Detach and Adapt: Learning Cross-Domain Disentangled Deep Representation

Yen-Cheng Liu[1], Yu-Ying Yeh[1], Tzu-Chien Fu[2], Sheng-De Wang[1],
Wei-Chen Chiu[3], Yu-Chiang Frank Wang[1]
[1]Department of Electrical Engineering, National Taiwan University, Taiwan
[2]Department of Electrical Engineering & Computer Science, Northwestern University, USA
[3]Department of Computer Science, National Chiao Tung University, Taiwan
{r04921003, b99202023}@ntu.edu.tw, tcfu@u.northwestern.edu, sdwang@ntu.edu.tw,
walon@cs.nctu.edu.tw, ycwang@ntu.edu.tw

Abstract

While representation learning aims to derive interpretable features for describing visual data, representation disentanglement further results in such features so that particular image attributes can be identified and manipulated. However, one cannot easily address this task without observing ground truth annotation for the training data. To address this problem, we propose a novel deep learning model of Cross-Domain Representation Disentangler (CDRD). By observing fully annotated source-domain data and unlabeled target-domain data of interest, our model bridges the information across data domains and transfers the attribute information accordingly. Thus, cross-domain feature disentanglement and adaptation can be jointly performed. In the experiments, we provide qualitative results to verify our disentanglement capability. Moreover, we further confirm that our model can be applied for solving classification tasks of unsupervised domain adaptation, and performs favorably against state-of-the-art image disentanglement and translation methods.

1. Introduction

The development of deep neural networks benefits a variety of areas such as computer vision, machine learning, and natural language processing, which results in promising progresses in realizing artificial intelligence environments. However, as pointed out in [1], it is fundamental and desirable for understanding the observed information around us. To be more precise, the above goal is achieved by identifying and disentangling the underlying explanatory factors hidden in the observed data and the derived learning models. Therefore, the challenge of representation learning is to have the learned latent element explanatory and disentangled from the derived abstract representation.

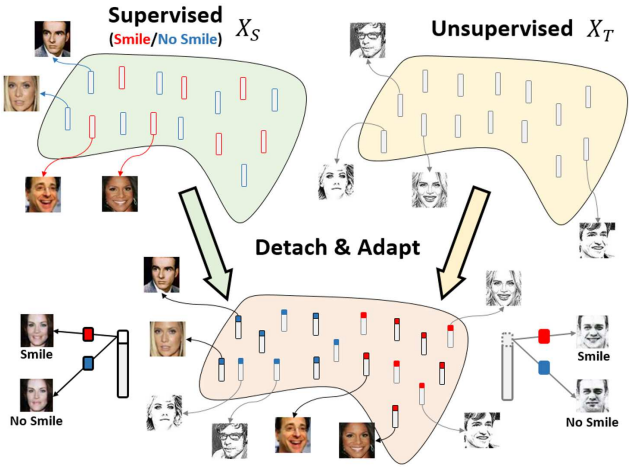


Figure 1: Illustration of cross-domain representation disentanglement. With attributes observed only in the source domain, we are able to disentangle, adapt, and manipulate the data across domains with particular attributes of interest.

With the goal of discovering the underlying factors of data representation associated with particular attributes of interest, representation disentanglement is the learning task which aims at deriving a latent feature space that decomposes the derived representation so that the aforementioned attributes (e.g., face identity/pose, image style, etc.) can be identified and described. Several works have been proposed to tackle this task in unsupervised [3, 10], semi-supervised [14, 24], or fully supervised settings [16, 25]. Once attribute of interest properly disentangled, one can produce the output images with particular attribute accordingly.

However, like most machine learning algorithms, representation disentanglement is not able to achieve satisfactory performances if the data to be described/manipulated are very different from the training ones. This is known as the problem of *domain shift* (or *domain/dataset bias*), and re-

quires the advance of transfer learning [26] or domain adaptation [27] techniques to address this challenging yet practical problem. Similarly, learning of deep neural networks for interpretable and disentangled representation generally requires a large number of annotated data, and also suffers from the above problem of domain shift.

To adapt cross-domain data for transferring the desirable knowledge such as label or attribute, one typically utilizes pre-collected or existing annotated data as source-domain training data, together with unlabeled data in the target domain of interest, for deriving the learning model. Since only unlabeled target-domain data is observed in the above scenario, it is considered as *unsupervised domain adaptation*. For example, given face images with expression annotation as source-domain data X_S, and facial sketches X_T without any annotation as target-domain data of interest (see Figure 1 for illustration), the goal of *cross-domain feature disentanglement* is to distinguish the latent feature corresponding to the expression by observing both X_S and X_T.

In this paper, we propose a novel deep neural networks architecture based on generative adversarial networks (GAN) [9]. As depicted in Figure 2, our proposed network observes cross-domain data with partial supervision (i.e., only annotation in X_S is available), and performs representation learning and disentanglement in the resulting shared latent space. It is worth noting that this can be viewed as a novel learning task of joint representation disentanglement and domain adaptation in an unsupervised setting, since only unlabeled data is available in the target domain during the training stage. Later in the experiments, we will further show that the derived feature representation can be applied to describe data from both source and target domains, and classification of target-domain data can be achieved with very promising performances.

We highlight the contributions of this paper as follows:

- To the best of our knowledge, we are the *first* to tackle the problem of representation disentanglement for cross-domain data.
- We propose an *end-to-end* learning framework for joint representation disentanglement and adaptation, while only attribute supervision is available in the source domain.
- Our proposed model allows one to perform conditional cross-domain image synthesis and translation.
- Our model further addresses the domain adaptation task of attribute classification. This qualitatively verifies our capability in describing and recognizing cross-domain data with particular attributes of interest.

