Semantic Concentration for Domain Adaptation

Shuang Li[1] Mixue Xie[1] Fangrui Lv[1] Chi Harold Liu[1*] Jian Liang[2] Chen Qin[3] Wei Li[4]
[1]Beijing Institute of Technology [2]Alibaba Group [3]University of Edinburgh [4]Inceptio Tech.
shuangli@bit.edu.cn michellexie102@gmail.com fangruilv@bit.edu.cn liuchi02@gmail.com
liangjianzb12@gmail.com Chen.Qin@ed.ac.uk liweimcc@gmail.com

Abstract

Domain adaptation (DA) paves the way for label annotation and dataset bias issues by the knowledge transfer from a label-rich source domain to a related but unlabeled target domain. A mainstream of DA methods is to align the feature distributions of the two domains. However, the majority of them focus on the entire image features where irrelevant semantic information, e.g., the messy background, is inevitably embedded. Enforcing feature alignments in such case will negatively influence the correct matching of objects and consequently lead to the semantically negative transfer due to the confusion of irrelevant semantics. To tackle this issue, we propose Semantic Concentration for Domain Adaptation (SCDA), which encourages the model to concentrate on the most principal features via the pair-wise adversarial alignment of prediction distributions. Specifically, we train the classifier to class-wisely maximize the prediction distribution divergence of each sample pair, which enables the model to find the region with large differences among the same class of samples. Meanwhile, the feature extractor attempts to minimize that discrepancy, which suppresses the features of dissimilar regions among the same class of samples and accentuates the features of principal parts. As a general method, SCDA can be easily integrated into various DA methods as a regularizer to further boost their performance. Extensive experiments on the cross-domain benchmarks show the efficacy of SCDA.

1. Introduction

Deep neural network (DNN) has achieved great success in diverse machine learning problems [17, 3, 33]. Unfortunately, the impressive performance gain heavily relies on the access to massive well-labeled training data. And it is often time and cost prohibitive to manually annotate sufficient training data in practice. Besides, another drawback of conventional deep learning is the poor generalization on a new dataset, due to the domain shift issue [2, 29, 1]. Hence, there is a strong motivation to utilize the knowledge of a label-rich domain (i.e., source domain) to assist the learning in a related but unlabeled domain (i.e., target domain), which is often referred to as domain adaptation (DA).

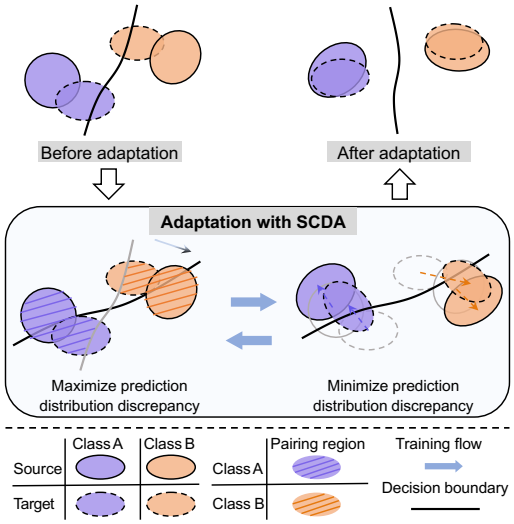


Figure 1. Illustration of the adversarial process of SCDA at the macro level. Classifier is trained to maximize the prediction distribution discrepancy of samples in the pairing region, which causes the decision boundary to pass through the high density area of the pairing region. While the feature extractor tries to minimize that discrepancy, which pushes the features away from the decision boundary. Finally, well-aligned features can be obtained through the adversarial game between the classifier and feature extractor.

To alleviate the domain shift problem, the common practice of DA is to reduce the cross-domain distribution discrepancy by learning domain-invariant feature representations. Generally, these DA methods can be roughly categorized as the discrepancy-based methods [23, 26, 7, 12], which align the domain distributions by minimizing a well-designed statistic metric, and the adversarial-based methods [8, 24, 41, 37, 20, 31], where the domain discriminator is designed to distinguish between source and target samples and the feature extractor tries to confuse the discriminator.

*Corresponding author.

Although these DA methods have admittedly achieved promising results, most of them use the features encoded without emphasis to match the feature distributions of two domains. In such case, irrelevant semantic information, e.g., the messy background is inevitably embedded, which may negatively influence the correct matching of objects and consequently lead to the semantically negative transfer.

To relieve this issue, we propose to achieve the Semantic Concentration for Domain Adaptation (SCDA) by leveraging the dark knowledge [49] (i.e., knowledge on the wrong predictions). Actually, SCDA is motivated by the findings in [53] that the class prediction made by the model depends on what it has concentrated on and the concentrated region for each class prediction can be located with the feature maps and corresponding classification weights. Thus, we expect to find the concentrated regions for wrong predictions and suppress the features of these regions when encoding the image into features.

For this purpose, we propose to class-wisely align the pair-wise prediction distributions in an adversarial manner, which is shown in Fig. 1. Samples of the same label from two domains compose the pairing region for each class. The paring of samples includes intra-domain paring (i.e., pairing within source domain) and inter-domain pairing (i.e., pairing between source and target samples). For any sample pair of the same label, the classifier is trained to maximize their prediction distribution discrepancy, while the feature extractor strives to minimize that discrepancy. From the micro perspective, when the feature extractor is fixed, maximizing the prediction distribution discrepancy of the sample pair will cause the classification weights for dark knowledge to be larger. Then to reduce that discrepancy, features of these dark knowledge have to be suppressed, since the classification weights for them became larger in the previous training of the classifier. From the macro perspective, to maximize the prediction discrepancy in the pairing region with the feature extractor fixed, the decision boundary will cross the high density area of the pairing region. Then, to reduce the discrepancy, features will be pushed away from the decision boundary. Finally, the model is able to concentrate on the most principal features and achieves well-aligned features class-wisely via the min-max game.

Our contributions are summarized as follows:

- This paper proposes a novel adversarial method for DA, i.e., the pair-wise adversarial alignment of prediction distribution discrepancy. Our method can suppress the irrelevant semantic information and accentuate the class object when encoding features, thus achieving the semantic concentration.
- As a simple and generic method, SCDA can be easily integrated as a regularizer into various DA methods and greatly improve their adaptation performances.
- Extensive experimental results and analysis demonstrate that SCDA greatly suppresses irrelevant semantics during the adaptation process, yielding state-of-the-art results on multiple cross-domain benchmarks.

2. Related Work

Feature Distribution Alignment. The distribution discrepancy between domains poses a great challenge for domain adaptation. To address this issue, the existing DA methods can be roughly divided into two categories. One is the statistical discrepancy based methods which aim to match various statistical moments across domains [25, 26, 51, 40, 18]. For instance, MDD [51] introduces the margin disparity discrepancy to reduce the distribution discrepancy with a rigorous generalization bound. And based on the Earth Mover's distance, [18] proposes an enhanced transport distance (ETD) to minimize the feature alignment loss.

