Domain Agnostic Learning with Disentangled Representations

Xingchao Peng 1  Zijun Huang 2  Ximeng Sun 1  Kate Saenko 1

Abstract
Unsupervised model transfer has the potential to greatly improve the generalizability of deep models to novel domains. Yet the current literature assumes that the separation of target data into distinct domains is known a priori. In this paper, we propose the task of Domain-Agnostic Learning (DAL): How to transfer knowledge from a labeled source domain to unlabeled data from arbitrary target domains? To tackle this problem, we devise a novel Deep Adversarial Disentangled Autoencoder (DADA) capable of disentangling domain-specific features from class identity. We demonstrate experimentally that when the target domain labels are unknown, DADA leads to state-of-the-art performance on several image classification datasets.

1. Introduction
Supervised machine learning assumes that training and testing data are sampled i.i.d from the same distribution, while in practice, the training and testing data are typically collected from related domains but under different distributions, a phenomenon known as domain shift (Quionero-Candela et al., 2009). To avoid the cost of annotating each new test domain, Unsupervised Domain Adaptation (UDA) tackles domain shift by aligning the feature distribution of the source domain with that of the target domain, resulting in domain-invariant features. However, current methods assume that target samples have domain labels and therefore can be isolated into separate homogeneous domains. For many practical applications, this is an overly strong assumption. For example, a hand-written character recognition system could encounter characters written by different people, on different materials, and under different lighting conditions; an image recognition system applied to images scraped from the web must handle mixed-domain data (e.g. paintings, sketches, clipart) without their domain labels.

In this paper, we consider Domain-Agnostic Learning (DAL), a more difficult but practical problem of knowledge transfer from one labeled source domain to multiple unlabeled target domains. The main challenges of domain-agnostic learning are that: (1) the target data has mixed domains, which hampers the effectiveness of mainstream learning methods (Long et al., 2015; Sun & Saenko, 2016; Saito et al., 2018), and (2) class-irrelevant information leads to negative transfer (Pan & Yang, 2010), especially when the target domain is highly heterogeneous.

Mainstream UDA methods align the source domain to the target domain by minimizing the Maximum Mean Discrepancy (Long et al., 2015; Tzeng et al., 2014), aligning high-order moments (Sun & Saenko, 2016; Zellinger et al., 2017), or adversarial training (Ganin & Lempitsky, 2015; Tzeng et al., 2017). However, these methods are designed for one-to-one domain alignment and do not account for multiple latent domains in the target. Multi-source domain adaptation (Peng et al., 2018; Xu et al., 2018; Mansour et al., 2009) considers adaptation between multiple sources and a single target domain and assumes domain labels on the source data. Continuous domain adaptation (Hoffman et al., 2014) aims to transfer knowledge to a continuously changing domain (e.g. cars in different decades), but in their scenario the target data are temporally related. Recently, domain generalization approaches (Li et al., 2018a; Carlucci et al., 2018b; Li et al., 2018b) have been introduced to adapt from multiple labeled source domains to an unseen target domain. All of the above models make a strong assumption that the target data are homogeneously sampled from the same distribution, unlike the scenario we consider here.

We postulate that a solution to domain-agnostic learning should not only learn invariance between source and target, but should also actively disentangle the class-specific features from the remaining information in the image. Deep neural networks are known to extract features in which multiple hidden factors are highly entangled (Bengio et al., 2013). Recent work attempts to disentangle features in the latent space of autoencoders with adversarial training (Cao et al.,...
Various methods have been proposed, including discrepancy-based approaches (Liu & Tuzel, 2016; Tzeng et al., 2017; Liu et al., 2018a), and reconstruction-based approaches (Yi et al., 2017; Zhu et al., 2017; Hoffman et al., 2018; Kim et al., 2017). These models are typically designed to tackle single source to single target adaptation. Compared with single source adaptation, multi-source domain adaptation (MSDA) assumes that training data are collected from multiple sources. Originating from the theoretical analysis in (Ben-David et al., 2010; Mansour et al., 2009; Crammer et al., 2008), MSDA has been applied to many practical applications (Xu et al., 2018; Duan et al., 2012; Peng et al., 2018). Specifically, Ben-David et al. (2010) introduce an H-Divergence between the weighted combination of source domains and a target domain. We propose a new and more practical learning paradigm, not yet considered in the UDA literature, where labeled data come from a single source domain but the testing data contain multiple unknown domains.

**Representation Disentanglement** The goal of learning disentangled representations is to model the factors of data variation. Recent works (Mathieu et al., 2016; Makhzani et al., 2016; Liu et al., 2018a; Odena et al., 2017) aim at learning an interpretable representation using generative adversarial networks (GANs) (Goodfellow et al., 2014; Kingma et al., 2014) and variational autoencoders (VAEs) (Rezende et al., 2014; Kingma & Welling, 2013). In a fully supervised setting, Lee et al. (2018) proposes to disentangle the feature representation into a domain-invariant content space and a domain-specific attribute space, producing diverse outputs without paired training images. Another work (Odena et al., 2017) proposes an auxiliary classifier GAN (AC-GAN) to achieve representation disentanglement. Despite promising performance, these methods focus on disentangling representation in a single domain. Liu et al. (2018a) introduces a unified feature disentangler to learn a domain-invariant representation from data across multiple domains. However, their model assumes that multiple source domains are available during training, which limits its practical application. In contrast, our model disentangles the representation based on one source domain and multiple unknown target domains, and proposes an improved approach to disentanglement that considers the class label and mutual information between features.

**Agnostic Learning** There are several prior studies of agnostic learning that are related to our work. Model-Agnostic Meta-Learning (MAML) (Finn et al., 2017) aims to train a model on a variety of learning tasks and solve a new task using only a few training examples. Different from MAML, our method mainly focuses on transferring knowledge to heterogeneous domains. Carlucci et al. (2018a) proposes a learning framework to seamlessly extend the knowledge from multiple source domain to an unseen target domain by
Domain Agnostic Learning with Disentangled Representations

3. DADA: Deep Adversarial Disentangled Autoencoder

We define the domain-agnostic learning task as follows: Given a source domain \( \mathcal{D}_s = \{(x^s_i, y^s_i)\}_{i=1}^{n_s} \) with \( n_s \) labeled examples, the goal is to minimize risk on \( N \) target domains \( \mathcal{D}_t = \{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_N\} \) without domain labels. We denote the target domains as \( \mathcal{D}_t = \{x^t_i\}_{i=1}^{n_t} \) with \( n_t \) unlabeled examples. Empirically, we want to minimize the target risk \( \epsilon_t(\theta) = \Pr_{(x, y) \sim \mathcal{D}_t}[\theta(x) \neq y] \), where \( \theta(x) \) is the classifier.

