Cooperative and Adversarial Learning: Co-enhancing Discriminability and Transferability in Domain Adaptation

Hui Sun, Zheng Xie, Xin-Ye Li, Ming Li

National Key Laboratory for Novel Software Technology,
Nanjing University, Nanjing 210023, China
{sunh,xiez,lixy,lim}@lamda.nju.edu.cn

Abstract

Discriminability and transferability are two goals of feature learning for domain adaptation (DA), as we aim to find the transferable features from the source domain that are helpful for discriminating the class label in the target domain. Modern DA approaches optimize discriminability and transferability by adopting two separate modules for the two goals upon a feature extractor, but lack fully exploiting their relationship. This paper argues that by letting the discriminative module and transfer module help each other, better DA can be achieved. We propose Cooperative and Adversarial LEarning (CALE) to combine the optimization of discriminability and transferability into a whole, provide one solution for making the discriminative module and transfer module guide each other. Specifically, CALE generates cooperative (easy) examples and adversarial (hard) examples with both discriminative module and transfer module. While the easy examples that contain the module knowledge can be used to enhance each other, the hard ones are used to enhance the robustness of the corresponding goal. Experimental results show the effectiveness of CALE for unifying the learning of discriminability and transferability, as well as its superior performance.

1 Introduction

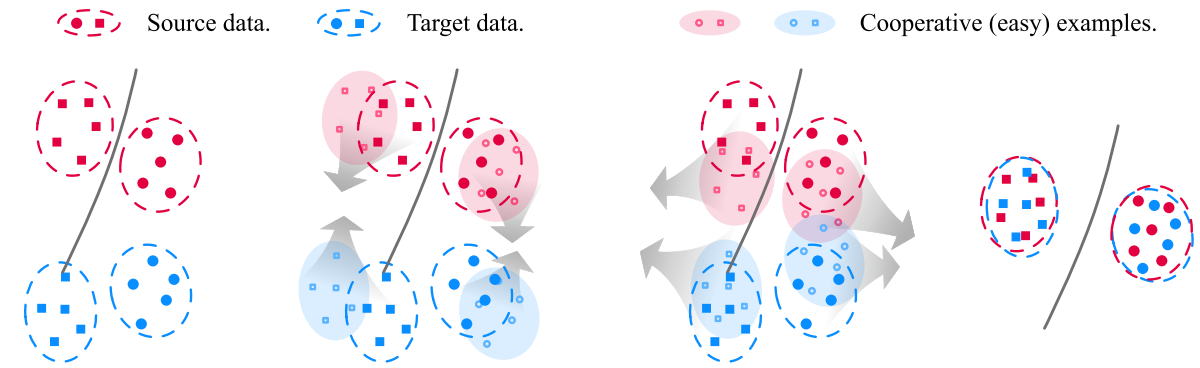
Transfer learning seeks to make the machine learning systems to perform well on new tasks by referring to the experience of some old tasks. Unsupervised domain adaptation (UDA), one of the most active subfields of transfer learning, shows a great ability to transfer knowledge across tasks (Pan and Yang 2010). It aims to build a model on an unlabeled target domain with the help of the source domain data, and has been applied to various tasks, e.g., computer vision (Liu, Zhang, and Wang 2021; Liu, Wang, and Long 2021), natural language processing (Lekhtman, Ziser, and Reichart 2021; Zhu et al. 2021), information retrieval (Li and Caragea 2020; Chen et al. 2020c; Kanagawa et al. 2019), reinforcement learning (Chen et al. 2021; Rao et al. 2020), etc.

Due to the potent power of representation learning in the deep neural networks, most of the modern UDA methods are dedicated to designing elaborated mechanisms for learning domain invariant feature representations, such that the distributions of source and target domain are well aligned in the representation space (Tzeng et al. 2014; Ganin et al. 2016; Tan et al. 2018). Such features bridge the source domain and target domain for transferring knowledge across domains.

A successful feature representation for domain adaptation requires two properties, *discriminability* and *transferability*, which form the fundamental goals of UDA feature learning (Ben-David et al. 2010; Chen et al. 2019b). Discriminability refers to the effectiveness of the representation to be used for the discriminative task, e.g., classification. Transferability refers to the ability of the representation to capture the invariance across the domains so that the different domains can be well aligned under the learned representation.

The modern deep UDA methods generally consist of a discriminative module (e.g., a classifier) and a transfer module (e.g., a domain discriminator based on Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) or a distribution distance function based on the statistical moment) upon the feature extractor. The two modules will be trained by the discriminability and transferability loss correspondingly. However, existing literature treats them as two separate terms in the loss function, produced by two modules that are *optimized separately* (Tzeng et al. 2014; Ganin et al. 2016; Long et al. 2018). Since the discriminability and transferability can be partly conflicting, without considering optimizing them together, the two goals may adversely impact each other, leading to a degradation in the performance or even a negative transfer to some extent (Wang et al. 2019b; Chen et al. 2019a; Kontar, Raskutti, and Zhou 2020).

