Collaborative Learning with Disentangled Features for Zero-shot Domain Adaptation

Won Young Jhoo, Jae-Pil Heo*
Sungkyunkwan University

{jhoowy, jaepilheo}@skku.edu

Abstract

Typical domain adaptation techniques aim to transfer the knowledge learned from a label-rich source domain to a label-scarce target domain in the same label space. However, it is often hard to get even the unlabeled target domain data of a task of interest. In such a case, we can capture the domain shift between the source domain and target domain from an unseen task and transfer it to the task of interest, which is known as zero-shot domain adaptation (ZSDA). Most of existing state-of-the-art methods for ZSDA attempted to generate target domain data. However, training such generative models causes significant computational overhead and is hardly optimized. In this paper, we propose a novel ZSDA method that learns a task-agnostic domain shift by collaborative training of domain-invariant semantic features and task-invariant domain features via adversarial learning. Meanwhile, the spatial attention map is learned from disentangled feature representations to selectively emphasize the domain-specific salient parts of the domain-invariant features. Experimental results show that our ZSDA method achieves state-of-the-art performance on several benchmarks.

1. Introduction

Recent deep learning methods achieved success in various computer vision tasks including image classification, segmentation, and object detection. However, such deep-learned models often suffer from severe performance degradation when the training data and testing data are from different domains due to the domain shift [12]. For example, a scene segmentation model trained on synthetic images performs worse when applied to real-world images, and vice versa. To tackle this problem, domain adaptation techniques that aim to transfer the knowledge learned from a label-rich source domain to the desired target domain, also known as "transfer learning", are actively being studied.

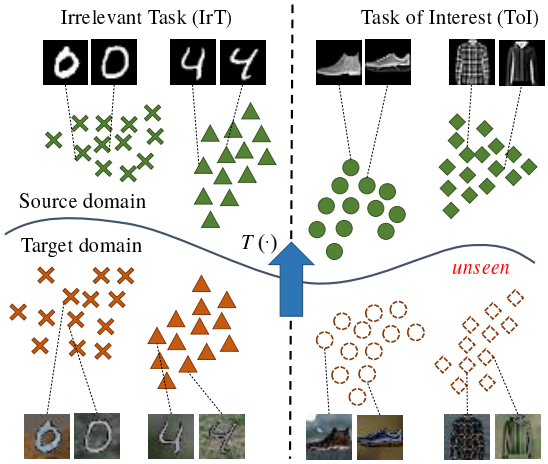


Figure 1: An example of zero-shot-domain adaptation. In this scenario, IrT is the digit image analysis and ToI is the fashion image analysis. The source and target domains are grayscale and color image domains, respectively. The objective of zero-shot domain adaptation is to train a model for ToI in the target domain, which is unseen during training time, by learning a task-agnostic domain shift $T(\cdot)$.

Typical domain adaptation methods assume that the target domain data is available at the training phase. However, in real-world applications, it is often not feasible to get the unlabeled target domain data that shares an identical label space with the source domain data of interest. Such a situation refers to a new transfer learning task, known as zero-shot domain adaptation (ZSDA) [25]. The objective of the zero-shot domain adaptation task is to transfer the domain shift to a task of interest (ToI) from an irrelevant task (IrT) as shown in Fig 1. This domain adaptation technique suggests a novel approach for various data scarcity problems. Suppose that we have a scene text dataset and a synthetic

*Corresponding author

text dataset to develop a scene text detection model. With ZSDA, we can also have a model that supports other languages by adding easily producible synthetic datasets.

Prior ZSDA techniques can be categorized into two approaches based on their strategies; 1) generating the target domain samples of ToI and 2) learning domain invariant feature representations over different domains. The methods in the first approach typically use generative models such as generative adversarial networks (GAN) [11] or variational autoencoder (VAE) [17] to reconstruct the target domain distribution and train a model with the generated samples. While this strategy is intuitive, generative models such as GANs and VAEs are prone to problems such as collapsing modes and label flipping. These problems can be made worse by samples that are used in domain adaptation.

The other approach for ZSDA is to learn domain invariant feature representations. Compared to the sample generation approach, those methods can avoid the aforementioned overhead and undesirable risks of reconstructing target domain data. Learning domain invariant feature representations have been actively studied for a long time to solve other types of transfer learning problems such as unsupervised domain adaptation [10, 27], partial domain adaptation [3], and few-shot domain adaptation [24]. However, all the listed techniques are not directly applicable to ZSDA, since they strictly require the unlabeled target domain data to have the same label space with the source domain. When the source and target domains have different labels, the discriminative features for IrT are overestimated and the features for ToI are underestimated while handling target domain data. Because of this difference among label distributions in domains, this results in a negative transfer effect.

One promising approach for ZSDA is to learn disentangled representations of domain-relevant features and task-relevant features. Recent techniques [26, 20] have a feature disentangler to learn a domain-invariant representation from multiple domains via adversarial learning. However, domain-invariant features alone cannot be sufficiently discriminative to deal with ToI distributions of the target domain, since the feature extractor never sees it in the training phase.

To address the aforementioned issue, we propose a more effective approach, which collaboratively learns class-agnostic domain feature representations and domain-invariant semantic feature representations. Our training scheme has two phases: disentanglement and refinement. In the disentanglement stage, we extend the domain adversarial adaptation approaches [9, 29] to learn class-agnostic domain features and domain-invariant features simultaneously. In the refinement stage, a domain feature is transformed into a spatial attention map. The spatial attention map selectively emphasizes the domain-specific salient parts of the domain-invariant semantic feature. This enhances the discriminative power of the imperfect semantic features.