The other category is inspired by the generative adversarial network (GAN) [10], which aims to learn domain-invariant features by playing a two-player min-max game [8, 24, 20, 41, 37, 5, 52]. For example, DANN [8] and CDAN [24] introduce a domain discriminator to play the min-max game where the domain discriminator strives to distinguish source samples from target samples while the feature extractor tries to confuse the domain discriminator.

However, these methods focus on the alignment of the entire image features. The irrelevant semantic information e.g., messy backgrounds, may predominate the adaptation process, leading to samples of different categories misaligned or samples in the same category unaligned.

Concentration Mechanism. There have been recent efforts toward boosting the adaptation performance via applying different degrees of concentration on distinct image regions [28, 42]. Several approaches leverage the attention-based methods to weight features at the pixel level, which facilitates the model concentrating on and transferring more principal semantic information across domains. [54, 16, 44] utilize attention mechanism to transfer features with high correlations across two distributions. DUCDA [54] develops an attention transfer mechanism for DA, which transfers the knowledge of discriminative patterns of source images to target. Differently, instead of exploring the space attention knowledge, DCAN [19] explores the low-level domain-dependent knowledge in the channel attention.

Although these attention-based DA methods can also suppress features of irrelevant semantics, most of them need to elaborately design a complex network architecture to derive the appropriate concentrations, greatly limiting their versatility. By contrast, our method leverages the pair-wise adversarial alignment on prediction space to achieve the concentration, which is easy to implement and can be used as a plug-and-play regularizer to various DA methods to further boost their performance.

Dark Knowledge. Numerous DA methods have explored the prediction space to boost the feature generation [37, 24, 4], while most of them only focus on the correct class prediction. To fully leverage the prediction information, we introduce the concept of dark knowledge [14], i.e., the knowledge on wrong predictions made by DNNs. In fact, dark knowledge is firstly proposed in knowledge distillation [14], where the knowledge is transferred from a powerful teacher model to a student [49, 50, 46]. For DA, the dark knowledge is also leveraged by some methods [15, 21] to excavate information contained in non-target labels. MCC [15] exploits the dark knowledge to formulate the tendency that a classifier confuses the predictions between the correct and ambiguous classes, and then minimizes the confusion. BCDM [21] proposes a novel metric using the dark knowledge of bi-classifiers to measure their discrepancy, where the classifiers are forced to produce more consistent predictions in a class-wise manner.

In this paper, we directly leverage the correspondence between the dark knowledge and its activated feature regions. By suppressing these features of dark knowledge via our proposed pair-wise adversarial alignment of predictions, we can effectively avoid the negative effect caused by irrelevant semantics in the adaptation process.

3. Method

3.1. Preliminaries and Motivation

In DA, there are two domains accessible: a labeled source domain with N_s samples, denoted as $\mathcal{S} = \{(\boldsymbol{x}_i^s, y_i^s)\}_{i=1}^{N_s}$ where $y_i^s \in \{1, 2, ..., C\}$ is the corresponding label of source sample $\boldsymbol{x}_i^s$, and an unlabeled target domain with N_t samples, denoted as $\mathcal{T} = \{\boldsymbol{x}_j^t\}_{j=1}^{N_t}$. Source and target domains share the same label space, but differ in the data probability distributions. Such distribution discrepancy often leads to the performance degradation when the network trained on source domain is directly applied to target domain. In this paper, we denote the network by $\mathcal{F}$ which is composed of a feature extractor $\mathcal{G}$ and a classifier $\mathcal{C}$. The goal of DA is to adapt the network $\mathcal{F}$ from source to target by fully exploring the knowledge of labeled source data and unlabeled target data during the training procedure.

Most DA methods are based on the feature distribution alignment where entire image features are considered. But irrelevant semantics, e.g., messy backgrounds may also be embedded into the entire features and thus the predictions for the wrong classes may be relatively high without suppression for these features, which may result in the semantically negative transfer. Hence, it is necessary to find these concentrated regions for dark knowledge and suppress the features of these regions. Motivated by the close relationship among the prediction, classification weights and the features shown in [53], we propose Semantically Concentration for Domain Adaptation (SCDA), which leverages the pair-wise adversarial alignment of prediction distribution to suppress the features of dark knowledge and thus accentuates the features of principal parts for correct class. Briefly, we take the classifier and the feature extractor as the two players in the adversarial game. The classifier tries to increase the classification weights for wrong classes by maximizing the pair-wise prediction distribution discrepancy. While the feature extractor strives to suppress the features for the wrong classes to reduce that discrepancy. Via the min-max game, we can suppress the influence of irrelevant semantics on the feature alignment of two domains.

3.2. Revisit the Class Activation Map

In this section, we revisit the class activation map in [53] to show the close relationship among the prediction, classification weights and features. For a particular class, its corresponding class activation map reflects which image region the model has concentrated on to make its prediction.

For a given image, let $a_h(u, v)$ denote the activation at spatial location (u, v) of the h-th channel of the feature maps in the last convolutional layer. Then performing the global average pooling (GAP) on the h-th channel, we obtain f_h, i.e., $f_h = \frac{1}{HW}\sum_{u,v} a_h(u, v)$, where H and W are the height and width of the feature map. For class c, the logit score z_c given by the model is $\sum_h w_h^c f_h$, where w_h^c is the classification weight (essentially the importance) of h-th feature map for class c. Here we omit the bias term, since it has no impact on the classification performance. Finally, the softmax score for class c is $p_c = \frac{\exp(z_c)}{\sum_c \exp(z_c)}$.

Plugging $f_h = \frac{1}{HW}\sum_{u,v} a_h(u, v)$ into the expression of z_c, we can obtain

$$\begin{aligned} z_c &= \sum_h w_h^c \frac{1}{HW}\sum_{u,v} a_h(u, v) \\ &= \frac{1}{HW}\sum_{u,v}\sum_h w_h^c a_h(u, v) \\ &= \frac{1}{HW}\sum_{u,v} A_c(u, v), \end{aligned} \tag{1}$$

where $A_c(u, v) = \sum_h w_h^c a_h(u, v)$. For a given model, HW is a constant. Thus, $A_c(u, v)$ directly reflects the importance of the activation at the spatial location (u, v) of the class activation map A_c when classifying an image to class c. Finally, by upsampling the class activation map to the size of the original image, we can locate the regions concentrated on by the model for a particular class.

