Rethinking Triplet Loss for Domain Adaptation

Weijian Deng, Liang Zheng, Yifan Sun, and Jianbin Jiao

***Abstract*—The gap in data distribution motivates domain adaptation research. In this area, image classification intrinsically requires the source and target features to be co-located if they are of the same class. However, many works only take a global view of the domain gap. That is, to make the data distributions globally overlap; and this does not necessarily lead to feature co-location at the class level.**

To resolve this problem, we study metric learning in the context of domain adaptation. Specifically, we introduce a similarity guided constraint (SGC). In the implementation, SGC takes the form of a triplet loss. The triplet loss is integrated into the network as an additional objective term. Here, an image triplet consists of two images of the same class and another image of a different class. Albeit simple, the working mechanism of our method is interesting and insightful. Importantly, images in the triplets are sampled from the source and target domains. From a micro perspective, by enforcing this constraint on every possible triplet, images from different domains but of the same class are mapped nearby, and those of different classes are far apart. From a macro perspective, our method ensures that cross-domain similarities are preserved, leading to intra-class compactness and inter-class separability. Extensive experiment on four datasets shows our method yields significant improvement over the baselines and has a competitive accuracy with the state-of-the-art results.

***Index Terms*—Domain adaptation, triplet loss, semantic alignment**

I. Introduction

In many real-world computer vision applications, the training and testing data distributions are often different because of *dataset bias* [1]. The distribution discrepancy decreases the generalization capability of the learned visual representations. An example is that a model trained on synthetic images may fail to perform well on real-world images. To reduce dataset bias, a commonly mentioned strategy is unsupervised domain adaptation (UDA). In UDA, we are usually provided with a labeled source dataset and an unlabeled target dataset. The goal is to learn a model from these data to minimize the test error on the target dataset.

UDA methods [2]–[7] usually learn a shared feature space where embeddings are domain-invariant. These methods usually minimize some measurement of *global* domain variance [3], [6], [7], such as the correlation distance [8]) and adversarial loss functions [2], [4], [5]. However, under global distribution alignment, images with different labels from the two domains may still be located nearby in the embedding space. This semantic misalignment problem is detrimental to classifier accuracy on the target domain. In literature, some methods [9]–[11] are proposed to address this problem. These methods propose to consider the class information while reducing the distribution discrepancy. In our attempt, we focus on the class-level alignment, and further study the semantic similarity and the semantic dissimilarity between source domain and target domain.

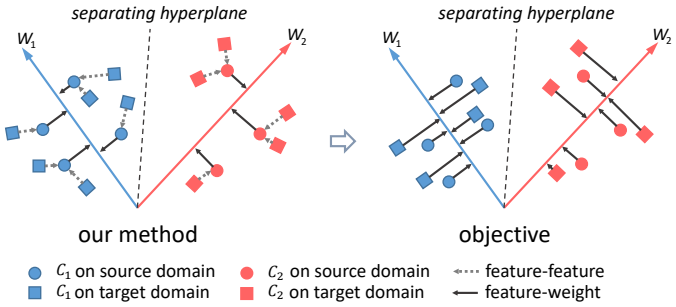


Fig. 1. Illustration of SGC effect. The overall objective is to make source and target features surround their corresponding weight vectors in the classifier, so as to ensure accurate classification. To this end, SGC aims to pull target features to source features with the same class labels. Meanwhile, labeled source images are directly pulled to the expected weight vectors by cross-entropy loss. Thus, SGC leads to more desirable embeddings, and improves the accuracy on the target dataset. In this figure, different colors denote different classes. Weight vectors W_1 and W_2 are corresponding to class C_1 and class C_2, respectively.

Under classification mode, the cross-entropy loss encourages embeddings to be close to their corresponding weight vectors of the classifier; and this ensures the embeddings can be classified correctly. Thus, if target and source features of the same class are close in the embedding space, the classifier learned on source images would have a high accuracy on the target dataset. Motivated by this, we propose a similarity guided constraint (SGC). SGC is enforced on the feature embeddings, the input of classifier, to ensure source and target embeddings have maximal intra-class distance than minimal inter-class distance. Here, we note that enforcing SGC requires labels of the target images, but naturally the target domain is unlabeled. As such, we assign high-confidence pseudo labels to some target images. As shown in Fig. 1, SGC pulls the target images to the source images with the same class labels. This *indirectly* enforces target images to surround their corresponding weight vectors, and thus leads to accuracy improvement on the target images. Importantly, SGC is robust to incorrect pseudo-labels. If we *directly* pull wrongly pseudo-labeled images toward their corresponding weight vectors (as cross-entropy does), the weight vectors would be biased by

W. Deng and L. Zheng are with the Research School of Computer Science, Australian National University, CBR, Australia. E-mail: dengwj16@gmail.com, liangzheng06@gmail.com.

Y. Sun is with Tsinghua University, Beijing, China. E-mail: sunyf15@mails.tsinghua.edu.cn.

J. Jiao is with the University of Chinese Academy of Sciences, Beijing, China. E-mail: jiaojb@ucas.ac.cn.

Corresponding authors: J. Jiao and L. Zheng.

them and dramatic accuracy drop would be caused. Instead, SGC imposes on feature embeddings rather than classifier, thus wrongly labeled target images *have no* direct impact on the weight vectors of classifier, as to be detailed in Section III-C.

In practice, SGC is implemented by using a triplet loss function. By minimizing the triplet loss, *SGC reduces the distance between semantically similar images and increases that of semantically dissimilar images.* This constraint aids to achieve class-level distribution alignment, and thus alleviates the semantic misalignment problem.

To summarize, this paper proposes a similarity guided constraint for unsupervised domain adaptation. With the joint supervision of the triplet loss and some existing constraints, our model can reduce the data distribution gap at both the class level and domain level. Our method is simple and effectively improves over several baselines. We also report competitive accuracy on four benchmarks.

II. Related Work

Unsupervised domain adaptation. Unsupervised domain adaptation methods attempt to minimize the shift between source and target data distributions. There are some methods focusing on learning a mapping function between source and target distributions [12]–[15]. In [15], Correlation Alignment is proposed to match the two distributions. In [14], the source and target domains are aligned in a subspace described by Eigenvectors. It is also effective seek to find a shared space for source and target features [3], [4], [6], [7]. Long *et al.* [7] and Tzeng *et al.* [6] utilize the maximum mean discrepancy (MMD) metric [16] to learn a share feature representation. Several recent methods [2], [4], [5], [10], [17], [18] adopt adversarial learning [19] to learn representations that are not distinguishable between domains. In these methods, a discriminator is trained to tell whether the image feature is from source domain or target domain, while the feature extractor is trained to fool the discriminator. Another strategy is to use adversarial learning to learn transformations in the pixel space from one domain to another [20]–[23]. For example, CYCADA [20] maps samples across domains at both pixel level and feature level.

