

# Multi-source Domain Adaptation in the Deep Learning Era: A Systematic Survey

Sicheng Zhao<sup>1</sup>, Bo Li<sup>1</sup>, Colorado Reed<sup>1</sup>, Pengfei Xu<sup>2</sup>, Kurt Keutzer<sup>1</sup>

<sup>1</sup>University of California, Berkeley, USA    <sup>2</sup>Didi Chuxing, China

schzhao@gmail.com, drlodian@gmail.com, cjr@cs.berkeley.edu,  
xupengfeipf@didiglobal.com, keutzer@berkeley.edu

## Abstract

In many practical applications, it is often difficult and expensive to obtain enough large-scale labeled data to train deep neural networks to their full capability. Therefore, transferring the learned knowledge from a separate, labeled source domain to an unlabeled or sparsely labeled target domain becomes an appealing alternative. However, direct transfer often results in significant performance decay due to *domain shift*. Domain adaptation (DA) addresses this problem by minimizing the impact of domain shift between the source and target domains. Multi-source domain adaptation (MDA) is a powerful extension in which the labeled data may be collected from multiple sources with different distributions. Due to the success of DA methods and the prevalence of multi-source data, MDA has attracted increasing attention in both academia and industry. In this survey, we define various MDA strategies and summarize available datasets for evaluation. We also compare modern MDA methods in the deep learning era, including latent space transformation and intermediate domain generation. Finally, we discuss future research directions for MDA.

## 1 Background and Motivation

The availability of large-scale labeled training data, such as ImageNet, has enabled deep neural networks (DNNs) to achieve remarkable success in many learning tasks, ranging from computer vision to natural language processing. For example, the classification error of the “Classification + localization with provided training data” task in the Large Scale Visual Recognition Challenge has reduced from 0.28 in 2010 to 0.0225 in 2017<sup>1</sup>, outperforming even human classification. However, in many practical applications, obtaining labeled training data is often expensive, time-consuming, or even impossible. For example, in fine-grained recognition, only the experts can provide reliable labels [Gebru *et al.*, 2017]; in semantic segmentation, it takes about 90 minutes to label each Cityscapes image [Cordts *et al.*, 2016]; in autonomous

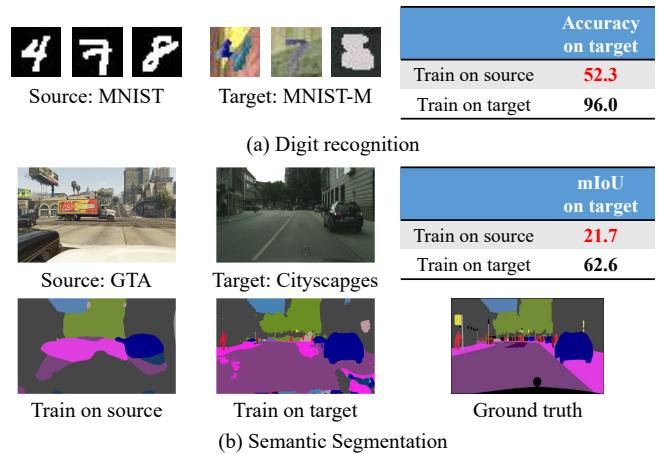


Figure 1: An example of *domain shift* in the single-source scenario. The models trained on the labeled source domain do not perform well when directly transferring to the target domain.

driving, it is difficult to label point-wise 3D LiDAR point clouds [Wu *et al.*, 2019].

One potential solution is to transfer a model trained on a separate, labeled source domain to the desired unlabeled or sparsely labeled target domain. But as Figure 1 demonstrates, the **direct transfer of models across domains leads to poor performance**. Figure 1(a) shows that even for the simple task of digit recognition, training on the MNIST source domain [LeCun *et al.*, 1998] for digit classification in the MNIST-M target domain [Ganin and Lempitsky, 2015] leads to a digit classification accuracy decrease from 96.0% to 52.3% when training a LeNet-5 model [LeCun *et al.*, 1998]. Figure 1(b) shows a more realistic example of training a semantic segmentation model on a synthetic source dataset GTA [Richter *et al.*, 2016] and conducting pixel-wise segmentation on a real target dataset Cityscapes [Cordts *et al.*, 2016] using the FCN model [Long *et al.*, 2015a]. If we train on the real data, we obtain a mean intersection-over-union (mIoU) of 62.6%; but if we train on synthetic data, the mIoU drops significantly to 21.7%.

The poor performance from directly transferring models across domains stems from a phenomenon known as *domain shift* [Torralba and Efros, 2011; Zhao *et al.*, 2018b]: whereby

<sup>1</sup><http://image-net.org/challenges/LSVRC/2017>

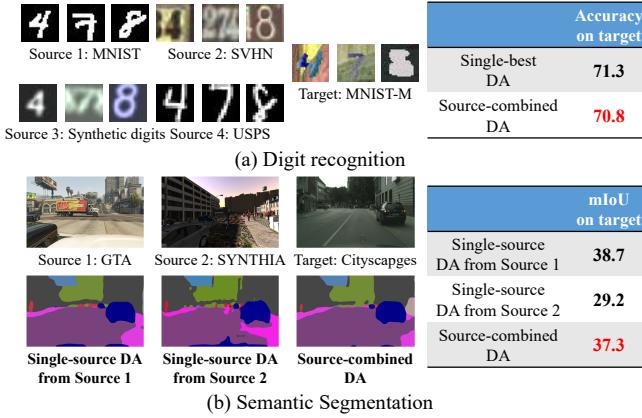


Figure 2: An example of *domain shift* in the multi-source scenario. Combining multiple sources into one source and directly performing single-source domain adaptation on the entire dataset does not guarantee better performance compared to just using the best individual source domain.

the joint probability distributions of observed data and labels are different in the two domains. Domain shift exists in many forms, such as from dataset to dataset, from simulation to real-world, from RGB images to depth, and from CAD models to real images.