From the expressions of $z_c = \frac{1}{HW}\sum_{u,v} A_c(u, v)$ and $p_c = \frac{\exp(z_c)}{\sum_c \exp(z_c)}$, we can see that the prediction distribution of an image depends on the class activation maps, while the class activation maps reflect what the model has concentrated on. This motivates us to leverage the class activation maps of wrong predictions to find the regions that the model

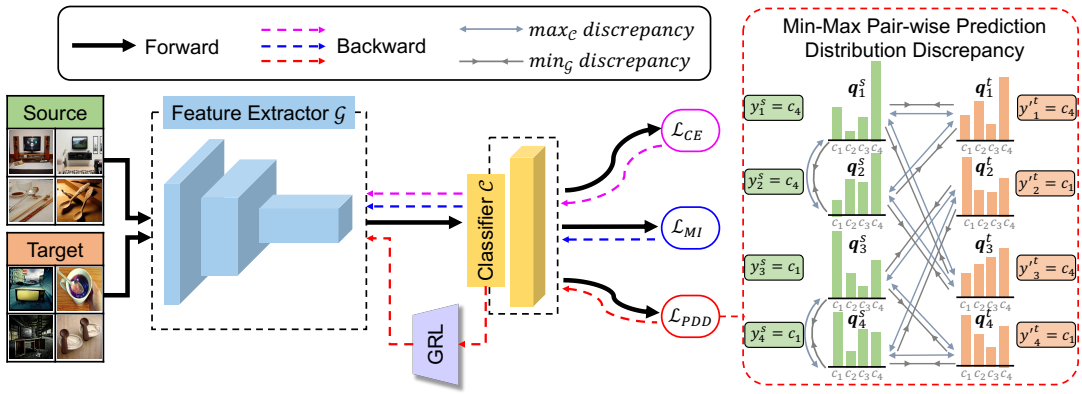


Figure 2. Overview of SCDA. $\{\boldsymbol{q}_i^s\}_{i=1}^4$ and $\{\boldsymbol{q}_j^t\}_{j=1}^4$ are the soften softmax predictions of a batch of source and target samples, respectively. GRL is the gradient reverse layer. $\mathcal{L}_{CE}$ is the cross-entropy loss on source domain. $\mathcal{L}_{MI}$ is the mutual information maximization loss on target domain. $\mathcal{L}_{PDD}$ is the pair-wise adversarial alignment loss of prediction distributions. The pairing of samples is shown in the right of the figure. The classifier is trained to maximize the prediction distribution discrepancy of each sample pair, while the feature extractor tries to minimize that discrepancy. Note that we use ground-truth labels for source samples, while pseudo labels for unlabeled target samples.

should not concentrate on and then suppress the features of these regions. Below we will describe how to achieve this idea via the pair-wise adversarial alignment of prediction distributions, which is the main component of our work.

3.3. Amplify Concentrations on Irrelevant Regions

Firstly, we describe the construction of our sample pairs, which is shown in Fig. 2. Samples of the same label from two domains compose the paring region for corresponding class. Since target domain is unlabeled, we employ the pseudo label predicted by the model for each target sample, i.e., $y'^t_j = \arg\max_c p_j^{t\,(c)}$ where $p_j^{t\,(c)}$ is the c-th element of the softmax outputs of target sample $\boldsymbol{x}_j^t$. Two samples are considered as a pair if their labels are same. For each class, there exist two kinds of sample pairs, i.e., intra-domain sample pairs [1] (pairing within source domain) and inter-domain sample pairs (pairing between source and target domains).

To amplify the concentrations on irrelevant regions, we train the classifier to maximize the prediction distribution discrepancy of each sample pair. Since we have two kinds of sample pairs, the total loss of prediction distribution discrepancy includes the intra-domain and inter-domain parts, i.e., $\mathcal{L}_{PDD_{s,s}}$ and $\mathcal{L}_{PDD_{s,t}}$, which are denoted as

$$\begin{aligned}\max_{\mathcal{C}}\ & \mathcal{L}_{PDD_{s,s}} + \mathcal{L}_{PDD_{s,t}} \\ &= \frac{1}{M_{s,s}} T^2 \sum_{y_i^s = y_k^s} JS(\boldsymbol{q}_i^s, \boldsymbol{q}_k^s) \\ &+ \frac{1}{M_{s,t}} T^2 \sum_{y_i^s = y'^t_j} JS(\boldsymbol{q}_i^s, \boldsymbol{q}_j^t). \end{aligned} \tag{2}$$

Here, we use *Jensen–Shannon* (JS) divergence to measure the discrepancy between a pair of predictions, due to its symmetry and finiteness compared with *Kullback-Leibler* divergence. $\boldsymbol{q}_i^s = \mathrm{softmax}(\mathcal{F}(\boldsymbol{x}_i^s)/T)$, where T is the temperature scaling parameter. To avoid the gradient vanishing, we multiply T^2 to maintain the magnitudes of gradients. $M_{s,s}$ and $M_{s,t}$ represent the number of samples satisfying $y_i^s = y_k^s$ and $y_i^s = y'^t_j$, respectively.

When the feature extractor is fixed, the class activation map only depends on the classification weights of the classifier. Since the sample pair belongs to the same class and the predictive scores for this class are both high, to maximize the prediction distribution discrepancy of the sample pair, the classification weights for other wrong classes increase. Thus, the irrelevant regions concentrated on by the model become more activated. Taking the "Bike" class for example, one image describes that a boy wearing a helmet is riding a bike and the other image describes a bike with flowers in the basket. For these two images, predictive scores for "Bike" are both high, such as with the prediction distributions of $[0.01, 0.79, 0.20]$ and $[0.15, 0.84, 0.01]$ respectively in the class order of "Flowers", "Bike" and "Helmet". The prediction distribution discrepancy mainly exists in the predictive scores for "Flowers" and "Helmet". To maximize the discrepancy, the former image will increase the predictive score for "Helmet", while the latter image will increase the score for "Flowers", which will cause the region of "Helmet" and "Flowers" with more concentration for the two images, respectively. By doing so, we amplify the concentrations on irrelevant regions.

3.4. Suppress Features of Irrelevant Semantics

In the previous section, we have found the regions that the model has concentrated on for the predictions of irrelevant classes. Now, we expect to suppress the features of these regions for a purer knowledge transfer in DA. To this end, we train the feature extractor to minimize the prediction distribution discrepancy of sample pairs, the loss of which is expressed as

[1] Here, we do not conduct intra-domain pairing within target domain, since target data have no ground-truth labels.

$$\min_{\mathcal{G}} \mathcal{L}_{PDD_{s,s}} + \mathcal{L}_{PDD_{s,t}}$$
$$= \frac{1}{M_{s,s}} T^2 \sum_{y_i^s = y_k^s} JS(\boldsymbol{q}_i^s, \boldsymbol{q}_k^s)$$
$$+ \frac{1}{M_{s,t}} T^2 \sum_{y_i^s = y'^t_j} JS(\boldsymbol{q}_i^s, \boldsymbol{q}_j^t). \quad (3)$$

Since the classification weights for wrong classes increased in the previous training of the classifier, to reduce the prediction distribution discrepancy, the feature extractor has to suppress the features of these irrelevant semantics and accentuate the features of similar parts in the sample pair. In the adversarial manner, for the intra-domain sample pairs, we can achieve the extraction of the most principal features for each class, which serves as good *teachers* for target domain. For the inter-domain sample pairs, the negative influence of domain shift is reduced and more emphasis is laid on the transfer of common knowledge across two domains.