In this paper, we also attempt to reduce the distribution discrepancy, and we are more concerned with preserving the similarities among the source and target images. In the community, some methods [9]–[11], [24] consider the class information for domain alignment. Pinheiro [11] proposes to classify an image by computing its similarity to prototype representations of each category. The authors of [10] consider the class information and propose the adversarial network for every class. Xie *et al.* [9] align the centers of source and target features of the same class to alleviate the semantic misalignment problem. The authors of [24] propose metric-based domain adaptation method by using triplet loss to train source images. Different from these method, this paper directly enforces metric-based constraint on both source and target images to achieve intra-class compactness and inter-class separability.

Closely to our work, Motiian *et al.*also pairs labeled source and target images to align distributions at class level. Our work is different from [25] in two aspects, 1) the setting of [25] is supervised domain adaptation, where the labeled target images are available; 2) they do not consider the domain-level alignment, while our work benefits from both domain-level alignment and class-level alignment

Self-training. Our method is related to self-training, a strategy in which the predictions of a classifier on unlabeled data are used to retrain the classifier [17], [26]–[34]. The assumption of self-training is that an image with the high predicted score is more likely to be classified correctly. In unsupervised domain adaptation, some methods [17], [35], [36] use pseudo-labeled images to improve classifier accuracy on the target dataset. Zhang *et al.* [17] propose a method named iCAN to progressively select pseudo-labeled images for training the classifier. Chen *et al.* [35] use two classifiers to assign labels for target images. Saito [36] adopt three asymmetric classifiers to improve the quality of pseudo labels. Unlike these methods, we leverage the selected images with their pseudo-labels for class-level alignment instead of retraining the classifier. This practice provides a new way to utilize unlabeled data for learning feature embeddings.

Deep Metric learning. Deep metric learning [37]–[42] aims to learn discriminative embeddings such that similar samples are nearer and different samples are further apart. Discriminative embeddings are also required in other tasks [43]–[48]. For example, Ben *et al.* [48] propose a coupled patch alignment (CPA) algorithm for cross-view gait recognition, which requires the intra-class compactness and inter-class separability across different views. In the community, the most widely used loss functions for deep metric learning are contrastive loss [37] and triplet loss [41]. The problem settings of these works are different from ours. We aim to reduce the distribution discrepancy and utilize the triplet loss [41] to preserve cross-domain similarities.

III. Methodology

A. Overview

In UDA, we are provided with a set of labeled images from the source dataset and a set of unlabeled images from the target dataset, where the data distributions of the two datasets are different. For the source dataset, we denote it as $\mathcal{D}_s = \{(\mathbf{x}_i^s, \mathbf{y}_i^s)\}_{i=1}^{n_s}$, where $\mathbf{x}_i^s$ is the i-th source image, $\mathbf{y}_i^s \in \{0, 1, ..., K-1\}$ is its label, and n_s is the total number of images. Similarly, we denote the target dataset as $\mathcal{D}_t = \{\mathbf{x}_j^t\}_{j=1}^{n_t}$, where $\mathbf{x}_i^t$ is the i-th target image and n_t is the total number of images. The goal is to use labeled source images and unlabeled target images to learn a classifier that generalizes well on the target dataset.

This paper utilizes the deep convolution neural network to learn the classifier. As shown in Fig. 2, our network is an end-to-end trained classification net. It mainly consists of two parts, feature extractor and classifier (the last fully-connected layer). SGC is imposed on the embeddings, the output of feature extractor, to learn an aligned embedding space. Meanwhile, the classifier takes the embeddings as input. Since the learned embeddings are intra-class compactness and inter-class separability, the classifier can work effectively on the target images.

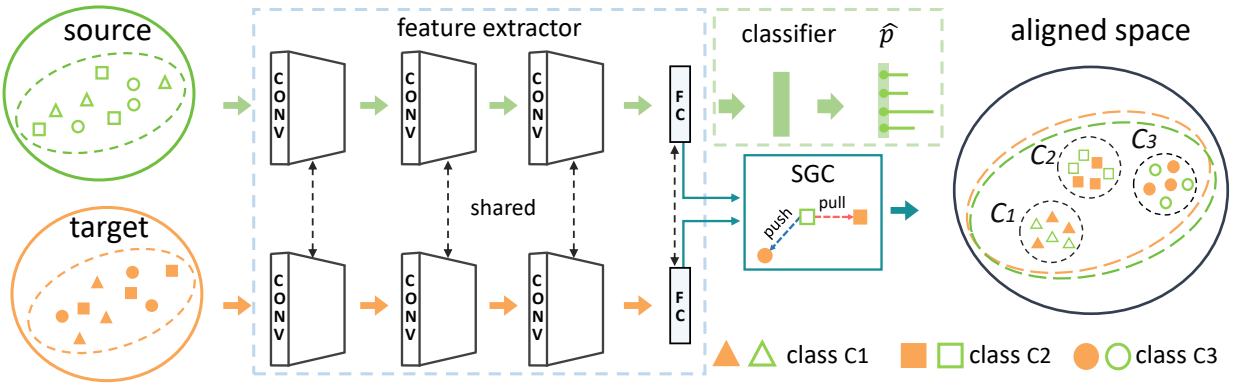


Fig. 2. Framework of the similarity guided constraint (SGC) method. With the supervision of SGC, our network has the ability to align the distributions at class level. Thus, images from different domains but have the same class label are expected to be aligned nearby, and vice versa. Since the target dataset is unlabeled, we assign pseudo labels to the target images. In this figure, different colors denote different domain distributions and different shapes represent different classes. Best viewed in color.

Moreover, SGC can be further combined with the existing deep domain-level alignment constraints. With the help of two types of constraints, our model has the ability to align data distributions at domain level and class level.

In the rest of this section, we introduce the similarity-constrained scheme in Section III-B1. In Section III-B2, we describe the similarity constrained alignment network with the joint supervision of triplet loss and domain-level alignment loss. In Section III-C, we have a discussion about our method.

B. Similarity Guided Constraint

1) Similarity-constrained Scheme: SGC takes a closer look at the class level to address the distribution discrepancy. Specifically, SGC requires the images regardless of their domains to follow two types of semantic relations:

- *Semantic similarity.* Images from the same class are semantically similar, thereby should be mapped nearby in the embedding space.
- *Semantic dissimilarity.* Images from different classes are semantically dissimilar, thereby should be mapped far apart in the embedding space.

Similarity-Guided loss function. SGC explicitly defines the inter-class and intra-class relations among source and target images. To preserve the two semantic relations, we naturally implement the constraint by the triplet loss [41].

Let $D_{i,j} = \|g_\theta(x_i) - g_\theta(x_j)\|_2^2$ measures the distance between two images in the feature space, where $g_\theta(\cdot)$ is the feature extractor. If x_i and x_j are with the same label, we want $D_{i,j}$ to be small, corresponding to the semantic similarity. If x_i and x_j are with different labels, we want $D_{i,j}$ to be large, corresponding to the semantic dissimilarity.

For an image triplet consisting of two semantically similar images x_a and x_p, and a semantically dissimilar image x_n, the triplet loss minimizes:

$$\mathcal{L}_s(\theta) = \sum_{\substack{a,p,n \\ y_a=y_p\neq y_n}} [m + D_{a,p} - D_{a,n}]_+ , \qquad (1)$$

where x_a and x_p is a positive pair (their labels y_a and y_p are same), x_a and x_n is a negative pair (their labels y_a and y_n are different). m is the margin that is enforced between positive and negative pairs.