2. Related Works

Representation Disentanglement

Disentangling the latent factors from the image variants has led to the understanding of the observed data [16, 25, 14, 24, 3, 10]. For example, by training from a sufficient amount of fully annotated data, Kulkarni *et al.* [16] proposed to learn interpretable and invertible graphics code when rendering image 3D models. Odena *et al.* [25] augmented the architecture of generative adversarial networks (GAN) with an auxiliary classifier. Given ground truth label/attribute information during training, the model enables the synthesized images to be conditioned on the desirable latent factors. Kingma *et al.* [14] extended variational autoencoder (VAE) [15] to achieve semi-supervised learning for disentanglement. Chen *et al.* [3] further tackles this task in an unsupervised manner by maximizing the mutual information between pre-specified latent factors and the rendered images; however, the semantic meanings behind the disentangled factors cannot be explicitly obtained. Despite the promising progress in the above methods on deep representation disentanglement, most existing works only focus on handling and manipulating data from a single domain of interest. In practical scenarios, such settings might not be of sufficient use. This is the reason why, in this work, we aim at learning and disentangling representation across data domains in an unsupervised setting (i.e., only source-domain data are with ground truth annotation).

Adaptation Across Visual Domains

Domain adaptation addresses the same learning task from data across domains. It requires one to transfer information from one (or multiple) domain(s) to another, while the domain shift is expected. In particular, unsupervised domain adaptation (UDA) deals with the task that no label supervision is available during training in the target domain. For existing UDA works, Long *et al.* [22] learned cross-domain projection for mapping data across domains into a common subspace. Long *et al.* [23] further proposed to reweight the instances across domains to alleviate the domain bias, and Ghifary *et al.* [6] presented scatter component analysis to maximize the separability of classes and minimize the mismatch across domains. Zhang *et al.* [33] utilized coupled dimension reduction across data domains to reduce the geometrical and distribution differences.

Inspired by the adversarial learning scheme [9], several deep learning based methods have been proposed for solving domain adaptation tasks. For example, Ganin *et al.* [5] introduced a domain classifier in a standard architecture of convolutional neural networks (CNN), with its gradient reversal layer serving as a domain-adaptive feature extractor despite the absence of labeled data in the target domain. Similarly, Tzeng *et al.* [30] utilized the domain confu-

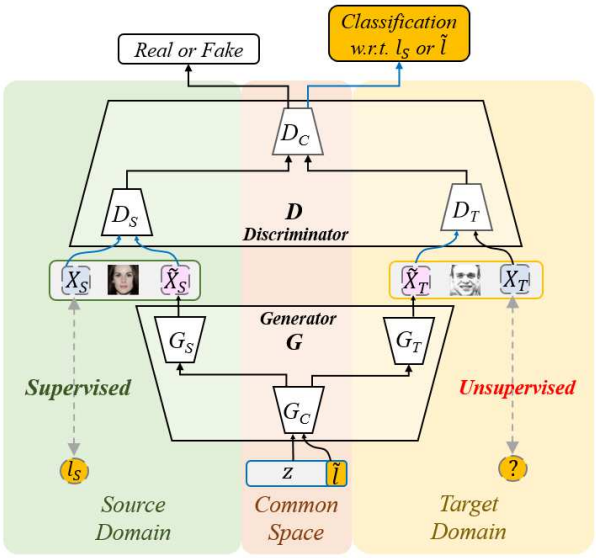


Figure 2: The network architecture of Cross-Domain Representation Disentangler (CDRD). Note that while source and target-domain data are presented during training, only attribute supervision is available in the source domain, and no cross-domain data pair is needed.

sion loss to learn shared representation for describing cross-domain data. Tzeng *et al.* [31] applied the architecture of weight sharing layers between feature extractors of source and target domains, which allows the learning of domain-specific feature embedding by utilizing such domain adversarial training strategies.

With the goal of converting images in one style to another, image–to–image translation can be viewed as another learning task that handles cross-domain visual data. For example, Isola *et al.* [11] approached this task by applying pairs of images for learning GAN-based models. Taigman *et al.* [29] performed such tasks by employing feature consistency across domains.Without the need to observe cross-domain image pairs, Zhu *et al.* [34] learned the dual domain mappings with a cycle consistency loss. Similar ideas can be found in [13] and [32]. Coupled GAN (CoGAN) [20] ties high-level information between two image domains for simultaneously rendering corresponding cross-domain images, and UNIT [19] is considered as an extended version of CoGAN, which integrates VAE and GAN to learn image translation in an unsupervised manner.

It is worth pointing out that, although approaches based on image translation are able to convert images from one domain to another, they do not exhibit the ability in learning and disentangling desirable latent representation (as ours does). As verified later in the experiments, the latent representation derived by image translation models cannot produce satisfactory classification performance for domain adaptation either.

3. Proposed Method

The objective of our proposed model, *Cross-Domain Representation Disentangler (CDRD)*, is to perform joint representation disentanglement and domain adaptation (as depicted in Figure 2). With only label supervision available in the source domain, our CDRD derives deep *disentangled feature representation* z with a corresponding *disentangled latent factor* $\tilde{l}$ for describing cross-domain data and their attributes, respectively. We now detail our proposed architecture of CDRD in the following subsections.

3.1. Cross-Domain Representation Disentangler

Since both AC-GAN [25] and InfoGAN[3] are known to learn interpretable feature representation using deep neural networks (in supervised and unsupervised settings, respectively), it is necessary to briefly review their architecture before introducing ours. Based on the recent success of GAN [9], both AC-GAN and InfoGAN take noise and additional class/condition as the inputs to the generator, while the label prediction is additionally performed at the discriminator for the purpose of learning disentangled features. As noted above, since both AC-GAN and InfoGAN are not designed to learn/disentangle representation for data across different domains, they cannot be directly applied for **cross-domain** representation disentanglement.

