.9	8.9	N/A	7.9	42.1	26.1	23.0
qdr	14.2	1.6	4.4	N/A	8.5	10.1	7.8
rel	37.4	11.5	33.3	10.1	N/A	26.4	23.7
skt	39.1	8.8	28.2	13.9	36.2	N/A	25.2
Avg.	27.6	8.0	21.0	10.5	29.9	21.2	**19.7**

JAN	*clp*	*inf*	*pnt*	*qdr*	*rel*	*skt*	Avg.
clp	N/A	7.8	24.5	14.3	38.1	25.7	22.1
inf	17.6	N/A	18.7	8.7	28.1	15.3	17.7
pnt	27.5	8.2	N/A	7.1	43.1	23.9	22.0
qdr	17.8	2.2	7.4	N/A	8.1	10.9	9.3
rel	33.5	9.1	32.5	7.5	N/A	21.9	20.9
skt	35.3	8.2	27.7	13.3	36.8	N/A	24.3
Avg.	26.3	7.1	22.2	10.2	30.8	19.5	**19.4**

DANN	*clp*	*inf*	*pnt*	*qdr*	*rel*	*skt*	Avg.
clp	N/A	9.1	23.2	13.7	37.6	28.6	22.4
inf	17.9	N/A	16.4	2.1	27.8	13.3	15.5
pnt	29.1	8.6	N/A	5.1	41.5	24.7	21.8
qdr	16.8	1.8	4.8	N/A	9.3	10.2	8.6
rel	36.5	11.4	33.9	5.9	N/A	24.5	22.4
skt	37.9	8.2	26.3	12.2	35.3	N/A	24.0
Avg.	27.6	7.8	20.9	7.8	30.3	20.3	**19.1**

RTN	*clp*	*inf*	*pnt*	*qdr*	*rel*	*skt*	Avg.
clp	N/A	8.1	21.1	13.1	36.1	26.5	21.0
inf	15.6	N/A	15.3	3.4	25.1	12.8	14.4
pnt	26.8	8.1	N/A	5.2	40.6	22.6	20.7
qdr	15.1	1.8	4.5	N/A	8.5	8.9	7.8
rel	35.3	10.7	31.7	7.5	N/A	22.9	21.6
skt	34.1	7.4	23.3	12.6	32.1	N/A	21.9
Avg.	25.4	7.2	19.2	8.4	28.4	18.7	**17.9**

ADDA	*clp*	*inf*	*pnt*	*qdr*	*rel*	*skt*	Avg.
clp	N/A	11.2	24.1	3.2	41.9	30.7	22.2
inf	19.1	N/A	16.4	3.2	26.9	14.6	16.0
pnt	31.2	9.5	N/A	8.4	39.1	25.4	22.7
qdr	15.7	2.6	5.4	N/A	9.9	11.9	9.1
rel	39.5	14.5	29.1	12.1	N/A	25.7	24.2
skt	35.3	8.9	25.2	14.9	37.6	N/A	25.4
Avg.	28.2	9.3	20.1	8.4	31.1	21.7	**19.8**

MCD	*clp*	*inf*	*pnt*	*qdr*	*rel*	*skt*	Avg.
clp	N/A	14.2	26.1	1.6	45.0	33.8	24.1
inf	23.6	N/A	21.2	1.5	36.7	18.0	20.2
pnt	34.4	14.8	N/A	1.9	50.5	28.4	26.0
qdr	15.0	3.0	7.0	N/A	11.5	10.2	9.3
rel	42.6	19.6	42.6	2.2	N/A	29.3	27.2
skt	41.2	13.7	27.6	3.8	34.8	N/A	24.2
Avg.	31.4	13.1	24.9	2.2	35.7	23.9	**21.9**