We propose to solve the task by learning domain-invariant features that are discriminative of the class. Figure 1 shows the proposed model. The feature generator \( G \) maps the input image to a feature vector \( f_G \), which has many highly entangled factors. The disentangler \( D \) is responsible for disentangling the features \( f_G \) into domain-invariant features \( f_{ci} \), domain-specific features \( f_{di} \), and class-irrelevant features \( f_{ci} \). The feature reconstructor \( R \) aims to recover \( f_G \) from either \( f_{di}, f_{ds} \) or \( f_{di}, f_{ci} \). \( D \) and \( R \) are implemented as the encoder and decoder in a Variational Autoencoder. A mutual information minimizer is applied between \( f_{ds} \) and \( f_{ci} \), as well as between \( f_{di} \) and \( f_{ds} \), to enhance the disentanglement. Adversarial training via a domain identifier aligns the source domain and the heterogeneous target domain in the \( f_{di} \) space. A class identifier \( C \) is trained on the labeled source domain to predict the class distribution \( f_C \) and to adversarially extract class-irrelevant features \( f_{ci} \). We next describe each component in detail.

Variational Autoencoders VAEs (Kingma & Welling, 2013) are a class of deep generative models that simultaneously train both a probabilistic encoder and decoder. The encoder is trained to generate latent vectors that roughly follow a Gaussian distribution. In our case, we learn each part of our disentangled representations by applying a VAE architecture with the following objective function:

\[
\mathcal{L}_{vae} = \|f_G - f_C\|^2_F + KL(q(z|f_C)||p(z)),
\]

where the first term aims at recovering the original features extracted by \( G \), and the second term calculates Kullback-Leibler divergence which penalizes the deviation of latent features from the prior distribution \( p(z) \) (as \( z \sim \mathcal{N}(0, I) \)). However, this property cannot guarantee that domain-invariant features are well disentangled from the domain-specific features or from class-irrelevant features, as the loss function in Equation 1 only aligns the latent features to a normal distribution.

Class Disentanglement To address the above problem, we employ class disentanglement to remove class-irrelevant features, such as background, in an adversarial way. First, we train the disentangler \( D \) and the \( K \)-way class identifier \( C \) to correctly predict the labels, supervised by the cross-entropy loss:

\[
\mathcal{L}_{ce} = -\mathcal{E}_{(x, y) \sim \mathcal{D}_t} \sum_{k=1}^{K} 1[k = y_s] \log(C(f_D))
\]

where \( f_D \in \{f_{di}, f_{ci}\} \).

In the second step, we fix the class identifier and train the disentangler \( D \) to fool the class identifier by generating class-irrelevant features \( f_{ci} \). This can be achieved by minimizing the negative entropy of the predicted class distribution:

\[
\mathcal{L}_{cnt} = -\frac{1}{n_s} \sum_{j=1}^{n_s} \log C(f_{ci}^j) - \frac{1}{n_t} \sum_{j=1}^{n_t} \log C(f_{ci}^j)
\]
where the first term and the second term indicate minimizing the entropy on the source domain and on heterogeneous target, respectively. The above adversarial training process forces the corresponding disentangler to extract class-irrelevant features.

**Domain Disentanglement** To tackle the domain agnostic learning task, disentangling class-irrelevant features is not enough, as it fails to align the source domain with the target. To achieve better alignment, we further propose to disentangle the learned features into domain-specific and domain-invariant and to thus align the source with the target domain in the domain-invariant latent space. This is achieved by exploiting adversarial domain classification in the resulting latent space. Specifically, we leverage a domain identifier $DI$, which takes the disentangled feature $(f_{di}, f_{ds})$ as input and outputs the domain label $l_f$ (source or target). The objective function of the domain identifier is as follows:

$$L_{DI} = -E[l_f \log P(l_f)] + E(1 - l_f)[\log P(1 - l_f)]. \quad (4)$$

Then the disentangler is trained to fool the domain identifier $DI$ to extract domain-invariant features.

**Mutual Information Minimization** To better disentangle the features, we minimize the mutual information between domain-invariant and domain-specific features $(f_{di}, f_{ds})$, as well as domain-invariant and class-irrelevant features $(f_{di}, f_{ci})$:

$$I(D_x; D_{f_{di}}) = \int \log \frac{dP_{XZ}}{dP_X \otimes dP_Z} dP_{XZ}, \quad (5)$$

where $x \in \{f_{ds}, f_{ci}\}$, $P_{XZ}$ is the joint probability distribution of $(D_x, D_{f_{di}})$, and $P_X = \int z dP_{XZ}$ and $P_Z = \int x dP_{XZ}$ are the marginals. Despite being a pivotal measure across different domains, the mutual information is only tractable for discrete variables, or for a limited family of problems where the probability distributions are unknown (Belghazi et al., 2018). The computation incurs a complexity of $O(n^2)$, which is undesirable for deep CNNs.

Is this paper, we adopt the Mutual Information Neural Estimator (MINE) (Belghazi et al., 2018)

$$I(X; Z) = \sup_{\theta \in \Theta} \mathbb{E}_{x \sim x} [T_{\theta}] - \log \mathbb{E}_{x \sim x} [e^{T_{\theta}}]. \quad (6)$$

which provides unbiased estimation of mutual information on n i.i.d samples by leveraging a neural network $T_{\theta}$.