To address this problem, we propose a novel network architecture of cross-domain representation disentangler (CDRD). As depicted in Figure 2, our CDRD model consists of two major components: Generators $\{G_S, G_T, G_C\}$, and Discriminators $\{D_S, D_T, D_C\}$. Similar to AC-GAN and InfoGAN, we have an auxiliary classifier attached at the end of the network, which shares all the convolutional layers with the discriminator D_C, followed by a fully connected layer to predict the label/attribute outputs. Thus, we regard our discriminator as a multi-task learning model, which not only distinguishes between synthesized and real images but also recognizes the associated image attributes.

To handle cross-domain data with only supervision from the source domain, we choose to share weights in higher layers in G and D, aiming at bridging the gap between high/coarse-level representations of cross-domain data. To be more precise, we split G and D in CDRD into multiple sub-networks specialized for describing data in the source domain $\{G_S, D_S\}$, target domain $\{G_T, D_T\}$, and the common latent space $\{G_C, D_C\}$ (see the green, yellow, and red-shaded colors in Figure 2, respectively).

Following the challenging setting of unsupervised domain adaptation, each input image X_S in the source domain is associated with a ground truth label l_S, while unsupervised learning is performed in the target domain. Thus, the common latent representation z in the input of CDRD together with a randomly assigned attribute $\tilde{l}$ would be the inputs for the generator. For the synthesized images $\tilde{X}_S$ and

Algorithm 1: Learning of CDRD

Data: Source domain: X_S and l_S; Target domain: X_T
Result: Configurations of CDRD

$\theta_G, \theta_D \leftarrow$ initialize
for *Iters. of whole model* **do**
  $z \leftarrow$ sample from $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$
  $\tilde{l} \leftarrow$ sample from attribute space
  $\tilde{X}_S, \tilde{X}_T \leftarrow$ sample from (1)
  $X_S, X_T \leftarrow$ sample mini-batch
  $\mathcal{L}_{adv}, \mathcal{L}_{dis} \leftarrow$ calculate by (2), (3)
  for *Iters. of updating generator* **do**
    $\theta_G \overset{+}{\leftarrow} -\Delta_{\theta_G}(-\mathcal{L}_{adv} + \lambda\mathcal{L}_{dis})$
  for *Iters. of updating discriminator* **do**
    $\theta_D \overset{+}{\leftarrow} -\Delta_{\theta_D}(\mathcal{L}_{adv} + \lambda\mathcal{L}_{dis})$

$\tilde{X}_T$, we have:

$$\tilde{X}_S \sim G_S(G_C(z, \tilde{l})), \tilde{X}_T \sim G_T(G_C(z, \tilde{l})) \tag{1}$$

The objective functions for adversarial learning in source and target domain are now defined as follows:

$$\mathcal{L}_{adv}^S = \mathbb{E}[\log(D_C(D_S(X_S)))] + \mathbb{E}[\log(1 - D_C(D_S(\tilde{X}_S)))],$$

$$\mathcal{L}_{adv}^T = \mathbb{E}[\log(D_C(D_T(X_T)))] + \mathbb{E}[\log(1 - D_C(D_T(\tilde{X}_T)))],$$

$$\mathcal{L}_{adv} = \mathcal{L}_{adv}^S + \mathcal{L}_{adv}^T. \tag{2}$$

Let $P(l|X)$ be a probability distribution over labels/attributes l calculated by the discriminator in CDRD. The objective functions for cross-domain representation disentanglement are defined below:

$$\mathcal{L}_{dis}^S = \mathbb{E}[\log P(l = \tilde{l}|\tilde{X}_S)] + \mathbb{E}[\log P(l = l_S|X_S)],$$

$$\mathcal{L}_{dis}^T = \mathbb{E}[\log P(l = \tilde{l}|\tilde{X}_T)] \quad ,$$

$$\mathcal{L}_{dis} = \mathcal{L}_{dis}^S + \mathcal{L}_{dis}^T. \tag{3}$$

With the above loss terms determined, we learn our CDRD by alternatively updating Generator and Discriminator with the following gradients:

$$\begin{aligned} \theta_G &\overset{+}{\leftarrow} -\Delta_{\theta_G}(-\mathcal{L}_{adv} + \lambda\mathcal{L}_{dis}) \\ \theta_D &\overset{+}{\leftarrow} -\Delta_{\theta_D}(\mathcal{L}_{adv} + \lambda\mathcal{L}_{dis}) \end{aligned} \tag{4}$$

We note that the hyperparameter λ is used to control the disentanglement ability. We will show its effect on the resulting performances in the experiments.

Similar to the concept in InfoGAN [3], the auxiliary classifier in D_C is to maximize the mutual information between the assigned label $\tilde{l}$ and the synthesized images in the source and target domains (i.e., $G_S(G_C(z, \tilde{l}))$ and $G_T(G_C(z, \tilde{l}))$). With network weights in high-level layers shared between source and target domains in both G and D, the disentanglement ability is introduced to the target domain by updating the parameters in G_T according to $\mathcal{L}_{dis}^T$ during the training process.

