Towards Category and Domain Alignment: Category-Invariant Feature Enhancement for Adversarial Domain Adaptation

Yuan Wu
Carleton University
yuan.wu3@carleton.ca

Diana Inkpen
University of Ottawa
Diana.Inkpen@uottawa.ca

Ahmed El-Roby
Carleton University
Ahmed.ElRoby@carleton.ca

Abstract

Adversarial domain adaptation has made impressive advances in transferring knowledge from the source domain to the target domain by aligning feature distributions of both domains. These methods focus on minimizing domain divergence and regard the adaptability, which is measured as the expected error of the ideal joint hypothesis on these two domains, as a small constant. However, these approaches still face two issues: (1) Adversarial domain alignment distorts the original feature distributions, deteriorating the adaptability; (2) Transforming feature representations to be domain-invariant needs to sacrifice domain-specific variations, resulting in weaker discriminability. In order to alleviate these issues, we propose category-invariant feature enhancement (CIFE), a general mechanism that enhances the adversarial domain adaptation through optimizing the adaptability. Specifically, the CIFE approach introduces category-invariant features to boost the discriminability of domain-invariant features with preserving the transferability. Experiments show that the CIFE could improve upon representative adversarial domain adaptation methods to yield state-of-the-art results on five benchmarks.

1. Introduction

In typical supervised machine learning algorithms, the training data and the test data are assumed to stem from the same distribution [9, 27]. Unfortunately, in many real-world cases, a shortage of labeled data in the interested domain is not uncommon. Thus, it is of great significance to investigate how to apply knowledge learned from a label-dense (source) domain to a label-scarce (target) domain. As the distributions of these two domains are often different, deep neural networks trained on the source domain are inclined to make spurious predictions on the target domain [18, 25, 28].

To address the above issue, domain adaptation is proposed to learn transferable representations across domains such that a model trained on the source domain can simultaneously perform well on the target domain [26]. Early domain adaptation methods reweigh the source instances based on their associations to the target domain with regard to human-engineered features [8]. Motivated by the domain adaptation theory [2, 1], which suggests that the expected error on the target domain is bounded by three elements: (1) the expected error on the source domain; (2) the divergence between the two domains; (3) the adaptability. Recent methods focus on minimizing the domain divergence and explore two possible strategies for aligning different domains. The first one is to minimize some measures of domain distance, such as maximum mean discrepancy (MMD) [29] and correlation distances [22]. The second one is adversarial domain adaptation, which employs a two-player minimax game similar to generative adversarial networks (GANs) [10]. In this paradigm, a domain discriminator is trained against a feature extractor, the domain discriminator aims to distinguish the source features from the target features while the feature extractor tries to confuse the discriminator. When these two components reach equilibrium, the learned features can be regarded as domain-invariant. These adversarial domain adaptation methods [16, 26] have yielded state-of-the-art results.

For most existing domain adaptation methods, the adaptability is assumed to be a small constant that never varies in the process of domain alignment [1]. However, this assumption is often violated in practical. Given feature representations, the adaptability can be explicitly quantified as the expected error of the ideal joint hypothesis over the source and target domains. When the adaptability is poor, good domain adaptation models can not be expected. As transforming features to be domain-invariant will inevitably distort the original feature distributions and enlarge the error of the ideal joint hypothesis, a good adaptability can not be fully guaranteed. Moreover, in the process of learning domain-invariant features, the transferability is enhanced at the expense of sacrificing discriminability [4]. In this paper, we propose a novel category-invariant feature enhancement (CIFE) mechanism, which introduces category-invariant features, to address the above two issues. Similar to domain alignment, the generation of category-invariant features can also be formulated as a two-player game: a feature extractor is trained against a category discriminator, the category discriminator tries to distinguish different labels, while the feature extractor aims to fool the category discriminator. By adversarially training the feature extractor and the category discriminator, we can make the learned features transferable across categories, i.e. category-invariant. We term this process as category alignment. The category-invariant features are supposed to represent domain-specific information and boost the model adaptability by complementing the discriminability of the domain-invariant features. Experiments show that our method enables existing adversarial domain adaptation models to learn transferable feature representations without sacrificing much discriminability, and yield state-of-the-art results on five benchmarks. The contributions of our paper are summarized as follows:

- We propose a category-invariant feature enhancement (CIFE) mechanism, which enhances the discriminability of the domain-invariant features by introducing the category-invariant features. The proposed CIFE improves the system performance by optimizing the adaptability, rather than further reducing the domain divergence.
- To evaluate the efficacy of CIFE, we embed CIFE into two existing adversarial domain adaptation methods and evaluate them on five benchmarks. Our proposed CIFE significantly improves upon these two methods by yielding state-of-the-art results.
- Further experiments are conducted to validate the feasibility of advancing domain adaptation by optimizing the adaptability, and explore how the hyperparameter influences the performance of the model.