Practically, MINE (6) can be computed as $I(X; Z) = \int \int P_{XZ}(x, z) T(x, z, \theta) - \log (\int \int P_{XZ}(x, z) e^{T(x, z, \theta)})$. Additionally, to avoid computing the integrals, we leverage Monte-Carlo integration:

$$I(X, Z) = \frac{1}{n} \sum_{i=1}^{n} T(x, z, \theta) - \log (\frac{1}{n} \sum_{i=1}^{n} e^{T(x, z, \theta)}) \quad (7)$$

where $(x, z)$ are sampled from the joint distribution and $z'$ is sampled from the marginal distribution. We implement a neural network to perform the Monte-Carlo integration defined in Equation 7.

**Ring-style Normalization** Conventional batch normalization (Ioffe & Szegedy, 2015) diminishes internal covariate shift by subtracting the batch mean and dividing by the batch standard deviation. Despite promising results on domain adaptation, batch normalization alone is not enough to guarantee that the embedded features are well normalized in the scenario of heterogeneous domains. The target data are sampled from multiple domains and their embedded features are scattered irregularly in the latent space. Zheng et al. (2018) proposes a ring-style norm constraint to maintain a balance between the angular classification margins of multiple classes. Its objective is as follows:

$$L_{ring} = 1 - \frac{1}{2n} \sum_{i=1}^{n} (||T(x_i)||^2 - R)^2 \quad (8)$$

where $R$ is the learned norm value. However, ring loss is not robust and may cause mode collapse if the learned $R$ is small. Instead, we incorporate the ring loss into a Geman-McClure model and minimize the following loss function:

$$L_{GM} = \frac{\sum_{i=1}^{n} (||T(x_i)||^2 - R)^2}{2n\beta + \sum_{i=1}^{n} (||T(x_i)||^2 - R)^2} \quad (9)$$

where $\beta$ is the scale factor of the Geman-McClure model.
Domain Agnostic Learning with Disentangled Representations

Figure 2. We demonstrate the effectiveness of DADA on three datasets: Digit-Five, Office-Caltech10 (Gong et al., 2012) and DomainNet (Peng et al., 2018) dataset. The Digit-Five dataset includes: MNIST (mt), MNIST-M (mm), SVHN (sv), Synthetic (syn), and USPS (up). The Office-Caltech10 dataset contains: Amazon (A), Caltech (C), DSLR (D), and Webcam (W). The DomainNet dataset includes: clipart (clp), infograph (inf), painting (pnt), quickdraw (qdr), real (rel), and sktech (skt).

Optimization Our model is trained in an end-to-end fashion. We train the class and domain disentanglement component, MINE and the reconstruction component iteratively with Stochastic Gradient Descent (Kiefer et al., 1952) or Adam (Kingma & Ba, 2014) optimizer. We employ the popular neural networks (e.g. LeNet, AlexNet, or ResNet) as our feature generator G. The detailed training procedure is presented in Algorithm 1.

4. Experiments

We compare the DADA model to state-of-the-art domain adaptation algorithms on the following tasks: digit classification (MNIST, SVHN, USPS, MNIST-M, Synthetic Digits) and image recognition (Office-Caltech10 (Gong et al., 2012), DomainNet (Peng et al., 2018)). Sample images of these datasets can be seen in Figure 2. Table 6 (supplementary material) shows the detailed number of images we use in our experiments. In the main paper, we only report major results; more implementation details are provided in the supplementary material. All of our experiments are implemented in the PyTorch1 platform.

4.1. Experiments on Digit Recognition

Digit-Five This dataset is a collection of five benchmarks for digit recognition, namely MNIST (LeCun et al., 1998), Synthetic Digits (Ganin & Lempitsky, 2015), MNIST-M (Ganin & Lempitsky, 2015), SVHN, and USPS. In our experiments, we take turns setting one domain as the source domain and the rest as the mixed target domain (discarding both the class and the domain labels), leading to five transfer tasks. To explore the effectiveness of each component in our model, we propose four different ablations, i.e. model I: with class disentanglement; model II: I + domain disentanglement; model III: II + ring loss; model IV: III + reconstruction loss. The detailed architecture of our model can be seen in Table 5 (supplementary material).

We compare our model to state-of-the-art baselines: Deep Adaptation Network (DAN) (Long et al., 2015), Domain Adversarial Neural Network (DANN) (Ganin & Lempitsky, 2015), Adversarial Discriminative Domain Adaptation (ADDA) (Tzeng et al., 2017), Maximum Classifier Discrepancy (MCD) (Saito et al., 2018), and Unified Feature Disentangler Network (UFDN) (Liu et al., 2018a). Specifically, DAN applies MMD loss (Gretton et al., 2007) to align the source domain with the target domain in reproducing kernel Hilbert space. DANN and ADDA align the source domain with target domain by adversarial loss. MCD is a domain adaptation framework which incorporates two classifiers. UFDN employs a variational autoencoder (Kingma & Welling, 2013) to disentangle domain-invariant representations. When conducting the baseline experiments, we utilize the code provided by the authors and keep the original experimental settings.

Results and Analysis The experimental results on the “Digit-Five” dataset are shown in Table 1. From these, we can make the following observations. (1) Model IV achieves 62.3% average accuracy, significantly outperforming other baselines on most of the domain-agnostic tasks. (2) The results of model I and II demonstrate the effectiveness of class disentanglement and domain disentanglement. Without minimizing the mutual information between disentangled features, UFDN performs poorly on this task. (3) In model III, the ring loss boost the performance by three percent, demonstrating that feature normalization is essential in domain-agnostic learning.

To dive deeper into the disentangled features, we plot in Figure 3(a)-3(d) the t-SNE embeddings of the feature representations learned on the sv→mm,mt,up,sv task with source-only features, UFDN features, MCD features, and DADA features, respectively. We observe that the features derived by our model are more separated between classes than UFDN and MCD features.