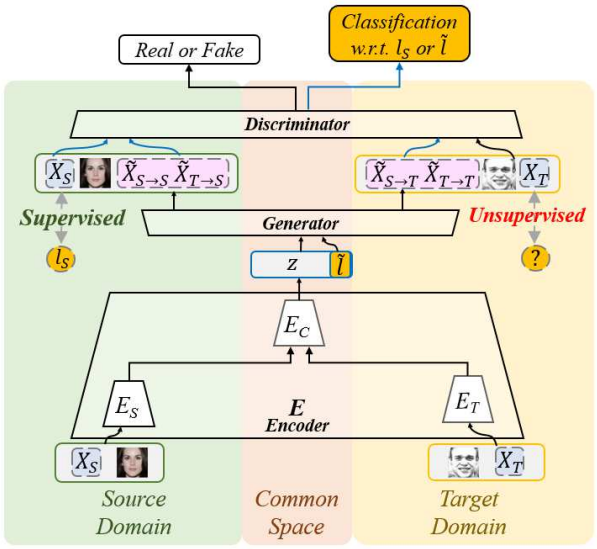


Figure 3: Our proposed architecture of Extended Cross-Domain Representation Disentangler (E-CDRD), which jointly performs cross-domain representation disentanglement and image translation.

3.2. Extended CDRD (E-CDRD)

Our CDRD can be further extended to perform joint image translation and disentanglement by adding an additional component of Encoder $\{E_S, E_T, E_C\}$ prior to the architecture of CDRD, as shown in Figure 3. Such Encoder-Generator pairs can be viewed as VAE models [15] for directly handling image variants in accordance with $\tilde{l}$.

It is worth noting that, as depicted in Figure 3, the Encoder $\{E_S, E_C\}$ and the Generator $\{G_S, G_C\}$ constitute a VAE module for describing source-domain data. Similar remarks can be applied for $\{E_T, E_C\}$ and $\{G_T, G_C\}$ in the target domain. It can be seen that, the components E_S and E_T first transform input real images X_S and X_T into a common feature, which is then encoded by E_C as latent representation:

$$\begin{aligned} z_S &\sim E_C(E_S(X_S)) = q_S(z_S|X_S), \\ z_T &\sim E_C(E_T(X_T)) = q_T(z_T|X_T). \end{aligned} \tag{5}$$

Once the latent representations z_S and z_T are obtained, the remaining architecture is the standard CDRD, which can be applied to recover the images with the assigned $\tilde{l}$ in the associated domains, i.e. $\tilde{X}_{S\to S}$ and $\tilde{X}_{T\to T}$:

$$\tilde{X}_{S\to S} \sim G_S(G_C(z_S, \tilde{l})), \tilde{X}_{T\to T} \sim G_T(G_C(z_T, \tilde{l})). \tag{6}$$

The VAE regularizes the Encoder by imposing a prior over the latent distribution $p(z)$. Typically we have $z \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. In E-CDRD, we advance the objective functions of VAE for each data domain as follows:

$$\mathcal{L}_{vae}^S = \|\Phi(X_S) - \Phi(\tilde{X}_{S\to S})\|_F^2 + KL(q_S(z_S|X_S)||p(z))$$

Algorithm 2: Learning of E-CDRD

Data: Source domain: X_S and l_S; Target domain: X_T
Result: Configurations of E-CDRD

$\theta_E, \theta_G, \theta_D \leftarrow$ initialize
for *Iters. of whole model* **do**
  $z \leftarrow$ sample from $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$
  $X_S, X_T \leftarrow$ sample mini-batch
  $z_S, z_T \leftarrow$ sample from (5)
  $\tilde{l} \leftarrow$ sample from attribute space
  $\tilde{X}_S, \tilde{X}_T \leftarrow$ sample from (1)
  $\tilde{X}_{S\to S}, \tilde{X}_{T\to T}, \tilde{X}_{S\to T}, \tilde{X}_{T\to S} \leftarrow$ sample from (6), (8)
  $\mathcal{L}_{vae}, \mathcal{L}_{adv}, \mathcal{L}_{dis} \leftarrow$ calculate by (7), (9), (10)
  for *Iters. of updating encoder* **do**
    $\theta_E \xleftarrow{+} -\Delta_{\theta_E}(\mathcal{L}_{vae})$
  for *Iters. of updating generator* **do**
    $\theta_G \xleftarrow{+} -\Delta_{\theta_G}(\mathcal{L}_{vae} - \mathcal{L}_{adv} + \lambda\mathcal{L}_{dis})$
  for *Iters. of updating discriminator* **do**
    $\theta_D \xleftarrow{+} -\Delta_{\theta_D}(\mathcal{L}_{adv} + \lambda\mathcal{L}_{dis})$

$$\mathcal{L}_{vae}^T = \|\Phi(X_T) - \Phi(\tilde{X}_{T\to T})\|_F^2 + KL(q_T(z_T|X_T)||p(z))$$
$$\mathcal{L}_{vae} = \mathcal{L}_{vae}^S + \mathcal{L}_{vae}^T \tag{7}$$

where the first term denotes the *perceptual loss* [12], which calculates the reconstruction error between the synthesized output $\tilde{X}_{S\to S}$ (or $\tilde{X}_{T\to T}$) and its original input X_S (or X_T) with network transformation Φ (the similarity metric of [17] is applied in our work). On the other hand, the second term indicates *Kullback-Leibler divergence* over the auxiliary distribution $q_S(z_S|X_S)$, $q_T(z_T|X_T)$ and the prior $p(z)$.