2. Related Work

The main objective of domain adaptation is to transfer the knowledge learned from the source domain to the target domain. Unsupervised domain adaptation (UDA) tackles a more challenging scenario where there is no direct access to the label information of the target domain. As deep neural networks can automatically extract feature representations from massive data, deep neural network-based methods have been widely studied for UDA. The deep adaptation network (DAN) applies maximum mean discrepancy (MMD) to layers embedded in a reproducing kernel hilbert space, effectively matching higher-order statistics of the two distributions [15]. The joint adaptation network (JAN) learns a learner by aligning the joint distributions of multiple domain-specific layers across different domains based on a joint MMD criterion [17]. The deep correlation alignment (CORAL) proposes to match the means and covariances of two domains [22].

Inspired by the success of generative adversarial networks (GANs), [7] proposes domain discriminative neural networks (DANNs) which could learn domain-invariant features by deploying adversarial learning between a domain discriminator and a feature extractor. DANN projects the source and target domains into a shared latent space and adversarially performs domain alignment to reduce the domain divergence in the domain adaptation theory [1]. Based on adversarial domain adaptation, a line of works further reduce the domain divergence through improving the domain discriminator or the procedure of adversarial learning. The adversarial discriminative domain adaptation (ADDA) uses asymmetric feature extractors for the two domains to conduct the alignment [23]. The multi-adversarial domain adaptation (MADA) captures multi-mode structures by re-weighting features with category predictions [19]. The cycle-consistent adversarial domain adaptation (CyCADA) implements domain adaptation at both pixel-level and feature-level by using cycle-consistent adversarial training [12]. The conditional adversarial domain adaptation (CDAN) conditions the domain discriminator on discriminative information by multiplicative interactions between feature representations and predictions [16]. The batch spectral penalization (BSP) penalizes the largest singular values to strengthen other eigenvectors of the learned domain-invariant features to boost the discriminability [4]. The dynamic adversarial adaptation network (DAAN) dynamically learns domain-invariant representations while quantitatively evaluating the relative importance of global and local domain distributions [30]. The batch nuclear-norm maximization (BNM) enlarges the nuclear-norm of the batch output matrix to make the learned domain-invariant features more discriminative [5]. The enhanced transport distance (ETD) exploits the attention scores estimating the similarity between samples to weigh the transport distance between domains and learned discriminative features by reducing the weighed distance [14]. The label propagation with augmented anchors (A^{2}LP) improves the label propagation via generation of unlabeled virtual samples with high confidence label prediction [31]

Most previous UDA methods concentrate on minimizing the domain divergence, few works explore optimizing the adaptability to improve the generalization ability of the model. In this paper, our CIFE approach tries to improve domain adaptation methods through optimizing the adaptability.

3. Method

In this work, we consider unsupervised domain adaptation in the following setting. There exist abundant labeled instances in the source domain, $D^s = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{n_s}$ with $\mathbf{x}_i^s \in \mathcal{X}$ and $y_i^s \in \mathcal{Y}$, and a set of unlabeled instances in the target domain, $D^t = \{\mathbf{x}_i^t\}_{i=1}^{n_t}$ with $\mathbf{x}_i^t \in \mathcal{X}$. The data in the two domains are drawn from different distributions $\mathcal{S}$ and $\mathcal{T}$, but share the same label space. The main objective is to learn a model $h : \mathcal{X} \to \mathcal{Y}$ that has a good capacity of generalizing on both source and target domains.

3.1. Adversarial Domain Adaptation

The key idea of adversarial domain adaptation is to learn domain-invariant features that can be generalized across domains. Starting from the domain adversarial neural networks (DANNs) [7], adversarial learning has been widely adopted to learn feature representations to bridge the domain divergence in UDA. Considering DANN as an example, the base network is composed of three components: a feature extractor F, a task-specific classifier C, and a binary domain discriminator D. The feature extractor $F : \mathcal{X} \to \mathbb{R}^m$ maps one input instance $\mathbf{x}$ from the input space $\mathcal{X}$ into a shared latent space $F(\mathbf{x}) \in \mathbb{R}^m$. The classifier $C : \mathbb{R}^m \to \mathcal{Y}$ transforms a feature vector in the shared latent space to the label space $\mathcal{Y}$. The domain discriminator $D : \mathbb{R}^m \to [0, 1]$ separates the source features (with domain index 0) from the target ones (with domain index 1) in the latent space. By adversarially training F to confuse D, DANN can learn transferable features across domains. Moreover, the feature extractor F and the classifier C are trained simultaneously to minimize the classification error on the source labeled data. This makes the learned features discriminative across categories. Formally, DANN can be formulated as:

$$\min_{F,C} \max_{D} \mathcal{L}_c(F, C) + \lambda_d \mathcal{L}_d(F, D) \tag{1}$$

$$\mathcal{L}_c(F, C) = \mathbb{E}_{(\mathbf{x}^s, y^s)\sim\mathcal{S}} \ell(C(F(\mathbf{x}^s)), y^s) \tag{2}$$

$$\begin{aligned} \mathcal{L}_d(F, D) =& \mathbb{E}_{\mathbf{x}^s\sim\mathcal{S}} \log[D(F(\mathbf{x}^s))] \\ &+ \mathbb{E}_{\mathbf{x}^t\sim\mathcal{T}} \log[1 - D(F(\mathbf{x}^t))] \end{aligned} \tag{3}$$

Where $\ell(\cdot, \cdot)$ is the canonical cross-entropy loss function, and λ_d is a trade-off hyperparameter.