The phenomenon of domain shift motivates the research on domain adaptation (DA), which aims to learn a model from a labeled source domain that can generalize well to a different, but related, target domain. Existing DA methods mainly focus on the single-source scenario. In the deep learning era, recent single-source DA (SDA) methods usually employ a co-joined architecture with two approaches to respectively represent the models for the source and target domains. One approach aims to learn a task model based on the labeled source data using corresponding task losses, such as cross-entropy loss for classification. The other approach aims to deal with the domain shift by aligning the target and source domains. Based on the alignment strategies, deep SDA methods can be classified into four categories:

1. *Discrepancy-based methods* try to align the features by explicitly measuring the discrepancy on corresponding activation layers, such as maximum mean discrepancy (MMD) [Long *et al.*, 2015b], correlation alignment [Sun *et al.*, 2017], and contrastive domain discrepancy [Kang *et al.*, 2019].
2. *Adversarial generative methods* generate fake data to align the source and target domains at pixel-level based on Generative Adversarial Network (GAN) [Goodfellow *et al.*, 2014] and its variants, such as CycleGAN [Zhu *et al.*, 2017; Zhao *et al.*, 2019b].
3. *Adversarial discriminative methods* employ an adversarial objective with a domain discriminator to align the features [Tzeng *et al.*, 2017; Tsai *et al.*, 2018].
4. *Reconstruction based methods* aim to reconstruct the target input from the extracted features using the source task model [Ghifary *et al.*, 2016].

In practice, the labeled data may be collected from multiple sources with different distributions [Sun *et al.*, 2015; Bhatt *et al.*, 2016]. In such cases, the aforementioned SDA methods could be trivially applied by combining the sources into a single source: an approach we refer to as *source-combined DA*. However, source-combined DA oftentimes results in a poorer performance than simply using one of the sources and discarding the others. As illustrated in Figure 2, the accuracy on the best single source digit recognition adaptation using DANN [Ganin *et al.*, 2016] is 71.3%, while the source-combined accuracy drops to 70.8%. For segmentation adaptation using CyCADA [Hoffman *et al.*, 2018b], the mIoU of source-combined DA (37.3%) is also lower than that of SDA from GTA (38.7%). Because the domain shift not only exists between each source and target, but also exists among different sources, the source-combined data from different sources may interfere with each other during the learning process [Riemer *et al.*, 2019]. Therefore, multi-source domain adaptation (MDA) is needed in order to leverage all of the available data.

The early MDA methods mainly focus on shallow models [Sun *et al.*, 2015], either learning a latent feature space for different domains [Sun *et al.*, 2011; Duan *et al.*, 2012] or combining pre-learned source classifiers [Schweikert *et al.*, 2009]. Recently, the emphasis on MDA has shifted to deep learning architectures. In this paper, we systematically survey recent progress on deep learning based MDA, summarize and compare similarities and differences in the approaches, and discuss potential future research directions.

## 2 Problem Definition

In the typical MDA setting, there are multiple source domains  $S_1, S_2, \dots, S_M$  ( $M$  is the number of sources) and one target domain  $T$ . Suppose the observed data and corresponding labels<sup>2</sup> in the  $i^{\text{th}}$  source  $S_i$  are drawn from distribution  $p_i(\mathbf{x}, \mathbf{y})$  are  $\mathbf{X}_i = \{\mathbf{x}_i^j\}_{j=1}^{N_i}$  and  $\mathbf{Y}_i = \{\mathbf{y}_i^j\}_{j=1}^{N_i}$ , respectively, where  $N_i$  is the number of source samples. Let  $\mathbf{X}_T = \{\mathbf{x}_T^j\}_{j=1}^{N_T}$  and  $\mathbf{Y}_T = \{\mathbf{y}_T^j\}_{j=1}^{N_T}$  denote the target data and corresponding labels drawn from the target distribution  $P_T(\mathbf{x}, \mathbf{y})$ , where  $N_T$  is the number of target samples.

Suppose the number of labeled target samples is  $N_{TL}$ , the MDA problem can be classified into different categories:

- *unsupervised MDA*, when  $N_{TL} = 0$ ;
- *fully supervised MDA*, when  $N_{TL} = N_T$ ;
- *semi-supervised MDA*, otherwise.

Suppose  $\mathbf{x}_i^j \in \mathbb{R}^{d_i}$  and  $\mathbf{x}_T^j \in \mathbb{R}^{d_T}$  are an observation in source  $S_i$  and target  $T$ , we can classify MDA into:

- *homogeneous MDA*, when  $d_1 = \dots = d_M = d_T$ ;
- *heterogeneous MDA*, otherwise.

Suppose  $\mathcal{C}_i$  and  $\mathcal{C}_T$  are the label set for source  $S_i$  and target  $T$ , we can define different MDA strategies:

- *closed set MDA*, when  $\mathcal{C}_1 = \dots = \mathcal{C}_M = \mathcal{C}_T$ ;

<sup>2</sup>The label could be any type, such as object classes, bounding boxes, semantic segmentation, etc.