Moreover, similar to the task of image translation, our E-CDRD also outputs images in particular domains accordingly, i.e. $\tilde{X}_{S\to T}$ is translated from the source to target domain, and $\tilde{X}_{T\to S}$ is the output from the target to source domain:

$$\tilde{X}_{S\to T} \sim G_T(G_C(z_S, \tilde{(l)})), \tilde{X}_{T\to S} \sim G_S(G_C(z_T, \tilde{(l)})). \tag{8}$$

With the above observations, the objective functions with adversarial learning for E-CDRD are modified as follows:

$$\begin{aligned}\mathcal{L}_{adv}^S = \mathbb{E}[\log(D_C(D_S(X_S)))] &+ \mathbb{E}[\log(1 - D_C(D_S(\tilde{X}_S)))] \\ &+ \mathbb{E}[\log(1 - D_C(D_S(\tilde{X}_{S\to S})))] \\ &+ \mathbb{E}[\log(1 - D_C(D_S(\tilde{X}_{T\to S})))],\end{aligned}$$

$$\begin{aligned}\mathcal{L}_{adv}^T = \mathbb{E}[\log(D_C(D_T(X_T)))] &+ \mathbb{E}[\log(1 - D_C(D_T(\tilde{X}_T)))] \\ &+ \mathbb{E}[\log(1 - D_C(D_T(\tilde{X}_{T\to T})))] \\ &+ \mathbb{E}[\log(1 - D_C(D_T(\tilde{X}_{S\to T})))],\end{aligned}$$

$$\mathcal{L}_{adv} = \mathcal{L}_{adv}^S + \mathcal{L}_{adv}^T. \tag{9}$$

Similarly, we revise the objective functions for representation disentanglement as follows:

$$\begin{aligned}\mathcal{L}_{dis}^S = &\mathbb{E}[\log P(l = \tilde{l}|\tilde{X}_S)] + \mathbb{E}[\log P(l = l_S|X_S)] \\ &+ \mathbb{E}[\log P(l = \tilde{l}|\tilde{X}_{S\to S})] + \mathbb{E}[\log P(l = \tilde{l}|\tilde{X}_{T\to S})]\end{aligned}$$

$$\begin{aligned}\mathcal{L}_{dis}^T = &\mathbb{E}[\log P(l = \tilde{l}|\tilde{X}_T)] \\ &+ \mathbb{E}[\log P(l = \tilde{l}|\tilde{X}_{T\to T})] + \mathbb{E}[\log P(l = \tilde{l}|\tilde{X}_{S\to T})]\end{aligned}$$

$$\mathcal{L}_{dis} = \mathcal{L}_{dis}^S + \mathcal{L}_{dis}^T. \tag{10}$$

To train our E-CDRD, we alternatively update Encoder, Generator and Discriminator with the following gradients:

$$\begin{aligned}\theta_E &\xleftarrow{+} -\Delta_{\theta_E}(\mathcal{L}_{vae}) \\ \theta_G &\xleftarrow{+} -\Delta_{\theta_G}(\mathcal{L}_{vae} - \mathcal{L}_{adv} + \lambda\mathcal{L}_{dis}) \\ \theta_D &\xleftarrow{+} -\Delta_{\theta_D}(\mathcal{L}_{adv} + \lambda\mathcal{L}_{dis})\end{aligned} \tag{11}$$

It is worth noting that, by jointly considering the above objective functions of VAE and those for adversarial and disentanglement learning, our E-CDRD can be applied for conditional cross-domain image synthesis and translation. Similar to CDRD, the hyperparameter λ controls the ability of E-CDRD for performing disentanglement (and will be analyzed in the experiments).

Finally, the pseudo code for training our CDRD and E-CDRD are summarized in Algorithms 1 and 2, respectively. Implementation details of our network architectures will be presented in the supplementary materials.

4. Experiments

We now evaluate the performance of our proposed method, which is applied to perform cross-domain representation disentanglement and adaptation simultaneously. As noted in Section 3.1, the discriminator in our CDRD (or E-CDRD) is augmented with an auxiliary classifier, which classifies images with respect to the disentangled latent factor l. With only supervision from the source-domain data, such a classification task is also considered as the task of unsupervised domain adaptation (UDA) for cross-domain visual classification. We will also provide quantitative UDA results to further support the use of our model for describing, manipulating, and recognizing cross-domain data with particular attributes of interest.

4.1. Datasets

We consider three different types of datasets, including digit, face, and scene, for performance evaluation:

Digits. *MNIST*, *USPS* and *Semeion* [18] are hand-written digit image datasets, which are viewed as different data domains. MNIST contains 60K/10K instances for training/testing, and USPS consists of 7291/2007 instances for training/testing. Semeion contains 1593 handwritten digits provided by about 80 persons, stretched in a rectangular box

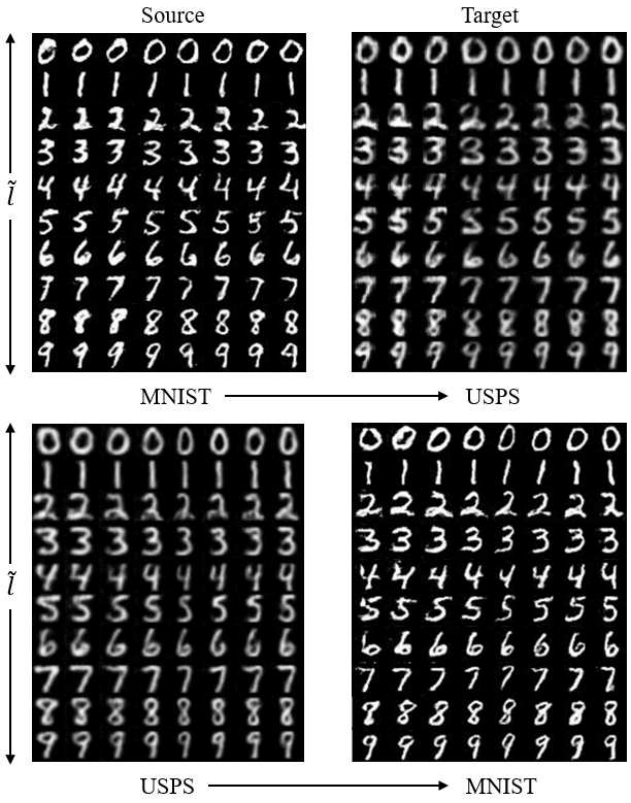


Figure 4: Cross-domain conditional image synthesis for MNIST → USPS and USPS → MNIST with the attribute $\tilde{l}$ as *digits*.