Our main contribution can be summarized as follows: (1) We propose an end-to-end framework for zero-shot domain adaptation which does not need any additional information or assumptions in the problem definition. (2) We propose a novel collaborative feature refinement with disentangled feature representations that can prevent negative transfer effects during zero-shot domain adaptation. (3) Our proposed method achieves state-of-the-art performance in extensive experiments on various benchmarks for zero-shot domain adaptation tasks.

2. Related Works

Domain-invariant Representation Domain adaptation aims to transfer the knowledge learned from the source domain to the target domain where labeled data is sparse or non-existent. Domain adaptation typically involves domain-invariant representation, which involves various methods. Some methods use Maximum Mean Discrepancy (MMD) [22, 23] as a way to minimize feature distribution discrepancies among different domains. With the development of generative adversarial networks (GAN) [11], many adversarial strategies such as gradient reversal layer [9], domain adversarial loss [29], and classifier discrepancy loss [19, 27] have been successfully used for various domain adaptation tasks. Domain-agnostic learning (DAL) [26] proposes a method to disentangle the feature representation via adversarial learning. However, all the aforementioned methods are not directly applicable to the zero-shot domain adaptation problem, since they utilize the unlabeled target domain data in the training phase.

Zero-shot Domain Adaptation Zero-shot domain adaptation (ZSDA) assumes that the label space of the given target domain data is different from the task of interest. Although various methods have been explored recently, ZSDA has not been well addressed due to negative transfer. Partial domain adaptation methods [3, 2] propose a way to mitigate negative transfer, but these methods also rely on the target domain data of ToI. In a similar problem setting, *d*-SNE [40] proposes a metric-based few-shot domain adaptation but the performance degradation is significant in zero-shot settings.

Some existing works on ZSDA utilize additional information to capture the accurate domain shift that is able to be generalized for ToI. Ishii *et al*. [15] utilizes the known attribute information of the domain (e.g., position of the camera). Yang *et al*. [41] uses multiple sources and target domain data determined by a vector of continuous variables. Zero-shot deep domain adaptation (ZDDA) [25] does not need such assumptions, but paired dual-domain samples are required during the training time. These restrictions help

align source and target domain representations but make it difficult to adapt in real-world applications.

Existing state-of-the-arts approaches to ZSDA [33, 34, 35] utilizes generative models such as coupled generative adversarial networks (CoGAN) [21] and variational autoencoder (VAE) [17], which aim to reconstruct ToI samples in target domain distribution. These methods use shared layers to capture the semantic concepts of ToI samples and coupled networks to generate dual-domain samples. However, these generative approaches produce significant overhead by generating target domain samples and inherit the difficulties of the data generation task. Instead, we propose a method to learn a refined semantic representation by using an attention mechanism.

Attention Mechanism Recently, attention mechanism has been widely applied to various neural network architectures that capture relevant characteristics of human perception. It enables the models to focus on the salient part of a given feature, and it is shown that the mechanism can improve model performance in various computer vision tasks including image segmentation [4, 8], image classification [37, 32], and image generation [42, 39]. In particular, Vaswani *et al*. [30] propose a self-attention mechanism that computes global dependencies given inputs and achieves state-of-the-art results in machine translation. There are also attempts at applying attention mechanism to unsupervised domain adaptation by learning a transferable attention [36] and temporal alignment [5].

3. Methodology

In the zero-shot domain adaptation (ZSDA), we have two different tasks, a task of interest (ToI) and an irrelevant task (IrT). ToI and IrT have different label spaces $\mathcal{C}_r$ and $\mathcal{C}_{ir}$, respectively. We also have two domains, a source domain $\mathcal{D}_s$ and a target domain $\mathcal{D}_t$. The data samples from the source domain of each task are denoted by $X_s^r = \{x_s^r, y_s^r\}$ and $X_s^{ir} = \{x_s^{ir}, y_s^{ir}\}$, where $y_s^r \in \mathcal{C}_r$ and $y_s^{ir} \in \mathcal{C}_{ir}$. Similarly, the data samples from the target domain of each task are defined as X_t^r and X_t^{ir}. The goal of ZSDA task is to learn a model for the unseen data X_t^r by using three labeled datasets X_s^r, X_s^{ir} and X_t^{ir}.

We propose to solve the ZSDA problem by learning two different features, class-agnostic domain features f_d and domain-invariant semantic features f_c. These two features f_d and f_c are extracted from different feature extractors G_d and G_c, respectively. Such two types of features f_d and f_c are derived from feature disentanglement by eliminating class-relevant and domain-relevant information from a shared representation, respectively. In typical unsupervised domain adaptation tasks, the domain-invariant feature f_c from feature disentanglement can be sufficient since the discrimination on X_t^r can be retained by the availability of unlabeled data samples $\{x_t^r\}$. However, in the ZSDA setting, the semantic features for ToI in f_c are often aligned with the IrT feature distribution due to the discrepancy between $\mathcal{D}_s$ and $\mathcal{D}_t$ during the training phase.