| Area | Task                     | Dataset                  | Reference   | #D | #S               | Labels         | Short description  |
|------|--------------------------|--------------------------|---|----|------------------|----------------|--|
| CV   | object classification    | digit recognition        | Digits-five (D)<br>LeCun <i>et al.</i> ; Netzer <i>et al.</i><br>Hull; Ganin and Lempitsky                | 5  | 145,298          | 10 classes     | handwritten, synthetic, and street-image digits                      |
|      |                          | Office-31 (O)            | Saenko <i>et al.</i>  | 3  | 4,110            | 31 classes     | images from amazon and taken by different cameras                    |
|      |                          | Office-Caltech (OC)      | Gong <i>et al.</i>  | 4  | 2,533            | 10 classes     | overlapping categories from Office-31 and C                          |
|      |                          | Office-Home (OH)         | Venkateswara <i>et al.</i>  | 4  | 15,500           | 65 classes     | artistic, clipart, product, and real objects                         |
|      |                          | ImageCLEF (IC)           | Challenge <sup>3</sup>  | 3  | 1,800            | 12 classes     | shared categories from 3 datasets                                    |
|      |                          | PACS (P)                 | Li <i>et al.</i>  | 4  | 9,991            | 7 classes      | photographic, artistic, cartoon, and sketchy objects                 |
|      |                          | DomainNet (DN)           | Peng <i>et al.</i>  | 6  | 600,000          | 345 classes    | clipart, infographic, artistic, quickdraw, real, and sketchy objects |
|      |                          | sentiment classification | SentiImage (SI)<br>Machajdik and Hanbury<br>You <i>et al.</i> ; You <i>et al.</i> ; Borth <i>et al.</i>   | 4  | 25,986           | 2 classes      | artistic and social images on visual sentiment                       |
|      |                          | vehicle counting         | WebCamT (W)<br>Zhang <i>et al.</i>  | 8  | 16,000           | vehicle counts | each camera used as one domain                                       |
|      |                          | semantic segmentation    | Sim2RealSeg (S2R)<br>Cordts <i>et al.</i> ; Yu <i>et al.</i><br>Richter <i>et al.</i> ; Ros <i>et al.</i> | 4  | 49,366           | 16 classes     | simulation-to-real adaptation for pixel-wise predictions             |
| NLP  | sentiment classification | AmazonReviews (AR)       | Chen <i>et al.</i>  | 4  | $\approx 12,000$ | 2 classes      | reviews on four kinds of products                                    |
|      |                          | MediaReviews (MR)        | Liu <i>et al.</i>   | 5  | 6897             | 2 classes      | reviews on products and movies                                       |
|      | part-of-speech tagging   | SANCL (S)                | Petrov and McDonald   | 5  | 5250             | tags           | part-of-speech tagging in 5 web domains                              |

Table 1: Released and freely available datasets for MDA, where ‘#D’ and ‘#S’ represent the number of domains and the total number of samples usually used for MDA, respectively.

- *open set MDA*, for at least one  $\mathcal{C}_i$ ,  $\mathcal{C}_i \cap \mathcal{C}_T \subset \mathcal{C}_T$ ;
- *partial MDA*, for at least one  $\mathcal{C}_i$ ,  $\mathcal{C}_T \subset \mathcal{C}_i$ ;
- *universal MDA*, when no prior knowledge of the label sets is available;

where  $\cap$  and  $\subset$  indicate the intersection set and proper subset between two sets.

Suppose the number of labeled source samples is  $N_{iL}$  for source  $S_i$ , the MDA problem can be classified into:

- *strongly supervised MDA*, when  $N_{iL} = N_i$  for  $i = 1 \dots M$ ;
- *weakly supervised MDA*, otherwise.

When adapting to multiple target domains simultaneously, the task becomes multi-target MDA. When the target data is unavailable during training [Yue *et al.*, 2019], the task is often called multi-source domain generalization or zero-shot MDA.

### 3 Datasets

The datasets for evaluating MDA models usually contain multiple domains with different styles, such as *synthetic* vs. *real*, *artistic* vs. *sketchy*, which impose large domain shift among different domains. Here we summarize the commonly employed datasets in both computer vision (CV) and natural language processing (NLP) areas, as shown in Table 1.

**Digit recognition.** Digits-five includes 5 digit image datasets sampled from different domains, including *handwritten MNIST (mt)* [LeCun *et al.*, 1998], *combined MNIST-M (mm)* [Ganin and Lempitsky, 2015] from MNIST and randomly extracted color patches, *street image SVHN (sv)* [Netzer *et al.*, 2011], *Synthetic Digits (sy)* [Ganin and Lempitsky, 2015] generated from Windows fonts by various conditions, and *handwritten USPS (up)* [Hull, 1994]. Usually, 25,000 images are sampled for training and 9,000 for testing in **mt**, **mm**, **sv**, and **sy**. The entire 9,298 images in **up** are selected.

**Object classification.** Office-31 [Saenko *et al.*, 2010] contains 4,110 images in 31 categories collected from office environments in 3 domains: Amazon (**A**) with 2,817 images downloaded from amazon.com, Webcam (**W**) and DSLR (**D**) with 795 and 498 images taken by web camera and digital SLR camera with different photographic settings.

Office-Caltech [Gong *et al.*, 2013] consists of the 10 overlapping categories shared by Office-31 [Saenko *et al.*, 2010]

and Caltech-256 (**C**) [Griffin *et al.*, 2007]. Totally there are 2,533 images.

Office-Home [Venkateswara *et al.*, 2017] consists of about 15,500 images from 65 categories of everyday objects in office and home settings. There are 4 different domains: Artistic images (**Ar**), Clip Art (**Cl**), Product images (**Pr**) and Real-World images (**Rw**).

ImageCLEF, originated from ImageCLEF 2014 domain adaptation challenge<sup>3</sup>, consists of 12 object categories shared by ImageNet ILSVRC 2012 (**I**), Pascal VOC 2012 (**P**), and **C**. Totally there are 600 images for each domain with 50 for each category.

PACS [Li *et al.*, 2017] contains 9,991 images of 7 object categories extracted from 4 different domains: Photo (**P**), Art paintings (**A**), Cartoon (**C**) and Sketch (**S**).

DomainNet [Peng *et al.*, 2019], the largest DA dataset to date for object classification, contains about 600K images from 6 domains: Clipart, Infograph, Painting, Quickdraw, Real, and Sketch. There are totally 345 object categories.

**Sentiment classification of images.** SentiImage [Lin *et al.*, 2020] is a DA dataset with 4 domains for binary sentiment classification of images: *social* Flickr and Instagram (**FI**) [You *et al.*, 2016], *artistic* ArtPhoto (**AP**) [Machajdik and Hanbury, 2010], *social* Twitter I (**TI**) [You *et al.*, 2015], and *social* Twitter II (**TII**) [Borth *et al.*, 2013]. There are 23,308, 806, 1,269, and 603 images in these 4 domains, respectively.