In this paper, we argue that the discriminability and transferability should be considered jointly. By letting the discriminative module and transfer module help each other, better domain adaptation can be achieved. We propose Cooperative and Adversarial LEarning (CALE) for UDA, which exploits the knowledge of optimizing one property to guide the another. Specifically, for cooperative learning (CLE), we generate easy examples in terms of discriminability loss and transferability loss, and then the discriminative and transfer module regularize their outputs to be consistent with the easy examples for their complementary modules. The easy examples contain the module knowledge of better feature representation, by exchanging the easy examples, the two modules take the guidance from each other, makes the UDA model consider one property while optimizing another. Figure 1 demonstrates how the easy examples generated by the



(a) Before adaptation. (b) Discriminability guides Transferability. (c) Transferability guides discriminability. (d) After adaptation.

Figure 1: Demonstration of cooperative learning (CLE) in CALE. The cooperative examples of discriminability and transferability help the feature learning when optimizing the other goal.

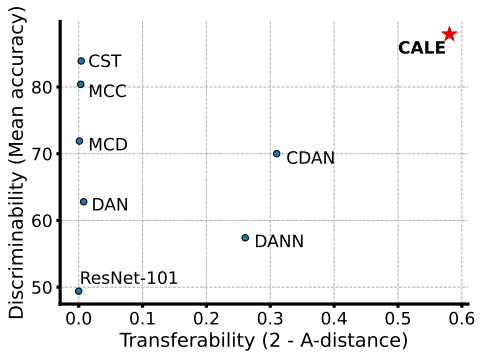


Figure 2: Discriminability and transferability.

two modules can guide each other. For adversarial learning (ALE), we regularize the output consistency of the two modules on their own hard examples. It further enhances the robustness of the feature learning in terms of discriminability and transferability. Figure 2 shows the target domain data's discriminability and transferability of several recent UDA approaches, in which we can observe that CALE not only achieve good discriminability in the target domain, but also maintains great transferability. Detailed explanation and experiment protocol can be found in Section 4.2. Besides, the conceptual framework of CALE is shown in Figure 3.

We emphasize our contributions in two aspects:

1. We argue that the discriminability and transferability in domain adaptation should be learned jointly instead of separately, and empirically validate our claim.
2. We propose a CALE framework to unify the learning of the discriminability and transferability. The CALE model not only achieves good performance but also keeps great transferability across domains.

2 Preliminaries

Unsupervised Domain Adaptation (UDA). Let $\mathcal{X}$ be some feature space and $\mathcal{Y}$ be some label space. In this paper, we focus on the problem of UDA, where we want to build a model for a target domain $\mathcal{T}$ defined over $\mathcal{X} \times \mathcal{Y}$ with the knowledge from a source domain $\mathcal{S}$. Formally, only a labeled set $\mathcal{D}_s = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$ with n_s labeled samples from the source domain $\mathcal{S}$ and an unlabeled set $\mathcal{D}_t = \{x_i^t\}_{i=1}^{n_t}$ with n_t unlabeled samples from the target domain $\mathcal{T}$ are available. As the two domains do not share the same distribution, the core problem of UDA is to handle the *domain shift* between $\mathcal{S}$ and $\mathcal{T}$ (Pan and Yang 2010).

Revisit the Current State-of-the-Art. Most modern domain adaptation approaches are based on the seminal theory proposed by Ben-David et al. (2010):

Theorem 1. *Let $\mathcal{H}$ be the hypothesis space, and given a source domain $\mathcal{S}$ and a target domain $\mathcal{T}$. The upper bound of the expected error on the target domain is:*

$$\epsilon_{\mathcal{T}}(h) \leq \epsilon_{\mathcal{S}}(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T}) + C \tag{1}$$

where $\epsilon_{\mathcal{S}}(h)$ is the expected error on the source domain, $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T})$ is a discrepancy metric of cross-domain distributions, and C is an ideal error defined as $C = \min_{h^* \in \mathcal{H}} [\epsilon_{\mathcal{S}}(h^*) + \epsilon_{\mathcal{T}}(h^*)]$.

From the perspective of representation learning, $\epsilon_{\mathcal{S}}(h)$ reflects the discriminability of $\mathcal{S}$, and $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T})$ reflects the transferability between $\mathcal{S}$ and $\mathcal{T}$. Noteworthily, most prior work believe that the third term C measures the discriminability especially in the target domain $\mathcal{T}$ (Xie et al. 2018; Chen et al. 2019b).

Inspired by this, recent deep UDA approaches generally consist of a discriminative module and a transfer module upon a shared feature extractor. The discriminative module always minimizes the supervised loss on the labeled source domain data, corresponding to the first term in Theorem 1, which optimizes the discriminability on $\mathcal{S}$. The transfer module guides the model to learn domain invariant feature representations through two mainstream technologies: moment matching (Tzeng et al. 2014; Long et al. 2015; Xie et al. 2018) and adversarial confusing (Ganin et al. 2016; Long et al. 2018), which optimizes the transferability across domains, i.e., the second term above. Moreover, to optimize the discriminability of target domain $\mathcal{T}$, some approaches

minimize the third term C by employing the predictions of $\mathcal{T}$ from the discriminative module, e.g., Shu et al. (2018) adopts them for self-training the discriminative module.