This loss encourages the distance between x_a and positive image x_p to be smaller than the distance between x_a and negative x_n by the margin m. By enforcing the loss on every image triplets, we can meet the requirement of SGC, *i.e.,* intra-class compactness and inter-class separability.

Label estimation for target images. When optimizing the network with the triplet loss, we are faced with one challenging issue: constructing images triplets requires labels of target images, but the target dataset is totally unlabeled. To address this issue, we propose to assign the pseudo labels to the target images: given the current network, we use the classifier to predict labels of target images.

To ensure the accuracy of the pseudo label, we adopt three tactics. (a) *Threshold* T. Intuitively, the image with the high predicted score is more likely to be classified correctly. Thus, the threshold T directly controls the quality of pseudo label. In practice, we only select target images with predicted scores above a high threshold T for building the semantic relations. Note that we set the threshold T a constant during training. (b) *Progressive selection.* During training, the classifier will gradually improve its accuracy on the target dataset, so we re-assign pseudo labels every several iterations. In this way, more and more target images will be *progressively* selected for training. (c) *domain-level alignment.* By reducing the distribution discrepancy, domain-level alignment can improve the accuracy of classifier on the target dataset. Thus, it can naturally improve the quality of the pseudo label.

During experimentation, we observe that triplet loss has the tolerance to incorrect labels, such that we only use the above three tactics. Many other sample prediction methods might also be helpful, such as consistency-based semi-supervised learning [26], co-training [33], and model fusion [49]. Moreover, adaptive threshold method proposed in iCAN [17] could also be useful. The effect of the aforementioned methods on our system can be validated in the future.

2) Collaborative Distribution Alignment: The similarity guided constraint can be integrated into existing domain-level alignment network. With the collaborative supervision of triplet loss and domain-level alignment loss, our full system

can align data distributions at both domain level and class level. Here, we name our full system as similarity-constrained alignment (SCA) network.

Domain-level alignment. Following the practice in [3], we adopt the JMMD metric for the domain-level alignment. The JMMD formally reduces the discrepancy in the joint distributions of the activations in domain-specific layers $\mathcal{L}$, i.e. $P(\mathbf{Z}^{s1}, \ldots, \mathbf{Z}^{s|\mathcal{L}|})$ and $Q(\mathbf{Z}^{t1}, \ldots, \mathbf{Z}^{t|\mathcal{L}|})$. Thus, the loss function of domain-level alignment is written as,

$$\begin{aligned}\mathcal{L}_d = &\frac{2}{n}\sum_{i=1}^{n/2}\left(\prod_{\ell\in\mathcal{L}} k^{\ell}(\mathbf{z}_{2i-1}^{s\ell}, \mathbf{z}_{2i}^{s\ell}) + \prod_{\ell\in\mathcal{L}} k^{\ell}(\mathbf{z}_{2i-1}^{t\ell}, \mathbf{z}_{2i}^{t\ell})\right) \\ &- \frac{2}{n}\sum_{i=1}^{n/2}\left(\prod_{\ell\in\mathcal{L}} k^{\ell}(\mathbf{z}_{2i-1}^{s\ell}, \mathbf{z}_{2i}^{t\ell}) + \prod_{\ell\in\mathcal{L}} k^{\ell}(\mathbf{z}_{2i-1}^{t\ell}, \mathbf{z}_{2i}^{s\ell})\right),\end{aligned} \tag{2}$$

where $n = n_s$, $\mathbf{z}^{t\ell}$ denotes the activations of the target image in the layer ℓ, and $\mathbf{z}^{s\ell}$ denotes the activations of the source image in the layer ℓ. k^{ℓ} is the kernel function in a reproducing kernel Hilbert space (RKHS).

Similarity constrained alignment objective. In our network, we adopts the cross-entropy loss as K-way classification loss function, this is corresponding to,

$$\mathcal{L}_c = \frac{1}{n_s}\sum_{i=1}^{n_s} L\left(f\left(g_\theta(\mathbf{x}_i^s)\right), \mathbf{y}_i^s\right), \tag{3}$$

where $L(\cdot,\cdot)$ is the cross-entropy loss, $g_\theta(\cdot)$ is the feature extractor, and $f(\cdot)$ is the classifier followed by a softmax over the K classes.

Finally, the final objective of the collaborative distribution alignment is written as,

$$\mathcal{L}_{sca} = \mathcal{L}_c + \alpha\mathcal{L}_d + \beta\mathcal{L}_s, \tag{4}$$

where $\mathcal{L}_c$ is the classification loss, $\mathcal{L}_d$ is the the domain confusion loss, and $\mathcal{L}_s$ is the triplet loss. The α and the β control the relative importance of domain-level alignment and similarity guided constraint, respectively.

C. Discussion

Why prefer the triplet loss function? As described in Section I, both the cross-entropy loss and the triplet loss have potential to enforce cross-domain similarity constraint. Since the triplet loss has **higher resistance** against incorrect pseudo-labels, we adopt it for implementing the proposed SGC. We illustrate this point in Fig. 3, with focus on a wrongly pseudo-labeled target image x in the embedding space.

In the case of Fig. 3 (a), the target image x is with wrong pseudo-label C_2 (its ground-truth is C_1). Under the classification mode, the cross-entropy loss makes the weight vector W_2 (corresponding to C_2) of classifier towards the wrongly pseudo-labeled x. This reduces the margin between W_1 and W_2, and thus diminishes the discriminative ability of classifier (especially when distinguishing between C_1 and C_2 images). In contrast, the triplet loss constructs an image triplet consisting of wrongly pseudo-labeled x and two source images (a C_1 image and a C_2 image). For the image triplet, the triplet loss aims to pull x_1 closer to the C_2 image and to push x far away from the C_1 image. This is already satisfied

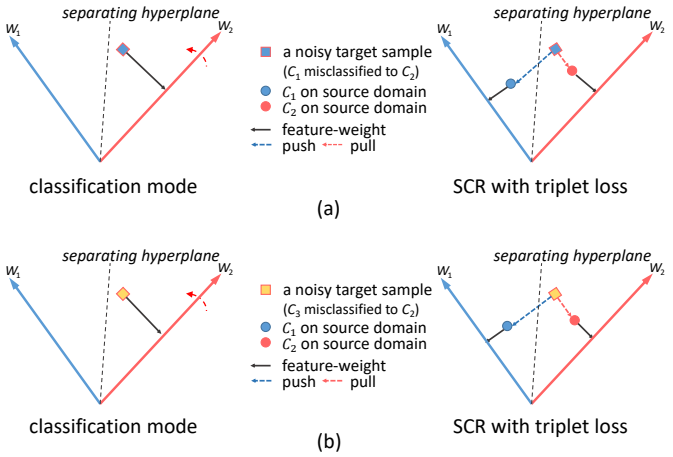


Fig. 3. SGC with triplet loss gains higher robustness against incorrect pseudo-labels. (a) A target image of class C_1 is wrongly assigned with a pseudo label C_2. Under classification mode, the wrongly pseudo-labeled image pulls the weight vector W_2 closer toward itself. This reduces the margin between W_1 and W_2, and leads to dramatic accuracy drop. In contrast, the triplet loss function requires the wrongly pseudo-labeled image to be closer to a C_2 source image than a C_1 source image. This requirement is roughly satisfied, which corresponds to *zero* or *slight* triplet loss. (b) More generally, wrongly pseudo-labeled image belongs to a third class in the triplet. Similarly, the requirement of triplet loss is also satisfied in this case. Thus, the wrongly pseudo-labeled image has little negative impact on the classifier.

in the case of Fig. 3 (a), so the wrongly pseudo-labeled x has almost no negative impact on the model.