SE	*clp*	*inf*	*pnt*	*qdr*	*rel*	*skt*	Avg.
clp	N/A	9.7	12.2	2.2	33.4	23.1	16.1
inf	10.3	N/A	9.6	1.2	13.1	6.9	8.2
pnt	17.1	9.4	N/A	2.1	28.4	15.9	14.6
qdr	13.6	3.9	11.6	N/A	16.4	11.5	11.4
rel	31.7	12.9	19.9	3.7	N/A	26.3	18.9
skt	18.7	7.8	12.2	7.7	28.9	N/A	15.1
Avg.	18.3	8.7	13.1	3.4	24.1	16.7	**14.1**

Table 3. **Single-source baselines on the DomainNet dataset.** Several single-source adaptation baselines are evaluated on the DomainNet dataset, including AlexNet [18], DAN [25], JAN [27], DANN [8], RTN [26], ADDA [40], MCD [38], SE [7]. In each sub-table, the column-wise domains are selected as the source domain and the row-wise domains are selected as the target domain. The green numbers represent the average performance of each column or row. The red numbers denote the average accuracy for all the 30 (source, target) combinations.

Standards	Models	A,C,D →W	A,C,W →D	A,D,W →C	C,D,W →A	Avg
Source Combine	Source only	99.0	98.3	87.8	86.1	92.8
	DAN [25]	99.3	98.2	89.7	**94.8**	95.5
Multi-Source	Source only	99.1	98.2	85.4	88.7	92.9
	DAN [25]	99.5	99.1	89.2	91.6	94.8
	DCTN [45]	99.4	99.0	90.2	92.7	95.3
	JAN [27]	99.4	**99.4**	91.2	91.8	95.5
	MEDA [44]	99.3	99.2	91.4	92.9	95.7
	MCD [38]	99.5	99.1	91.5	92.1	95.6
	M^{3}SDA (ours)	99.4	99.2	91.5	94.1	96.1
	M^{3}SDA-β (ours)	**99.5**	99.2	**92.2**	94.5	**96.4**

Table 4. **Results on Office-Caltech10 dataset**. A,C,W and D represent *Amazon*, *Caltech*, *Webcam* and *DSLR*, respectively. All the experiments are based on ResNet-101 pre-trained on ImageNet.

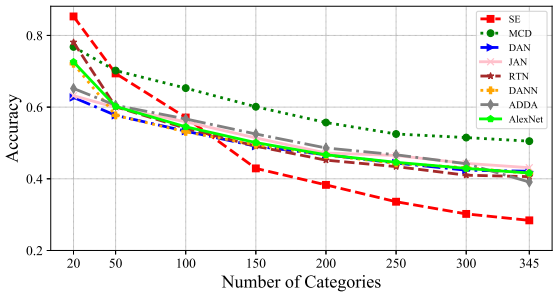

Figure 4. **Accuracy *vs.* Number of categories.** This plot shows the *painting*→*real* scenario. More plots with similar trend can be accessed in Figure 5 (see Appendix).

know, our models achieve the best performance among all the results ever reported on this dataset. We have also tried AlexNet, but it did not work as well as ResNet-101.

5.3. Experiments on DomainNet

Single-Source Adaptation To demonstrate the intrinsic difficulty of DomainNet, we evaluate multiple state-of-the-art algorithms for single-source adaptation: Deep Alignment Network (**DAN**) [25], Joint Adaptation Network (**JAN**) [27], Domain Adversarial Neural Network (**DANN**) [8], Residual Transfer Network (**RTN**) [26], Adversarial Deep Domain Adaptation (**ADDA**) [40], Maximum Classifier Discrepancy (**MCD**) [38], and Self-Ensembling (**SE**) [7]. As the DomainNet dataset contains 6 domains, experiments for 30 different (sources, target) combinations are performed for each baseline. For each domain, we follow a 70%/30% split scheme to participate our dataset into training and testing trunk. The detailed statistics can be viewed in Table 8 (see Appendix). All other experimental settings (neural network, learning rate, stepsize, etc.) are kept the same as in the original papers. Specifically, DAN, JAN, DANN, and RTN are based on AlexNet [18], ADDA and MCD are based on ResNet-101 [13], and SE is based on ResNet-152 [13]. Table 3 shows all the source-only and experimental results. (Source-only results for ResNet-101 and ResNet-152 are in Appendix, Table 7). The results show that our dataset is challenging, especially for the *infograph* and *quickdraw* domain. We argue that the difficulty is mainly introduced by the large number of categories in our dataset.