16x16 in a gray scale of 256 values. We resize these images to 28x28 pixels to match the resolution of the images in MNIST and USPS. For UDA, we follow the same protocol in [22, 23] to construct source and target-domain data with the digits as the attributes.

Faces. We consider facial *photo* and *sketch* images as data in different domains. For facial photo images, we consider the CelebFaces Attributes dataset (CelebA) [21], which is a large-scale face image dataset including more than $200K$ celebrity photos annotated with 40 facial attributes. We randomly select half of the dataset as the photo domain, then convert the other half into sketch images based on the procedure used in [11] (which thus results in our sketch domain data). For simplicity, among the 40 attributes of face images, we choose *"glasses"* and *"smiling"* as the attributes of interest. The common rule of thumb 80/20 is used for the training/testing dataset split.

Scenes. We have *photo* and *paint* images as scene image data in different domains. We collect 1,098 scene photos from Google Image Search and Flickr. We randomly select half of the photo collection as the photo domain and apply the style transfer method in [34] on the rest half to produce the painting images. Each image is manually labeled as *"night"*, i.e. day/night, and *"season"*, i.e. winter/summer,

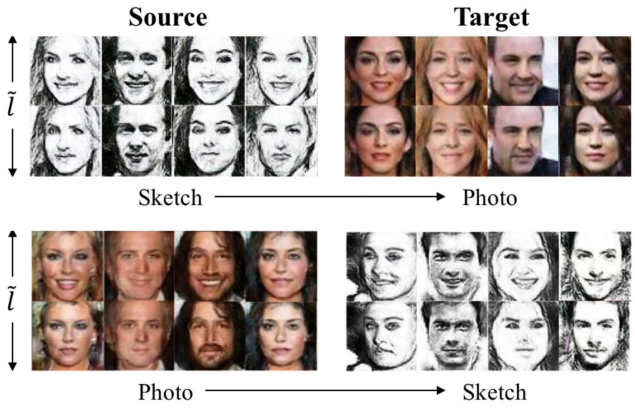


Figure 5: Cross-domain conditional image synthesis for Sketch → Photo and Photo → Sketch with $\tilde{l}$ as *smiling*. Note that the identities are different across image domains.

for attribute of interest. We use 80% of all the data in each domain for training and the rest 20% for testing.

It is worth repeating that, while the image data in both domains are presented during training, we do not require any paired cross-domain image pairs to learn our models (neither do [34, 20, 19]). And, for fair comparisons, the ground truth attribute is only available for the source-domain data for all the methods considered.

4.2. Cross-Domain Representation Disentanglement and Translation

We first conduct *conditional image synthesis* to evaluate the effectiveness of CDRD for representation disentanglement. Recall that the architecture of our CDRD allows one to freely control the disentangled factor $\tilde{l}$ via (1) with randomly sampled z to produce the corresponding output.

Single Source Domain vs. Single Target Domain. Considering a pair of data domains from each image type (i.e., digit, face, or scene images), we plot the results of conditional image synthesis in Figures 4 and 5. From these results, we have a random vector z as the input in each column, and verify that the images at either domain can be properly synthesized and manipulated (based on the attribute of interest).

Single Source vs. Multiple Target Domains. We now extend our CDRD to perform cross-domain representation disentanglement, in which a single source domain and multiple target domains are of use. From the results shown in Figure 6, we see that our CDRD can be successfully applied for this challenging task even with only attribute supervision from the single source-domain data. This confirms our design of high-level sharing weights in CDRD.

Next, we evaluate our performance for *conditional image–to–image translation*. This requires the use of our E-CDRD for joint cross-domain representation disentangle-

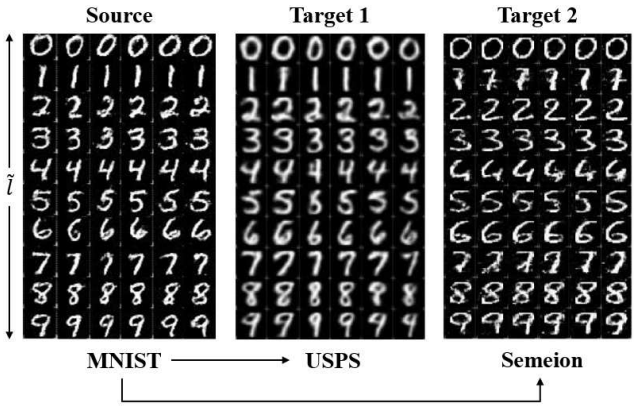


Figure 6: Cross-domain conditional image synthesis from a single source to multiple target domains: MNIST → USPS and Semeion with $\tilde{l}$ as *digits*.