Limitations of Current Approaches. Current approaches mainly optimize discriminability and transferability separately, which has the following limitations. (1) When the discriminative module (discriminability) is optimized independently, it may destroy the transferability of representations. For instance, to improve the discriminability on the source domain, the model extracts too many source-specific features, so the independent transfer module is difficult to transfer the source-specific knowledge to the target. (2) The transfer module optimizes transferability independently without caring about the discriminability of features. This may make the extracted features to be non-discriminable. For example, in image classification, if the backgrounds are similar across domains, the information of backgrounds can easily minimize $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S},\mathcal{T})$ and enhance the transferability of representations. However, background in image classification always is non-discriminable and even deteriorate UDA.

3 Cooperative and Adversarial Learning

In this section, we introduce Cooperative and Adversarial LEarning (CALE), which enhances the discriminability and transferability synchronously from an overall perspective.

CALE is a general framework that combines the optimizations of discriminability and transferability into a whole. It works upon the common domain adaptation framework that consists of three parts: a feature extractor $z = F(x)$, a discriminative module $G(z)$ for enhancing the discriminability, and a transfer module $H(z)$ for enhancing the transferability (Tzeng et al. 2014; Long et al. 2015; Xie et al. 2018; Ganin et al. 2016; Bousmalis et al. 2016). In UDA for classification, the discriminative module is typically a category classifier $\hat{y} = G(z)$ trained with supervised (and sometimes with self-training) classification loss. The transfer module is usually implemented through moment matching (Tzeng et al. 2014; Long et al. 2015) or adversarial confusing (Ganin et al. 2016; Long et al. 2018). We denote the loss of $G(z)$ as discriminability loss $\mathcal{L}_{disc}$ and the loss of $H(z)$ as transferability loss $\mathcal{L}_{tran}$. Most existing UDA approaches are generally trained by minimizing the overall loss:

$$\mathcal{L} = \mathcal{L}_{\text{disc}} + \mathcal{L}_{\text{tran}}\,. \tag{2}$$

Though discriminability loss and transferability loss are both minimized, the two modules $G(z)$ and $H(z)$ are independent, focusing on different goals during the optimization. As discussed before (Section 2), it leads to unsatisfactory feature learning. To tackle this problem, CALE (see Figure 3) provides a mechanism to bridge the discriminability module and the transferability module. Specifically, inspired by adversarial examples learning (Goodfellow, Shlens, and Szegedy 2014; Miyato et al. 2018), CALE generates cooperative (easy) examples and adversarial (hard) examples for both discriminability and transferability. Then, the easy examples are used to guide each other, and the hard ones are used to enhance their own robustness.

In the remainder of this section, we first introduce the cooperative learning (CLE) and the adversarial learning (ALE) of CALE, which together can be regarded as a general regularization framework for domain adaptation. Then, we provide an instantiated model based on the CALE framework.

3.1 Cooperative Learning

CALE leverages a cooperative learning (CLE) mechanism between the discriminative module G and the transfer module H, as shown in Figure 3. To make the discriminative module and the transfer module guide each other, we try to let the discriminative module provide a clue of the feature learning direction for the transfer module to search transferable features in a more discriminable way, and vice versa. We achieve this by generating easy examples for training each other, as shown in Figure 1.

Discriminability guides Transferability. During the training, the easy examples for the discriminative module provide a clue of more discriminable directions for the transfer module. As shown in Figure 1b, CLE generates more discriminable examples by adding cooperative (in opposite to adversarial) perturbations to the original examples, and then minimizes the consistency loss of the transfer module on such discriminable examples:

$$\begin{aligned}\ell_{\text{CLE\_tran}}(x) &= \text{Dist.}\left[H(F(x))\|H(F(x_{\text{disc}}))\right];\\ x_{\text{disc}} &= \operatorname*{argmin}_{x';\|x-x'\|\le\epsilon} \ell_{\text{disc}}(x')\,,\end{aligned} \tag{3}$$

where ℓ_{disc} is a discriminability loss function for $\mathcal{L}_{\text{disc}}$, and Dist. is a distance function for measuring the discrepancy between two distributions, e.g., KL-divergence or cross-entropy. It drives the transfer module to use more discriminable features to regularize (teach) the original features.