1http://pytorch.org
Domain Agnostic Learning with Disentangled Representations

Table 1. Accuracy on “Digit-Five” dataset with domain agnostic learning protocol. DADA achieves 62.3% accuracy, significantly outperforming other baselines. We incrementally add each component to our model, aiming to study their effectiveness on the final results. (model I: with class disentanglement; model II: I + domain disentanglement; model III: II + ring loss; model IV: III + reconstruction loss. mt, up, sv, sy, mm are abbreviations for MNIST, USPS, SVHN, Synthetic Digits, MNIST-M.)

| Models         | mt→mm,sv,up | mm→mt,sv,up | sv→mt,mm,sv,up | sy→mt,mm,sv,up | up→mt,mm,sv,up | Avg |
|----------------|-------------|-------------|----------------|----------------|----------------|-----|
| Source Only    | 20.5±1.2    | 53.5±0.9    | 62.9±0.3       | 77.9±0.4       | 22.6±0.4       | 47.5|
| DAN (Long et al., 2015) | 21.7±1.0    | 55.3±0.7    | 63.2±0.5       | 79.3±0.2       | 40.2±0.4       | 51.9|
| DANN (Ganin & Lempitsky, 2015) | 22.8±1.1    | 45.2±0.6    | 61.8±0.2       | 79.3±0.3       | 38.7±0.6       | 49.6|
| ADDA (Tzeng et al., 2017) | 23.4±1.3    | 54.8±0.8    | 63.5±0.4       | 79.6±0.3       | 43.5±0.5       | 52.9|
| UFDN (Liu et al., 2018a) | 20.2±1.5    | 41.6±0.7    | 64.5±0.4       | 60.7±0.3       | 44.6±0.2       | 46.3|
| MCD (Saito et al., 2018) | 28.7±1.3    | 43.8±0.8    | 75.1±0.3       | 78.9±0.3       | 55.3±0.4       | 56.4|
| DADA+class (I) | 28.9±1.2    | 50.1±0.9    | 65.4±0.2       | 79.8±0.1       | 50.4±0.3       | 54.9|
| DADA+domain (II) | 34.1±1.7    | 57.1±0.4    | 71.3±0.4       | 82.5±0.3       | 45.4±0.4       | 57.5|
| DADA+ring (III) | 35.3±1.5    | 57.5±0.6    | 80.1±0.3       | 82.9±0.2       | 46.2±0.3       | 60.4|
| DADA+rec (IV) | **39.4±1.4** | **61.1±0.7** | **80.1±0.4**   | **83.7±0.2**   | **47.2±0.4**   | **62.3**|

Figure 3. Feature visualization: t-SNE plot of source features, UFDN (Liu et al., 2018a) features, MCD (Saito et al., 2018) features and DADA features on agnostic target domain in sv → mm, mt, up, sy setting. We use different markers and different colors to denote different categories. (Best viewed in color.)

4.2. Experiments on Office-Caltech10

Office-Caltech10 (Gong et al., 2012) This dataset includes 10 common categories shared by Office-31 (Saenko et al., 2010) and Caltech-256 datasets (Griffin et al., 2007). It contains four domains: Caltech (C), which are sampled from Caltech-256 dataset, Amazon (A), which contains images collected from amazon.com, Webcam (W) and DSLR (D), which are images taken by web camera and DSLR camera under office environment. In our experiments, we take turns to set one domain as the source domain and the rest as the heterogeneous target domain, leading to four DAL tasks.

In our experiments, we leverage two popular networks, AlexNet (Krizhevsky et al., 2012) and ResNet (He et al., 2016), as the backbone of the feature generator G. Both the networks are pre-trained on ImageNet (Deng et al., 2009). Other components are randomly initialized with normal distribution. In the optimization procedure, we set the learning rate of randomly initialized parameters ten times of the pre-trained parameters. The architecture of other components can be seen in Table 7 (supplementary material).

| Models         | mt→mm,sv,up | mm→mt,sv,up | sv→mt,mm,sv,up | sy→mt,mm,sv,up | up→mt,mm,sv,up | Avg |
|----------------|-------------|-------------|----------------|----------------|----------------|-----|
| Source Only    | 20.5±1.2    | 53.5±0.9    | 62.9±0.3       | 77.9±0.4       | 22.6±0.4       | 47.5|
| DAN (Long et al., 2015) | 21.7±1.0    | 55.3±0.7    | 63.2±0.5       | 79.3±0.2       | 40.2±0.4       | 51.9|
| DANN (Ganin & Lempitsky, 2015) | 22.8±1.1    | 45.2±0.6    | 61.8±0.2       | 79.3±0.3       | 38.7±0.6       | 49.6|
| ADDA (Tzeng et al., 2017) | 23.4±1.3    | 54.8±0.8    | 63.5±0.4       | 79.6±0.3       | 43.5±0.5       | 52.9|
| UFDN (Liu et al., 2018a) | 20.2±1.5    | 41.6±0.7    | 64.5±0.4       | 60.7±0.3       | 44.6±0.2       | 46.3|
| MCD (Saito et al., 2018) | 28.7±1.3    | 43.8±0.8    | 75.1±0.3       | 78.9±0.3       | 55.3±0.4       | 56.4|
| DADA+class (I) | 28.9±1.2    | 50.1±0.9    | 65.4±0.2       | 79.8±0.1       | 50.4±0.3       | 54.9|
| DADA+domain (II) | 34.1±1.7    | 57.1±0.4    | 71.3±0.4       | 82.5±0.3       | 45.4±0.4       | 57.5|
| DADA+ring (III) | 35.3±1.5    | 57.5±0.6    | 80.1±0.3       | 82.9±0.2       | 46.2±0.3       | 60.4|
| DADA+rec (IV) | **39.4±1.4** | **61.1±0.7** | **80.1±0.4**   | **83.7±0.2**   | **47.2±0.4**   | **62.3**|

In addition to the baselines mentioned in Section 4.1, we add three baselines: Residual Transfer Network (RTN) (Long et al., 2016), Joint Adaptation Network (JAN) (Long et al., 2017) and Self Ensembling (French et al., 2018). Specifically, RTN employs residual layer (He et al., 2016) for better knowledge transfer, based on DAN (Long et al., 2015). JAN leverages a joint MMD-loss layer to align the features in two consecutive layers. SE applies self-ensembling learning based on a teacher-student model and was the winner of the Visual Domain Adaptation Challenge. We do not apply these methods in digit recognition because the LeNet-based model (LeCun et al., 1989) is too simple to add a residual or joint training layer. We also omit ADDA and UFDN baselines as these models fail to converge while training on Office-Caltech10 under the domain-agnostic setup.