3.5. Overall Formulation

Different from previous work [37, 36] that use alternate updating to achieve the adversarial manner, we leverage the gradient reverse layer (GRL) in Fig. 2 to achieve the optimization of all network parameters with the stochastic gradient descent. The overall loss function is defined as

$$\mathcal{L}_{SCDA} = \mathcal{L}_{CE} - \alpha \mathcal{L}_{PDD} - \beta \mathcal{L}_{MI}, \quad (4)$$

where α and β are two positive trade-off parameters.

$\mathcal{L}_{CE}$ is the standard cross-entropy loss to supervise the learning on source domain, which is denoted as

$$\min_{\mathcal{F}} \mathcal{L}_{CE} = \frac{1}{N_s} \sum_{i=1}^{N_s} \mathcal{E}(\mathcal{F}(\boldsymbol{x}_i^s), y_i^s), \quad (5)$$

where $\mathcal{E}(\cdot, \cdot)$ is the cross-entropy loss function.

$\mathcal{L}_{PDD}$ is our proposed adversarial loss of the prediction distribution discrepancy to achieve the semantic concentration for DA, the expression of which is denoted as

$$\min_{\mathcal{G}} \max_{\mathcal{C}} \mathcal{L}_{PDD} = \mathcal{L}_{PDD_{s,s}} + \mathcal{L}_{PDD_{s,t}}. \quad (6)$$

To avoid the tedious updating steps in alternate updating, we leverage the gradient reverse layer in [8] to achieve the adversarial training by one back-propagation.

$\mathcal{L}_{MI}$ is the mutual information maximization loss on target domain, which is introduced to improve the quality of pseudo labels. The expression of $\mathcal{L}_{MI}$ is

$$\max_{\mathcal{F}} \mathcal{L}_{MI} = H(\hat{Y}) - H(\hat{Y}|X)$$
$$= -\sum_{c=1}^{C} \hat{p}^{(c)} \log \hat{p}^{(c)} + \frac{1}{N_t} \sum_{j=1}^{N_t} \langle \boldsymbol{p}_j^t, \log \boldsymbol{p}_j^t \rangle, \quad (7)$$

where $\boldsymbol{p}_j^t$ is the softmax prediction of target sample $\boldsymbol{x}_j^t$, $\hat{p}^{(c)}$ is the c-th element of $\hat{\boldsymbol{p}} = \frac{1}{N_t} \sum_{j=1}^{N_t} \boldsymbol{p}_j^t$ and $\langle \cdot, \cdot \rangle$ is the inner product operation. Actually, the second term of $\mathcal{L}_{MI}$ is equivalent to the entropy minimization [11], which is a generic technique used in DA methods to enhance the discriminability of the model for target data, e.g., [52, 25, 39]. However, the entropy minimization may result into collapsed trivial solutions [45]. To avoid this, we introduce the first term of $\mathcal{L}_{MI}$ to ensure the diversity of predictions. Besides, we also set a threshold of 0.8 to select target samples with relatively correct classification, i.e., only $\{\boldsymbol{x}_j^t | \max_c p_j^{t\,(c)} \geq 0.8\}$ participate in inter-domain pairing.

The effects of different loss terms will be analyzed in details in the ablation study.

3.6. Regularizer to Existing DA Methods

As a simple but powerful method, SCDA is orthogonal to most existing DA methods and can be easily integrated into them as a regularizer to bring remarkable improvements by simply adding a gradient reverse layer. Taking CDAN [24] as an example, the integrated loss is formulated as:

$$\mathcal{L}_{SCDA} + \gamma \mathcal{L}_{adv}, \quad (8)$$

where γ is the trade-off parameter and $\mathcal{L}_{adv}$ is the domain-adversarial loss for the domain discriminator in CDAN. We suggest that readers refer to [24] for the detailed formulation of $\mathcal{L}_{adv}$. The adversarial process in [24] is that domain discriminator strives to correctly classify the domain labels of samples while the feature extractor aims to generate features that can deceive the domain discriminator. In addition, our method can also be plugged into other DA methods, such as statistical discrepancy based methods [51]. We will show the effects of SCDA as a regularizer in experiments.

4. Experiment

4.1. Experimental Setting

DomainNet [32] is the largest and the most challenging dataset for DA so far. It contains about 0.6 million images of 345 categories drawn from six diverse domains: Clipart (**clp**), Infograph (**inf**), Painting (**pnt**), Quickdraw (**qdr**), Real (**rel**) and Sketch (**skt**). Permuting the six domains, we build 30 adaptation tasks: **clp**→**inf**, ..., **skt**→**rel**.

Office-Home [43] is a more challenging benchmark dataset for visual domain adaptation, which includes 15,500 images of 65 categories spreading in four distinct domains: Artistic images (**Ar**), Clip Art (**Cl**), Product images (**Pr**) and Real-World images (**Rw**). 12 adaptation tasks are constructed to evaluate our method, i.e., **Ar**→**Cl**, ..., **Rw**→**Pr**.

Office-31 [35] is a classical real-world benchmark dataset for DA. It contains 4,110 images of 31 classes shared by three distinct domains: Amazon (**A**), Webcam (**W**) and DSLR (**D**). We construct 6 adaptation tasks to evaluate our method, i.e., **A**→**W**, ..., **D**→**W**.

Implementation details. Following the standard protocol for DA [8, 24, 22], we use all the labeled source

Table 1. Accuracy(%) on DomainNet for UDA (ResNet-101). In each sub-table, the column-wise domains are selected as the source domain and the row-wise domains are selected as the target domain. [† Implement according to source code.]

ADDA [41]	clp	inf	pnt	qdr	rel	skt	Avg.
clp	-	11.2	24.1	3.2	41.9	30.7	22.2
inf	19.1	-	16.4	3.2	26.9	14.6	16.0
pnt	31.2	9.5	-	8.4	39.1	25.4	22.7
qdr	15.7	2.6	5.4	-	9.9	11.9	9.1
rel	39.5	14.5	29.1	12.1	-	25.7	24.2
skt	35.3	8.9	25.2	14.9	37.6	-	25.4
Avg.	28.2	9.3	20.1	8.4	31.1	21.7	19.8

DANN [8]	clp	inf	pnt	qdr	rel	skt	Avg.
clp	-	15.5	34.8	9.5	50.8	41.4	30.4
inf	31.8	-	30.2	3.8	44.8	25.7	27.3
pnt	39.6	15.1	-	5.5	54.6	35.1	30.0
qdr	11.8	2.0	4.4	-	9.8	8.4	7.3
rel	47.5	17.9	47.0	6.3	-	37.3	31.2
skt	47.9	13.9	34.5	10.4	46.8	-	30.7
Avg.	35.7	12.9	30.2	7.1	41.4	29.6	26.1

MIMTFL [9]	clp	inf	pnt	qdr	rel	skt	Avg.
clp	-	15.1	35.6	10.7	51.5	43.1	31.2
inf	32.1	-	31.0	2.9	48.5	31.0	29.1
pnt	40.1	14.7	-	4.2	55.4	36.8	30.2
qdr	18.8	3.1	5.0	-	16.0	13.8	11.3
rel	48.5	19.0	47.6	5.8	-	39.4	32.1
skt	51.7	16.5	40.3	12.3	53.5	-	34.9
Avg.	38.2	13.7	31.9	7.2	45.0	32.8	28.1