Transferability guides Discriminability. Similar to the previous part, CLE generates more transferable examples to provide guidance for the discriminative module, so that the discriminative module will be more likely to use transferable features for the discriminative task, i.e., Figure 1c. CLE generates transferable examples, and then minimizes the consistency loss of discriminative module on such transferable examples:

$$\begin{aligned}\ell_{\text{CLE\_disc}}(x) &= \text{Dist.}\left[G(F(x))\|G(F(x_{\text{tran}}))\right];\\ x_{\text{tran}} &= \operatorname*{argmin}_{x';\|x-x'\|\le\epsilon} \ell_{\text{tran}}(x')\,,\end{aligned} \tag{4}$$

where ℓ_{tran} is a transferability loss function for $\mathcal{L}_{\text{tran}}$.

To calculate the cooperative examples, we approximate the minimizers above by move x towards its negative gradient of the loss, i.e., $x' = x - \nabla_x \ell(x)/\|\nabla_x \ell(x)\|_2$.

CLE bridges the discriminable module G and the transferable module H by exchanging their easy examples, which makes the two modules learn each other's virtues. By learning from discriminative examples, the transfer module H pays more attention to the discriminative features when trying to align domains. Similarly, the discriminative module G focuses more on using domain invariant features for the discriminative task. Consequently, the two modules progress together to learn feature representations that are both transferable and discriminable.

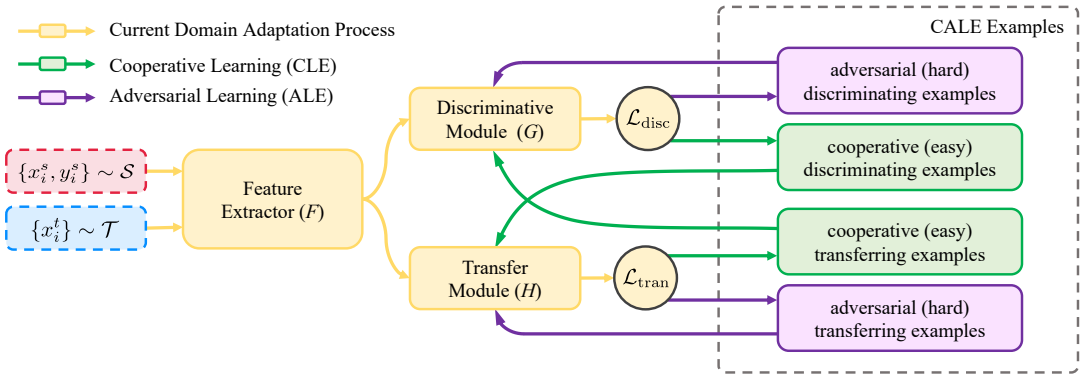


Figure 3: Cooperative and adversarial learning framework. The discriminative module G enhances the feature discriminability by minimizing a classification loss. The transfer module H enhances the feature transferability by minimizing domain distribution moments (moment matching) or maximizing an error of a domain discriminator (adversarial confusing).

3.2 Adversarial Learning

In addition to exchanging the easy examples between the transfer module H and the discriminative module G, CALE further generates hard examples for themselves to enhance the robustness of the feature learning in terms of discriminability and transferability, i.e., adversarial learning (ALE).

The adversarial regularization loss for discriminability can be formulated as:

$$\begin{aligned} \ell_{\text{ALE\_disc}}(x) &= \text{Dist.}\left[G(F(x)) \| G(F(x_{\text{non-disc}}))\right]; \\ x_{\text{non-disc}} &= \underset{x';\|x-x'\|\leq\epsilon}{\text{argmax}} \ \ell_{\text{disc}}(x'). \end{aligned} \tag{5}$$

Similarly, the adversarial regularization loss for transferability can be formulated as:

$$\begin{aligned} \ell_{ALE\_\text{tran}}(x) &= \text{Dist.}\left[H(F(x)) \| H(F(x_{\text{non-disc}}))\right]; \\ x_{\text{non-tran}} &= \underset{x';\|x-x'\|\leq\epsilon}{\text{argmax}} \ \ell_{\text{tran}}(x'). \end{aligned} \tag{6}$$

To calculate the adversarial examples, we move x towards its gradient of the losses, i.e., $x' = x + \nabla_x \ell(x) / \|\nabla_x \ell(x)\|_2$.

3.3 The CALE Model

We provide an instantiated classification domain adaptation model based on our CALE framework.

Transfer Module. The transfer module has two mainstream technologies for domain alignment: moment matching and adversarial confusing. While CALE can be applied to both of them, in this paper, we focus on the adversarial confusing. Let $H(z)$ be a binary classification network for domain discrimination, the training of model would be a min-max game between feature extractor F and domain discriminator H: H aims to distinguish the domain label of feature representation while F aims to confuse H. Formally, the transferability loss can be written as:

$$\begin{aligned} \mathcal{L}_{\text{tran}}(\theta_F, \theta_H) = \mathbb{E}_{x_i^s \sim \mathcal{D}_s} \log\left[H(F(x_i^s))\right] \\ + \mathbb{E}_{x_i^t \sim \mathcal{D}_t} \log\left[1 - H(F(x_i^t))\right]. \end{aligned} \tag{7}$$