To alleviate this problem, we introduce a feature refinement stage to learn a domain-specific spatial attention map from the class-agnostic domain features f_d that guides where to attend in the f_c. By applying this domain-specific attention map to the f_c, the negative transfer effect is reduced while the positive transfer effect is enhanced. Note that, we describe our method based on an assumption that the given tasks are those for image classification.

3.1. Feature Disentanglement

The goal of the feature disentanglement process is to remove domain-relevant information from the $f_c = G_c(x)$ and task-relevant information from the $f_d = G_d(x)$. Our feature disentanglement method adopts an adversarial learning strategy with a domain discriminator D, and two classifiers C_r and C_{ir} for ToI and IrT, respectively. The overall procedure of the feature disentanglement procedure is illustrated in Fig. 2a.

Domain-invariant Feature The domain-invariant semantic feature f_c is learned via adversarial learning between D and G_c. The domain discriminator D is asked to distinguish the domain label from given features f_d or f_c. Specifically, the loss function for the domain discriminator $\mathcal{L}_D$ is defined as follows:

$$\mathcal{L}_D^{f_d} = -\mathbb{E}_{x_s \sim \mathcal{D}_s}[\log(D(f_d))] - \mathbb{E}_{x_t \sim \mathcal{D}_t}[\log(1 - D(f_d))] \tag{1}$$

$$\mathcal{L}_D^{f_c} = -\mathbb{E}_{x_s \sim \mathcal{D}_s}[\log(D(f_c))] - \mathbb{E}_{x_t \sim \mathcal{D}_t}[\log(1 - D(f_c))] \tag{2}$$

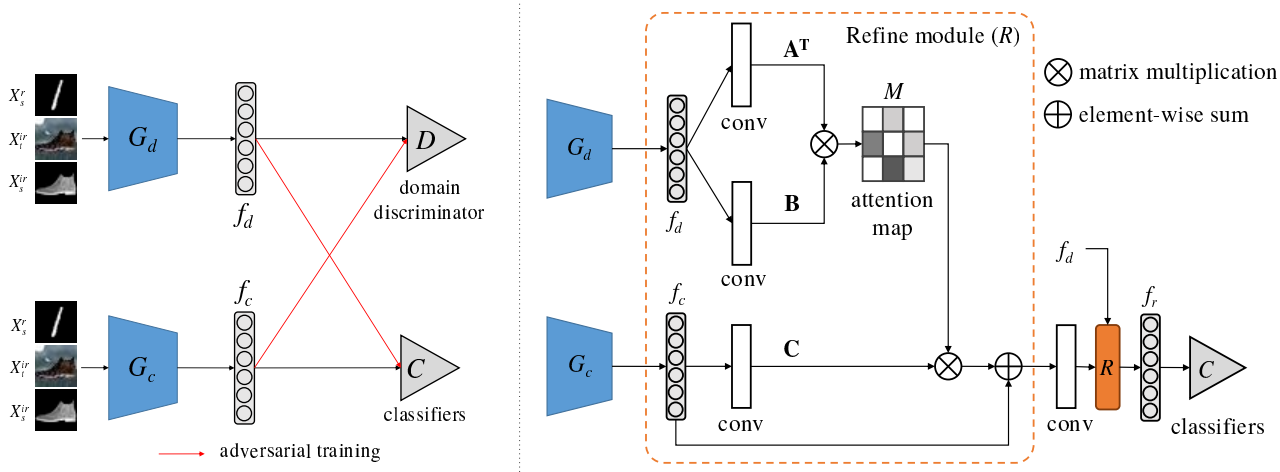
$$\mathcal{L}_D = \mathcal{L}_D^{f_d} + \mathcal{L}_D^{f_c} \tag{3}$$

On the other hand, the feature extractor G_c is trained with two classifiers C_r and C_{ir} to preserve the semantic information of f_c. This is achieved by minimizing the classification errors. Thus, the loss functions of two classifiers are defined as follows:

$$\mathcal{L}_{C_r}^{f_c} = -\mathbb{E}_{(x,y) \sim p_{ToI}}[\ell(y, C_r(f_c))] \tag{4}$$

$$\mathcal{L}_{C_{ir}}^{f_c} = -\mathbb{E}_{(x,y) \sim p_{IrT}}[\ell(y, C_{ir}(f_c))], \tag{5}$$

where $\ell(\cdot)$ is the cross-entropy loss, and p^r, p^{ir} are the probability distributions of ToI and IrT, respectively. Meanwhile, the semantic feature extractor G_c is trained to fool



(a) Disentanglement stage (b) Collaborative learning stage

Figure 2: Overview of our method. (a) The G_c extracts a semantic feature f_c and G_d extracts a domain feature f_d. Then the domain and the class information are removed from f_c and f_d via adversarial training, respectively. (b) After the disentanglement stage, the refinement module is trained to produce a refined feature f_r based on the attention mechanism. Both processes are repeated in every iteration. Note that, the orange-colored box (R) represents the refine module.