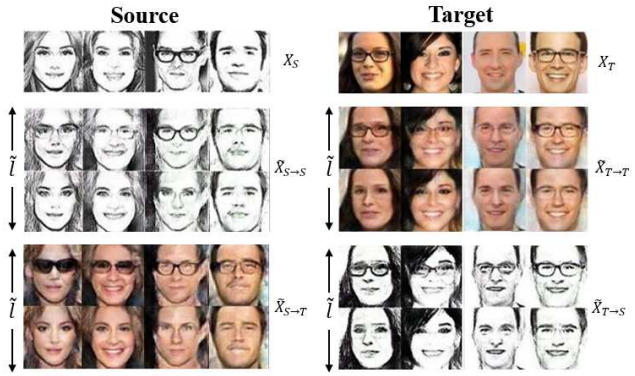


Figure 7: Cross-domain conditional image translation for facial Sketch → Photo with $\tilde{l}$ as *glasses*.

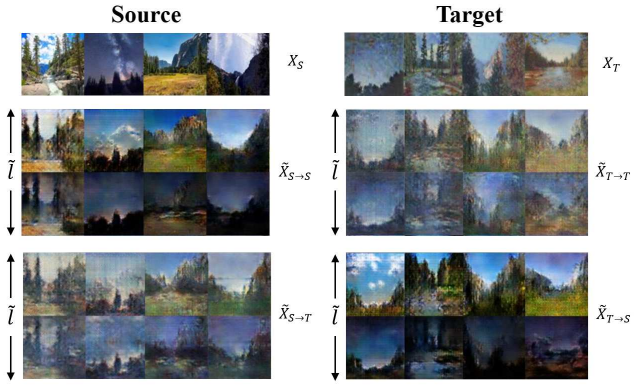
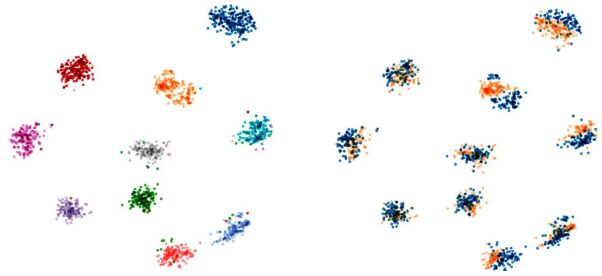


Figure 8: Cross-domain conditional image translation for scene images: Photo → Paint with $\tilde{l}$ as *night*.

ment and translation, i.e., a representation encoded from one domain can be translated to another with the specified attribute value $\tilde{l}$.

We utilize face and scene image data, as shown in Figures 7 and 8, respectively. Take Figure 7 as example, when having a facial sketch as the input in source domain, our E-CDRD is able to manipulate the attribute of glasses not only for the output image in the same domain (i.e., sketch), but also for the facial photo in the target domain (e.g., photo).



(a) Colorize w.r.t. **class 0-9**. (b) Colorize w.r.t. **source/target**.

Figure 9: t-SNE visualization of the handwritten image features for MNIST→USPS. Difference colors indicate (a) attribute and (b) domain information.

The above results support the use of our CDRD and E-CDRD for learning disentangled feature representation from cross-domain data, and confirm its effectiveness in producing or translating images with manipulated attributes in either domain.

As an additional remark, for conditional image translation, one might consider an alternative solution, which first performs conditional image synthesis using source-domain data, followed by using existing off-the-shelf image-to-image translation frameworks to convert such outputs into the images in the target domain. However, such integrated approaches cannot guarantee that proper disentangled representation can be obtained in the shared feature space.

4.3. Unsupervised Domain Adaptation

Finally, we verify that our model can be applied to image classification, i.e., use of the discriminator in our model for recognizing images with particular attribute l. As noted above, this can be viewed as the task of UDA since only supervision is available in the source domain during training.

Digits. For UDA with digit images, we consider MNIST→USPS and USPS→MNIST, and we evaluate the classification accuracy for target-domain images. Table 1 lists and compares the performances of recent UDA methods. We can see that a significant improvement was achieved by our CDRD.

It is worth noting that, while UNIT [19] reported 0.9597 for M→U and 0.9358 for U→M, UPDAG [2] achieved 0.9590 for M→U, they considered much larger datasets (UNIT required 60000/7291 images for MNIST/USPS, and UPDAG required 50000/6562 for MNIST/USPS). We follow the same protocol in [22, 23] for reporting our results in Table 1.

In addition, we extract latent features from the last shared

Table 1: UDA accuracy (%) for recognizing target-domain images with the attribute of digits (0-9). Take M → U for example, we set MNIST and USPS as source and target domains, respectively.

	GFK [8]	JDA [22]	SA [4]	TJM [23]	SCA [6]	JGSA [33]	DC [30]	GR [5]	CoGAN [20]	ADDA [31]	DRCN [7]	ADGAN [28]	CDRD
M → U	67.22	67.28	67.78	63.28	65.11	80.44	79.10	77.10	91.20	89.40	91.80	92.50	**95.05**
U → M	46.45	59.65	48.80	52.25	48.00	68.15	66.50	73.00	89.10	90.10	73.67	90.80	**94.35**
Average	56.84	63.47	58.29	57.77	56.55	74.30	72.80	75.05	90.15	89.75	82.74	91.65	**94.70**

Table 2: UDA accuracy (%) of cross-domain classification on face and scene images.

(a) Faces.

Domain	$\tilde{l}$	CoGAN	UNIT	CDRD	E-CDRD
sketch ($\mathcal{S}$)	smiling	89.50	90.10	90.19	90.01
photo ($\mathcal{T}$)	-	78.90	81.04	**87.61**	**88.28**
sketch ($\mathcal{S}$)	glasses	96.63	97.65	97.06	97.19
photo ($\mathcal{T}$)	-	81.01	79.89	**94.49**	**94.84**

(b) Scenes.