Results The experimental results on Office-Caltech10 dataset are shown in Table 2. For fair comparison, we utilize the same backbone as the baselines and separately show the results. From these results, we make the following observations. (1) Our model achieves 89.8% accuracy with...
Table 2. Accuracy on Office-Caltech10 dataset with DAL protocol. The methods in the above table are based on “AlexNet” backbone and the methods below are based on the “ResNet” backbone. For both backbones, our model outperforms other baselines.

| Method                        | A → C,D,W | C → A,D,W | D → A,C,W | W → A,C,D | Average |
|-------------------------------|-----------|-----------|-----------|-----------|---------|
| AlexNet (Krizhevsky et al., 2012) | 83.1±0.2  | 88.9±0.4  | 86.7±0.4  | 82.2±0.3  | 85.2    |
| DAN (Long et al., 2015)       | 82.5±0.3  | 86.2±0.4  | 75.7±0.5  | 80.4±0.2  | 81.2    |
| RTN (Long et al., 2016)       | 85.2±0.4  | 89.8±0.3  | 81.7±0.3  | 83.7±0.4  | 85.1    |
| JAN (Long et al., 2017)       | 83.5±0.3  | 88.5±0.2  | 80.1±0.3  | 85.9±0.4  | 84.5    |
| DANN (Ganin & Lempitsky, 2015)| 85.9±0.4  | 90.5±0.3  | 88.6±0.4  | 90.4±0.2  | 88.9    |
| DADA (Ours)                   | 86.3±0.3  | 91.7±0.4  | 89.9±0.3  | 91.3±0.3  | 89.8    |
| ResNet (He et al., 2016)      | 90.5±0.3  | 94.3±0.2  | 88.7±0.4  | 82.5±0.3  | 89.0    |
| SE (French et al., 2018)      | 90.3±0.4  | 94.7±0.4  | 88.5±0.3  | 85.3±0.4  | 89.7    |
| MCD (Saito et al., 2018)      | 91.7±0.4  | 95.3±0.3  | 89.5±0.2  | 84.3±0.2  | 90.2    |
| DANN (Ganin & Lempitsky, 2015)| 91.5±0.4  | 94.3±0.4  | 90.5±0.3  | 86.3±0.3  | 90.6    |
| DADA (Ours)                   | 92.0±0.4  | 95.1±0.3  | 91.3±0.4  | 93.1±0.3  | 92.9    |

Figure 4. Empirical analysis: (a) A-Distance of ResNet, MCD and DADA features on two different tasks; (b) training errors and accuracy on C→A,D,W task. (c)-(d) confusion matrices of MCD, and DADA models on W→A,C,D task.

an AlexNet backbone (Krizhevsky et al., 2012) and 92.9% accuracy with a ResNet backbone, outperforming the corresponding baselines on most shifts. (2) The adversarial method (DANN) works better than the feature alignment methods (DAN, RTN, JAN). More interestingly, negative transfer (Pan & Yang, 2010) occurs for feature alignment methods. This is somewhat expected, as these models align the entangled features directly, including the class-irrelevant features. (3) From the ResNet results, we observe limited improvements for the baselines from the source-only model, especially for boosting-based SE method. This phenomenon suggests that the boosting procedure works poorly when the target domain is heterogeneously distributed.

To better analyze the error modes, we plot the confusion matrices for MCD (84.3% accuracy) and DADA (93.1% accuracy) on W→A,C,D task in Figure 4(c)-4(d). The figures illustrate MCD mainly confuses “calculator” vs. “keyboard”, “backpack” vs. “headphones”, and “monitor” vs. “projector”, while DADA is able to distinguish them with disentangled features.

A-Distance Ben-David et al. (2010) suggests A-distance as a measure of domain discrepancy. Following Long et al. (2015), we calculate the approximate A-distance \( d_A = 2(1 - 2\epsilon) \) for W→A,C,D and D→A,C,W tasks, where \( \epsilon \) is the generalization error of a two-sample classifier (kernel SVM) trained on the binary problem to distinguish input samples between the source and target domains. Figure 4(a) displays \( d_A \) for the two tasks with raw ResNet features, MCD features, and DADA features, respectively. We observe that the \( d_A \) of both MCD features and DADA features are smaller than ResNet features, and the \( d_A \) on DADA features is smaller than \( d_A \) on MCD features, which is in consistent with the quantitative results, demonstrating the effectiveness of our disentangled features.

Convergence Analysis As DADA involves multiple losses and a complex learning procedure including adversarial learning and disentanglement, we analyze the convergence performance for the C→A,D,W task, as showed in Figure 4(b) (lines are smoothed for easier analysis). We plot the cross-entropy loss on the source domain, ring loss defined by Equation 9, mutual information defined by Equation 7, and the accuracy in the figure. Figure 4(b) illustrates that the training losses gradually converge and the accuracy become steady after about 20 epochs of training.

Convergence Analysis As DADA involves multiple losses and a complex learning procedure including adversarial learning and disentanglement, we analyze the convergence performance for the C→A,D,W task, as showed in Figure 4(b) (lines are smoothed for easier analysis). We plot the cross-entropy loss on the source domain, ring loss defined by Equation 9, mutual information defined by Equation 7, and the accuracy in the figure. Figure 4(b) illustrates that the training losses gradually converge and the accuracy become steady after about 20 epochs of training.
Table 3. Accuracy on the DomainNet dataset (Peng et al., 2018) dataset with DAL protocol. The table below shows the results based on AlexNet (Krizhevsky et al., 2012) backbone and the below are the results of ResNet (He et al., 2016) backbone. For both setting, our model outperforms other baselines.