the domain discriminator D to learn domain-invariant features. As a result, the aggregated loss functions of the feature extractors during the domain-disentanglement process are represented as follows:

$$\mathcal{L}_{G_c} = \mathcal{L}_{C_r}^{f_c} + \mathcal{L}_{C_{ir}}^{f_c} - \mathcal{L}_{D}^{f_c} \tag{6}$$

$$\mathcal{L}_{G_d} = \mathcal{L}_{D}^{f_d} \tag{7}$$

Task-invariant Feature Once the domain information is disentangled from f_c, the class-relevant semantic information is removed from the domain-specific feature f_d via adversarial learning between the classifiers and the domain feature extractor G_d. At first, the classifiers C_r and C_{ir} are fixed and only G_d is updated to disable the classification capability from f_d. This is achieved by maximizing the entropy of the predicted class distribution. The loss function is defined as follows:

$$\mathcal{L}_{G_d} = -\frac{1}{n_r}\sum_{j=1}^{n_r}\log C_r(f_d^j) - \frac{1}{n_{ir}}\sum_{j=1}^{n_{ir}}\log C_{ir}(f_d^j), \tag{8}$$

where n_r and n_{ir} are the numbers of data samples in ToI and IrT, respectively. Once the domain feature extractor G_d is updated, the classifiers are then trained to identify the class-relevant features from f_d while G_d is fixed. Similar to Eq. 4, the loss functions for two classifiers are defined as follows:

$$\mathcal{L}_{C_r}^{f_d} = -\mathbb{E}_{(x,y)\sim p^r}[\ell(y, C_r(f_d))] \tag{9}$$

$$\mathcal{L}_{C_{ir}}^{f_d} = -\mathbb{E}_{(x,y)\sim p^{ir}}[\ell(y, C_{ir}(f_d))] \tag{10}$$

3.2. Collaborative Learning

As discussed earlier, the semantic features optimized for IrT is prone to mislead the classifier for ToI C_r, even though the contextual and domain information are successfully disentangled. To address this, we propose a collaborative refinement for the feature map to highlight the important parts of the target domain via a domain-specific attention map from f_d. Fig 2b shows the overall process of the refinement step. Specifically, we model our attention method upon the transformer architecture [30].

Given a domain feature $f_d \in \mathbb{R}^{C\times N}$, first feed it into convolution layers and transform it into two features $\mathbf{A}$ and $\mathbf{B}$, where $\{\mathbf{A}, \mathbf{B}\} \in \mathbb{R}^{\bar{C}\times N}$. Additionally, C, $\bar{C}$, and N are the numbers of channels, reduced channels, and spatial locations of a feature map, respectively. Note that, we consistently set $\bar{C} = C/8$ in all the experiments. The resulting attention map M is computed as follows:

$$m_{ji} = \frac{\exp(\mathbf{A}_i{}^T \cdot \mathbf{B}_j)}{\sum_{i=1}^{N}\exp(\mathbf{A}_i{}^T \cdot \mathbf{B}_j)}, \tag{11}$$

where m_{ji} represents the i^{th} position's importance on j^{th} position. Meanwhile, the semantic feature f_c is also fed into a convolution layer and transformed into a feature $\mathbf{C} \in$

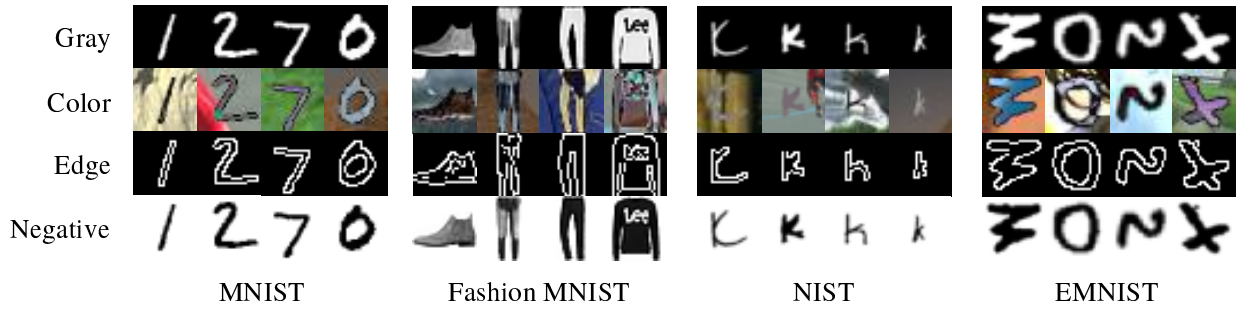


Figure 3: The sample images of synthetic domains of X-NIST.

$\mathbb{R}^{C\times N}$. Then the output of a refinement module O is:

$$O_j = \gamma \sum_{i=1}^{N} (m_{ji}\mathbf{C}_i) + F_j \tag{12}$$

where $F = f_c$. The γ is a learnable scalar initialized at zero and gradually learns to assign more weight to the domain-specific attended feature map. The above equation (Eq. 12) is computed at both refinement modules as illustrated in Fig. 2b to produce the refined feature f_r. Note that all refinement modules share the same domain feature f_d as an input. This overall refinement process is trained by the following classification losses,

$$\begin{aligned}\mathcal{L}^{f_r} = &- \lambda_r \mathbb{E}_{(x,y)\sim p^{ir}}[\ell(y, C_{ir}(f_r))] \\ &- \lambda_r \mathbb{E}_{(x,y)\sim p^{r}}[\ell(y, C_{r}(f_r))],\end{aligned} \tag{13}$$

and the refined feature f_r is utilized for the final classification result. λ_r is a hyper-parameter to balance the losses of the disentanglement stage and the collaborative loss $\mathcal{L}^{f_r}$. In practice, we set $\lambda_r = 3.0$.

4. Experiments

We evaluate our proposed method against the state-of-the-art ZSDA techniques on both synthetic and real datasets.