Domain	$\tilde{l}$	CoGAN	UNIT	CDRD	E-CDRD
photo ($\mathcal{S}$)	night	98.04	98.49	97.06	97.14
paint ($\mathcal{T}$)	-	65.18	67.81	**84.21**	**85.58**
photo ($\mathcal{S}$)	season	86.74	85.64	86.21	88.92
paint ($\mathcal{T}$)	-	65.94	66.09	**79.87**	**80.03**

layer (prior to the auxiliary classifier) in Discriminator. We visualize such projected features via t-SNE, and show the results in Figure 9. From Figure 9a, we see that the image features of each class of digits were well separated, while the features of the same class but from different domains were properly clustered (see Figure 9b). This confirms the effectiveness of our model in describing and adapting cross-domain image data.

Faces and Scenes. Tables 2a and 2b show our UDA performance and comparisons using cross-domain face and scene images, respectively. It can be seen that, neither CoGAN nor UNIT were able to produce satisfactory performances, as the performance gaps between source and target-domain images were from about 10% to 30%. In contrast, the use of our E-CDRD reported much smaller performance gaps, and confirms that our model is preferable for translating and adapting cross-domain images with particular attributes of interest.

It is worth noting that our E-CDRD reported further improved results than CDRD, since joint disentanglement and translation is performed when learning E-CDRD, which results in improved representation for describing cross-domain data. Another remark is that, by observing synthesized data with given assigned label $\tilde{l}$, our classifier is able to observe target domain data together with assigned attribute information. This is different from traditional domain adaptation methods, as our method breaks the limitation of lacking of ground truth attributes in target domain.

4.4. Sensitivity Analysis

As noted in Section 3, we have a hyperparameter λ in (4) controlling the disentanglement ability of our model. In order to analyze its effect on the performance, we vary λ from 0.00 to 1000 and plot the corresponding disentangled results in Figure 10. From this figure, we see that smaller λ values were not able to manipulate images with different attributes, while extremely large λ would result in degraded image quality (due to the negligence of the image adversarial loss). Thus, the choice of λ between 0.5 and 10 would be preferable (we set and fix $\lambda = 1$ in all our experiments).

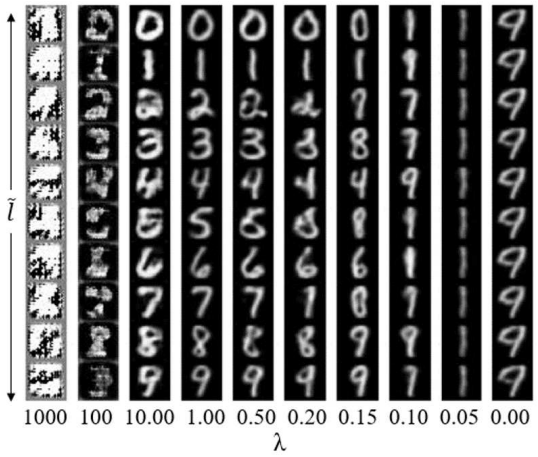


Figure 10: Sensitivity analysis on λ for M→U. Each column shows synthesized USPS images by a λ choice, while the elements in each column are expected to be associated with $\tilde{l} = 0 - 9$.

5. Conclusions

We presented a deep learning framework of Cross-Domain Representation Disentangler (CDRD) and its extension (E-CDRD). Our models perform joint representation disentanglement and adaption of cross-domain images, while only attribute supervision is available in the source domain. We successfully verified that our models can be applied to conditional cross-domain image synthesis, translation, and the task of unsupervised domain adaptation.

Acknowledgments This work was supported in part by the Ministry of Science and Technology of Taiwan under grants MOST 107-2636-E-009-001 and 107-2634-F-002-010.

References

[1] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(8):1798–1828, 2013.

[2] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[3] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[4] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013.

[5] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2015.

[6] M. Ghifary, D. Balduzzi, W. B. Kleijn, and M. Zhang. Scatter component analysis: A unified framework for domain adaptation and domain generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 39(7):1414–1430, 2017.

[7] M. Ghifary, W. B. Kleijn, M. Zhang, D. Balduzzi, and W. Li. Deep reconstruction-classification networks for unsupervised domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[8] B. Gongg, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[10] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. β-VAE: Learning basic visual concepts with a constrained variational framework. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

[11] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial nets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[12] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[13] T. Kim, M. Cha, H. Kim, J. Lee, and J. Kim. Learning to discover cross-domain relations with generative adversarial networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2017.

[14] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[15] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

[16] T. D. Kulkarni, W. F. Whitney, P. Kohli, and J. Tenenbaum. Deep convolutional inverse graphics network. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[17] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther. Autoencoding beyond pixels using a learned similarity metric. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2016.

[18] M. Lichman. UCI machine learning repository, 2013.

[19] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[20] M.-Y. Liu and O. Tuzel. Coupled generative adversarial networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[21] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[22] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu. Transfer feature learning with joint distribution adaptation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013.

[23] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu. Transfer joint matching for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[24] A. Makhzani, J. Shlens, N. Jaitly, and I. Goodfellow. Adversarial autoencoders. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.

[25] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier GANs. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2017.

[26] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 22(10):1345–1359, 2010.

[27] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE Signal Processing Magazine*, 32(3):53–69, 2015.

[28] S. Sankaranarayanan, Y. Balaji, C. D. Castillo, and R. Chellappa. Generate to adapt: Aligning domains using generative adversarial networks. *arXiv preprint arXiv:1704.01705*, 2017.

[29] Y. Taigman, A. Polyak, and L. Wolf. Unsupervised cross-domain image generation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

[30] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[31] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[32] Z. Yi, H. Zhang, P. T. Gong, et al. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[33] J. Zhang, W. Li, and P. Ogunbona. Joint geometrical and statistical alignment for visual domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[34] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networkss. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.