**Vehicle counting.** WebCamT [Zhang *et al.*, 2017] is a vehicle counting dataset from large-scale city camera videos with low resolution, low frame rate, and high occlusion. Totally there are 60,000 frames with vehicle bounding box and count annotations. For MDA, 8 cameras located in different intersections are selected, each with more than 2,000 labeled images. We can view each camera as a domain.

**Scene segmentation.** Sim2RealSeg contains 2 synthetic datasets (GTA, SYNTHIA) and 2 real datasets (Cityscapes, BDDS) for segmentation. Cityscapes (**CS**) [Cordts *et al.*, 2016] contains vehicle-centric urban street images collected from a moving vehicle in 50 cities from Germany and neighboring countries. There are 5,000 images with pixel-wise annotations into 19 classes. BDDS [Yu *et al.*, 2018]

<sup>3</sup><http://imageclef.org/2014/adaptation>

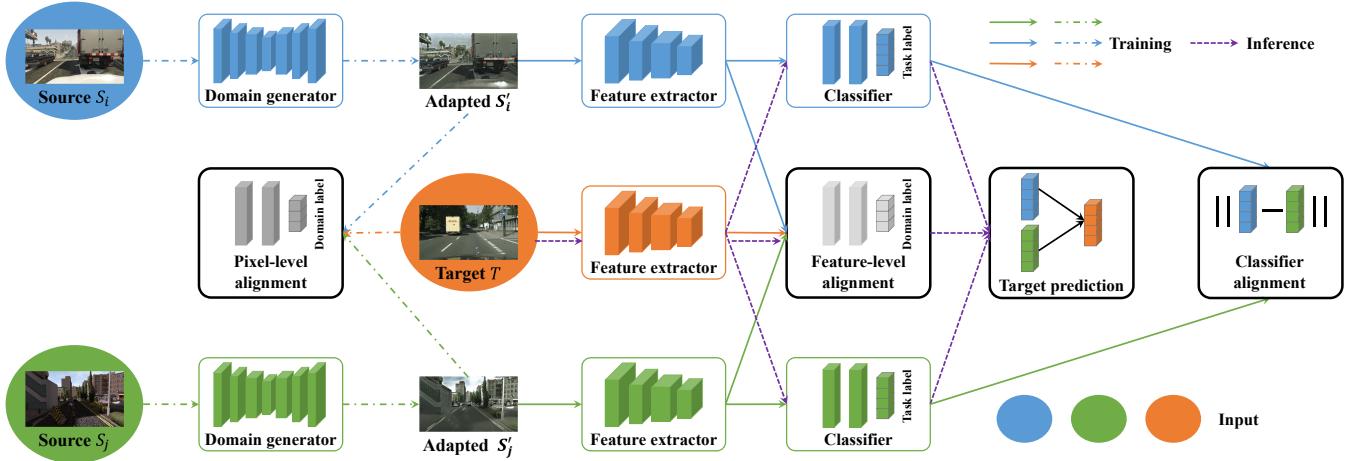


Figure 3: Illustration of widely employed framework for MDA. The solid arrows and dashed dot arrows indicate the training of latent space transformation and intermediate domain generation, respectively. The dashed arrows represent the reference process. Most existing MDA methods can be obtained by employing different component details, enforcing some constraints, or slightly changing the architecture. Best viewed in color.

contains 10,000 real-world dash cam video frames with a compatible label space with Cityscapes. GTA [Richter *et al.*, 2016] is a vehicle-egocentric image dataset collected in the high-fidelity rendered computer game GTA-V. It contains 24,966 images (video frames) with 19 classes as Cityscapes. SYNTHIA [Ros *et al.*, 2016] is a large synthetic dataset. To pair with Cityscapes, a subset, named SYNTHIA-RANDCITYSCAPES, is designed with 9,400 images which are automatically annotated with 16 compatible Cityscapes classes, one void class, and some unnamed classes. The common 16 classes are used for MDA.

**Sentiment classification of natural languages.** Amazon Reviews [Chen *et al.*, 2012] is a dataset of reviews on four kinds of products: Books (**B**), DVDs (**D**), Electronics (**E**), and Kitchen appliances (**K**). Reviews are encoded as 5,000 dimensional feature vectors of unigrams and bigrams and are labeled with binary sentiment. Each source has 2,000 labeled examples, and the target test set has 3,000 to 6,000 examples.

Media Reviews [Liu *et al.*, 2017] contains 16 domains of product reviews and movie reviews for binary sentiment classification. 5 domains with 6,897 labeled samples are usually employed for MDA, including Apparel, Baby, Books, Camera taken from Amazon and MR from Rotten Tomato.

**Part-of-speech tagging.** The SANCL dataset [Petrov and McDonald, 2012] contains part-of-speech tagging annotations in 5 web domains: Emails (**E**), Weblogs (**W**), Answers (**A**), Newsgroups (**N**), and Reviews (**R**). 750 sentences from each source are used for training.

Unless otherwise specified, each domain is selected as the target and the rest domains are considered as the sources. For WebCamT, 2 domains are randomly selected as the target. For Sim2RealSeg, MDA is often performed using the simulation-to-real setting [Zhao *et al.*, 2019a], *i.e.* from synthetic GTA, SYNTHIA to real Cityscapes, BDDS. For SANCL, **N**, **R**, and **A** are used as target domains, while **E** and **W** are used as target domains [Guo *et al.*, 2018].

## 4 Deep Multi-source Domain Adaptation

Existing methods on deep MDA primarily focus on the unsupervised, homogeneous, closed set, strongly supervised, one target, and target data available settings. That is, there is one target domain, the target data is unlabeled but available during the training process, the source data is fully labeled, the source and target data are observed in the same data space, and the label sets of all sources and the target are the same. In this paper, we focus on MDA methods under these settings.