Similarly, in Fig. 3 (b), the target image x is with wrong pseudo-label C_2 (its ground-truth is C_3). It is a new class image *w.r.t.* weight vectors W_1 and W_2. Under classification mode, x also deteriorates the discriminative ability of classifier, especially when classifying C_2 and C_3 images. In contrast, when the triplet loss function is imposed on an image triplet consists of x and two source images (a C_1 image and a C_2 image), its requirement is also satisfied. Thus, the resulting triplet loss turns to be negligible.

In short, the cross-entropy loss function extensively exposes the model to incorrect pseudo-labels, while the triplet loss function enables the model to have relatively high resistance against them. Thus, the triplet loss is more suitable for implementing the proposed SGC.

Collaborative alignment. Both domain-level alignment and SGC are significant during the distribution alignment. First, domain-level alignment makes source and target distributions globally overlapped, and thus reduces the difficulty of class-level alignment. Moreover, by reducing the global distribution discrepancy, domain-level alignment improves the accuracy of classifier on the target dataset. This ensures the quality of pseudo-labeled target images, and thus benefits SGC. Second, SGC locally aligns distributions, and leads to more suitable domain-level alignment.

During training, the two supervisions work collaboratively to align distributions at both domain level and class level. Moreover, the learned domain-invariant feature is characterized by intra-class compactness and inter-class separability. Based on this, the classifier learned on source images works effectively on the target dataset.

Progressive label estimation. To ensure the quality of pseudo labels, we only select target images with their scores above a high threshold T. At the early stage of training, the classifier is relatively poor on the target dataset, so only a few target images have predicted scores above the threshold T. With the help of the SGC, the classifier will gradually improve itself during training. Thus, we re-assign labels for the target images every several iterations (K). This strategy can be regarded as easy-to-hard sample selection. By doing so, more target images will meet the condition of threshold T, and be progressively selected for constructing image triplets. Thus, we keep the threshold T fixed during training.

IV. Experimental Evaluation

A. Datasets

We evaluate the proposed unsupervised domain adaptation method on four datasets: Office-31 [52], ImageCLEF-DA, Office-Home [53], and MNIST-USPS-SVHN.

Office-31 contains 4,652 images and 31 categories collected from three distinct domains: *Amazon* (**A**), *Webcam* (**W**) and *DSLR* (**D**). The images in DSLR are captured with a digital SLR camera and have high resolution. Amazon consists of images downloaded from online merchants (www.amazon.com). These images are of products at medium resolution. The images in Webcam are collected by a web camera, which are of low resolution. We evaluate the proposed method across six transfer tasks **A** $\rightarrow$ **W**, **D** $\rightarrow$ **W**, **W** $\rightarrow$ **D**, **A** $\rightarrow$ **D**, **D** $\rightarrow$ **A** and **W** $\rightarrow$ **A**.

ImageCLEF-DA is a benchmark dataset for ImageCLEF 2014 domain adaptation challenge. It contains three subsets, including *Caltech-256* (**C**), *ImageNet ILSVRC 2012* (**I**), and *Pascal VOC 2012* (**P**), and each subset is considered as a domain. There are 12 categories and each categories contains 50 images. We use all domain combinations and build 6 transfer tasks: **I** $\rightarrow$ **P**, **P** $\rightarrow$ **I**, **I** $\rightarrow$ **C**, **C** $\rightarrow$ **I**, **C** $\rightarrow$ **P**, and **P** $\rightarrow$ **C**. We report the results following the protocol in [3].

Office-Home is a large-scale benchmark for testing domain adaptation methods. There are 15,500 images in Office-Home. It contains four distinct domains, each corresponding to 65 different categories. The domains are: *Artistic images*(**Ar**), *Clip Art* (**Cl**), *Product images* (**Pr**) and *Real-World images* (**Rw**). We report the results following the protocol in [53].

MNIST-USPS-SVHN is a challenging domain adaptation task of three digits datasets: MNIST [54], USPS and SVNH [55]. MNIST-USPS-SVHN makes a good complement to previous datasets for diverse domain adaptation scenarios. We conduct experiments on three tasks: SVHN $\rightarrow$ MNIST, MNIST $\rightarrow$ USPS, and USPS $\rightarrow$ MNIST. We report the results following the protocol in [56].

B. Implementation Details

We implement our method on pytorch framework. For Office-31, ImageCLEF-DA, and Office-Home, we fine-tune from ResNet-50 model [50]. During training, all the images are resized to 256×256; for digits datasets, we adopt LeNet [54] as network structure. During training, all the images are resized to 32×32.

We adopt random flipping and random cropping as data augmentation methods. We use SGD for optimization, and adopt the same INV learning rate strategy as in RevGrad [4]. The learning rate decreases gradually after each iteration from 0.001, the momentum is set to 0.9, and the weight decay is set to 0.0004. We set the parameters $\alpha = 1$ and $\beta = 1$ in Eq. 4. For triplet loss, we follow the sampling *batch-hard* strategy in [57]. Specifically, we randomly select C classes and randomly select K images of each class from source images. Similarly, we select $C \cdot K$ pseudo-labeled target images. Thus, we get a mini-batch of $2C \cdot K$ training images. For each image, we sample the hard negative and hard positive within a mini–batch, yielding $2C \cdot K$ triplets. In the experiment, we set K=7 and C=4. In addition, we randomly select 32 source images and 32 target images for domain-level alignment loss and cross-entropy loss.

We adopt a two-stage training procedure: we first train the classifier and the feature extractor by minimizing Eq. 2 and Eq. 3, and then train the entire system (SCA) by minimizing Eq. 4. For the first stage, we train the network for 6000 iterations. For stage two, we train for anther 30000 iterations. We set threshold $T = 0.9$, and assign pseudo labels for target images every 2000 iterations.

C. Comparison with State-of-the-art Methods

Compared methods. For Office-31, ImageCLEF-DA, and Office-Home, we compare the proposed method with several state-of-the-art methods, including DAN [7], RTN [51], JAN [3], RevGrad [4], MADA [10], SimNet [11], and iCAN [17]. These methods are all based on the deep neural network (ResNet-50 [50]) to learn domain-invariant embeddings. For MNIST-USPS-SVHN, we compare the proposed method with RevGrad [4], ADDA [56], MSTN [9], Cycada [20], PixelDA [21], and M-ADDA [24].

For the fair comparison, the results of these methods are directly reported from their original papers.