3.2. Category-Invariant Feature

The batch spectral penalization [4] reveals that the feature representations can be decomposed into eigenvectors with importance quantified by the corresponding singular values from a spectral analysis viewpoint. The feature transferability mainly resides in the eigenvectors with top singular values, the eigenvectors with low singular values embody domain-specific variations and should be discouraged. In contrast, the feature discriminability depends on all eigenvectors because the rich discriminative structures can not be fully expressed by only a few eigenvectors. Therefore, there exists a contraction between transferability and

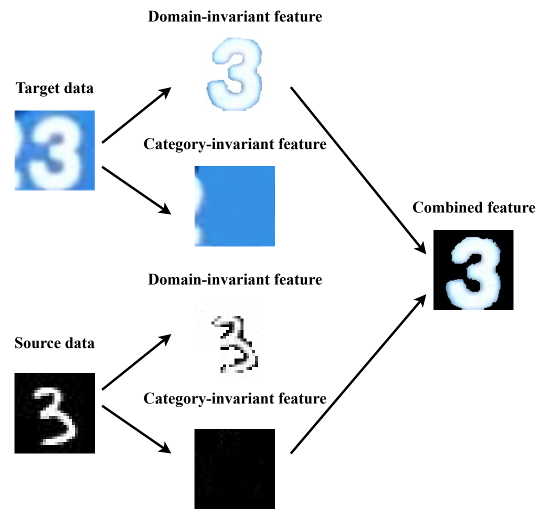


Figure 1. The illustration of the CIFE mechanism.

discriminability. In the process of learning transferable feature representations, we need to sacrifice some discriminability by suppressing domain-specific variations. As the target domain has no label, this sacrifice mainly resides in the target data. In order to complement the sacrificed domain-specific information, we introduce the category-invariant features. The category-invariant features can be yielded through category alignment and are supposed to be category-irrelevant and domain-specific. In practice, by replacing the domain discriminator D with a category discriminator D_t, we can use adversarial training between the feature extractor F and the category discriminator D_t to obtain the category-invariant features by processing the training instances and their corresponding labels. In UDA, as we have no access to the label information of the target domain, the category-invariant features can only be learned from the source domain, which can be encoded as follows:

$$\mathcal{L}_d^c(F, D_t) = \mathbb{E}_{(\mathbf{x}^s, y^s)\sim\mathcal{S}} \ell(D_t(F(\mathbf{x}^s)), y^s) \tag{4}$$

As the category-invariant features are domain-specific, which has been demonstrated to be beneficial to their own domain and detrimental to other domains [3], it is challenging to apply the source category-invariant features to boost the classification accuracy of the target data.

3.3. Category-Invariant Feature Enhancement

In this paper, we assume that the essential information of an image can be decomposed into two independent sets: (1) the domain-invariant information, that is discriminative across domains; (2) the domain-specific information, that is transferable across categories for one domain. For one specific image, the combination of these two types of information is expected to represent each individual characteristic of that image. For a source image and a target im-

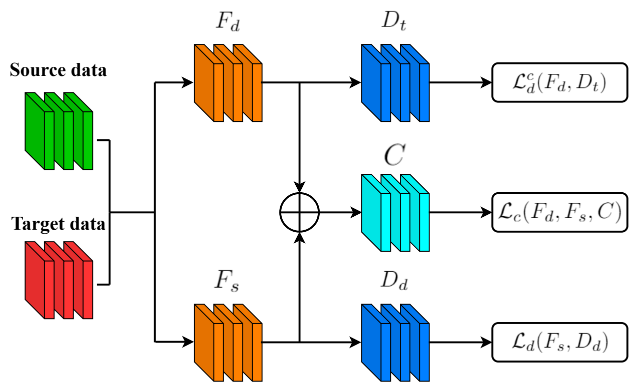


Figure 2. The architecture of CIFE+DANN where CIFE introduces category-invariant features to enhance the discriminability of the domain-invariant features learned by DANN. CIFE can be easily plugged into any adversarial domain adaptation method, which is end-to-end trainable. The CIFE+DANN model consists of five components: the domain-specific feature extractor F_d, which is used to capture the category-invariant features; the domain-invariant feature extractor F_s, which aims to learn the domain-invariant features; the category discriminator D_t, which identifies different labels of the input features; the domain discriminator D_d, which differentiates source features from target features; the classifier C, which is used to conduct the classification. $\mathcal{L}_c(F_d, F_s, C)$ is the typical cross-entropy loss, $\mathcal{L}_d(F_s, D_d)$ and $\mathcal{L}_d^c(F_d, D_t)$ are adversarial loss functions that guide the domain-invariant feature generation and the category-invariant feature generation, respectively.

age that share the same label, we can obtain their domain-invariant features and category-invariant features. If these features contain no noise, e.g. the domain-invariant features are not contaminated by the domain-specific information while the category-invariant features carry no discriminative information. When we combine the category-invariant features of the source image and the domain-invariant features of the target image, the combination is expected to contain all essential information of the source image, as illustrated in Figure 1. Therefore, it is feasible to feed the combinations of source category-invariant features and target domain-invariant features to a classifier trained on the source domain and obtain accurate predictions.