There are some theoretical analysis to support existing MDA algorithms. Most theories are based on the seminal theoretical model [Blitzer *et al.*, 2008; Ben-David *et al.*, 2010]. Mansour *et al.* [2009] assumed that the target distribution can be approximated by a mixture of the  $M$  source distributions. Therefore, weighted combination of source classifiers has been widely employed for MDA. Moreover, tighter cross domain generalization bound and more accurate measurements on domain discrepancy can provide intuitions to derive effective MDA algorithms. Hoffman *et al.* [2018a] derived a novel bound using DC-programming and calculated more accurate combination weights. Zhao *et al.* [2018a] extended the generalization bound of seminal theoretical model to multiple sources under both classification and regression settings. Besides the domain discrepancy between the target and each source [Hoffman *et al.*, 2018a; Zhao *et al.*, 2018a], Li *et al.* [2018] also considered the relationship between pairwise sources and derived a tighter bound on weighted multi-source discrepancy. Based on this bound, more relevant source domains can be picked out.

Typically, some task models (*e.g.* classifiers) are learned based on the labeled source data with corresponding task loss, such as cross-entropy loss for classification. Meanwhile, specific alignments among the source and target domains are conducted to bridge the domain shift so that the learned task models can be better transferred to the target domain. Based on the different alignment strategies, we can classify MDA

| Reference                       | Feature extractor | Feature alignment method | Feature alignment loss    | Feature alignment domains | Classifier alignment | #C | Classifier weight | Task backbone | Dataset   | Result           |
|---------------------------------|-------------------|--------------------------|---------------------------|---------------------------|----------------------|----|-------------------|---------------|-----------|------------------|
| [Mancini <i>et al.</i> , 2018]  | shared            | —                        | —                         | target and each source    | CT loss              | 1  | —                 | AlexNet       | O, OC, P  | 83.6, 91.8, 85.3 |
| [Guo <i>et al.</i> , 2018]      | shared            | discrepancy              | MMD                       | target and each source    | —                    | M  | PoS metric        | AlexNet       | AR, S     | 84.8, 90.1       |
| [Hoffman <i>et al.</i> , 2018a] | shared            | discrepancy              | Renyi-divergence         | target and each source    | CT loss              | 1  | —                 | AlexNet       | O         | 87.6             |
| [Zhu <i>et al.</i> , 2019]      | shared            | discrepancy              | MMD                       | target and each source    | $\mathcal{L}_1$ loss | M  | uniform           | ResNet-50     | O, OH, IC | 90.2, 89.4, 74.1 |
| [Rakshit <i>et al.</i> , 2019]  | unshared          | discrepancy              | $\mathcal{L}_2$ distance  | pairwise all domains      | CT loss              | 1  | —                 | ResNet-50     | O, OC, IC | 88.3, 97.5, 91.2 |
| [Peng <i>et al.</i> , 2019]     | shared            | discrepancy              | moment distance           | pairwise all domains      | $\mathcal{L}_1$ loss | M  | relative error    | LeNet-5       | D         | 87.7             |
| [Guo <i>et al.</i> , 2020]      | shared            | discrepancy              | mixture distance          | target and each source    | CT loss              | 1  | —                 | ResNet-101    | OC        | 96.4             |
| [Xu <i>et al.</i> , 2018]       | shared            | discriminator            | GAN loss                  | target and each source    | —                    | M  | perplexity score  | AlexNet       | D, O, IC  | 74.2, 83.8, 80.8 |
| [Li <i>et al.</i> , 2018]       | shared            | discriminator            | Wasserstein               | pairwise all domains      | CT loss              | 1  | —                 | AlexNet       | D         | 79.9             |
| [Zhao <i>et al.</i> , 2018a]    | shared            | discriminator            | $\mathcal{H}$ -divergence | target and each source    | CT loss              | 1  | —                 | BiLSTM        | AR        | 82.7             |
| [Wang <i>et al.</i> , 2019]     | shared            | discriminator            | Wasserstein               | pairwise all domains      | CT loss              | 1  | —                 | AlexNet       | D         | 76.6             |
| [Zhao <i>et al.</i> , 2020]     | unshared          | discriminator            | Wasserstein               | target and each source    | —                    | M  | Wasserstein       | FCN8s         | W         | 1.4              |
|                                 |                   |                          |                           |                           |                      |    |                   | LeNet-5       | D         | 88.1             |
|                                 |                   |                          |                           |                           |                      |    |                   | AlexNet       | O         | 84.2             |

Table 2: Comparison of different latent space transformation methods for MDA, where ‘#C’, ‘CT loss’, and ‘MMD’ are short for the number of classifiers during reference ( $M$  is the number of source domains), combined task loss, and maximum mean discrepancy, respectively. ‘Result’ is the average performance of all target domains measured by accuracy for classification and counting error for vehicle counting.

into different categories. *Latent space transformation* tries to align the latent space (*e.g.* features) of different domains based on optimizing the discrepancy loss or adversarial loss. *Intermediate domain generation* explicitly generates an intermediate adapted domain for each source that is indistinguishable from the target domain. The task models are then trained on the adapted domain. Figure 3 summarizes the common overall framework of existing MDA methods.

#### 4.1 Latent Space Transformation

The two common methods for aligning the latent spaces of different domains are discrepancy-based methods and adversarial methods. We discuss these two methods below, and Table 2 summarizes key examples of each method.

**Discrepancy-based methods** explicitly measure the discrepancy of the latent spaces (typically features) from different domains by optimizing some specific discrepancy losses, such as maximum mean discrepancy (MMD) [Guo *et al.*, 2018; Zhu *et al.*, 2019], Renyi-divergence [Hoffman *et al.*, 2018a],  $\mathcal{L}_2$  distance [Rakshit *et al.*, 2019], and moment distance [Peng *et al.*, 2019]. Guo *et al.* [2020] claimed that different discrepancies or distances can only provide specific estimates of domain similarities and that each distance has its pathological cases. Therefore, they consider the mixture of several distances [Guo *et al.*, 2020], including  $\mathcal{L}_2$  distance, Cosine distance, MMD, Fisher linear discriminant, and Correlation alignment. Minimizing the discrepancy to align the features among the source and target domains does not introduce any new parameters that must be learned.