Discriminative Module. The discriminative module is a classifier that usually minimize the classification error on source domain to enhance the discriminability of it. Besides, to enhance the discriminability of target domain, self-training has been widely adopted, e.g., Liu, Wang, and Long (2021); Kumar, Ma, and Liang (2020); Cai et al. (2021); Chen et al. (2020b). Therefore, the discriminability loss in our model is then formulated as follows:

$$\begin{aligned} \mathcal{L}_{\text{disc}}(\theta_F, \theta_G) = \mathbb{E}_{(x_i^s, y_i^s) \sim \mathcal{D}_s} \ell_{ce}\left(G(F(x_i^s)), y_i^s\right) \\ + \mathbb{E}_{x_i^t \sim \mathcal{D}_t} \mathbb{I}_{[\max(\hat{y}_i^t) \geq \tau]} \ell_{ce}\left(G(F(x_i^t)), \text{PL}(\hat{y}_i^t)\right) \end{aligned} \tag{8}$$

Where $\ell_{ce}(p, q) = -\sum_i q_i \log(p_i)$ is the cross-entropy loss function, θ_G means the parameters of the category classifier G, and $\text{PL}(\hat{y}) = \text{onehot}(\text{argmax}(\hat{y}))$ means the pseudo label of prediction.

Unified Cooperative and Adversarial Learning. Specifically, cooperative regularization losses for discriminability and transferability are calculated by:

$$\begin{aligned} \ell_{\text{CLE\_disc}}(x) &= \text{Dist.}\left[G(F(x)) \| G(F(x_{tran}))\right]; \\ x_{\text{tran}} &= \underset{x';\|x-x'\|\leq\epsilon}{\text{argmin}} \ \ell_{ce}(H(F(x')), H(F(x))), \\ \ell_{\text{CLE\_tran}}(x) &= \text{Dist.}\left[H(F(x)) \| H(F(x_{disc}))\right]; \\ x_{\text{disc}} &= \underset{x';\|x-x'\|\leq\epsilon}{\text{argmin}} \ \ell_{ce}(G(F(x')), G(F(x))). \end{aligned} \tag{9}$$

Moreover, adversarial regularization losses are:

$$\begin{aligned} \ell_{\text{ALE\_disc}}(x) &= \text{Dist.}\left[G(F(x)) \| G(F(x_{\text{non-disc}}))\right]; \\ x_{\text{non-disc}} &= \underset{x';\|x-x'\|\leq\epsilon}{\text{argmax}} \ \ell_{ce}(G(F(x')), G(F(x))). \\ \ell_{\text{ALE\_tran}}(x) &= \text{Dist.}\left[H(F(x)) \| H(F(x_{\text{non-tran}}))\right]; \\ x_{\text{non-tran}} &= \underset{x';\|x-x'\|\leq\epsilon}{\text{argmax}} \ \ell_{ce}(H(F(x')), H(F(x))). \end{aligned} \tag{10}$$

Hence, the cooperative and adversarial regularization terms for discriminability are summarized as:

$$\begin{aligned} \mathcal{R}_{\text{disc}}(\theta_F, \theta_G) = \mathbb{E}_{x_i \sim \mathcal{D}_s \cup \mathcal{D}_t} \left[\ell_{\text{CLE\_disc}}(x_i; \theta_F, \theta_G) \right. \\ \left. + \ell_{\text{ALE\_disc}}(x_i; \theta_F, \theta_G)\right], \end{aligned} \tag{11}$$

Algorithm 1: Cooperative and Adversarial Learning (CALE)

Require: Source domain dataset $\mathcal{D}_s = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$ and target domain dataset $\mathcal{D}_t = \{x_i^t\}_{i=1}^{n_t}$.
Ensure: A model with parameters $\theta = (\theta_F, \theta_G, \theta_H)$
1: Initialize parameters $\theta = (\theta_F, \theta_G, \theta_H)$ randomly.
2: **for** $iter = 0$ to $MaxIteration$ **do**
3: Calculate the transferability loss $\mathcal{L}_{\text{tran}}$ and discriminability loss $\mathcal{L}_{\text{disc}}$ (Equ 7 and 8).
4: Calculate cooperative (easy) examples $\{x_{\text{disc}}\}$ and $\{x_{\text{tran}}\}$ (Equ 9).
5: Calculate adversarial (hard) examples $\{x_{\text{non-disc}}\}$ and $\{x_{\text{non-tran}}\}$ (Equ 10).
6: Calculate discriminability reg. $\mathcal{R}_{\text{disc}}(\theta_F, \theta_G)$ with $\{x_{\text{tran}}\}$ and $\{x_{\text{non-disc}}\}$ (Equ 11).
7: Calculate transferability reg. $\mathcal{R}_{\text{tran}}(\theta_F, \theta_H)$ with $\{x_{\text{disc}}\}$ and $\{x_{\text{non-tran}}\}$ (Equ 12).
8: Updating parameters by gradient backpropagation: $\theta \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}$ (Equ 13).