As stated above, the source category-invariant features can be used to complement the target domain-invariant features. Our proposed CIFE can easily be embedded into existing adversarial domain adaptation approaches, such as DANN [7] and conditional adversarial domain adaptation (CDAN) [16]. The architecture of CIFE+DANN is illustrated in Figure 2, which consists of five components: a domain-specific feature extractor F_d, a category discriminator D_t, a domain-invariant feature extractor F_s, a domain discriminator D_d, and a classifier C. The domain-specific feature extractor F_d aims to learn the category-invariant features, while the domain-invariant feature extractor F_s captures the domain-invariant features. With the CIFE applied to the DANN, there exist two two-player minimax games: The first one is played between the F_d and the D_t, aiming to extract category-invariant features from the source domain; The second one is the typical adversarial learning, which is deployed between the F_s and the D_d, trying to capture domain-invariant features from both the source and target domains. By incorporating these two types of adversarial learning, the transferability of the domain-invariant features can be preserved as much as possible, which is originally tailored for the domain adaptation. While the discriminability of the domain-invariant features can be enhanced by complementing with domain-specific information. The source data should be involved in both minimax objectives, while the target data only participate in the domain alignment. In our work, we concatenate category-invariant features and domain-invariant features, feeding the concatenations to the classifier C as the input to conduct classification. Thus, the classification loss should be rewritten as:

$$\mathcal{L}_c(F_d, F_s, C) = \mathbb{E}_{(\mathbf{x}^s, y^s)\sim\mathcal{S}}\ell(C([F_d(\mathbf{x}^s), F_s(\mathbf{x}^s)]), y^s) \tag{5}$$

where $[\cdot,\cdot]$ indicates the concatenation of two vectors. To learn representations with both transferability and discriminability, the dual adversarial learning of CIFE+DANN is formulated as:

$$\min_{F_d, F_s, C}\max_{D_t, D_d}\mathcal{L}_c(F_d, F_s, C) + \lambda_d\mathcal{L}_d(F_s, D_d) + \lambda_c\mathcal{L}_d^c(F_d, D_t) \tag{6}$$

where λ_d and λ_c are hyperparameters that trade-off different loss functions.

3.4. Training and Predicting Procedure

The training algorithm of CIFE+DANN, which uses mini-batch stochastic gradient descent, is presented in Algorithm 1. In each iteration, the source and target samples are fed into the model to generate category-invariant and domain-invariant features. The category-invariant features

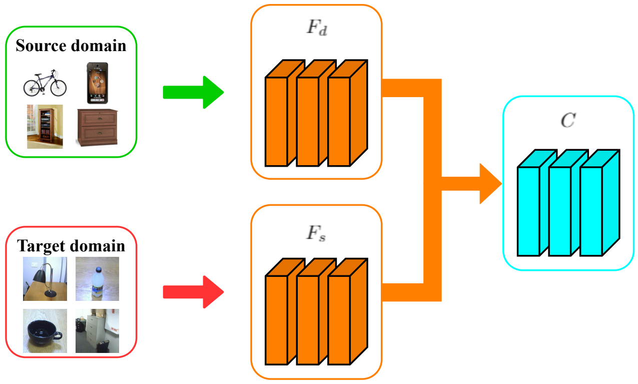


Figure 3. The predicting process of the CIFE+DANN method.

Algorithm 1 Stochastic gradient descent training algorithm of CIFE+DANN

1: **Input:** Source domain: D^s, target domain: D^t, and batch size: N.
2: **Output:** Configurations of CIFE+DANN
3: **Initialize** λ_d and λ_c
4: **for** number of training iterations **do**
5: $(\mathbf{x}^s, y^s) \leftarrow$ RANDOMSAMPLE(D^s, N)
6: $(\mathbf{x}^t) \leftarrow$ RANDOMSAMPLE(D^t, N)
7: Calculate $l_D = \lambda_d \mathcal{L}_d(F_s, D_d) + \lambda_c \mathcal{L}_d^c(F_d, D_t)$; Update D_d and D_t by ascending along gradients ∇l_D.
8: Calculate $loss = \mathcal{L}_c(F_d, F_s, C) + \lambda_d \mathcal{L}_d(F_s, D_d) + \lambda_c \mathcal{L}_d^c(F_d, D_t)$; Update F_s, F_d and C by descending along gradients $\nabla loss$.
9: **end for**

are obtained by adversarially training the domain-specific feature extractor F_d and the category discriminator D_t, while the domain-invariant features are yielded by letting the domain-invariant feature extractor F_s compete with the domain discriminator D_d. The concatenations of category-invariant features and domain-invariant features are supposed to be both transferable and discriminative. λ_d and λ_c are hyperparameters balancing different losses. When evaluating the target test data, some source data are required to provide category-invariant features. Therefore, we randomly draw samples from the source training set to fulfill this requirement. The predicting process is illustrated in Figure 3.