**Adversarial methods** try to align the features by making them indistinguishable to a discriminator. Some representative optimized objectives include GAN loss [Xu *et al.*, 2018],  $\mathcal{H}$ -divergence [Zhao *et al.*, 2018a], Wasserstein distance [Li *et al.*, 2018; Wang *et al.*, 2019; Zhao *et al.*, 2020]. These methods aim to confuse the discriminator’s ability to distinguishing whether the features from multiple sources were drawn from the same distribution. Compared with GAN loss and  $\mathcal{H}$ -divergence, Wasserstein distance can provide more stable gradients even when the target and source distributions do not overlap [Zhao *et al.*, 2020]. The discriminator is often

implemented as a network, which leads to new parameters that must be learned.

There are many modular implementation details for both types of methods, such as how to align the target and multiple sources, whether the feature extractors are shared, how to select the more relevant sources, and how to combine the multiple predictions from different classifiers.

**Alignment domains.** There are different ways to align the target and multiple sources. The most common method is to pairwise align the target with each source [Xu *et al.*, 2018; Guo *et al.*, 2018; Zhao *et al.*, 2018a; Hoffman *et al.*, 2018a; Zhu *et al.*, 2019; Zhao *et al.*, 2020; Guo *et al.*, 2020]. Since domain shift also exists among different sources, several methods enforce pairwise alignment between every domain in both the source and target domains [Li *et al.*, 2018; Rakshit *et al.*, 2019; Peng *et al.*, 2019; Wang *et al.*, 2019].

**Weight sharing of feature extractor.** Most methods employ shared feature extractors to learn domain-invariant features. However, domain invariance may be detrimental to discriminative power. On the contrary, Rakshit *et al.* [2019] adopted one feature extractor for each source and target pair with unshared weights, while Zhao *et al.* [2020] first pre-trained one feature extractor for each source and then mapped the target into the feature space of each source. Correspondingly, there are  $M$  and  $2M$  feature extractors. Although unshared feature extractors can better align the target and sources in the latent space, this substantially increases the number of parameters in the model.

**Classifier alignment.** Intuitively, the classifiers trained on different sources may result in misaligned predictions for the target samples that are close to the domain boundary. By minimizing specific classifier discrepancy, such as  $\mathcal{L}_1$  loss [Zhu *et al.*, 2019; Peng *et al.*, 2019], the classifiers are better aligned, which can learn a generalized classification boundary for target samples mentioned above. Instead of explicitly training one classifier for each source, many methods focus on training a compound classifier based on specific combined task loss, such as normalized activations [Mancini *et al.*, 2018] and bandit controller [Guo *et al.*, 2020].

**Target prediction.** After aligning the features of target

| Reference                    | Domain generator      | Pixel alignment domains      | Feature alignment loss | Feature alignment domains | #C  | Classifier weight | Task backbone | Dataset            | Task | Result       |
|------------------------------|-----------------------|------------------------------|------------------------|---------------------------|-----|-------------------|---------------|--------------------|------|--------------|
| [Russo <i>et al.</i> , 2019] | CoGAN shared          | target and each source       | GAN loss               | target and each source    | $M$ | uniform           | DeepLabV2     | S2R-CS             | seg  | 42.8         |
| [Zhao <i>et al.</i> , 2019a] | CycleGAN shared       | target and aggregated source | GAN loss               | target and each source    | 1   | —                 | FCN8s         | S2R-CS<br>S2R-BDDS | seg  | 41.4<br>36.3 |
| [Lin <i>et al.</i> , 2020]   | VAE+CycleGAN unshared | target and combined source   | —                      | —                         | 1   | —                 | ResNet-18     | SI                 | cls  | 68.1         |

Table 3: Comparison of different intermediate domain generation methods for MDA, where ‘#C’, ‘seg’, and ‘cls’ are short for the number of classifiers during reference ( $M$  is the number of source domains), segmentation, and classification, respectively. ‘Result’ is the average performance of all target domains measured by accuracy for classification and mean intersection-over-union (mIoU) for segmentation.

and source domains in the latent space, the classifiers trained based on the labeled source samples can be used to predict the labels of a target sample. Since there are multiple sources, it is possible that they will yield different target predictions. One way to reconcile these different predictions is to uniformly average the predictions from different source classifiers [Zhu *et al.*, 2019]. However, different sources may have different relationships with the target, *e.g.* one source might better align with the target, so a non-uniform, weighted averaging of the predictions leads to better results. Weighting strategies, known as a *source selection process*, include uniform weight [Zhu *et al.*, 2019], perplexity score based on adversarial loss [Xu *et al.*, 2018], point-to-set (PoS) metric using Mahalanobis distance [Guo *et al.*, 2018], relative error based on source-only accuracy [Peng *et al.*, 2019], and Wasserstein distance based weights [Zhao *et al.*, 2020].

Besides the source importance, Zhao *et al.* [2020] also considered the sample importance, *i.e.* different samples from the same source may still have different similarities from the target samples. The source samples that are closer to the target are distilled (based on a manually selected Wasserstein distance threshold) to fine-tune the source classifiers. Automatically and adaptively selecting the most relevant training samples for each source remains an open research problem.

## 4.2 Intermediate Domain Generation

Feature-level alignment only aligns high-level information, which is insufficient for fine-grained predictions, such as pixel-wise semantic segmentation [Zhao *et al.*, 2019a]. Generating an intermediate adapted domain with pixel-level alignment, typically via GANs [Goodfellow *et al.*, 2014], can help address this problem.