Comparison on the Office-31 dataset. We compare the proposed method with the recent state-of-the-art methods in Table I. The baseline is the network that we modify from ResNet-50, and it does not adopt any domain adaptation technique. In this paper, we adopt JMMD [3] for the domain-level alignment, and the result of "Basel. + D" is on par with the experiment in [3].

Compared with "Basel.", "Basel. + D" achieves higher performance, which indicates that it has ability to reduce the distribution discrepancy. Moreover, only adopting the SGC can also improve the baseline performance: it gains $+4.5\%$ improvement over the baseline in average accuracy. This indicates that the proposed constraint has ability to alleviate the distribution discrepancy problem. The working mechanism of SGC is that it ensures the cross-domain similarities are preserved, and thus aligns source and target distributions at class level.

Moreover, our full system (SCA) achieves 87.7% in average accuracy. This is the best performance on the Office-31 dataset. SCA achieves the highest performance on three tasks (W $\rightarrow$ A, W $\rightarrow$ D, and W $\rightarrow$ A). SCA is higher than MADA [10] (87.7%

TABLE I
COMPARISON OF VARIOUS METHODS FOR UNSUPERVISED DOMAIN ADAPTATION ON THE OFFICE-31 DATASET IN TERMS OF TEST ACCURACY (%). THE BEST RESULTS ARE IN **BOLD**. "BASEL." DENOTES THE BASELINE TRAINED ONLY THE SOURCE DATASET, " S" REPRESENTS THE SIMILARITY GUIDED CONSTRAINT, AND "D" DENOTES THE DOMAIN-LEVEL ALIGNMENT. SCA IS THE FULL SYSTEM ("BASEL. + D + S").

Method	A → W	A → D	W → A	W → D	D → A	D → W	Avg.
ResNet-50 [50]	68.2 ± 0.2	68.9 ± 0.2	60.7 ± 0.3	99.3 ± 0.1	62.5 ± 0.3	96.7 ± 0.1	76.7
DAN [7]	80.5 ± 0.4	78.6 ± 0.2	62.8 ± 0.2	99.6 ± 0.1	63.6 ± 0.3	97.1 ± 0.2	80.4
RTN [51]	84.5 ± 0.2	77.5 ± 0.3	64.8 ± 0.3	99.4 ± 0.1	66.2 ± 0.2	96.8 ± 0.1	81.6
JAN [3]	85.4 ± 0.3	84.7 ± 0.3	70.0 ± 0.4	99.8 ± 0.2	68.6 ± 0.3	97.4 ± 0.2	84.3
RevGrad [4]	82.0 ± 0.4	79.7 ± 0.4	67.4 ± 0.5	99.1 ± 0.1	68.2 ± 0.4	96.9 ± 0.2	82.2
MADA [10]	90.0 ± 0.2	87.8 ± 0.2	66.4 ± 0.3	99.6 ± 0.1	70.3 ± 0.3	97.4 ± 0.1	85.2
SimNet [11]	88.6 ± 0.5	85.3 ± 0.3	71.8 ± 0.6	99.7 ± 0.2	**73.4** ± 0.8	98.2 ± 0.2	86.2
iCAN [17]	92.5	**90.1**	69.9	**100.0**	72.1	**98.8**	87.2
Basel.	76.5 ± 0.3	78.0 ± 0.2	64.0 ± 0.3	99.0 ± 0.1	65.0 ± 0.2	94.8 ± 0.1	79.6
Basel. + D	87.2 ± 0.3	84.9 ± 0.2	69.8 ± 0.3	99.2 ± 0.1	67.8 ± 0.3	96.5 ± 0.1	84.2
Basel. + S	85.0 ± 0.2	87.0 ± 0.2	67.2 ± 0.3	99.4 ± 0.1	67.5 ± 0.4	98.2 ± 0.1	84.1
SCA	**93.6** ± 0.1	89.5 ± 0.1	**72.4** ± 0.3	**100.0** ± .0	72.6 ± 0.3	98.0± 0.2	**87.7**

TABLE II
COMPARISON OF VARIOUS METHODS FOR UNSUPERVISED DOMAIN ADAPTATION ON THE IMAGECLEF-DA DATASET IN TERMS OF TEST ACCURACY (%). THE BEST RESULTS ARE IN **BOLD**."BASEL." DENOTES THE BASELINE TRAINED ONLY THE SOURCE DATASET, " S" REPRESENTS THE SIMILARITY GUIDED CONSTRAINT, AND "D" DENOTES THE DOMAIN-LEVEL ALIGNMENT. SCA IS THE FULL SYSTEM ("BASEL. + D + S").

Method	I → P	P → I	I → C	C → I	C → P	P → C	Avg.
ResNet-50 [50]	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 [7]	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
RTN [51]	75.6 ± 0.3	86.8 ± 0.1	95.3 ± 0.1	86.9 ± 0.3	72.7 ± 0.3	92.2 ± 0.4	84.9
JAN [3]	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 [10]	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
iCAN [17]	**79.5**	**89.7**	94.7	89.9	**78.5**	92.0	87.4
Basel.	74.3 ± 0.3	83.6 ± 0.2	91.8 ± 0.2	76.0 ± 0.3	64.0 ± 0.2	87.8 ± 0.2	79.6
Basel. + D	75.6 ± 0.2	86.8 ± 0.3	95.8 ± 0.2	86.2 ± 0.2	74.6 ± 0.2	91.6 ± 0.3	85.1
Basel. + S	75.3 ± 0.3	86.2 ± 0.2	94.0 ± 0.2	82.2 ± 0.3	67.8 ± 0.3	91.3 ± 0.2	82.8
SCA	78.1 ± 0.3	89.2 ± 0.1	**96.8**± 0.2	**91.3** ± 0.2	78.2 ± 0.2	**94.0** ± 0.3	**87.9**

TABLE III
COMPARISON OF VARIOUS METHODS FOR UNSUPERVISED DOMAIN ADAPTATION ON THE OFFICE-HOME DATASET IN TERMS OF TEST ACCURACY (%). THE BEST RESULTS ARE IN **BOLD**.

Method	Source Target	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 [50]		34.9	50.0	58.0	37.4	41.9	46.2	38.5	31.2	60.4	53.9	41.2	59.9	46.1
DAN [7]		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
RevGrad [4]		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 [3]		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
Basel.		35.5	57.3	65.0	41.8	50.4	52.4	41.8	30.6	64.3	59.1	38.2	64.2	50.1
Basel. + D		44.5	61.4	70.0	50.1	62.3	64.5	50.8	45.5	72.0	63.0	54.0	77.7	59.7
Basel. + S		43.5	62.0	71.2	52.3	60.0	59.7	47.0	37.2	71.4	65.5	46.2	77.0	57.8
SCA		**46.7**	**64.6**	**71.3**	**53.1**	**65.3**	**65.2**	**54.6**	**47.2**	**72.7**	**68.2**	**56.0**	**80.2**	**62.1**

vs. 85.2%). Moreover, our method outperforms SimNet and JAN by 1.5% and 3.4%, respectively.