Multi-Source Domain Adaptation DomainNet contains six domains. Inspired by Xu et al [45], we introduce two MSDA standards: (1) *single best*, reporting the single best-performing source transfer result on the test set, and (2) *source combine*, combining the source domains to a single domain and performing traditional single-source adaptation. The first standard evaluates whether MSDA can improve the best single source UDA results; the second testify whether MSDA is necessary to exploit.

Baselines For both *single best* and *source combine* experiment setting, we take the following state-of-the-art methods as our baselines: Deep Alignment Network (**DAN**) [25], Joint Adaptation Network (**JAN**) [27], Domain Adversarial Neural Network (**DANN**) [8], Residual Transfer Network (**RTN**) [26], Adversarial Deep Domain Adaptation (**ADDA**) [40], Maximum Classifier Discrepancy (**MCD**) [38], and Self-Ensembling (**SE**) [7]. For multi-source domain adaptation, we take Deep Cocktail Network (**DCTN**) [45] as our baseline.

Results The experimental results of multi-source domain

Standards	Models	*inf,pnt,qdr, rel,skt→clp*	*clp,pnt,qdr, rel,skt→inf*	*clp,inf,qdr, rel,skt→pnt*	*clp,inf,pnt, rel,skt→qdr*	*clp,inf,pnt, qdr,skt →rel*	*clp,inf,pnt, qdr,rel →skt*	Avg
Single Best	Source Only	39.6±0.58	8.2±0.75	33.9 ± 0.62	11.8 ± 0.69	41.6 ± 0.84	23.1±0.72	26.4 ± 0.70
	DAN [25]	39.1±0.51	11.4±0.81	33.3±0.62	**16.2**±0.38	42.1±0.73	29.7±0.93	28.6±0.63
	RTN [26]	35.3±0.73	10.7±0.61	31.7±0.82	13.1±0.68	40.6±0.55	26.5±0.78	26.3±0.70
	JAN [27]	35.3±0.71	9.1±0.63	32.5±0.65	14.3±0.62	43.1±0.78	25.7±0.61	26.7±0.67
	DANN [8]	37.9±0.69	11.4±0.91	33.9±0.60	13.7±0.56	41.5±0.67	28.6±0.63	27.8±0.68
	ADDA [40]	39.5±0.81	14.5±0.69	29.1±0.78	14.9±0.54	41.9±0.82	30.7±0.68	28.4±0.72
	SE [7]	31.7±0.70	12.9±0.58	19.9±0.75	7.7±0.44	33.4±0.56	26.3±0.50	22.0±0.66
	MCD [38]	42.6±0.32	19.6±0.76	42.6±0.98	3.8±0.64	50.5±0.43	33.8±0.89	32.2±0.66
Source Combine	Source Only	47.6±0.52	13.0±0.41	38.1±0.45	13.3±0.39	51.9±0.85	33.7±0.54	32.9±0.54
	DAN [25]	45.4±0.49	12.8±0.86	36.2±0.58	15.3±0.37	48.6±0.72	34.0±0.54	32.1±0.59
	RTN [26]	44.2±0.57	12.6±0.73	35.3±0.59	14.6±0.76	48.4±0.67	31.7±0.73	31.1±0.68
	JAN [27]	40.9±0.43	11.1±0.61	35.4±0.50	12.1±0.67	45.8±0.59	32.3±0.63	29.6±0.57
	DANN [8]	45.5±0.59	13.1±0.72	37.0±0.69	13.2±0.77	48.9±0.65	31.8±0.62	32.6±0.68
	ADDA [40]	47.5±0.76	11.4±0.67	36.7±0.53	14.7±0.50	49.1±0.82	33.5±0.49	32.2±0.63
	SE [7]	24.7±0.32	3.9±0.47	12.7±0.35	7.1±0.46	22.8±0.51	9.1±0.49	16.1±0.43
	MCD [38]	54.3±0.64	22.1±0.70	45.7±0.63	7.6±0.49	58.4±0.65	43.5±0.57	38.5±0.61
Multi-Source	DCTN [45]	48.6±0.73	23.5±0.59	48.8±0.63	7.2±0.46	53.5±0.56	47.3±0.47	38.2±0.57
	M^{3}SDA* (ours)	57.0±0.79	22.1±0.68	50.5±0.45	4.4± 0.21	62.0±0.45	48.5±0.56	40.8± 0.52
	M^{3}SDA (ours)	57.2±0.98	24.2±1.21	51.6±0.44	5.2±0.45	61.6±0.89	**49.6**±0.56	41.5±0.74
	M^{3}SDA-β (ours)	**58.6**±0.53	**26.0**± 0.89	**52.3**±0.55	6.3±0.58	**62.7**±0.51	49.5±0.76	**42.6**±0.64
Oracle Results	AlexNet	65.5±0.56	27.7±0.34	57.6±0.49	68.0±0.55	72.8±0.67	56.3±0.59	58.0±0.53
	ResNet101	69.3±0.37	34.5±0.42	66.3±0.67	66.8±0.51	80.1±0.59	60.7±0.48	63.0±0.51
	ResNet152	71.0±0.63	36.1±0.61	68.1 ± 0.49	69.1±0.52	81.3±0.49	65.2±0.57	65.1±0.55