ResNet-101 [13]	clp	inf	pnt	qdr	rel	skt	Avg.
clp	-	19.3	37.5	11.1	52.2	41.0	32.2
inf	30.2	-	31.2	3.6	44.0	27.9	27.4
pnt	39.6	18.7	-	4.9	54.5	36.3	30.8
qdr	7.0	0.9	1.4	-	4.1	8.3	4.3
rel	48.4	22.2	49.4	6.4	-	38.8	33.0
skt	46.9	15.4	37.0	10.9	47.0	-	31.4
Avg.	34.4	15.3	31.3	7.4	40.4	30.5	26.6

CDAN† [24]	clp	inf	pnt	qdr	rel	skt	Avg.
clp	-	20.4	36.6	9.0	50.7	42.3	31.8
inf	27.5	-	25.7	1.8	34.7	20.1	22.0
pnt	42.6	20.0	-	2.5	55.6	38.5	31.8
qdr	21.0	4.5	8.1	-	14.3	15.7	12.7
rel	51.9	23.3	50.4	5.4	-	41.4	34.5
skt	50.8	20.3	43.0	2.9	50.8	-	33.6
Avg.	38.8	17.7	32.8	4.3	41.2	31.6	27.7

MDD† [51]	clp	inf	pnt	qdr	rel	skt	Avg.
clp	-	20.5	40.7	6.2	52.5	42.1	32.4
inf	33.0	-	33.8	2.6	46.2	24.5	28.0
pnt	43.7	20.4	-	2.8	51.2	41.7	32.0
qdr	18.4	3.0	8.1	-	12.9	11.8	10.8
rel	52.8	21.6	47.8	4.2	-	41.2	33.5
skt	54.3	17.5	43.1	5.7	54.2	-	35.0
Avg.	40.4	16.6	34.7	4.3	43.4	32.3	28.6

SCDA	clp	inf	pnt	qdr	rel	skt	Avg.
clp	-	18.6	39.3	5.1	55.0	44.1	32.4
inf	29.6	-	34.0	1.4	46.3	25.4	27.3
pnt	44.1	19.0	-	2.6	56.2	42.0	32.8
qdr	30.0	4.9	15.0	-	25.4	19.8	19.0
rel	54.0	22.5	51.9	2.3	-	42.5	34.6
skt	55.6	18.5	44.7	6.4	53.2	-	35.7
Avg.	42.6	16.7	37.0	3.6	47.2	34.8	**30.3**

CDAN +SCDA	clp	inf	pnt	qdr	rel	skt	Avg.
clp	-	19.5	40.4	10.3	56.7	46.0	34.6
inf	35.6	-	36.7	4.5	50.3	29.9	31.4
pnt	45.6	20.0	-	4.2	56.8	41.9	33.7
qdr	28.3	4.8	11.5	-	20.9	19.2	17.0
rel	55.5	22.8	53.7	3.2	-	42.1	35.5
skt	58.4	21.1	47.8	10.6	56.5	-	38.9
Avg.	44.7	17.6	38.0	6.6	48.2	35.8	**31.8**

MDD +SCDA	clp	inf	pnt	qdr	rel	skt	Avg.
clp	-	20.4	43.3	15.2	59.3	46.5	36.9
inf	32.7	-	34.5	6.3	47.6	29.2	30.1
pnt	46.4	19.9	-	8.1	58.8	42.9	35.2
qdr	31.1	6.6	18.0	-	28.8	22.0	21.3
rel	55.5	23.7	52.9	9.5	-	45.2	37.4
skt	55.8	20.1	46.5	15.0	56.7	-	38.8
Avg.	44.3	18.1	39.0	10.8	50.2	37.2	**33.3**

Table 2. Accuracy (%) on Office-Home for UDA (ResNet-50).

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 [13]	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
DANN [8]	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 [26]	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
MCD [37]	48.9	68.3	74.6	61.3	67.6	68.8	57.0	47.1	75.1	69.1	52.2	79.6	64.1
ETD [18]	51.3	71.9	**85.7**	57.6	69.2	73.7	57.8	51.2	79.3	70.2	57.5	82.1	67.3
SymNets [52]	47.7	72.9	78.5	64.2	71.3	74.2	64.2	48.8	79.5	74.5	52.6	82.7	67.6
TADA [44]	53.1	72.3	77.2	59.1	71.2	72.1	59.7	53.1	78.4	72.4	60.0	82.9	67.6
GVB-GD [5]	57.0	74.7	79.8	64.6	74.1	74.6	65.2	55.1	81.0	74.6	59.7	84.3	70.4
SCDA	57.5	76.9	80.3	65.7	74.9	74.5	65.5	53.6	79.8	74.5	59.6	83.7	70.5
CDAN [24]	50.7	70.6	76.0	57.6	70.0	70.0	57.4	50.9	77.3	70.9	56.7	81.6	65.8
CDAN+SCDA	57.1	75.9	79.9	66.2	76.7	75.2	65.3	55.6	81.9	74.7	**62.6**	84.5	71.3
MDD [51]	54.9	73.7	77.8	60.0	71.4	71.8	61.2	53.6	78.1	72.5	60.2	82.3	68.1
MDD+SCDA	58.9	77.2	81.0	66.6	75.5	75.9	64.1	56.3	82.2	73.3	61.5	84.3	71.4
MCC [15]	55.1	75.2	79.5	63.3	73.2	75.8	66.1	52.1	76.9	73.8	58.4	83.6	69.4
MCC+SCDA	57.1	**79.1**	82.7	67.7	75.3	77.6	66.3	52.5	81.9	**74.9**	60.1	**85.0**	71.7
DCAN [19]	54.5	75.7	81.2	67.4	74.0	76.3	67.4	52.7	80.6	74.1	59.1	83.5	70.5
DCAN+SCDA	**60.7**	76.4	82.8	**69.8**	**77.5**	**78.4**	**68.9**	**59.0**	**82.7**	**74.9**	61.8	84.5	**73.1**

data and unlabeled target data as training data and evaluate on unlabeled target data. We implement our approach in PyTorch framework[30]. To fairly compare with existing methods, we use the same backbone networks, i.e., ResNet-50 [13] pre-trained on ImageNet [34] for datasets: Office-31 and Office-Home, and ResNet-101 [13] pre-trained on ImageNet [34] for DomainNet [32]. In experiments, the input image size is cropped to 224 $\times$ 224. We employ the mini-batch stochastic gradient descent (SGD) optimizer with momentum of 0.9 and the learning rate strategy as described in [8] for network optimization. To reduce the effect of unreliable predictions in the early training stage, we simply let the hyper-parameter $\alpha = \alpha_0\rho$, where ρ is the training progress changing from 0 to 1. And following MCC [15], we use Deep Embedded Validation (DEV) [48] to select the hyper-parameters and find that $T = 10$, $\alpha_0 = 1.0$, $\beta = 0.1$ works well on all datasets. Besides, the parameter sensitivity analysis is provided in section 4.3 to test the robustness of SCDA. Each adaptation task is evaluated by averaging the results of three random trials. Code of SCDA is available at https://github.com/BIT-DA/SCDA.