4.1. Datasets

To evaluate our ZSDA method for classification tasks, we use MINST (D_M) [18], Fashion-MNIST (D_F) [38], NIST (D_N) [13], EMNIST (D_E) [6], and Office-Home [31] datasets. We denote the set of character datasets, $\{D_M, D_F, D_N, D_E\}$ as X-NIST.

MNIST is a hand-written digit image dataset. It contains 60,000 training and 10,000 testing images. Each example is a 28 $\times$ 28 size grayscale image, associated with a label from 10 classes.

Fashion-MNIST contains silhouettes of fashion images. It has the same number of training and testing samples as MNIST. Also, each example is a 28 $\times$ 28 size grayscale image, associated with a label from 10 classes as is MNIST.

NIST is a hand-written letters dataset. We selected 52 classes from this dataset which are upper and lower case letters. It contains 387,361 training and 23,941 testing images, and each is a 128 $\times$ 128 size grayscale image.

EMNIST is a hand-written letters dataset derived from NIST and converted to a 28 $\times$ 28 size image format as like MNIST. In this paper, we use EMNIST letters split, which merges the upper and lower case letters. In total, it contains 124,800 training and 20,800 testing images from 26 classes.

Office-Home dataset is a more challenging domain adaptation benchmark crawled through several search engines and online image directories. It consists of images from 4 different domains: Artistic images (**Ar**), Clip Art (**Cl**), Product images (**Pr**), and Real-World images (**Rw**). The dataset contains images of 65 object categories for each domain, and the total number of images in the dataset is approximately 15,500.

The datasets D_M, D_f, D_N, and D_E are all in a gray domain (domain G). To evaluate our method, we follow the same protocol with CoCoGAN [33] that creates color (domain C), edge (domain E), and negative domain (domain N). The color domain is created by using Ganin's method [9], blending the original image with randomly extracted patches from the BSDS500 dataset [1]. The edge domain image is obtained by using the canny edge detector, and the negative domain image I_n is obtained by applying $I_n = 255 - I$ for a given grayscale image I. The sample images of the generated domains are shown in Fig 3.

4.2. Implementation Details

We implement our method using PyTorch. In all experiments, the discriminator D is implemented with two fully connected layers, and classifiers C_r and C_{ir} are implemented as a single fully connected layer. We use two

Domains	Methods (ToI / IrT)	MNIST (D_M)			FashionMNIST (D_F)			NIST (D_N)		EMNIST (D_E)	
		D_F	D_N	D_E	D_M	D_N	D_E	D_M	D_F	D_M	D_F
G → C	ZDDA	73.2	92.0	94.8	51.6	43.9	65.3	34.3	21.9	71.2	47.0
	CoCoGAN	78.1	92.4	95.6	56.8	56.7	66.8	41.0	44.9	75.0	54.8
	Wang *et al.*	81.2	93.3	95.0	57.4	58.7	62.0	44.6	**45.5**	72.4	58.9
	Ours (No Refine)	68.6	86.7	96.6	57.3	61.2	73.3	31.3	17.1	81.9	71.7
	Ours	**93.3**	**97.0**	**97.9**	**67.7**	**72.6**	**76.3**	**45.7**	31.3	**86.4**	**74.1**
G → E	ZDDA	72.5	91.5	93.2	54.1	54.0	65.8	42.3	28.4	73.6	50.7
	CoCoGAN	79.6	94.9	95.4	61.5	57.5	71.0	48.0	36.3	77.9	58.6
	Wang *et al.*	81.4	93.5	96.3	63.2	58.7	72.4	49.9	38.6	78.2	61.1
	Ours (No Refine)	84.7	89.2	94.3	54.2	39.0	63.2	46.7	33.8	68.5	67.0
	Ours	**92.9**	**95.5**	**98.9**	**65.0**	**60.7**	**74.4**	**53.4**	**46.9**	**91.1**	**82.9**
G → N	ZDDA	77.9	82.4	90.5	61.4	47.4	62.7	37.8	38.7	76.2	53.4
	CoCoGAN	80.3	87.5	93.1	66.0	52.2	69.3	45.7	53.8	81.1	56.5
	Wang *et al.*	-	-	-	-	-	-	-	-	-	-
	Ours (No Refine)	88.5	95.8	99.0	60.6	82.6	**84.4**	58.2	52.9	93.0	**91.0**
	Ours	**97.7**	**97.0**	**99.2**	**81.3**	**85.4**	84.2	**58.7**	**59.0**	**93.4**	89.9
C → G	ZDDA	67.4	85.7	87.6	55.1	49.2	59.5	39.6	23.7	75.5	52.0
	CoCoGAN	73.2	89.6	94.7	61.1	50.7	70.2	47.5	57.7	80.2	67.4
	Wang *et al.*	73.7	91.0	93.4	62.4	53.5	71.5	50.6	58.1	83.5	70.9
	Ours (No Refine)	98.7	98.0	99.2	88.9	86.6	89.0	61.2	64.2	90.9	92.1
	Ours	**98.9**	**99.1**	**99.3**	**89.3**	**89.1**	**89.6**	**69.0**	**69.1**	**92.8**	**93.3**
N → G	ZDDA	78.5	90.7	87.6	56.6	57.1	67.1	34.1	39.5	67.7	45.5
	CoCoGAN	80.1	92.8	93.6	63.4	61.0	72.8	47.0	43.9	78.8	58.4
	Wang *et al.*	82.6	94.6	95.8	67.0	68.2	77.9	51.1	44.2	79.7	62.2
	Ours (No Refine)	89.8	97.2	98.9	61.7	82.7	82.1	52.6	**53.8**	92.9	**91.4**
	Ours	**94.9**	**98.5**	**99.2**	**83.4**	**84.0**	**86.3**	**58.4**	51.0	**93.3**	91.3

Table 1: Experimental results on the synthetic domains. The domain G, C, E, and N refer to gray, color, edge, and negative domains. Ours (No Refine) represents the results without the refinement stage. The best results are marked in **bold**. The baseline results are taken from the papers [25, 33, 34].

refinement modules for all experiments.