and the cooperative and adversarial regularization terms for transferability are summarized as:

$$\begin{aligned}\mathcal{R}_{\text{tran}}(\theta_F, \theta_H) =& \mathbb{E}_{x_i \sim \mathcal{D}_s \cup \mathcal{D}_t} \big[\ell_{\text{CLE\_tran}}(x_i; \theta_F, \theta_H) \\ &+ \ell_{\text{ALE\_tran}}(x_i; \theta_F, \theta_H) \big] .\end{aligned} \tag{12}$$

To sum up, the min-max paradigm can be written as

$$\begin{aligned}\min_{\theta_F, \theta_G} \max_{\theta_H} & \mathcal{L}_{\text{disc}}(\theta_F, \theta_G) + \mathcal{L}_{\text{tran}}(\theta_F, \theta_H) \\ &+ \lambda \left[\mathcal{R}_{\text{disc}}(\theta_F, \theta_G) + \mathcal{R}_{\text{tran}}(\theta_F, \theta_H) \right] ,\end{aligned} \tag{13}$$

and the pseudocode is shown in Algorithm 1.

4 Experiments

We conduct extensive experiments to show the effectiveness of the proposed CALE framework.

4.1 Experimental Protocol

Datasets. We evaluate the CALE model and the baselines on three benchmark visual datasets: **Office-31** contains 4110 images from 31 categories of three distant domains, including Amazon (**A**), Webcam (**W**), and DSLR (**D**). **Office-Home**, a more challenging dataset, consists of 15588 images of 65 object classes in office and home environments, forming four extremely dissimilar domains: Artistic (**Ar**), Clip Art (**Cl**), Product (**Pr**), and Real World (**Rw**). **VisDA-2017**, a large dataset with 152397 **Synthetic** 3D rendered images and 55388 **Real-world** photos across 12 categories.

Baselines. We compare CALE with state-of-the-art deep domain adaptation approaches: Deep Adaptation Network (DAN) (Long et al. 2015), Domain Adversarial Neural Network (DANN) (Ganin et al. 2016), Conditional Domain Adversarial Network (CDAN) (Long et al. 2018), Maximum Classifier Discrepancy (MCD) (Saito et al. 2018b), Minimum Class Confusion (MCC) (Jin et al. 2019), Transferable Attention for Domain Adaptation (TADA) (Wang et al. 2019a), Transferable Adversarial Training (TAT) (Liu et al. 2019), Batch Spectral Penalization (BSP) (Chen et al. 2019b), and Cycle Self-Training (CST) (Liu, Wang, and Long 2021). Following the original implementation, all methods are with the ResNet (He et al. 2016) as their backbone. Due to the success of the Transformer (Vaswani et al. 2017; Dosovitskiy et al. 2021), we further compare CALE with other baselines that use Data-efficient image Transformers (DeiT) (Touvron et al. 2021) backbone, including CDTrans (Xu et al. 2021) and our implemented DeiT version of some representative baselines from above.

Implementation Details. We employ ResNet (He et al. 2016) and DeiT (Touvron et al. 2021) pretrained on ImageNet1K (Deng et al. 2009) as the backbone and attach a bottleneck layer with 256 units as the feature extractor F. We use a single fully-connected layer as the discriminative module G, and a three-layered conditional domain discriminator (Long et al. 2018) as the transfer module H. The network architecture is commonly used architecture of several recent domain adaptation approaches (e.g. (Long et al. 2018; Chen et al. 2019b; Jin et al. 2019; Jiang et al. 2020)) for fair comparison. We adopt SGD with the learning rate annealed from 0.01 for training like Long et al. (2018); Jiang et al. (2020), and leverage FixMatch (Sohn et al. 2020) for self-trining in $\mathcal{L}_{disc}$, which has been widely adopted in prior UDA work (e.g., (Liu, Wang, and Long 2021)). Throughout the experiments, the trade-off parameter of CALE regularization λ is set to 1, and the self-training threshold τ in Equation 8 is set to 0.95. The distance function Dist. is set as cross-entropy in *Office-31* and *Office-Home*, KL-divergence in *VisDA-2017*. More details about code and datasets can be found at https://github.com/sunh-23/CALE.

4.2 Results and Analysis

Comparison with state-of-the-art. Experimental results on *Office-Home* and *VisDA-2017* are shown in Table 1, and Table 2, respectively. Due to space limitations, the results on the *Office-31*, the simplest one among three datasets, are reported in appendix. The bolded and underlined numbers denote the best and the second best performance. The results marked with * are based on our reproduction, while all the other results are from their original paper. The results reveal several insightful observations: (1) With both backbones (ResNet and DeiT), CALE outperforms the compared SOTA approaches. Especially on the challenging datasets that have large domain shift: *OfficeHome* (ResNet-50: +1.8%, DeiT-Base: +1.0%) and *VisDA-2017* (ResNet-101: +3.0%, DeiT-Base: +2.4%), the enhancements are significant. (2) TADA and TAT try to let the transfer module asymmetrically guide the discriminative module. BSP attempts to penalize the transferability for enhancing discriminability heuristically. CALE outperforms TADA, TAT, and BSP, which verifies that the proposed idea of bidirectional co-enhancing the discriminability and transferability is helpful for building a better domain adaptation model, compared to the unidirectional (TADA and TAT) or imbalanced (BSP) enhancement of discriminability or transferability.