**Domain generator.** Since the original GAN is highly under-constrained, some improved versions are employed, such as Coupled GAN (CoGAN) in [Russo *et al.*, 2019] and CycleGAN in MADAN [Zhao *et al.*, 2019a]. Instead of directly taking the original source data as input to the generator [Russo *et al.*, 2019; Zhao *et al.*, 2019a], Lin *et al.* [2020] used a variational autoencoder to map all source and target domains to a latent space and then generated an adapted domain from the latent space. Russo *et al.* [2019] then tried to align the target and each adapted domain, while Lin *et al.* [2020] aligned the target and combined adapted domain from the latent space. Zhao *et al.* [2019a] proposed to aggregate different adapted domains using a sub-domain aggregation discriminator and cross-domain cycle discriminator, where the pixel-level alignment is then conducted between the aggregated and target domains. Zhao *et al.* [2019a] and

Lin *et al.* [2020] showed that the semantics might change in the intermediate representation, and that enforcing a semantic consistency before and after generation can help preserve the labels.

**Feature alignment and target prediction.** Feature-level alignment is often jointly considered with pixel-level alignment. Both alignments are usually achieved by minimizing the GAN loss with a discriminator. One classifier is trained on each adapted domain [Russo *et al.*, 2019] and the multiple predictions for a given target sample are averaged. Only one classifier is trained on the aggregated domain [Zhao *et al.*, 2020] or on the combined adapted domain [Lin *et al.*, 2020] which is obtained by a unique generator from the latent space for all source domains. The comparison of these methods are summarized in Table 3.

## 5 Conclusion and Future Directions

In this paper, we provided a survey of recent MDA developments in the deep learning era. We motivated MDA, defined different MDA strategies, and summarized the datasets that are commonly used for performing MDA evaluation. Our survey focused on a typical MDA setting, *i.e.* unsupervised, homogeneous, closed set, and one target MDA. We classified these methods into different categories, and compared the representative ones technically and experimentally. We conclude with several open research directions:

**Specific MDA strategy implementation.** As introduced in Section 2, there are many types of MDA strategies, and implementing an MDA strategy according to the specific problem requirement would likely yield better results than a one-size-fits-all MDA approach. Further investigation is needed to determine which MDA strategies work the best for which types of problems. Also, real-world applications may have a small amount of labeled target data; determining how to include this data and what fraction of this data is needed for a certain performance remains an open question.

**Multi-modal MDA.** The labeled source data may be of different modalities, such as LiDAR, radar, and image. Further research is needed to find techniques for fusing different data modalities in MDA. A further extension of this idea is to have varied modalities in different sources as well as partially labeled, multi-modal sources.

**Incremental and online MDA.** Designing incremental and online MDA algorithms remains largely unexplored and may provide great benefit for real-world scenarios, such as updating deployed MDA models when new source or target data becomes available.