In the experiments on the X-NIST dataset, we use three convolutional layers to implement both feature extractor G_d and G_c. The input image size is resized to 28×28, and the feature dimensionalities of f_d, f_c, and f_r are $128 \times 7 \times 7$. The batch size is set to 64.

In the experiments on the Office-Home dataset, we use ResNet-50 [14] as the feature extractors, and the semantic feature extractor G_d is initialized with ImageNet [7] pre-trained weights. Note that, the previous works [34, 25] also utilize ImageNet samples or ImageNet pre-trained weights. The feature dimensionalities of f_d, f_c and f_r are $512 \times 7 \times 7$, and the batch size is set to 8.

We use an Adam optimizer [16] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$ for all experiments. The initial learning rate is set to 0.0002 and decayed by 0.1 two times through the entire training iterations. The number of epochs is set based on the ToI dataset sizes, 100 for $\{D_M, D_F, D_E\}$, 30 for D_N, and 50 for Office-Home dataset. In addition, we sample the equal number of training data from each of X_s^{ir}, X_t^{ir} and X_s^r for a mini-batch (64 for X-NIST, 8 for Office-Home). We also forced to match the labels for X_s^{ir} and X_t^{ir} within a mini-batch, and we observe that this sampling method helps to find the domain shift between the same class in the early stage of training and thus improves the performances on several domain adaptation settings. The effectiveness of the label-matching sampling method is described in supplementary the material.

Since the ZSDA problem assumes that the target-domain data of ToI is unavailable during the training phase, the way to report test set accuracy depends on whether or not a validation set has been provided. If a validation set has been provided such as the X-NIST dataset, then the test set accuracy that we report is when the sum of validation accuracies on X_s^{ir}, X_t^{ir} and X_s^r reaches the the highest. Otherwise, the test set accuracy is reported at the last epoch.

A. Source domain = Art, Clip Art

Source	Art (**Ar**)			Clip Art (**Cl**)		
Target	**Cl**	**Pr**	**Rw**	**Ar**	**Pr**	**Rw**
CoCoGAN	62.2	69.5	74.5	66.7	74.0	66.4
Wang *et al*.	62.7	71.9	76.3	72.6	75.1	73.9
Source Only	55.1±4.5	65.3±2.3	77.5±2.3	42.5±2.3	62.1±2.3	59.5±1.8
Ours	71.0±3.2	76.5±1.9	85.1±1.1	62.1±3.0	68.7±2.1	75.1±2.3

B. Source domain = Product, Real-World

Source	Product (**Pr**)			Real-World (**Rw**)		
Target	**Ar**	**Cl**	**Rw**	**Ar**	**Cl**	**Pr**
CoCoGAN	57.6	53.4	71.7	69.2	51.3	65.8
Wang *et al*.	70.3	60.8	74.8	72.2	61.4	72.2
Source Only	47.9±3.5	52.3±3.4	70.2±1.7	65.8±2.1	60.6±4.2	83.2±2.5
Ours	64.4±2.1	69.2±1.8	82.0±0.6	77.9±1.0	76.2±2.5	88.5±1.8

Table 2: Experimental results on Office-Home dataset. The results from the methods CoCoGAN, and Wang *et al*. are the accuracy when they use 10 random categories as the ToI. Our method and source only method report the average accuracy and standard error of the mean (SEM) over six different ToI/IrT splits. The baseline results are taken from the paper [34].

4.3. Results on X-NIST Dataset

Given four datasets, we conduct experiments on ten different pairs of ToI and IrT. Note that, we did not conduct experiments between D_E and D_N, since the two datasets cover the same task. As a result, we conduct experiments on five different source and target domain pairs which are (G $\rightarrow$ C), (G $\rightarrow$ E), (G $\rightarrow$ N), (C $\rightarrow$ G), and (N $\rightarrow$ G). Then we compare our method with three baselines: ZDDA [25], CoCoGAN [33], and Wang *et al*. [34].

Table 1 reports the classification accuracies on the unseen target domain data of ToI. Our method significantly outperforms the baseline methods. Especially in (C $\rightarrow$ G) and (N $\rightarrow$ G) tasks, our proposed method achieves 18.09% and 11.7% performance improvements on average compared to the respective state-of-the-art techniques. This confirms that our method effectively learns the domain-invariant semantic features from the ToI and captures the important regions at the target domain images. Meanwhile, the performance only drops in the (NIST, FashionMNIST) task pair and (gray, color) domain pair. This seems that the IrT-like objects of the background color image hinder identifying the ToI objects in a higher degree since the content size of NIST is smaller than others.