3.5. Discussion

For the representation learning of adversarial domain adaptation, both transferability and discriminability are crucial. Specifically, transferability can guarantee the knowledge learned from the source domain to generalize to the target domain and discriminability enables the model to identify different categories [18, 4]. Based on the domain adaptation theory [1], an essential prerequisite for domain adaptation is a good adaptability over the source and target domains. In domain adaptation, there exist three main research issues: (1) what to transfer, (2) how to transfer, and (3) when to transfer [18]. Adversarial domain adaptation approaches focus on the first challenge, trying to learn both transferable and discriminative features through minimizing domain divergence adversarially. However, as stated in [4], this class of techniques is risky in transforming features to be domain-invariant as it needs to suppress domain-specific variations of the original feature distributions, imposing detrimental effects to the discriminability of the learned features. In addition, most previous UDA methods regard the adaptability as a small constant and ignore its influence in domain alignment, few works investigate optimizing adaptability to improve the system performance.

In this paper, we introduce category-invariant features, that are obtained by performing category alignment. These features are supposed to be category-irrelevant and domain-specific, and can be applied to compensate for the sacrificed domain-specific variations of the domain-invariant features, boosting their discriminability. By incorporating category-invariant features with domain-invariant features, the adaptability can be essentially controlled to be small in the training process, yielding lower expected error on the target domain. In summary, our proposed CIFE can enable existing adversarial domain adaptation approaches to learn both transferable and discriminative feature representations, and the category-invariant features can make contributions to the ”what to transfer” research issue.

4. Experiments

4.1. Dataset

Office-31 [21] is a standard domain adaptation dataset. It contains images among 31 classes from 3 domains: Amazon (A) with 2817 images, which contains images down-

Table 1. Accuracy (%) on Office-31.

Method	A→W	D→W	W→D	A→D	D→A	W→A	Avg
ResNet-50 [11]	68.4±0.2	96.7±0.1	99.3±0.1	68.9±0.2	62.5±0.3	60.7±0.3	76.1
DAN [15]	80.5±0.4	97.1±0.2	99.6±0.1	78.6±0.2	63.6±0.3	62.8±0.2	80.4
DANN [6]	82.0±0.4	96.9±0.2	99.1±0.1	79.7±0.4	68.2±0.4	67.4±0.5	82.2
JAN [17]	85.4±0.3	97.4±0.2	99.8±0.2	84.7±0.3	68.6±0.3	70.0±0.4	84.3
MADA [19]	90.0±0.1	97.4±0.1	99.6±0.1	87.8±0.2	70.3±0.3	66.4±0.3	85.2
CDAN [16]	93.1±0.2	98.2±0.2	**100.0**±0.0	89.8±0.3	70.1±0.4	68.0±0.4	86.6
BSP [4]	93.3±0.2	98.2±0.2	**100.0**±0.0	93.0±0.2	73.6±0.3	72.6±0.3	88.5
ETD [14]	92.1	**100.0**	**100.0**	88.0	71.0	67.8	86.2
A^2LP [25]	87.7	98.1	99.0	87.8	75.8	**75.9**	87.4
BNM [5]	92.8	98.8	**100.0**	92.9	73.5	73.8	88.6
CIFE+DANN	90.7±0.3	99.0±0.1	**100.0**±0.0	90.0±0.5	71.0±0.3	69.9±0.3	86.8
CIFE+CDAN	**94.0**±0.2	99.3±0.1	**100.0**±0.0	**93.4**±0.2	**75.9**±0.2	74.3±0.3	**89.5**

Table 2. Accuracy (%) on ImageCLEF-DA.

Method	I→P	P→I	I→C	C→I	C→P	P→C	Avg
ResNet-50 [11]	74.8±0.3	83.9±0.1	91.5±0.3	78.0±0.2	65.5±0.3	91.2±0.3	80.7
DAN [15]	74.5±0.4	82.2±0.2	92.8±0.2	86.3±0.4	69.2±0.4	89.8±0.4	82.5
DANN [6]	75.0±0.6	86.0±0.3	96.2±0.4	87.0±0.5	74.3±0.5	91.5±0.6	85.0
JAN [17]	76.8±0.4	88.0±0.2	94.7±0.2	89.5±0.3	74.2±0.3	91.7±0.3	85.8
MADA [19]	75.0±0.3	87.9±0.2	96.0±0.3	88.8±0.3	75.2±0.2	92.2±0.3	85.8
CDAN [16]	76.7±0.3	90.6±0.3	97.0±0.4	90.5±0.4	74.5±0.3	93.5±0.4	87.1
DAAN [30]	78.5	91.3	94.4	88.4	74.0	94.3	86.8
ETD [14]	**81.0**	91.7	97.9	93.3	**79.5**	95.0	89.7
A^2LP [31]	79.3	91.8	96.3	91.7	78.1	96.0	88.9
CIFE+DANN	77.0±0.2	91.1±0.2	97.3±0.3	90.8±0.3	74.5±0.5	93.7±0.3	87.4
CIFE+CDAN	79.5±0.3	**93.0**±0.2	**98.2**±0.3	**93.6**±0.3	79.2±0.4	**96.1**±0.4	**90.0**

loaded from amazon.com, Webcam (W) with 795 images, and DSLR (D) with 498 images, containing images obtained by web camera and DSLR camera with different settings, respectively. We conduct evaluations on all 6 tasks: A→W, D→W, W→D, A→D, D→A, and W→A