Comparison on the ImageCLEF-DA dataset. The experimental results in Table IV show that our SGC improves the baseline accuracy (82.8% vs. 79.6 %). This indicates that SGC is effective for reducing distribution discrepancy. Moreover, our full system (SCA) also achieves competitive average accuracy in ImageCLEF-DA. The accuracy of SCA is 0.5% higher than the second best method iCAN [17]. This indicates that using pseudo-labeled images to align features instead of retraining classifier is beneficial domain adaptation. Moreover, SCA respectively outperforms the MADA [10] and JAN [3] by 2.1% and 2.1%. Specifically, our methods achieve the highest performance on three tasks (I→ C, C → I and P → C). The results further validate the benefit of enforcing SGC on the domain-level alignment method.

Comparison on the Office-Home dataset. In Table III, we also compare the proposed method with state-of-the-art methods, including DAN [7], RevGrad [4], and JAN [3]. As shown in Table III, SGC gains $+7.7\%$ improvements over the baseline in average accuracy. It indicates that SGC has ability to reduce the distribution gap. Moreover, with the collaborative supervision of SGC and the domain-level alignment, our full system (SCA) achieves competitive average accuracy (62.1%)

TABLE IV
COMPARISON OF VARIOUS METHODS FOR UNSUPERVISED DOMAIN ADAPTATION ON THE DIGITS DATASETS IN TERMS OF TEST ACCURACY (%). THE BEST RESULTS ARE IN **BOLD**. WE EVALUATE OUR METHOD USING THE SETUP IN [56].

Method	SVHN→MNIST	MNIST→USPS	USPS→MNIST
Source Only	60.1 ± 1.1	75.2 ± 1.6	57.1 ± 1.7
RevGrad [4]	73.9	77.1 ± 1.8	73.0 ± 2.0
ADDA [56]	76.0 ± 1.8	89.4 ± 0.2	90.1 ± 0.8
MSTN [9]	91.7 ± 1.5	92.9 ± 1.1	-
M-ADDA [24]	-	95.2	94.0
Cycada [20]	90.4 ± 0.4	95.6 ± 0.2	**96.5** ± 0.1
PixelDA [21]	-	95.9	-
SCA	**92.0** ± 1.6	**96.1** ± 1.4	95.5 ± 1.17

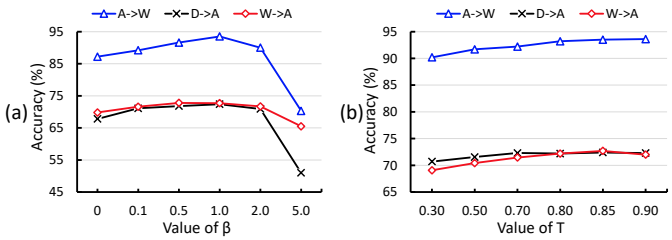


Fig. 4. The impact of the parameters of SGC on the accuracy. (a): sensitivity to parameter β (weight of the similarity guided constraint) in Eq. 4. A larger β means that the constraint has a greater impact on the distribution alignment. (b): sensitivity to parameter T (threshold for selecting target images). A larger T denotes that target images should have larger confidence score, and their pseudo labels are more likely to be correct. Thus, the T shows the sensitivity of triplet loss to label estimation error.

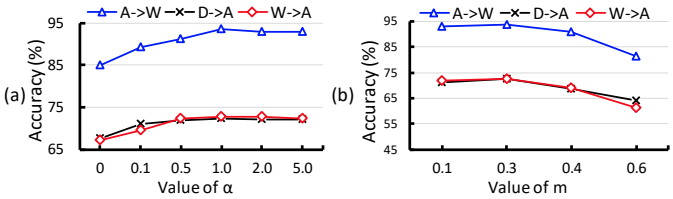


Fig. 5. The impact of the parameters of SGC on the accuracy. (a): sensitivity to parameter α (weight of the domain-level alignment) in Eq. 4. (b): sensitivity to parameter m in Eq. 1 (margin that is enforced between positive and negative pairs).

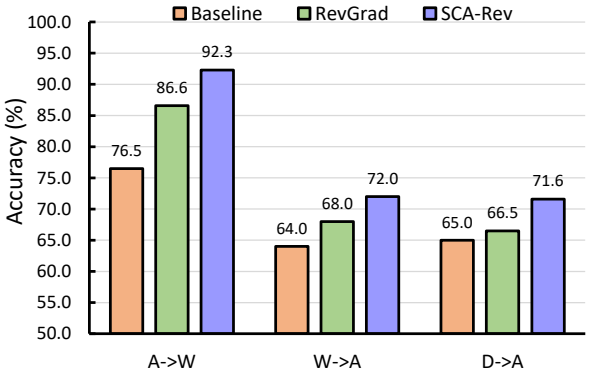


Fig. 6. Performance of various methods on three tasks (**A** → **W**, **W** → **A**, and **D** → **A**) of Office-31. Reverse Gradient (RevGrad) [4] is a domain-level alignment method based on adversarial learning. SCA-Rev denotes the similarity guided constraint is enforced on RevGrad.

on Office-Home.

Comparison on the SVHN-MNIST-USPS. In Table IV, we also report the results of SCA and compare with other state-of-the-art methods. On three digits datasets, we adopt RevGrad [4] as domain-level alignment method. SCA achieves 96.1% in MNIST → USPS, which outperforms M-ADDA by 0.9%. This indicates that enforcing metric-based constraint on both source and target images is beneficial for domain alignment. Moreover, SCA also surpasses MSTN by 3.2% and 0.3% in MNIST → USPS and SVHN → MNIST, respectively. SCA achieves best results on tasks (SVHN→MNIST and MNIST→USPS). The improvement show that our SGC can effectively achieve semantic alignment.

D. Component Analysis

In this section, we present step-by-step evaluation to analyze the effectiveness of the similarity guided constraint. The experiment is based on our full system (SCA).

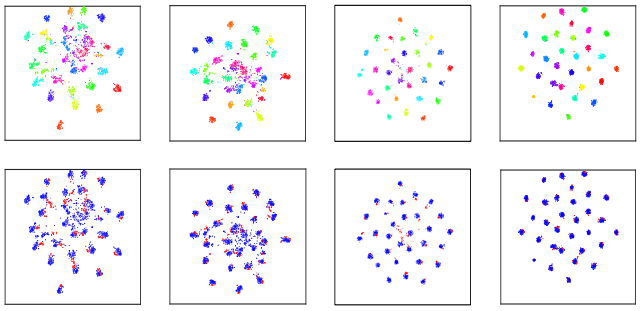
Weight of similarity guided constraint. In Fig. 4 (a), we demonstrate the transfer accuracy of SCA by varying the $\beta \in \{0, 0.1, 0.5, 1, 2, 5\}$ on three tasks, **A** → **W**, **W** → **A**, and **D** → **A**, where β in Eq. 4 controls the relative importance of the SGC. As shown in Fig. 4 (a), when β increases from 0 to 1, the performance on three tasks grow and reach the best at $\beta = 1$. However, when β is too large (β=5), the accuracy will drop by a large margin. Empirically, the best parameter β is between 0.5 to 2 in SCA.