Table 5. **Multi-source domain adaptation results on the DomainNet dataset.** Our model M^{3}SDA and M^{3}SDA-β achieves **41.5%** and **42.6%** accuracy, significantly outperforming all other baselines. M^{3}SDA* indicates the normal average of all the classifiers. When the target domain is *quickdraw*, the multi-source methods perform worse than single-source and source only baselines, which indicates negative transfer [31] occurs in this case. (*clp*: *clipart*, *inf*: *infograph*, *pnt*: *painting*, *qdr*: *quickdraw*, *rel*: *real*, *skt*: *sketch*.)

adaptation are shown in Table 5. We report the results of the two different weighting schemas and all the baseline results in Table 5. Our model M^{3}SDA achieves an average accuracy of **41.5%**, and M^{3}SDA-β boosts the performance to **42.6%**. The results demonstrate that our models designed for MSDA outperform the *single best* UDA results, the *source combine* results, and the multi-source baseline. From the experimental results, we make three interesting observations. (1)The performance of M^{3}SDA* is 40.8%. After applying the weight vector $\mathcal{W}$, M^{3}SDAimproves the mean accuracy by 0.7 percent. (2) In *clp,inf,pnt,rel,skt→qdr* setting, the performances of our models are worse than source-only baseline, which indicates that negative transfer [31] occurs. (3) In the *source combine* setting, the performances of DAN [25], RTN [26], JAN [27], DANN [8] are lower than the source only baseline, indicating the negative transfer happens when the training data are from multiple source domains.

Effect of Category Number To show how the number of categories affects the performance of state-of-the-art domain adaptation methods, we choose the *painting→real* setting in DomainNet and gradually increase the number of category from 20 to 345. The results are in Figure 4. An interesting observation is that when the number of categories is small (which is exactly the case in most domain adaptation benchmarks), all methods tend to perform well. However, their performances drop at different rates when the number of categories increases. For example, SE [7] performs the best when there is a limit number of categories, but worst when the number of categories is larger than 150.