To verify the effectiveness of our proposed attention mechanism, we also performed ablation experiments on the X-NIST dataset. We removed the feature refinement process during training and evaluated model performance only using domain-invariant feature f_c. The results without refinement module are reported in Table 1. In most cases, the refinement module significantly improves the performance compared to when it is not used. Specifically, the average performance over all experimental settings improves from 81.72% to 88.28%. Those results clearly demonstrate the benefits of our attention module that helps to prevent negative transfer in ZSDA.

4.4. Results on Office-Home Dataset

To evaluate our method in real-world domains, we also conducted experiments on the Office-Home dataset. Since this dataset does not provide an explicit split between ToI and IrT, the previous works on ZSDA (CoCoGAN [33] and Wang *et al*. [34]) used 10 random categories from 65 categories in Office-Home as the ToI, and the rest of them as the IrT. Since he training/test category splits are not explicit reported in those previous work, we conducted the experiments in the following manner to make the comparison as fair as possible:

First, we split the 65 categories of Office-Home into six different subsets while the number of each categories in each subset is fixed to 10. For each subset, we use the subset as the ToI and the rest as the IrT (i.e. 10 ToI categories and 55 IrT categories). We conducted experiments on all of the 12 different domain adaptation tasks, and the average accuracy and standard error of the mean (SEM) over six different ToI/IrT splits is reported. We also reported performance on ResNet-50 model trained only with the ToI data of source domain as *Source Only*.

Table 2 shows the experimental results on the Office-Home benchmark. Even if the domain shift between the source and the target domain is more ambiguous than that of the synthetic domains, our method significantly outperforms the *Source Only* results. Also, our method out-

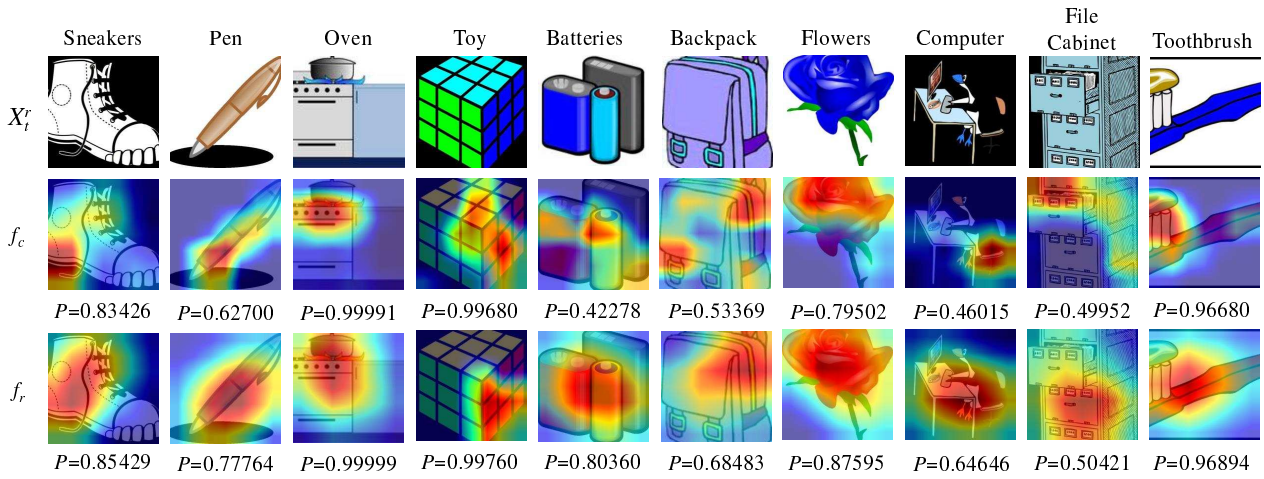


Figure 4: Grad-CAM [28] visualization results on Office-Home experiments (**Ar** $\rightarrow$ **Cl**). The f_c visualizations are from the last convolutional outputs of G_T, and the f_r visualizations are from the convolutional outputs after the first refinement module. The input images X_t^r are unseen ToI samples from each target domain, and P is the softmax score of the classification results corresponding to each feature.

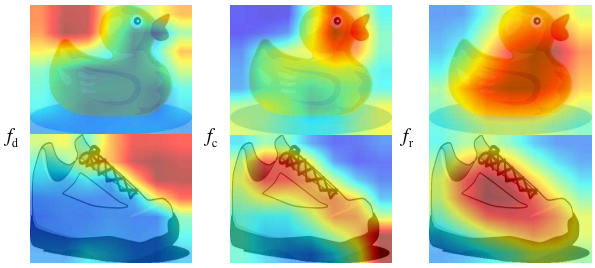


Figure 5: Visualization of channel-wise mean of each feature map from the Office-Home domain adaptation results.

performs state-of-the-art baselines by 4.38% on average. These results confirm the merits of our proposed method for ZSDA even in the most challenging real-world scenarios.

4.5. Visualization of Feature Refinement

Grad-CAM [28] is a visualization technique that produces a localization map highlighting attended regions in the image by utilizing gradients. We apply Grad-CAM on the two Office-Home domain adaptation results to qualitatively analyze the effect of the refinement module. Fig 4 shows the attended regions when features f_c and f_r are fed to the classifier C_r. We observe that less weighted parts of the target object in f_c become highlighted in the f_r.