4.2. Results

Results on DomainNet are presented in Table 1. Obviously, SCDA outperforms all the compared methods significantly in terms of the average accuracy. Particularly, applying SCDA to CDAN and MDD improves their prediction accuracy by 4.1% and 4.7% respectively. One interpretation

Table 3. Accuracy (%) on Office-31 for UDA (ResNet-50). [Avg‡: mean values except D↔W]

Method	A→W	D→W	W→D	A→D	D→A	W→A	Avg	Avg‡
ResNet-50 [13]	68.4	96.7	99.3	68.9	62.5	60.7	76.1	65.1
DANN [8]	82.0	96.9	99.1	79.7	68.2	67.4	82.2	74.3
JAN [26]	85.4	97.4	99.8	84.7	68.6	70.0	84.3	77.2
CAT [6]	91.1	98.6	99.6	90.6	70.4	66.5	86.1	79.7
ETD [18]	92.1	**100.0**	**100.0**	88.0	71.0	67.8	86.2	79.7
MCD [37]	88.6	98.5	**100.0**	92.2	69.5	69.7	86.5	80.0
SymNets [52]	90.8	98.8	**100.0**	93.9	74.6	72.5	88.4	83.0
TADA [44]	94.3	98.7	99.8	91.6	72.9	73.0	88.4	83.0
GVB-GD [5]	94.8	98.7	**100.0**	95.0	73.4	73.7	89.3	84.2
SCDA	94.2	98.7	99.8	95.2	75.7	76.2	90.0	85.3
CDAN [24]	94.1	98.6	**100.0**	92.9	71.0	69.3	87.7	81.8
CDAN+SCDA	94.7	98.7	**100.0**	95.4	77.1	76.0	90.3	85.8
MDD [51]	94.5	98.4	**100.0**	93.5	74.6	72.2	88.9	83.7
MDD+SCDA	95.3	99.0	**100.0**	95.4	77.2	75.9	**90.5**	**85.9**
MCC [15]	**95.5**	98.6	**100.0**	94.4	72.9	74.9	89.4	84.4
MCC+SCDA	93.7	98.6	**100.0**	**96.4**	76.5	76.0	90.2	85.7
DCAN [19]	95.0	97.5	**100.0**	92.6	77.2	74.9	89.5	84.9
DCAN+SCDA	94.8	98.2	**100.0**	94.6	**77.5**	**76.4**	90.3	85.8

Table 4. Ablation Study of SCDA on Office-31 (ResNet-50).

Method	A→W	D→W	W→D	A→D	D→A	W→A	Avg
ResNet-50	68.4	96.7	99.3	68.9	62.5	60.7	76.1
+ SCDA (w/o $\mathcal{L}_{PDD}$)	91.3	98.6	99.8	92.2	69.2	68.6	86.6
+ SCDA (w/o $\mathcal{L}_{PDD_{s,t}}$)	91.8	98.4	100.0	92.5	71.4	70.8	87.5
+ SCDA (w/o $\mathcal{L}_{PDD_{s,s}}$)	92.2	98.6	100.0	94.1	72.8	72.6	88.3
+ SCDA (w/o $\mathcal{L}_{MI}$)	92.6	98.7	100.0	94.4	74.1	73.4	88.9
+ SCDA	94.2	98.7	99.8	95.2	75.7	76.2	**90.0**

is that our method suppresses the features of irrelevant semantics that may confuse the alignment process of CDAN and MDD. The encouraging results demonstrate the superiority of SCDA in processing complex datasets and its universality to existing DA methods.

Results on Office-Home are shown in Table 2, where we achieve comparable and even better performance, compared with these state-of-art DA methods. Moreover, our method achieves extra gain of 5.5% and large improvements on **Cl** → **Ar**, **Cl**→ **Pr**, **Cl**→ **Rw** when applied to CDAN. The reason is that the images in **Cl** are rather complicated, while SCDA can purify the transferred knowledge by suppressing the features of irrelevant semantics. And DCAN+SCDA achieves the best performance of 73.1%. These improvements validate the effectiveness of SCDA.

Results on Office-31 are summarized in Table 3. Obviously, we substantially obtain superior prediction accuracy over other popular adaptation methods. Particularly, when applying SCDA to MDD, we achieve the highest accuracy of 90.5%. The outcomes show that SCDA is beneficial to promote the adaptation capability, especially on complex scenarios, e.g., **A** → **D**, **D** → **A** and **W** → **A**.

4.3. Analysis

Ablation Study. To investigate the efficacy of different components of SCDA, we conduct thorough ablation analysis on Office-31 based on ResNet-50: (1) SCDA (w/o $\mathcal{L}_{MI}$) denotes the variant of removing the mutual information maximization loss on target domain; (2) SCDA (w/o $\mathcal{L}_{PDD_{s,s}}$) and SCDA (w/o $\mathcal{L}_{PDD_{s,t}}$) respectively denote the variant of removing the pair-wise adversarial alignment of prediction distributions within source domain and cross domains; (3) SCDA (w/o $\mathcal{L}_{PDD}$) denotes the removal of both $\mathcal{L}_{PDD_{s,s}}$ and $\mathcal{L}_{PDD_{s,t}}$. The results are shown in Table 4, where we can obviously see that full method SCDA outperforms other variants. While SCDA (w/o $\mathcal{L}_{PDD_{s,t}}$) suffers a obvious degradation of 2.5%, which indicates the importance of transferring the common knowledge and suppressing domain-specific knowledge for DA problems by our loss $\mathcal{L}_{PDD_{s,t}}$. And SCDA is superior to SCDA (w/o $\mathcal{L}_{PPA_{s,s}}$), because $\mathcal{L}_{PPA_{s,s}}$ conduces to the constructing of good *teachers* for target samples by learning the most principal features for classification. Besides, through improving the quality of pseudo labels for the paring process, SCDA achieves better performance than SCDA (w/o $\mathcal{L}_{MI}$).