6. Conclusion

In this paper, we have collected, annotated and evaluated by far the largest domain adaptation dataset named DomainNet. The dataset is challenging due to the presence of notable domain gaps and a large number of categories. We hope it will be beneficial to evaluate future single- and multi-source UDA methods.

We have also proposed M^{3}SDA to align multiple source domains with the target domain. We derive a meaningful error bound for our method under the framework of cross-moment divergence. Further, we incorporate the moment matching component into deep neural network and train the model in an end-to-end fashion. Extensive experiments on multi-source domain adaptation benchmarks demonstrate that our model outperforms all the multi-source baselines as well as the best single-source domain adaptation method.

7. Acknowledgements

We thank Ruiqi Gao, Yizhe Zhu, Saito Kuniaki, Ben Usman, Ping Hu for their useful discussions and suggestions. We thank anonymous annotators for their hard work to label the data. This work was partially supported by NSF and Honda Research Institute. The authors also acknowledge support from CIFAR AI Chairs Program.

References

[1] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010. 1, 2, 3, 5, 12

[2] Shai Ben-David, John Blitzer, Koby Crammer, Fernando Pereira, et al. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, pages 137–144, 2007. 12

[3] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *Advances in Neural Information Processing Systems*, pages 343–351, 2016. 3

[4] Koby Crammer, Michael Kearns, and Jennifer Wortman. Learning from multiple sources. *Journal of Machine Learning Research*, 9(Aug):1757–1774, 2008. 1, 3, 12

[5] Carl De Boor, Carl De Boor, Etats-Unis Mathématicien, Carl De Boor, and Carl De Boor. *A practical guide to splines*, volume 27. Springer-Verlag New York, 1978. 4

[6] Lixin Duan, Dong Xu, and Shih-Fu Chang. Exploiting web images for event recognition in consumer videos: A multiple source domain adaptation approach. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1338–1345. IEEE, 2012. 2, 3

[7] Geoff French, Michal Mackiewicz, and Mark Fisher. Self-ensembling for visual domain adaptation. In *International Conference on Learning Representations*, 2018. 2, 7, 8, 13, 14

[8] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1180–1189, Lille, France, 07–09 Jul 2015. PMLR. 1, 2, 6, 7, 8

[9] Muhammad Ghifary, W Bastiaan Kleijn, and Mengjie Zhang. Domain adaptive neural networks for object recognition. In *Pacific Rim international conference on artificial intelligence*, pages 898–904. Springer, 2014. 2

[10] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li. Deep reconstruction-classification networks for unsupervised domain adaptation. In *European Conference on Computer Vision*, pages 597–613. Springer, 2016. 3

[11] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2066–2073. IEEE, 2012. 2, 6

[12] Arthur Gretton, Karsten M Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J Smola. A kernel method for the two-sample-problem. In *Advances in neural information processing systems*, pages 513–520, 2007. 2

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7, 14

[14] Judy Hoffman, Mehryar Mohri, and Ningshan Zhang. Algorithms and theory for multiple-source adaptation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 8246–8256. Curran Associates, Inc., 2018. 1, 2, 3, 5

[15] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. CyCADA: Cycle-consistent adversarial domain adaptation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1989–1998, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018. PMLR. 1, 3

[16] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1857–1865, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. 3

[17] Piotr Koniusz, Yusuf Tas, Hongguang Zhang, Mehrtash Harandi, Fatih Porikli, and Rui Zhang. Museum exhibit identification challenge for the supervised domain adaptation and beyond. In *The European Conference on Computer Vision (ECCV)*, September 2018. 2

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 7

[19] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 2, 6

[20] Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. Mmd gan: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems*, pages 2203–2213, 2017. 3

[21] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Deeper, broader and artier domain generalization. In *International Conference on Computer Vision*, 2017. 2

[22] Yujia Li, Kevin Swersky, and Rich Zemel. Generative moment matching networks. In *International Conference on Machine Learning*, pages 1718–1727, 2015. 3