Fig 5 shows the visualization of channel-wise mean of each feature map. In this example, we can see that the empty regions of the figures which are usually present in the clip art domain are highlighted in f_d, and the non-empty regions are highlighted in the refined feature f_r. These results demonstrate that our refinement module can effectively emphasize the under-estimated or missed ToI features while suppressing the irrelevant features.

5. Conclusion

In this paper, we propose an attention-based collaborative learning method with disentangled feature representations to solve the challenging ZSDA problem where the target domain data of ToI is unavailable. Our method first disentangles the given input to task-invariant and domain-invariant features based on adversarial learning. The refinement module collaboratively learns where to emphasize or suppress from the domain-invariant feature based on a task-agnostic attention map inferred from task-invariant features. To the best of our knowledge, this is the first attempt to use an attention mechanism in ZSDA. The ablation study verifies that the proposed attention-based refinement module significantly improves the overall performance of the zero-shot domain adaptation, and we provide visual explanations about the feature refinement results. Our future work contains extensions of the proposed method to other computer vision tasks such as scene text detection and recognition.

Acknowledgments This work was supported in part by MCST/KOCCA (No. R2020070002), MSIT/IITP (No. 2020-0-00973, 2019-0-00421, 2020-0-01821, and 2020-0-01550), MSIT/NRF (No. NRF-2020R1F1A1076602), and MSIT&KNPA/KIPoT (Police Lab 2.0, No. 210121M06).

References

[1] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):898–916, 2010. 5

[2] Zhangjie Cao, Kaichao You, Mingsheng Long, Jianmin Wang, and Qiang Yang. Learning to transfer examples for partial domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2985–2994, 2019. 2

[3] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2229–2238, 2019. 2

[4] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3640–3649, 2016. 3

[5] Min-Hung Chen, Zsolt Kira, Ghassan AlRegib, Jaekwon Yoo, Ruxin Chen, and Jian Zheng. Temporal attentive alignment for large-scale video domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6321–6330, 2019. 3

[6] Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. Emnist: Extending mnist to handwritten letters. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2921–2926. IEEE, 2017. 5

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[8] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3146–3154, 2019. 3

[9] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 2, 5

[10] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016. 2

[11] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014. 2

[12] A. Gretton, AJ. Smola, J. Huang, M. Schmittfull, KM. Borgwardt, and B. Schölkopf. *Covariate shift and local learning by distribution matching*, pages 131–160. MIT Press, Cambridge, MA, USA, 2009. 1

[13] Patrick Grother and Kayee Hanaoka. Nist special database 19 handprinted forms and characters 2nd edition. *National Institute of Standards and Technology, Tech. Rep*, 2016. 5

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[15] Masato Ishii, Takashi Takenouchi, and Masashi Sugiyama. Zero-shot domain adaptation based on attribute information. In *Asian Conference on Machine Learning*, pages 473–488. PMLR, 2019. 2

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[17] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2, 3

[18] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 5

[19] Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. Sliced wasserstein discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10285–10295, 2019. 2

[20] Alexander H Liu, Yen-Cheng Liu, Yu-Ying Yeh, and Yu-Chiang Frank Wang. A unified feature disentangler for multi-domain image translation and manipulation. *arXiv preprint arXiv:1809.01361*, 2018. 2

[21] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. *arXiv preprint arXiv:1606.07536*, 2016. 3

[22] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015. 2

[23] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. *arXiv preprint arXiv:1602.04433*, 2016. 2

[24] Saeid Motiian, Quinn Jones, Seyed Mehdi Iranmanesh, and Gianfranco Doretto. Few-shot adversarial domain adaptation. *arXiv preprint arXiv:1711.02536*, 2017. 2

[25] Kuan-Chuan Peng, Ziyan Wu, and Jan Ernst. Zero-shot deep domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 764–781, 2018. 1, 2, 6, 7

[26] Xingchao Peng, Zijun Huang, Ximeng Sun, and Kate Saenko. Domain agnostic learning with disentangled representations. In *International Conference on Machine Learning*, pages 5102–5112. PMLR, 2019. 2

[27] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3723–3732, 2018. 2

[28] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 8

[29] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017. 2

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 3, 4

[31] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017. 5

[32] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2017. 3

[33] Jinghua Wang and Jianmin Jiang. Conditional coupled generative adversarial networks for zero-shot domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3375–3384, 2019. 3, 5, 6, 7

[34] Jinghua Wang and Jianmin Jiang. Adversarial learning for zero-shot domain adaptation. In *European Conference on Computer Vision*, pages 329–344. Springer, 2020. 3, 6, 7

[35] Qian Wang and Toby P Breckon. Generalized zero-shot domain adaptation via coupled conditional variational autoencoders. *arXiv preprint arXiv:2008.01214*, 2020. 3

[36] Ximei Wang, Liang Li, Weirui Ye, Mingsheng Long, and Jianmin Wang. Transferable attention for domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5345–5352, 2019. 3

[37] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 3

[38] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. 5

[39] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018. 3

[40] Xiang Xu, Xiong Zhou, Ragav Venkatesan, Gurumurthy Swaminathan, and Orchid Majumder. d-sne: Domain adaptation using stochastic neighborhood embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2497–2506, 2019. 2

[41] Yongxin Yang and Timothy Hospedales. Zero-shot domain adaptation via kernel regression on the grassmannian. *arXiv preprint arXiv:1507.07830*, 2015. 2

[42] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International conference on machine learning*, pages 7354–7363. PMLR, 2019. 3