Impact of threshold T. Threshold T intuitively shows the sensitivity of triplet loss to label estimation error. Specifically, T controls the quality of the pseudo label: If T is small, some selected target images are with low predicted score; and these images are more likely to be classified incorrectly. When T becomes larger, the number of incorrect pseudo labels is smaller. To validate the impact of T, we conduct the experiment on the full system (SCA). We vary the $T \in \{0.3, 0.5, 0.7, 0.8, 0.85, 0.9, 0.95\}$ on tasks **A** → **W**, **W** → **A**, **D**→ **A**, and report results in Fig. 4 (b). When T grows from 0.3 to 0.8, SCA accuracy gradually increases. When T is larger ($T \in [0.85, 0.95]$), accuracy is stable. This indicates that the triplet loss has ability to tolerate incorrect pseudo labels.

Weight of domain-level alignment. We also show the impact of α in Eq. 4 in Fig. 5 (a). α controls the relative importance of domain-level alignment. We observe that accuracy is stable when $\alpha \in [0, 5]$, thus we empirically set $\alpha = 1$.

Impact of margin m. We demonstrate the accuracy of SCA by varying the $m \in 0.1, 0.3, 0.4, 0.6$ in Fig. 5 (b), where m is the margin of triplet loss (Eq. 1). When m increases from 0.1 to 0.3, the accuracy on three tasks grow and reach the best at $m = 0.3$. However, when m is too large (m=0.6), the accuracy will drop by a large margin. This is due to larger margins increase the difficulty of the feature learning.

Domain-level alignment method. In this paper, we adopt JMMD [3] for the domain-level alignment. We note that the proposed SGC can work collaboratively with other domain-level alignment methods. To validate this, we conduct the experiment on three tasks of Office-31: **A** → **W**, **W** → **A**,



(a) Basel. (b) Basel.+D (c) Basel.+S (d) SCA

Fig. 7. Visualization of cross-domain embeddings for task $\mathbf{A} \rightarrow \mathbf{W}$ on Office-31 [52]. From left to right: (a) baseline (source images only), (b) domain-level alignment based on JMMD [3], (c) class-level alignment with SGC, and (d) the full system (SCA). In the first row, different colors denote different object categories. In the second row, red points represent the samples in $\mathbf{W}$, and blue represents samples in $\mathbf{A}$. We clearly observe that SGC allows the two domains to be well aligned on the class level, and eventually leads to more suitable domain-level alignment.

and $\mathbf{D} \rightarrow \mathbf{A}$, and report the results in Fig. 6. We adopt an adversarial adaptation method named Reverse Gradient (RevGrad) [4] as domain-level alignment method. Based on the RevGrad, we construct the similarity constrained alignment network (SCA-Rev).

As shown in Fig. 6, SCA-Rev gains +5.7%, +4.0% and 5.1% improvements over RevGrad on $\mathbf{A} \rightarrow \mathbf{W}$, $\mathbf{W} \rightarrow \mathbf{A}$, and $\mathbf{D} \rightarrow \mathbf{A}$, respectively. The results indicate that our SGC can work collaboratively with other domain-level alignment methods. In addition, the results also demonstrate that aligning distributions at domain level and class level is significant.

Feature visualization. We provide visualization over task $\mathbf{A} \rightarrow \mathbf{W}$ by T-SNE [58]. Compared with baseline (Fig. 7(a)), domain-level alignment method JMMD (Fig. 7(b)) globally aligns the distributions. However, there is semantic misalignment problem, *i.e.*, embeddings of different classes are mixed up. Moreover, we observe that SGC can align the distributions at class-level (Fig. 7(c)), but it fails to map some embeddings well. We think that this is caused by some wrongly pseudo-labeled images. Benefiting from both SGC and domain-level alignment, our full system (SCA) produces more discriminative embeddings (shown in Fig. 7(d)). It preserves cross-domain similarities while aligning the distributions, thus source and target embeddings of the same class are co-located.

V. Conclusion

This paper focuses on class-level distribution alignment from the metric learning perspective. In our attempt, we present a similarity guided constraint (SGC) method. SGC enforces intra-class compactness and inter-class separability among source and target features. This is consistent with the intrinsic requirement of classification, and thus improves the accuracy on the target dataset. Moreover, SGC is able to work collaboratively with the existing domain-level alignment constraint. With the joint supervision of two constraints, our model can align distributions at both domain and class level, resulting in more discriminative embeddings. The experimental results on four benchmarks validate the effectiveness of SGC for class-level distribution alignment.

Acknowledgment

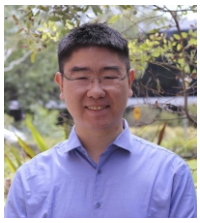
This research was conducted by the Australian Research Council Centre of Excellence for Robotic Vision (project number CE140100016). Dr Liang Zheng is the recipient of an Australian Research Council Discovery Early Career Award (project number DE200101283) funded by the Australian Government. The authors are very grateful for the support by NSFC under grant 61771447. We thank Xiaoxiao Sun and Joshua Marsh for their valuable comments.

References

[1] A. Torralba, A. A. Efros *et al.*, "Unbiased look at dataset bias." in *Proc. CVPR*, 2011, pp. 1521–1528.
[2] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. S. Lempitsky, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, pp. 59:1–59:35, 2016.
[3] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in *Proc. ICML*, 2017, pp. 2208–2217.
[4] Y. Ganin and V. S. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. ICML*, 2015, pp. 1180–1189.
[5] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. CVPR*, 2017, pp. 2962–2971.
[6] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," [Online]. Available: https://arxiv.org/abs/1412.3474.
[7] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," in *Proc. ICML*, 2015, pp. 97–105.
[8] B. Sun and K. Saenko, "Deep CORAL: correlation alignment for deep domain adaptation," in *Proc. ECCVW*, 2016, pp. 443–450.
[9] S. Xie, Z. Zheng, L. Chen, and C. Chen, "Learning semantic representations for unsupervised domain adaptation," in *Proc. ICML*, 2018, pp. 5419–5428.
[10] Z. Pei, Z. Cao, M. Long, and J. Wang, "Multi-adversarial domain adaptation," in *Proc. AAAI*, 2018, pp. 3934–3941.
[11] P. O. Pinheiro, "Unsupervised domain adaptation with similarity learning," in *Proc. CVPR*, 2018, pp. 8004–8013.
[12] B. Kulis, K. Saenko, and T. Darrell, "What you saw is not what you get: Domain adaptation using asymmetric kernel transforms," in *Proc. CVPR*, 2011, pp. 1785–1792.
[13] R. Gopalan, R. Li, and R. Chellappa, "Domain adaptation for object recognition: An unsupervised approach," in *Proc. ICCV*, 2011, pp. 999–1006.
[14] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, "Unsupervised visual domain adaptation using subspace alignment," in *Proc. ICCV*, 2013, pp. 2960–2967.
[15] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *Proc. AAAI*, 2016, pp. 2058–2065.
[16] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. J. Smola, "A kernel method for the two-sample-problem," in *Proc. NIPS*, 2006, pp. 513–520.
[17] W. Zhang, W. Ouyang, W. Li, and D. Xu, "Collaborative and adversarial network for unsupervised domain adaptation," in *Proc. CVPR*, 2018, pp. 3801–3809.
[18] Z. Cao, L. Ma, M. Long, and J. Wang, "Partial adversarial domain adaptation," in *Proc. ECCV*, 2018, pp. 139–155.
[19] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial networks," in *Proc. NIPS*, 2014, pp. 2672–2680.
[20] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," in *Proc. ICML*, 2018, pp. 1994–2003.
[21] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *Proc. CVPR*, 2017, pp. 95–104.
[22] M. Liu and O. Tuzel, "Coupled generative adversarial networks," in *Proc. NIPS*, 2016, pp. 469–477.