ImageCLEF-DA is a benchmark for ImageCLEF 2014 domain adaptation challenges. It is organized by selecting 12 common classes shared by three domains: Caltech-256 (C), ImageNet ILSVRC 2012 (I), and Pascal VOC 2012 (P). Each domain contains 600 images and 50 images for each class. The three domains in this dataset are of the same size, which is good complementation of the Office-31 dataset where different domains are of different sizes. We evaluate all methods on 6 tasks: I→P, P→I, I→C, C→I, C→P, and P→C.

VisDA-2017 [20] is a large simulation-to-real dataset with two domains. It contains over 280,000 images of 12 classes. The source domain is termed Synthetic which contains images obtained by rendering 3D models of the same object classes as in the real data from different angles and under different lighting conditions. The target domain is termed Real which comprises natural images. We evaluate the task: Synthetic→Real.

Digits. We investigate three digits datasets: MNIST, USPS, and Street View House Numbers (SVHN). Each dataset contains digit images of 10 classes (0-9). MNIST consists of grayscale handwritten digit images of size 28×28, USPS contains 16×16 grayscale images and SVHN composes 32×32 colored images which might contain more than one digit in each image. We adopt the experimental settings of CyCADA [12] with three tasks: MNIST to USPS (M→U), USPS to MNIST (U→M), and SVHN to MNIST (S→M). All input images should be resized to the size of 32×32.

Office-Home [24] is a more complicated dataset than Office-31, which consists of around 15500 images from 65 classes in office and home settings. There exist 4 domains in this dataset: Artistic Images (Ar) denotes artistic depictions for object images, Clip Art (Cl) shows picture collection of clipart, Product Images (Pr) presents object images with a clear background and is similar to Amazon category in Office-31, and Real-World Images (Rw) represents object images collected with a regular camera. We establish 12 transfer tasks by using all domain combinations.

4.2. Comparison Methods

We extend domain adversarial neural network (DANN) [7], conditional adversarial domain adaptation (CDAN)

Table 3. Accuracy (%) on Digits and VisDA-2017.

Method	M→U	U→M	S→M	Avg	Method	Synthetic→Real
No Adaptation [12]	82.2	69.6	67.1	73.0	ResNet-101 [11]	52.4
DANN [6]	90.4	94.7	84.2	89.8	DANN [6]	57.4
ADDA [23]	89.4	90.1	86.3	88.6	DAN [15]	61.1
CyCADA [12]	95.6	96.5	90.4	94.2	JAN [17]	65.7
CDAN [16]	93.9	96.9	88.5	93.1	CDAN [16]	73.7
BSP [4]	95.0	98.1	92.1	95.1	BSP [4]	75.9
CIFE+DANN	93.7	97.4	93.1	94.7	**CIFE+DANN**	74.4
CIFE+CDAN	**96.1**	**98.8**	**94.3**	**96.4**	**CIFE+CDAN**	**78.1**

Table 4. Accuracy (%) on Office-Home.

Method	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg
ResNet-50 [11]	34.9	50.0	58.0	34.7	41.9	46.2	38.5	31.2	60.4	53.9	41.2	59.9	46.1
DAN [15]	43.6	57.0	67.9	45.8	56.5	60.4	44.0	43.6	67.7	63.1	51.5	74.3	56.3
DANN [6]	45.6	59.3	70.1	47.0	58.5	60.9	46.1	43.7	68.5	63.2	51.8	76.8	57.6
JAN [17]	45.9	61.2	68.9	50.4	59.7	61.0	45.8	43.4	70.3	63.9	52.4	76.8	58.3
CDAN [16]	49.0	69.3	74.5	54.4	66.0	68.4	55.6	48.3	75.9	68.4	55.4	80.5	63.8
BSP [4]	52.0	68.6	76.1	58.0	70.3	70.2	**58.6**	50.2	77.6	**72.2**	**59.3**	81.9	66.3
CIFE+DANN	47.8	65.9	73.4	48.3	62.7	64.2	48.7	46.9	74.5	68.2	53.4	80.7	61.2
CIFE+CDAN	**52.3**	**71.0**	**78.3**	**58.9**	**71.8**	**72.3**	58.1	**52.4**	**79.7**	71.1	58.9	**83.2**	**67.4**

[16] with the proposed category-invariant feature enhancement (CIFE). We compare with a number of state-of-the-art methods: Deep adaptation network (DAN) [15], domain adversarial neural network (DANN) [7], joint adaptation network (JAN) [17], multi-adversarial domain adaptation (MADA) [19], conditional adversarial domain adaptation (CDAN) [16], adversarial discriminative domain adaptation (ADDA) [23], cycle-consistent adversarial domain adaptation (CyCADA) [12], batch spectral penalization (BSP) [4], dynamic adversarial domain adaptation (DAAN) [30], batch nuclear-norm maximization (BNM) [5], enhanced transport distance (ETD) [14] and label propagation with augmented anchors (A^2LP) [31].