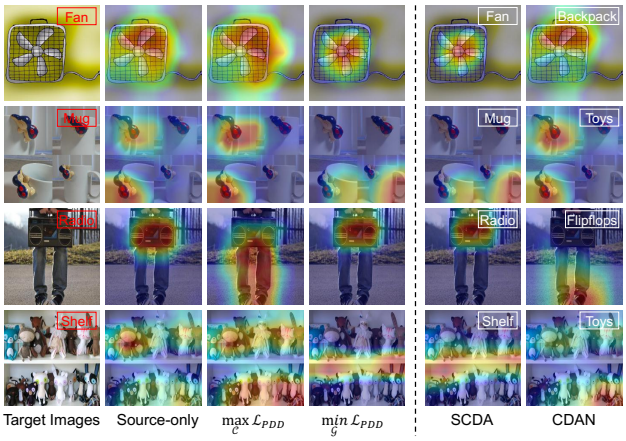
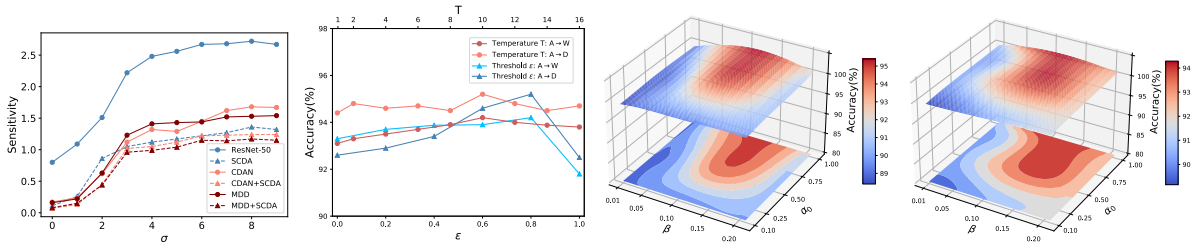


Figure 3. Concentration visualizations of the last convolutional layer of different methods on the task Rw → Ar of Office-Home. Here, the red font denotes the ground-truth labels, while the white font represents the pseudo labels predicted by different methods.

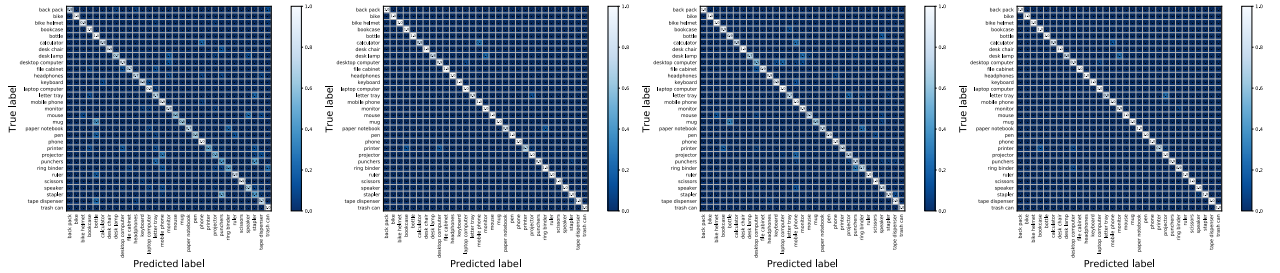
Visual Explanations for Semantic Concentration. In this section, we utilize the visualization technique in [38] to visualize which region SCDA has concentrated on in the adversarial process, which is shown in Fig. 3. We can observe that the concentration on irrelevant regions significantly increases after maximizing the prediction distribution discrepancy loss $\mathcal{L}_{PDD}$, and then, the features of irrelevant/principal regions are suppressed/accentuated by minimizing the discrepancy, which verifies the aforementioned micro explanations. Besides, the final results demonstrate that our method indeed achieves the semantic concentration for the critical parts in image classification.

Anti-jamming Ability Test. Since our method aims to suppress the features of irrelevant semantics and accentuate the features of principal parts, we conduct this experiment to test its anti-jamming ability by adding Gaussian noises with zero-mean to a batch of randomly selected input images and then testing the sensitivity of different



(a) W → A (b) A → D and A → W (c) A → D (d) A → W

Figure 4. (a) is anti-jamming ability test of different methods on task W → A of Office-31 as the variance σ of added Gaussian noise increasing from 0 to 10. (b) is the sensitivity of SCDA to parameters T and ϵ on tasks A → D and A → W. (c) and (d) are the sensitivity of SCDA to parameters α_0 and β on tasks A → D and A → W, respectively.



(a) ResNet-50 (b) SCDA (c) CDAN (d) CDAN+SCDA

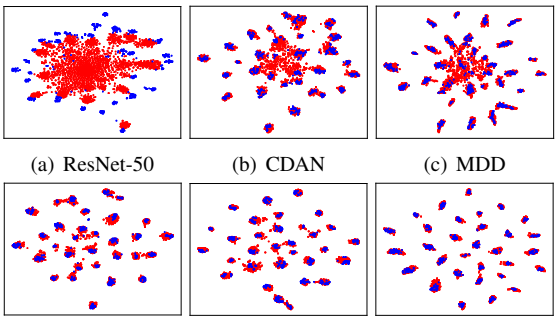
Figure 5. The confusion matrices on target domain of different methods on the task A → D of Office-31. (Zoom in for clear visualization.)

methods as [47]. The results are illustrated in Fig. 4(a). It can be clearly observed that the sensitivities of SCDA, CDAN+SCDA and MDD+SCDA (dashed lines) are much smaller and also grow more slowly compared to the corresponding baseline methods (solid lines). Such phenomenon reveals that SCDA can significantly suppress the features of irrelevant noises, further proving the superiority of SCDA.

Confusion Matrix. The confusion matrices of different methods are given in Fig. 5. For ResNet-50 and CDAN, there exist numerous wrong predictions appearing in the off-diagonal, e.g., most samples of "mug" are misclassified into "bottle". By contrast, we can clearly see quantitative improvements of SCDA and CDAN+SCDA, the reason of which can be explained as the pair-wise adversarial alignment in each class leads to more compact features and thus reduces the class confusion. The encouraging results further show the advantages of SCDA either as an independent method or as a regularizer integrated into existing methods.

t-SNE Visualization. Fig. 6 visualizes the feature representations learned by ResNet-50, CDAN, MDD, SCDA, CDAN+SCDA and MDD+SCDA with t-SNE [27]. We can clearly see that target data are not aligned well with source data using original methods, while SCDA can learn highly discriminative features and keep clear boundaries.

Parameter Sensitivity. Fig. 4(b), 4(c) and 4(d) show the sensitivity of SCDA to temperature T, threshold ϵ and two loss trade-offs α_0 and β on tasks A → D and A → W. The results in Fig. 4(c) and 4(d) show that SCDA is not that sensitive when $\alpha_0 \in \{0.5, 0.75, 1.0\}$ and $\beta \in \{0.1, 0.15\}$.



(a) ResNet-50 (b) CDAN (c) MDD

(d) SCDA (e) CDAN+SCDA (f) MDD+SCDA

Figure 6. The visualization of features learned by different methods on the task W → A of Office-31. Blue and red dots represent source and target features, respectively.

In Fig. 4(b), SCDA is not sensitive to T, but sensitive to ϵ (with $\epsilon = 0.8$ working best). Because unreliable pseudo labels will confuse the pairing if ϵ too small, and too large ϵ will lead to the insufficient knowledge transfer.

5. Conclusion

In this paper, we propose Semantic Concentration for Domain Adaptation (SCDA) to accentuate the features of principal parts and suppress the features of irrelevant semantics via the pair-wise adversarial alignment on the prediction space within source domain and across domains. Orthogonal to most DA methods, SCDA can be easily integrated as a regularizer to bring further improvements. Extensive experimental results verify the efficacy of SCDA.