[23] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *Proc. CVPR*, 2018, pp. 994–1003.
[24] I. H. Laradji and R. Babanezhad, "M-ADDA: unsupervised domain adaptation with deep metric learning," in *Proc. ICML Workshop of Domain Adaptation for Visual Understanding (DAVU)*, 2018.
[25] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto, "Unified deep supervised domain adaptation and generalization," in *Proc. ICCV*, 2017, pp. 5716–5726.
[26] G. French, M. Mackiewicz, and M. Fisher, "Self-ensembling for visual domain adaptation," in *Proc. ICLR*, 2017.
[27] L.-J. Li and L. Fei-Fei, "Optimol: automatic online picture collection via incremental model learning," *International journal of computer vision*, vol. 88, no. 2, pp. 147–168, 2010.
[28] X. Chen, A. Shrivastava, and A. Gupta, "NEIL: extracting visual knowledge from web data," in *Proc. ICCV*, 2013, pp. 1409–1416.
[29] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," [Online]. Available: https://arxiv.org/abs/1610.02242.
[30] I. Radosavovic, P. Dollár, R. B. Girshick, G. Gkioxari, and K. He, "Data distillation: Towards omni-supervised learning," in *Proc. CVPR*, 2018, pp. 4119–4128.
[31] G. Kang, L. Zheng, Y. Yan, and Y. Yang, "Deep adversarial attention alignment for unsupervised domain adaptation: the benefit of target expectation maximization," in *Proc. ECCV*, 2018, pp. 420–436.
[32] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Bian, and Y. Yang, "Progressive learning for person re-identification with one example," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2872–2881, 2019.
[33] X. Dong, L. Zheng, F. Ma, Y. Yang, and D. Meng, "Few-example object detection with model communication," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1641–1654, 2018.
[34] H. Fan, L. Zheng, C. Yan, and Y. Yang, "Unsupervised person re-identification: Clustering and fine-tuning," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 14, no. 4, p. 83, 2018.
[35] M. Chen, K. Q. Weinberger, and J. Blitzer, "Co-training for domain adaptation," in *Proc. NIPS*, 2011, pp. 2456–2464.
[36] K. Saito, Y. Ushiku, and T. Harada, "Asymmetric tri-training for unsupervised domain adaptation," in *Proc. ICML*, 2017, pp. 2988–2997.
[37] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. CVPR*, 2005, pp. 539–546.
[38] J. Goldberger, S. T. Roweis, G. E. Hinton, and R. Salakhutdinov, "Neighbourhood components analysis," in *Proc. NIPS*, 2004, pp. 513–520.
[39] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *Journal of Machine Learning Research*, vol. 10, pp. 207–244, 2009.
[40] H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *Proc. CVPR*, 2016, pp. 4004–4012.
[41] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. CVPR*, 2015, pp. 815–823.
[42] J. Hu, J. Lu, and Y. Tan, "Discriminative deep metric learning for face verification in the wild," in *CVPR*, 2014, pp. 1875–1882.
[43] X. Ben, P. Zhang, Z. Lai, R. Yan, X. Zhai, and W. Meng, "A general tensor representation framework for cross-view gait recognition," *Pattern Recognition*, vol. 90, pp. 87–98, 2019.
[44] X. Ben, C. Gong, P. Zhang, R. Yan, Q. Wu, and W. Meng, "Coupled bilinear discriminant projection for cross-view gait recognition," *IEEE Trans. Circuits Syst. Video Technol.*, 2019.
[45] Y. Guo, G. Ding, and J. Han, "Robust quantization for general similarity search," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 949–963, 2017.
[46] G. Wu, J. Han, Y. Guo, L. Liu, G. Ding, Q. Ni, and L. Shao, "Unsupervised deep video hashing via balanced code for large-scale video retrieval," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1993–2007, 2018.
[47] G. Ding, Y. Guo, K. Chen, C. Chu, J. Han, and Q. Dai, "Decode: deep confidence network for robust image classification," *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 3752–3765, 2019.
[48] X. Ben, C. Gong, P. Zhang, X. Jia, Q. Wu, and W. Meng, "Coupled patch alignment for matching cross-view gaits," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 3142–3157, 2019.
[49] X. Dong, Y. Yan, M. Tan, Y. Yang, and I. W. Tsang, "Late fusion via subspace search with consistency preservation," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 518–528, 2018.
[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.
[51] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Unsupervised domain adaptation with residual transfer networks," in *Proc. NIPS*, 2016, pp. 136–144.
[52] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *Proc. ECCV*, 2010, pp. 213–226.
[53] H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan, "Deep hashing network for unsupervised domain adaptation," in *Proc. CVPR*, 2017, pp. 5385–5394.
[54] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner *et al.*, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
[55] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," 2011.
[56] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. CVPR*, 2017, pp. 2962–2971.
[57] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," [Online]. Available: https://arxiv.org/abs/1703.07737.
[58] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

Weijian Deng received the M.S. degree from University of the Chinese Academy of Sciences, Beijing, China, in 2019. He is currently pursuing the Ph.D. degree in Computer Science at Australian National University, Canberra, Australia. His current research interests are visual understanding, pattern recognition, and person re-identification.



Liang Zheng is a Lecturer and a Computer Science Futures Fellow in the Research School of Computer Science, Australian National University. He received the Ph.D. degree in Electronic Engineering from Tsinghua University, China, in 2015, and the B.E. degree in Life Science from Tsinghua University, China, in 2010. He was a postdoc researcher in the Center for Artificial Intelligence, University of Technology Sydney, Australia. His research interests include image retrieval, classification, and person re-identification.

Yifan Sun received the B.E. degree in mechanical engineering from Tsinghua University, China, in 2005, and the M.S. degree in optical engineering from Tsinghua University, China, in 2008. He received the Ph.D. degree in information and communication engineering, from Tsinghua University, China, in 2019. His research interests are computer vision, person re-identification and deep embedding learning.

Jianbin Jiao (M'10) received the B.S., M.S., and Ph.D. degrees from the Harbin Institute of Technology (HIT), Harbin, China, in 1989, 1992, and 1995, respectively, all in mechanical and electronic engineering. From 1997 to 2005, he was an Associate Professor with HIT. Since 2006, he has been a Professor with the School of Electronic, Electrical, and Communication Engineering, University of the Chinese Academy of Sciences, Beijing, China. His current research interests include image processing, pattern recognition, and intelligent surveillance.