## References

- [Ben-David *et al.*, 2010] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. In *Machine Learning*, 2010.
- [Bhatt *et al.*, 2016] Himanshu S Bhatt, Arun Rajkumar, and Shourya Roy. Multi-source iterative adaptation for cross-domain classification. In *IJCAI*, 2016.
- [Blitzer *et al.*, 2008] John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. Learning bounds for domain adaptation. In *NIPS*, 2008.
- [Borth *et al.*, 2013] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *ACM MM*, 2013.
- [Chen *et al.*, 2012] Minmin Chen, Zhixiang Xu, Kilian Q Weinberger, and Fei Sha. Marginalized denoising autoencoders for domain adaptation. In *ICML*, 2012.
- [Cordts *et al.*, 2016] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016.
- [Duan *et al.*, 2012] Lixin Duan, Dong Xu, and Shih-Fu Chang. Exploiting web images for event recognition in consumer videos: A multiple source domain adaptation approach. In *CVPR*, 2012.
- [Ganin and Lempitsky, 2015] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015.
- [Ganin *et al.*, 2016] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Fleuret, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 2016.
- [Gebru *et al.*, 2017] Timnit Gebru, Judy Hoffman, and Li Fei-Fei. Fine-grained recognition in the wild: A multi-task domain adaptation approach. In *ICCV*, 2017.
- [Ghifary *et al.*, 2016] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li. Deep reconstruction-classification networks for unsupervised domain adaptation. In *ECCV*, 2016.
- [Gong *et al.*, 2013] Boqing Gong, Kristen Grauman, and Fei Sha. Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation. In *ICML*, 2013.
- [Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.
- [Griffin *et al.*, 2007] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007.
- [Guo *et al.*, 2018] Jiang Guo, Darsh Shah, and Regina Barzilay. Multi-source domain adaptation with mixture of experts. In *EMNLP*, 2018.
- [Guo *et al.*, 2020] Han Guo, Ramakanth Pasunuru, and Mohit Bansal. Multi-source domain adaptation for text classification via distancenet-bandits. In *AAAI*, 2020.
- [Hoffman *et al.*, 2018a] Judy Hoffman, Mehryar Mohri, and Ningshan Zhang. Algorithms and theory for multiple-source adaptation. In *NeurIPS*, 2018.
- [Hoffman *et al.*, 2018b] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*, 2018.
- [Hull, 1994] Jonathan J. Hull. A database for handwritten text recognition research. *TPAMI*, 1994.
- [Kang *et al.*, 2019] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *CVPR*, 2019.
- [LeCun *et al.*, 1998] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *PIEEE*, 1998.
- [Li *et al.*, 2017] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, 2017.
- [Li *et al.*, 2018] Yitong Li, Michael Andrew Murias, Samantha Major, Geraldine Dawson, and David E Carlson. Extracting relationships by multi-domain matching. In *NeurIPS*, 2018.
- [Lin *et al.*, 2020] Chuang Lin, Sicheng Zhao, Lei Meng, and Tat-Seng Chua. Multi-source domain adaptation for visual sentiment classification. In *AAAI*, 2020.
- [Liu *et al.*, 2017] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. Adversarial multi-task learning for text classification. In *ACL*, 2017.
- [Long *et al.*, 2015a] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.
- [Long *et al.*, 2015b] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *ICML*, 2015.
- [Machajdik and Hanbury, 2010] Jana Machajdik and Allan Hanbury. Affective image classification using features inspired by psychology and art theory. In *ACM MM*, 2010.
- [Mancini *et al.*, 2018] Massimiliano Mancini, Lorenzo Porzi, Samuel Rota Bulò, Barbara Caputo, and Elisa Ricci. Boosting domain adaptation by discovering latent domains. In *CVPR*, 2018.
- [Mansour *et al.*, 2009] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation with multiple sources. In *NIPS*, 2009.
- [Netzer *et al.*, 2011] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshops*, 2011.
- [Peng *et al.*, 2019] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *ICCV*, 2019.
- [Petrov and McDonald, 2012] Slav Petrov and Ryan McDonald. Overview of the 2012 shared task on parsing the web. In *SANCL*, 2012.
- [Rakshit *et al.*, 2019] Sayan Rakshit, Biplob Banerjee, Gemma Roig, and Subhasis Chaudhuri. Unsupervised multi-source domain adaptation driven by deep adversarial ensemble learning. In *GCPR*, 2019.
- [Richter *et al.*, 2016] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *ECCV*, 2016.
- [Riemer *et al.*, 2019] Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. Learning to learn without forgetting by maximizing transfer and minimizing interference. In *ICLR*, 2019.
- [Ros *et al.*, 2016] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*, 2016.
- [Russo *et al.*, 2019] Paolo Russo, Tatiana Tommasi, and Barbara Caputo. Towards multi-source adaptive semantic segmentation. In *ICIP*, 2019.
- [Saenko *et al.*, 2010] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *ECCV*, 2010.
- [Schweikert *et al.*, 2009] Gabriele Schweikert, Gunnar Rätsch, Christian Widmer, and Bernhard Schölkopf. An empirical analysis of domain adaptation algorithms for genomic sequence analysis. In *NIPS*, 2009.
- [Sun *et al.*, 2011] Qian Sun, Rita Chattopadhyay, Sethuraman Panchanathan, and Jieping Ye. A two-stage weighting framework for multi-source domain adaptation. In *NIPS*, 2011.
- [Sun *et al.*, 2015] Shiliang Sun, Honglei Shi, and Yuanbin Wu. A survey of multi-source domain adaptation. *Information Fusion*, 2015.
- [Sun *et al.*, 2017] Baochen Sun, Jiaoshi Feng, and Kate Saenko. Correlation alignment for unsupervised domain adaptation. In *Domain Adaptation in Computer Vision Applications*. 2017.
- [Torralba and Efros, 2011] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR*, 2011.
- [Tsai *et al.*, 2018] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *CVPR*, 2018.
- [Tzeng *et al.*, 2017] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, 2017.
- [Venkateswara *et al.*, 2017] Hemant Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *CVPR*, 2017.
- [Wang *et al.*, 2019] Haotian Wang, Wenjing Yang, Zhipeng Lin, and Yue Yu. Tmda: Task-specific multi-source domain adaptation via clustering embedded adversarial training. In *ICDM*, 2019.
- [Wu *et al.*, 2019] Bichen Wu, Xuanyu Zhou, Sicheng Zhao, Xiangyu Yue, and Kurt Keutzer. Squeezesegv2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a lidar point cloud. In *ICRA*, 2019.
- [Xu *et al.*, 2018] Ruijia Xu, Ziliang Chen, Wangmeng Zuo, Junjie Yan, and Liang Lin. Deep cocktail network: Multi-source unsupervised domain adaptation with category shift. In *CVPR*, 2018.
- [You *et al.*, 2015] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *AAAI*, 2015.
- [You *et al.*, 2016] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In *AAAI*, 2016.
- [Yu *et al.*, 2018] Fisher Yu, Wenyi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv:1805.04687*, 2018.
- [Yue *et al.*, 2019] Xiangyu Yue, Yang Zhang, Sicheng Zhao, Alberto Sangiovanni-Vincentelli, Kurt Keutzer, and Boqing Gong. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In *ICCV*, 2019.
- [Zhang *et al.*, 2017] Shanghang Zhang, Guanhong Wu, Joao P Costeira, and Jose MF Moura. Understanding traffic density from large-scale web camera data. In *CVPR*, 2017.
- [Zhao *et al.*, 2018a] Han Zhao, Shanghang Zhang, Guanhong Wu, José MF Moura, Joao P Costeira, and Geoffrey J Gordon. Adversarial multiple source domain adaptation. In *NeurIPS*, 2018.
- [Zhao *et al.*, 2018b] Sicheng Zhao, Xin Zhao, Guiuguang Ding, and Kurt Keutzer. Emotiongan: Unsupervised domain adaptation for learning discrete probability distributions of image emotions. In *ACM MM*, 2018.
- [Zhao *et al.*, 2019a] Sicheng Zhao, Bo Li, Xiangyu Yue, Yang Gu, Pengfei Xu, Runbo Hu, Hua Chai, and Kurt Keutzer. Multi-source domain adaptation for semantic segmentation. In *NeurIPS*, 2019.
- [Zhao *et al.*, 2019b] Sicheng Zhao, Chuang Lin, Pengfei Xu, Sendong Zhao, Yuchen Guo, Ravi Krishna, Guiuguang Ding, and Kurt Keutzer. Cyclemotiongan: Emotional semantic consistency preserved cyclegan for adapting image emotions. In *AAAI*, 2019.
- [Zhao *et al.*, 2020] Sicheng Zhao, Guangzhi Wang, Shanghang Zhang, Yang Gu, Yaxian Li, Zhichao Song, Pengfei Xu, Runbo Hu, Hua Chai, and Kurt Keutzer. Multi-source distilling domain adaptation. In *AAAI*, 2020.
- [Zhu *et al.*, 2017] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.
- [Zhu *et al.*, 2019] Yongchun Zhu, Fuzhen Zhuang, and Deqing Wang. Aligning domain-specific distribution and classifier for cross-domain classification from multiple sources. In *AAAI*, 2019.