| Models          | clp→inf,pnt | inf→clp,pnt, qdc,rel,skt | pnt→clp,inf, qdc,rel,skt | qdr→clp,inf, | rel→clp,inf, | pnt,qdc,rel,skt | Avg   |
|-----------------|-------------|---------------------------|---------------------------|--------------|--------------|-----------------|-------|
| AlexNet (Krizhevsky et al., 2012) | 22.5±0.4 | 15.3±0.2 | 21.2±0.3 | 6.0±0.2 | 17.2±0.3 | 21.8±0.3 | 17.3 |
| DAN (Long et al., 2015) | 23.7±0.3 | 14.9±0.4 | 22.7±0.2 | 7.6±0.3 | 19.4±0.4 | 23.4±0.5 | 18.6 |
| RTN (Long et al., 2016) | 21.4±0.3 | 14.2±0.3 | 21.0±0.4 | 7.7±0.2 | 17.8±0.3 | 20.8±0.4 | 17.2 |
| JAN (Long et al., 2017) | 21.1±0.4 | 16.5±0.2 | 21.6±0.3 | 9.9±0.1 | 15.4±0.2 | 22.5±0.3 | 17.8 |
| DANN (Ganin & Lempitsky, 2015) | 24.1±0.2 | 15.2±0.4 | 24.5±0.3 | 8.2±0.4 | 18.0±0.3 | 24.1±0.4 | 19.1 |
| DADA (Ours) | 23.9±0.4 | 17.9±0.4 | 25.4±0.5 | 9.4±0.2 | 20.5±0.3 | 25.2±0.4 | 20.4 |
| ResNet101 (He et al., 2016) | 25.6±0.2 | 16.8±0.3 | 25.8±0.4 | 9.2±0.2 | 20.6±0.5 | 22.3±0.1 | 20.1 |
| SE (French et al., 2018) | 21.3±0.2 | 8.5±0.1 | 14.5±0.2 | 13.8±0.4 | 16.0±0.4 | 19.7±0.2 | 15.6 |
| MCD (Saito et al., 2018) | 25.1±0.3 | 19.1±0.4 | 27.0±0.3 | 10.4±0.3 | 20.2±0.2 | 22.5±0.4 | 20.7 |
| DADA (Ours) | 26.1±0.4 | 20.0±0.3 | 26.5±0.4 | 12.9±0.4 | 20.7±0.4 | 22.8±0.2 | 21.5 |

Table 4. One-to-one (o-o) vs. one-to-many alignment (o-m). We only show the source domain in the table, the remaining five domains set as the target domain.

| Source | clp | inf | pnt | qdr | rel | skt | Avg |
|--------|-----|-----|-----|-----|-----|-----|-----|
| DAN (o-o) | 25.2 | 14.9 | 24.1 | 7.8 | 20.4 | 25.2 | 19.6 |
| DAN (o-m) | 23.7 | 14.9 | 22.7 | 7.6 | 19.4 | 23.4 | 18.6 |
| JAN (o-o) | 24.2 | 18.1 | 23.2 | 7.8 | 15.8 | 23.8 | 18.8 |
| JAN (o-m) | 21.1 | 16.5 | 21.6 | 9.9 | 15.4 | 22.5 | 17.8 |

4.3. Experiments on the DomainNet dataset

**DomainNet**

(Peng et al., 2018) This dataset contains approximately 0.6 million images distributed among 345 categories. It contains six distinct domains: **Clipart (clp)**, a collection of clipart images; **Infograph (inf)**, infographic images with specific object; **Painting (pnt)**, artistic depictions of object in the form of paintings; **Quickdraw (qdr)**, drawings from the worldwide players of game “Quick Draw!”

- **Real (rel)**, photos and real world images; and **Sketch (skt)**, sketches of specific objects. It is very large-scale and includes rich informative vision cues across different domains, providing a good testbed for DAL. Sample images can be seen from Figure 2. Following Section 4.2, we take turns to set one domain as the source domain and the rest as the heterogeneous target domain, leading to six DAL tasks.

**Results** The experimental results on DomainNet (Peng et al., 2018) are shown in Table 3. The results shows our model achieves **21.5%** accuracy with a ResNet backbone. Note that this dataset contains about 0.6 million images, and so a one percent accuracy improvement is not a trivial achievement. Our model gets comparable results with the best-performing baseline when the source domain is **pnt**, or **qdr** and outperforms other baselines for the rest of the tasks. From the experimental results, we make two interesting observations. (1) In DAL, the SE model (French et al., 2018) performs poorly when the number of categories is large, which is in consistent with results in (Peng et al., 2018). (2) The adversarial alignment method (DANN) performs better than feature alignment methods in DAL, a similar trend to that in Section 4.2.

**One-to-one vs. one-to-many alignment** In the DAL task, the UDA models are performing one-to-many alignment as the target data have no domain labels. However, traditional feature alignment methods such as DAN and JAN are designed for one-to-one alignment. To investigate the effectiveness of domain labels, we design a controlled experiment for DAN and JAN. First, we provide the domain labels and perform one-to-one unsupervised domain adaptation. Then we take away the domain labels and perform one-to-one unsupervised domain adaptation. These results further demonstrate that DAL is a more challenging task and that traditional feature alignment methods need to be re-thought for this problem.

5. Conclusion

In this paper, we first propose a novel domain agnostic learning (DAL) schema and demonstrate the importance of DAL in practical scenarios. Towards tackling DAL task, we have proposed a novel Deep Adversarial Disentangled Autoencoders (DADA) to disentangle domain-invariant features in the latent space. We have proposed to leveraging class disentanglement and mutual information minimizer to enhance the feature disentanglement. Empirically, we demonstrate that the ring-loss-style normalization boosts the performance of DADA in DAL task. An extensive empirical evaluation on DAL benchmarks demonstrate the efficacy of the proposed model against several state-of-the-art domain adaptation algorithms.
6. Acknowledgement

We thank Saito Kuniaki, Ben Usman, Ping Hu for their useful discussions and suggestions. We thank anonymous reviewers and area chairs for their useful insight to improve this work. This work was partially supported by NSF and Honda Research Institute.

References

Belghazi, M. I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., and Hjelm, D. Mutual information neural estimation. In Dy, J. and Krause, A. (eds.), Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research, pp. 531–540, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/belghazi18a.html.

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

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence, 35(8):1798–1828, 2013.

Cao, J., Katzir, O., Jiang, P., Lischinski, D., Cohen-Or, D., Tu, C., and Li, Y. Dida: Disentangled synthesis for domain adaptation. arXiv preprint arXiv:1805.08019, 2018.