Ablation Study. We conduct numerous ablation studies to validate the individual contributions of different compo-

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	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	43.6	57.0	67.9	45.8	56.5	60.4	44.0	43.6	67.7	63.1	51.5	74.3	56.3
DANN	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
CDAN	49.0	69.3	74.5	54.4	66.0	68.4	55.6	48.3	75.9	68.4	55.4	80.5	63.8
MCD*	51.6	72.7	77.6	62.5	68.6	70.4	62.7	52.1	78.2	74.4	57.9	82.2	67.6
MCC*	57.7	79.3	82.8	66.7	76.5	77.8	67.2	55.1	81.5	74.4	61.0	85.9	72.2
CST	59.0	**79.6**	**83.4**	68.4	77.1	76.7	68.9	56.4	83.0	75.3	62.2	85.1	73.0
TADA	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
TAT	51.6	69.5	75.4	59.4	69.5	68.6	59.5	50.5	76.8	70.9	56.6	81.6	65.8
BSP	52.0	68.6	76.1	58.0	70.3	70.2	58.6	50.2	77.6	72.2	59.3	81.9	66.3
CALE	**65.1**	75.3	80.8	**68.7**	**80.2**	**78.4**	**69.7**	**64.5**	**83.3**	**76.0**	**68.0**	**87.6**	**74.8**
DeiT-Base*	54.1	76.4	83.0	66.5	76.3	77.5	65.4	48.0	81.9	72.9	53.2	84.2	70.0
DAN*	56.7	76.0	83.2	68.4	75.7	78.6	66.3	50.6	81.3	74.8	56.4	84.7	71.1
DANN*	61.0	72.2	82.0	69.8	75.7	78.2	67.5	62.6	84.9	78.0	64.8	87.6	73.7
MCD*	60.2	77.8	83.9	72.4	73.2	75.5	68.6	59.2	82.8	80.7	62.3	86.4	73.6
MCC*	64.2	**85.8**	87.3	77.8	83.7	85.6	75.2	60.4	86.7	79.9	63.5	89.8	78.3
CST*	65.9	85.3	**88.0**	76.5	81.2	85.6	75.0	52.1	87.0	78.5	60.7	90.1	77.2
CDTrans	68.8	85.0	86.9	**81.5**	**87.1**	**87.3**	**79.6**	63.3	**88.2**	82.0	66.0	**90.6**	80.5
CALE	**71.5**	84.1	87.6	78.4	86.3	85.3	79.2	**70.7**	87.7	**82.4**	**74.2**	90.4	**81.5**

Table 1: Classification accuracy (%) on Office-Home with ResNet-50 (upper) and DeiT-Base (lower).

Method	Plane	Bicycle	Bus	Car	Horse	Knife	Motor	Person	Plant	Ski	Train	Truck	Avg
ResNet-101	72.3	6.1	63.4	91.7	52.7	7.9	80.1	5.6	90.1	18.5	78.1	25.9	49.4
DAN	68.1	15.4	76.5	87.0	71.1	48.9	82.3	51.5	88.7	33.2	88.9	42.2	62.8
DANN	81.9	77.7	82.8	44.3	81.2	29.5	65.1	28.6	51.9	54.6	82.8	7.8	57.4
CDAN	93.6	82.3	66.2	80.6	92.7	10.7	87.1	70.0	94.6	38.4	76.6	47.2	70.0
MCD	87.0	60.9	83.7	64.0	88.9	79.6	84.7	76.9	88.6	40.3	83.0	25.8	71.9
MCC	94.5	80.8	78.4	65.3	90.6	79.4	87.5	82.2	94.7	81.0	86.0	44.6	80.4
CST*	96.2	**89.9**	73.3	**92.7**	94.8	97.0	81.5	79.4	**96.2**	87.2	83.3	47.6	84.9
CALE	**97.3**	88.2	**87.6**	74.7	**96.2**	**97.9**	**92.5**	**84.1**	94.5	**92.9**	**91.4**	**58.0**	**87.9**
DeiT-Base*	97.4	56.3	75.2	47.2	80.5	43.0	92.3	6.1	69.0	51.9	91.5	27.1	61.5
DAN*	97.2	50.3	83.3	46.2	93.5	77.7	95.4	18.1	87.2	69.8	94.4	26.6	70.0
DANN*	95.5	72.7	81.9	41.5	86.3	41.8	88.5	75.6	85.2	75.8	92.7	40.4	73.2
MCD*	94.7	82.6	66.4	77.9	86.1	97.0	92.5	73.6	95.0	27.0	89.2	50.0	77.7
MCC*	98.1	90.7	81.9	76.5	96.0	97.8	90.8	55.1	95.1	86.1	93.0	**63.9**	85.4
CST*	98.1	**93.5**	81.3	**90.5**	**97.6**	96.7	90.5	62.3	**98.0**	93.8	92.2	59.6	87.8
CDTrans	97.1	90.5	82.4	77.5	96.6	96.1	93.6	**88.6**	97.9	86.9	90.3	62.8	88.4
CALE	**98.7**	89.9	**88.0**	87.8	97.4	**98.5**	**96.0**	86.5	97.7	**95.5**	**95.4**	58.5	**90.8**