[23] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*, pages 700–708, 2017. 1

[24] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In *Advances in neural information processing systems*, pages 469–477, 2016. 2

[25] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 97–105, Lille, France, 07–09 Jul 2015. PMLR. 1, 2, 6, 7, 8, 13

[26] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. In *Advances in Neural Information Processing Systems*, pages 136–144, 2016. 7, 8

[27] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I. Jordan. Deep transfer learning with joint adaptation networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 2208–2217, 2017. 1, 2, 6, 7, 8

[28] Yishay Mansour, Mehryar Mohri, Afshin Rostamizadeh, and A R. Domain adaptation with multiple sources. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1041–1048. Curran Associates, Inc., 2009. 1, 3

[29] Youssef Mroueh, Tom Sercu, and Vaibhava Goel. McGan: Mean and covariance feature matching GAN. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2527–2535, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. 3

[30] OA Muradyan and S Ya Khavinson. Absolute values of the coefficients of the polynomials in weierstrass's approximation theorem. *Mathematical notes of the Academy of Sciences of the USSR*, 22(2):641–645, 1977. 12

[31] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010. 6, 8, 14

[32] Xingchao Peng and Kate Saenko. Synthetic to real adaptation with generative correlation alignment networks. In *2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018, Lake Tahoe, NV, USA, March 12-15, 2018*, pages 1982–1991, 2018. 1, 2

[33] Xingchao Peng, Baochen Sun, Karim Ali, and Kate Saenko. Learning deep object detectors from 3d models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1278–1286, 2015. 2

[34] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017. 2

[35] Xingchao Peng, Ben Usman, Kuniaki Saito, Neela Kaushik, Judy Hoffman, and Kate Saenko. Syn2real: A new benchmark forsynthetic-to-real visual domain adaptation. *CoRR*, abs/1806.09755, 2018. 2

[36] Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. *Dataset Shift in Machine Learning*. The MIT Press, 2009. 1

[37] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010. 2, 6

[38] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 4, 5, 6, 7, 8, 14

[39] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *AAAI*, volume 6, page 8, 2016. 1, 2, 3, 6

[40] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 4, 2017. 1, 2, 6, 7, 8

[41] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014. 1, 2

[42] Vladimir N Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity*, pages 11–30. Springer, 2015. 13

[43] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *(IEEE) Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[44] Jindong Wang, Wenjie Feng, Yiqiang Chen, Han Yu, Meiyu Huang, and Philip S Yu. Visual domain adaptation with manifold embedded distribution alignment. In *ACM Multimedia Conference*, 2018. 2, 6, 7

[45] Ruijia Xu, Ziliang Chen, Wangmeng Zuo, Junjie Yan, and Liang Lin. Deep cocktail network: Multi-source unsupervised domain adaptation with category shift. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3964–3973, 2018. 2, 3, 6, 7, 8

[46] Zili Yi, Hao (Richard) Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*, pages 2868–2876, 2017. 3

[47] Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. Central moment discrepancy (CMD) for domain-invariant representation learning. *CoRR*, abs/1702.08811, 2017. 1, 3

[48] Zhen Zhang, Mianzhi Wang, Yan Huang, and Arye Nehorai. Aligning infinite-dimensional covariance matrices in reproducing kernel hilbert spaces for domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3437–3445, 2018. 3

[49] Han Zhao, Shanghang Zhang, Guanhang Wu, José MF Moura, Joao P Costeira, and Geoffrey J Gordon. Adversarial multiple source domain adaptation. In *Advances in Neural Information Processing Systems*, pages 8568–8579, 2018. 1, 2, 5

[50] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017. 1, 3

[51] Fuzhen Zhuang, Xiaohu Cheng, Ping Luo, Sinno Jialin Pan, and Qing He. Supervised representation learning: Transfer learning with deep autoencoders. In *IJCAI*, pages 4119–4125, 2015. 2