Carlucci, F. M., Russo, P., Tommasi, T., and Caputo, B. Agnostic domain generalization. CoRR, abs/1808.01102, 2018a. URL http://arxiv.org/abs/1808.01102.

Carlucci, F. M., Russo, P., Tommasi, T., and Caputo, B. Agnostic domain generalization. CoRR, abs/1808.01102, 2018b. URL http://arxiv.org/abs/1808.01102.

Crammer, K., Kearns, M., and Wortman, J. Learning from multiple sources. Journal of Machine Learning Research, 9(Aug):1757–1774, 2008.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pp. 248–255. IEEE, 2009.

Duan, L., Xu, D., and Chang, S.-F. 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, pp. 1338–1345. IEEE, 2012.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In Precup, D. and Teh, Y. W. (eds.), Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research, pp. 1126–1135, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/finn17a.html.

French, G., Mackiewicz, M., and Fisher, M. Self-ensembling for visual domain adaptation. In International Conference on Learning Representations, 2018. URL https://openreview.net/forum?id=rkpoTaxA-.

Ganin, Y. and Lempitsky, V. Unsupervised domain adaptation by backpropagation. In Bach, F. and Blei, D. (eds.), Proceedings of the 32nd International Conference on Machine Learning, volume 37 of Proceedings of Machine Learning Research, pp. 1180–1189, Lille, France, 07–09 Jul 2015. PMLR. URL http://proceedings.mlr.press/v37/ganin15.html.

Ghifary, M., Kleijn, W. B., and Zhang, M. Domain adaptive neural networks for object recognition. In Pacific Rim international conference on artificial intelligence, pp. 898–904. Springer, 2014.

Gong, B., Shi, Y., Sha, F., and Grauman, K. Geodesic flow kernel for unsupervised domain adaptation. In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pp. 2066–2073. IEEE, 2012.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In Advances in neural information processing systems, pp. 2672–2680, 2014.

Gretton, A., Borgwardt, K. M., Rasch, M., Schölkopf, B., and Smola, A. J. A kernel method for the two-sample-problem. In Advances in neural information processing systems, pp. 513–520, 2007.

Griffin, G., Holub, A., and Perona, P. Caltech-256 object category dataset. 2007.

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

Hoffman, J., Darrell, T., and Saenko, K. Continuous manifold based adaptation for evolving visual domains. In Computer Vision and Pattern Recognition (CVPR), 2014.
Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., Efros, A., and Darrell, T. CyCADA: Cycle-consistent adversarial domain adaptation. In Dy, J. and Krause, A. (eds.), Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research, pp. 1989–1998, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/hoffman18a.html.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In International conference on machine learning, pp. 448–456, 2015.

Kiefer, J., Wolfowitz, J., et al. Stochastic estimation of the maximum of a regression function. The Annals of Mathematical Statistics, 23(3):462–466, 1952.

Kim, T., Cha, M., Kim, H., Lee, J. K., and Kim, J. Learning to discover cross-domain relations with generative adversarial networks. In Precup, D. and Teh, Y. W. (eds.), Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research, pp. 1857–1865, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/kim17a.html.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013.

Kingma, D. P., Mohamed, S., Rezende, D. J., and Welling, M. Semi-supervised learning with deep generative models. In Advances in neural information processing systems, pp. 3581–3589, 2014.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pp. 1097–1105, 2012.

LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., and Jackel, L. Backpropagation applied to handwritten zip code recognition. Neural Computation, 1989.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998.

Lee, H.-Y., Tseng, H.-Y., Huang, J.-B., Singh, M., and Yang, M.-H. Diverse image-to-image translation via disentangled representations. In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y. (eds.), Computer Vision – ECCV 2018, pp. 36–52, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01246-5.

Li, H., Jialin Pan, S., Wang, S., and Kot, A. C. Domain generalization with adversarial feature learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5400–5409, 2018a.

Li, Y., Gong, M., Tian, X., Liu, T., and Tao, D. Domain generalization via conditional invariant representation, 2018b.

Liu, A. H., Liu, Y., Yeh, Y., and Wang, Y. F. A unified feature disentangler for multi-domain image translation and manipulation. CoRR, abs/1809.01361, 2018a. URL http://arxiv.org/abs/1809.01361.

Liu, M.-Y. and Tuzel, O. Coupled generative adversarial networks. In Advances in neural information processing systems, pp. 469–477, 2016.

Liu, Y.-C., Yeh, Y.-Y., Fu, T.-C., Wang, S.-D., Chiu, W.-C., and Wang, Y.-C. F. Detach and adapt: Learning cross-domain disentangled deep representation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018b.

Long, M., Cao, Y., Wang, J., and Jordan, M. Learning transferable features with deep adaptation networks. In Bach, F. and Blei, D. (eds.), Proceedings of the 32nd International Conference on Machine Learning, volume 37 of Proceedings of Machine Learning Research, pp. 97–105, Lille, France, 07–09 Jul 2015. PMLR. URL http://proceedings.mlr.press/v37/long15.html.

Long, M., Zhu, H., Wang, J., and Jordan, M. I. Unsupervised domain adaptation with residual transfer networks. In Advances in Neural Information Processing Systems, pp. 136–144, 2016.

Long, M., Zhu, H., Wang, J., and Jordan, M. I. 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, pp. 2208–2217, 2017. URL http://proceedings.mlr.press/v70/long17a.html.

Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., and Frey, B. Adversarial autoencoders. ICLR workshop, 2016.

Mansour, Y., Mohri, M., Rostamizadeh, A., and R, A. Domain adaptation with multiple sources. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L. (eds.), Advances in Neural Information Processing Systems 21, pp. 1041–1048. Curran Associates, Inc., 2009.
Mathieu, M. F., Zhao, J. J., Zhao, J., Ramesh, A., Sprechmann, P., and LeCun, Y. Disentangling factors of variation in deep representation using adversarial training. In *Advances in Neural Information Processing Systems*, pp. 5040–5048, 2016.

Odena, A., Olah, C., and Shlens, J. Conditional image synthesis with auxiliary classifier GANs. In *Advances in Neural Information Processing Systems*, pp. 2642–2651, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/odena17a.html.