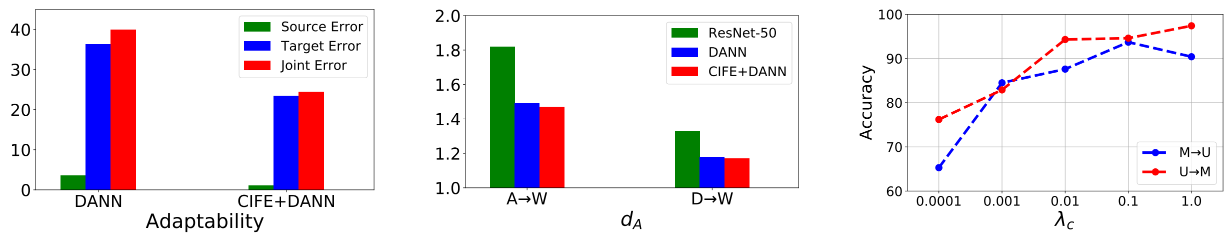
4.3. Implementation Details

The standard evaluation protocols [16] of unsupervised domain adaptation are followed in our experiments. All labeled source samples and unlabeled target samples are used in the training stage. In the testing stage, we randomly draw some samples from the training source dataset to provide the category-invariant features. The average classification accuracy based on three random experiments is reported. No data augmentation is used in any of the experiments to allow a fair comparison. For Office-31, ImageCLEF-DA, and Office-Home datasets, we use ResNet-50 [11] pre-trained on ImageNet [13] as the backbone. For VisDA-2017 dataset, we use ResNet-101 [11] pre-trained on ImageNet [13] as the backbone. For digits datasets, we adopt a modified version of Lenet architecture as the base network and train models from scratch. For each backbone network, we use all its layers up to the second last one as the domain-invariant feature extractor F_s. The domain-specific feature extractor F_d adopts the same architecture as F_s. The classifier C uses a single fully-connected layer whose input dimension should be the sum of the output dimensions of F_s and F_d. For domain discriminator D_d, we use the same architecture as DANN [6]. The architecture of the category discriminator D_t is similar to that of D_d, the only difference lies in the top layer, a softmax layer is used for D_t while a sigmoid layer is used for D_d.

We implement all experiments using PyTorch. We adopt mini-batch SGD with momentum of 0.9 and the learning rate annealing strategy as [7]: the learning rate is adjusted by $\eta_p = \frac{\eta_0}{(1+\theta p)^\beta}$, where p denotes the process of training epochs that is normalized to be in $[0, 1]$, and we set $\eta_0 = 0.01$, $\theta = 10$, $\beta = 0.75$, which are optimized to promote convergence and low errors on the source domain. λ_d is progressively changed from 0 to 1 by multiplying to $\frac{1-exp(-\delta p)}{1+exp(-\delta p)}$, where $\delta = 10$. For all experiments, we select λ_c in the range $\{0.0001, 0.001, 0.01, 0.1, 1.0\}$ via tuning on the unlabeled target data.

4.4. Results

The results on Office-31 are reported in Table 1, with experimental results of baselines directly reported from their original papers wherever available. It indicates that our proposed CIFE mechanism significantly improves the accuracies of DANN [7] and CDAN [16], and achieves state-of-the-art results. Specifically, CIFE+CDAN achieves the best accuracies not only on tasks: A→W, W→D, A→D, and D→A, but also on the average accuracy. Compared with CDAN, there is an obvious improvement in classification



(a) Adaptability (b) A-distance (c) λ_c

Figure 4. Analysis of adaptability, domain divergence and parameter sensitivity.

accuracies on relatively difficult tasks D→A and W→A where the source domain is quite small. Moreover, as reported in Table 2, 3, and 4, our method can also boost the performance of DANN and CDAN on ImageCLEF-DA, VisDA-2017, Digits and Office-Home. In particular, CIFE+CDAN exceeds baselines on 4 out of 6 domains and achieves the best performance in terms of the average accuracy for ImageCLEF-DA dataset. For Digits datasets, CIFE+CDAN can yield better results on all three tasks, it can also achieve the best average performance compared with other methods. For VisDA-2017 dataset, CIFE+CDAN produces the best average classification accuracy among all comparison methods and outperforms the baseline of ResNet-101 model pre-trained on ImageNet with a great margin. For Office-Home dataset, CIFE+CDAN outperforms other baselines on 9 out of 12 domains and yields the best average classification accuracy.

4.5. Analysis

Adaptability. In this study, we investigate how the category-invariant feature influences the adaptability. In order to compute the adaptability, we train a multi-layer perceptron (MLP) classifier over the feature representations learned by DANN [6] and CIFE+DANN on VisDA-2017. The MLP classifier is trained on all labeled data from both the source and target domains. It should be noted that the target labels are only used in this analysis. When training the MLP classifier, all feature extractors in DANN and CIFE+DANN should be fixed. As shown in Figure 4(a), we compare the error rates of the ideal joint hypothesis on the source domain, the target domain, and their sum. We observe that the adaptability of the CIFE+DANN is much lower than that of DANN. Obviously, a higher error rate indicates weaker discriminability, leading to poor adaptability as suggested by the domain adaptation theory [1].