References

[1] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *NeurIPS*, pages 137–144, 2007. 1

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

[3] Liangchieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 40(4):834–848, 2018. 1

[4] Safa Cicek and Stefano Soatto. Unsupervised domain adaptation via regularized conditional alignment. In *ICCV*, October 2019. 3

[5] Shuhao Cui, Shuhui Wang, Junbao Zhuo, Chi Su, Qingming Huang, and Qi Tian. Gradually vanishing bridge for adversarial domain adaptation. In *CVPR*, June 2020. 2, 6, 7

[6] Zhijie Deng, Yucen Luo, and Jun Zhu. Cluster alignment with a teacher for unsupervised domain adaptation. In *ICCV*, October 2019. 7

[7] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *ICML*, volume 32, pages 647–655. ACM, 2014. 1

[8] Yaroslav Ganin and Victor S. Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, pages 1180–1189, 2015. 1, 2, 5, 6, 7

[9] Jian Gao, Yang Hua, Guosheng Hu, Chi Wang, and Neil M Robertson. Reducing distributional uncertainty by mutual information maximisation and transferable feature learning. In *ECCV*, 2020. 6

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, pages 2672–2680, 2014. 2

[11] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *NeurIPS*, pages 529–536, 2004. 5

[12] Arthur Gretton, Karsten M Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J Smola. A kernel method for the two-sample-problem. In *NeurIPS*, pages 513–520, 2007. 1

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 6, 7

[14] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *Computer ence*, 14(7):38–39, 2015. 3

[15] Ying Jin, Ximei Wang, Mingsheng Long, and Jianmin Wang. Minimum class confusion for versatile domain adaptation. In *ECCV*, volume 12366, pages 464–480. Springer, 2020. 3, 6, 7

[16] Guoliang Kang, Liang Zheng, Yan Yan, and Yi Yang. Deep adversarial attention alignment for unsupervised domain adaptation: the benefit of target expectation maximization. In *ECCV*, September 2018. 2

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, pages 1097–1105, 2012. 1

[18] Mengxue Li, Yi-Ming Zhai, You-Wei Luo, Peng-Fei Ge, and Chuan-Xian Ren. Enhanced transport distance for unsupervised domain adaptation. In *CVPR*, June 2020. 2, 6, 7

[19] Shuang Li, Chi Harold Liu, Qiuxia Lin, Binhui Xie, Zhengming Ding, Gao Huang, and Jian Tang. Domain conditioned adaptation network. In *AAAI*, pages 11386–11393, 2020. 2, 6, 7

[20] Shuang Li, Chi Harold Liu, Binhui Xie, Limin Su, Zhengming Ding, and Gao Huang. Joint adversarial domain adaptation. In *ACM MM*, pages 729–737. ACM, 2019. 1, 2

[21] Shuang Li, Fangrui Lv, Binhui Xie, Chi Harold Liu, Jian Liang, and Chen Qin. Bi-classifier determinacy maximization for unsupervised domain adaptation. In *AAAI*, 2020. 3

[22] Hong Liu, Mingsheng Long, Jianmin Wang, and Michael Jordan. Transferable adversarial training: A general approach to adapting deep classifiers. In *ICML*, pages 4013–4022, 2019. 5

[23] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. In *ICML*, pages 97–105. ACM, 2015. 1

[24] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *NeurIPS*, pages 1647–1657, 2018. 1, 2, 3, 5, 6, 7

[25] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. In *NeurIPS*, pages 136–144, 2016. 2, 5

[26] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *ICML*, pages 2208–2217. ACM, 2017. 1, 2, 6, 7

[27] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(Nov):2579–2605, 2008. 8

[28] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. Recurrent models of visual attention. In *NeurIPS*, page 2204–2212. MIT Press, 2014. 2

[29] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *TKDE*, 22(10):1345–1359, 2010. 1

[30] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8024–8035, 2019. 6

[31] Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. Multi-adversarial domain adaptation. In *AAAI*, 2018. 1

[32] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *ICCV*, pages 1406–1415, 2019. 5, 6

[33] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. In *NeurIPS*, volume 2015, pages 91–99, 2015. 1

[34] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy,

Aditya Khosla, and Michael Bernstein. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 6

[35] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *ECCV*, pages 213–226, 2010. 5

[36] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Adversarial dropout regularization. In *ICLR*. OpenReview.net, 2018. 5

[37] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *CVPR*, pages 3723–3732, 2018. 1, 2, 3, 5, 6, 7

[38] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017. 7

[39] Rui Shu, Hung H. Bui, Hirokazu Narui, and Stefano Ermon. A DIRT-T approach to unsupervised domain adaptation. In *ICLR*, 2018. 5

[40] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV*, pages 443–450, 2016. 2

[41] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, pages 2962–2971, 2017. 1, 2, 6

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 2

[43] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *CVPR*, pages 5018–5027, 2017. 5

[44] Ximei Wang, Liang Li, Weirui Ye, Mingsheng Long, and Jianmin Wang. Transferable attention for domain adaptation. In *AAAI*, 2019. 2, 6, 7

[45] Xiaofu Wu, Suofei Zhang, Quan Zhou, Zhen Yang, Chunming Zhao, and Longin Jan Latecki. Entropy minimization vs. diversity maximization for domain adaptation. *ArXiv*, 2002.01690, 2020. 5

[46] Ting-Bing Xu and Cheng-Lin Liu. Data-distortion guided self-distillation for deep neural networks. In *AAAI*, pages 5565–5572. AAAI Press, 2019. 3

[47] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun S. Suggala, David I. Inouye, and Pradeep Ravikumar. On the (in)fidelity and sensitivity of explanations. In *NeurIPS*, dec 2019. 8

[48] Kaichao You, Ximei Wang, Mingsheng Long, and Michael I. Jordan. Towards accurate model selection in deep unsupervised domain adaptation. In *ICML*, volume 97, pages 7124–7133, 2019. 6

[49] Sukmin Yun, Jongjin Park, Kimin Lee, and Jinwoo Shin. Regularizing class-wise predictions via self-knowledge distillation. In *CVPR*, pages 13873–13882, 2020. 2, 3

[50] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *ICCV*, pages 3712–3721. IEEE, 2019. 3

[51] Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging theory and algorithm for domain adaptation. In *ICML*, pages 7404–7413, 2019. 2, 5, 6, 7

[52] Yabin Zhang, Hui Tang, Kui Jia, and Mingkui Tan. Domain-symmetric networks for adversarial domain adaptation. In *CVPR*, pages 5031–5040, 2019. 2, 5, 6, 7

[53] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, pages 2921–2929. IEEE Computer Society, 2016. 2, 3

[54] Junbao Zhuo, Shuhui Wang, Weigang Zhang, and Qingming Huang. Deep unsupervised convolutional domain adaptation. In *ACM MM*, MM '17, page 261–269. Association for Computing Machinery, 2017. 2