Pan, S. J. and Yang, Q. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

Peng, X. and Saenko, K. 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*, pp. 1982–1991, 2018. doi: 10.1109/WACV.2018.00219. URL https://doi.org/10.1109/WACV.2018.00219.

Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., and Wang, B. Moment matching for multi-source domain adaptation. *arXiv preprint arXiv:1812.01754*, 2018.

Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. *Dataset Shift in Machine Learning*. The MIT Press, 2009. ISBN 0262170051, 9780262170055.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 1278–1286, Beijing, China, 22–24 Jun 2014. PMLR. URL http://proceedings.mlr.press/v32/rezende14.html.

Romijnders, R., Meletis, P., and Dubbelman, G. A domain agnostic normalization layer for unsupervised adversarial domain adaptation. *CoRR*, abs/1809.05298, 2018. URL http://arxiv.org/abs/1809.05298.

Saenko, K., Kulis, B., Fritz, M., and Darrell, T. Adapting visual category models to new domains. In *European conference on computer vision*, pp. 213–226. Springer, 2010.

Saito, K., Watanabe, K., Ushiku, Y., and Harada, T. Maximum classifier discrepancy for unsupervised domain adaptation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Sun, B. and Saenko, K. Deep CORAL: correlation alignment for deep domain adaptation. *CoRR*, abs/1607.01719, 2016. URL http://arxiv.org/abs/1607.01719.

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

Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. Adversarial discriminative domain adaptation. In *Computer Vision and Pattern Recognition (CVPR)*, volume 1, pp. 4, 2017.

Xu, R., Chen, Z., Zuo, W., Yan, J., and Lin, L. Deep cocktail network: Multi-source unsupervised domain adaptation with category shift. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3964–3973, 2018.

Yi, Z., Zhang, H. R., Tan, P., and Gong, M. Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*, pp. 2868–2876, 2017.

Zellinger, W., Grubinger, T., Lughofer, E., Natschläger, T., and Saminger-Platz, S. Central moment discrepancy (CMD) for domain-invariant representation learning. *CoRR*, abs/1702.08811, 2017. URL http://arxiv.org/abs/1702.08811.

Zheng, Y., Pal, D. K., and Savvides, M. Ring loss: Convex feature normalization for face recognition. *arXiv preprint arXiv:1803.00130*, 2018.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.
Supplementary Materials

A. Model Architecture

We provide the detailed model architecture (Table 5 and Table 7) for each component in our model: Generator, Disentangler, Domain Classifier, Classifier and MINE.

Table 5. Model Architecture for ‘Digit-Five’. For each convolution layer, we list the input dimension, output dimension, kernel size, stride, and padding. For the fully-connected layer, we provide the input and output dimensions. For drop-out layers, we provide the probability of an element to be zeroed.

| layer | configuration |
|-------|---------------|
| Feature Generator | |
| 1 | Conv2D (3, 64, 5, 1, 2), BN, ReLU, MaxPool |
| 2 | Conv2D (64, 64, 5, 1, 2), BN, ReLU, MaxPool |
| 3 | Conv2D (64, 128, 5, 1, 2), BN, ReLU |
| Disentangler | |
| 1 | Dropout(0.5), FC (2048, 2048), BN, ReLU |
| 2 | Dropout(0.5), FC (2048, 2048), BN, ReLU |
| Domain Identifier | |
| 1 | FC (2048, 256), LeakyReLU |
| 2 | FC (256, 2), LeakyReLU |
| Class Identifier | |
| 1 | FC (2048, 10), BN, Softmax |
| Reconstructor | |
| 1 | FC (4096, 8192) |
| Mutual Information Estimator | |
| fc1_x | FC (2048, 512) |
| fc1_y | FC (2048, 512), LeakyReLU |
| 2 | FC (512,1) |

B. Details of datasets

We provide the detailed information of datasets (Table 6). For Digit-Five and the DomainNet dataset, we provide the train/test split for each domain and for Office-Caltech10, we provide the number of images in each domain.

Table 6. Detailed information for datasets

| Digit-Five | Splits | mnist | mnist_m | svhn | syn | usps | Total |
|------------|--------|-------|---------|------|-----|-----|-------|
| Train      | 55,000 | 55,000| 25,000  | 25,000 | 7,348 | 167,348 |
| Test       | 10,000 | 10,000| 14,549  | 9,000 | 1,860 | 37,309 |

Office-Caltech10

| Splits | amazon | caltech | dslr | webcam | Total |
|--------|--------|---------|------|--------|-------|
| Train  | 34,019 | 37,087  | 52,867 | 120,750 | 416,401 |
| Test   | 14,818 | 16,114  | 22,892 | 51,750 | 179,609 |

DomainNet

| Splits | clp | inf | pnt | qdr | rel | abt | Total |
|--------|-----|-----|-----|-----|-----|-----|-------|
| Train  | 34,019 | 37,087 | 52,867 | 120,750 | 122,563 | 49,115 | 416,401 |
| Test   | 14,818 | 16,114 | 22,892 | 51,750 | 52,764 | 21,271 | 179,609 |

Table 7. Model Architecture for ‘Office-Caltech10’ and ‘DomainNet’. For each convolution layer, we list the input dimension, output dimension, kernel size, stride, and padding. For the fully-connected layer, we provide the input and output dimensions. For drop-out layers, we provide the probability of an element to be zeroed.

| layer | configuration |
|-------|---------------|
| Feature Generator: ResNet101 or AlexNet | |
| Disentangler | |
| 1 | Dropout(0.5), FC (2048, 2048), BN, ReLU |
| 2 | Dropout(0.5), FC (2048, 2048), BN, ReLU |
| Domain Identifier | |
| 1 | FC (2048, 256), LeakyReLU |
| 2 | FC (256, 2), LeakyReLU |
| Class Identifier | |
| 1 | FC (2048, 10), BN, Softmax |
| Reconstructor | |
| 1 | FC (4096, 2048) |
| Mutual Information Estimator | |
| fc1_x | FC (2048, 512) |
| fc1_y | FC (2048, 512), LeakyReLU |
| 2 | FC (512,1) |