Distribution Discrepancy. As shown in the domain adaptation theory [1], the domain discrepancy and adaptability are two important factors that bound the generalization error on the target domain. The A-distance [1] is a measure of domain discrepancy, defined as $d_A = 2(1-2\epsilon)$, where ϵ is the error rate of the domain discriminator trained to distinguish source features from target features. In this study, we compare the A-distance of ResNet-50 [11], DANN [6], and CIFE+DANN on two tasks of Office-31 dataset: A→W and D→W. The results are shown in Figure 4(b). It can be noted that the DANN and CIFE+DANN yield smaller A-distances, indicating the adversarial training can effectively reduce the domain divergence. The A-distance of CIFE+DANN is close to that of DANN on both tasks, revealing that our CIFE improves system performance by optimizing the adaptability rather than further reducing the domain divergence.

Parameter Sensitivity Analysis. In this section, we discuss the sensitivity of CIFE+DANN to the values of the hyperparameter λ_c. We evaluate the influence of λ_c on Digits dataset, especially, the M→U and U→M tasks. λ_c is explored in the range $\{0.0001, 0.001, 0.01, 0.1, 1.0\}$. The results are shown in Figure 4(c). From Figure 4(c), we observe that the selection of λ_c has an influence on the system performance. For the task M→U, with an increase of λ_c, the accuracy increases rapidly and obtains the best value at $\lambda_c = 0.1$. After this point, the further increase of λ_c deteriorates the performance. For the task U→M, the accuracy increases from 0.0001 to 1.0 and reaches its best at $\lambda_c = 1.0$. This analysis suggests that a properly selected λ_c can effectively improve the system performance.

5. Conclusion

In this paper, we propose a novel category-invariant feature enhancement (CIFE) mechanism for adversarial domain adaptation. The CIFE incorporates category-invariant features to existing adversarial domain adaptation methods to boost the discriminability of the domain-invariant features. It improves the performance of UDA models by optimizing the adaptability. This approach provides an alternative to the mainstream UDA methods which focus on minimizing the divergence between two domains. We demonstrate that the category-invariant features learned from the source domain can be beneficial to the classification of the target domain. The CIFE approach is general and can be embedded into the existing adversarial domain adaptation methods to boost system performance.

References

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

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

[3] Xilun Chen and Claire Cardie. Multinomial adversarial networks for multi-domain text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1226–1240, 2018.

[4] Xinyang Chen, Sinan Wang, Mingsheng Long, and Jianmin Wang. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In *International Conference on Machine Learning*, pages 1081–1090, 2019.

[5] Shuhao Cui, Shuhui Wang, Junbao Zhuo, Liang Li, Qingming Huang, and Qi Tian. Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3941–3950, 2020.

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

[7] 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.

[8] Boqing Gong, Kristen Grauman, and Fei Sha. Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 222–230, 2013.

[9] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[11] 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.

[12] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. PMLR, 2018.

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

[14] Mengxue Li, Yi-Ming Zhai, You-Wei Luo, Peng-Fei Ge, and Chuan-Xian Ren. Enhanced transport distance for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13936–13944, 2020.

[15] 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.

[16] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pages 1640–1650, 2018.

[17] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, pages 2208–2217. PMLR, 2017.

[18] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[19] Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. Multi-adversarial domain adaptation. *arXiv preprint arXiv:1809.02176*, 2018.

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

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

[22] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer, 2016.

[23] 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.

[24] 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.

[25] Yuan Wu and Yuhong Guo. Dual adversarial co-learning for multi-domain text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6438–6445, 2020.

[26] Yuan Wu, Diana Inkpen, and Ahmed El-Roby. Dual mixup regularized learning for adversarial domain adaptation. In *European Conference on Computer Vision*, pages 540–555. Springer, 2020.

[27] Yuan Wu, Diana Inkpen, and Ahmed El-Roby. Conditional adversarial networks for multi-domain text classification. In

Proceedings of the Second Workshop on Domain Adaptation for NLP, pages 16–27, 2021.

[28] Yuan Wu, Diana Inkpen, and Ahmed El-Roby. Mixup regularized adversarial networks for multi-domain text classification. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7733–7737. IEEE, 2021.

[29] Hongliang Yan, Yukang Ding, Peihua Li, Qilong Wang, Yong Xu, and Wangmeng Zuo. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2272–2281, 2017.

[30] Chaohui Yu, Jindong Wang, Yiqiang Chen, and Meiyu Huang. Transfer learning with dynamic adversarial adaptation network. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 778–786. IEEE, 2019.

[31] Yabin Zhang, Bin Deng, Kui Jia, and Lei Zhang. Label propagation with augmented anchors: A simple semi-supervised learning baseline for unsupervised domain adaptation. In *European Conference on Computer Vision*, pages 781–797. Springer, 2020.