Table 2: Classification accuracy (%) on VisDA-2017 with ResNet-101 (upper) and DeiT-Base (lower).

nents. The ablation results are reported in Table 3, where $\mathcal{L}_{\text{tran}}$ denotes the transferability loss in Equation 7, *CALE* denotes the CALE regularization terms, and *ST* denotes the self-training term in Equation 8. In rows 4 and 5, we split the CALE regularization into the transferability regularization $\mathcal{R}_{\text{tran}}$ and the discriminability regularization $\mathcal{R}_{\text{disc}}$ as shown in the Equation 13. Compare rows 4, 5, and 8, it shows that both $\mathcal{R}_{\text{tran}}$ and $\mathcal{R}_{\text{disc}}$ are necessary, especially enhancing transferability through $\mathcal{R}_{\text{tran}}$ is critical. Moreover, in rows 5 and 6, we split the CALE regularization into cooperative learning (CLE) and adversarial learning (ALE) terms. The comparison of rows 5, 6, and 8 verifies that both the CLE and the ALE mechanisms are effective, and the enhancement from CLE is more remarkable. Last, it is worth noting that the result of row 8, our method is still the best on *Office-Home* even without the help of self-training.

Discriminability and Transferability. Figure 2 shows the discriminability of target domain and the transferability across domains, learned by CALE and other UDA approaches. We use test accuracy to quantify the discriminability on target domain. For transferability, we adopt $\mathcal{A}$-distance as a (negative) indicator. Ben-David et al. (2006) introduces $\mathcal{A}$-distance $= 2(1 - 2\epsilon)$ to measure the cross-domain discrepancy, where ϵ means the test error of a binary classifier to distinguish different domains on the learned feature representation. As shown in Figure 2, CALE is the only

	$\mathcal{L}_{\text{tran}}$	*CALE*		*ST*	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
1	-	-		-	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
2	-	-		✓	51.0	66.4	75.9	48.0	65.0	64.5	55.3	46.2	77.1	70.9	55.3	79.5	62.9
3	✓	-		-	49.0	69.3	74.5	54.4	66.0	68.4	55.6	48.3	75.9	68.4	55.4	80.5	63.8
		$\mathcal{R}_{\text{tran}}$	$\mathcal{R}_{\text{disc}}$														
4	✓	-	✓	-	55.5	66.1	71.7	57.9	72.1	71.5	58.4	55.3	76.2	66.7	61.9	82.3	66.3
5	✓	✓	-	-	60.3	71.7	77.1	64.8	76.9	74.2	66.4	61.1	79.7	74.2	65.1	84.9	71.4
		CLE	*ALE*														
6	✓	-	✓	-	61.6	72.7	77.7	65.4	77.4	76.3	62.8	59.4	80.4	72.4	65.3	85.0	71.4
7	✓	✓	-	-	62.5	74.9	78.6	65.3	77.9	76.7	65.4	60.3	80.8	73.7	65.7	85.6	72.3
8	✓	✓		-	63.2	75.9	79.9	68.1	79.3	77.3	68.4	64.0	82.1	74.8	67.1	86.7	73.9
9	✓	✓		✓	65.1	75.3	80.8	68.7	80.2	78.4	69.7	64.5	83.3	76.0	68.0	87.6	74.8

Table 3: Results of ablation study on Office-Home (ResNet-50).

Pareto optimal among compared approaches. It indicates that CALE can learn more discriminative features for the target, and transfer more knowledge from source to target.

In addition, hyper-parameter sensitivity analysis and feature distribution visualization are reported in the appendix.

5 Related Works

Domain adaptation aims to overcome the *domain-shift* for transferring knowledge from a label-rich source domain to a label-poor target domain (Pan and Yang 2010). In the early shallow regime, Bickel, Brückner, and Scheffer (2007); Sugiyama et al. (2007) re-weight source data for reusing in the target domain, Wang and Mahadevan (2008); Blitzer, McDonald, and Pereira (2006) learn representations invariant across domains, while Gao et al. (2008) transfer model based on shared parameters and Mihalkova, Huynh, and Mooney (2007) based on relational-knowledge.

The theoretical analysis of Ben-David et al. (2010) and the success of deep neural networks inspire the modern UDA approaches. Two mainstream technologies were rised: moment matching and adversarial confusing. The moment matching approaches design and minimize the cross-domain distribution distance based on the statistical moment. Sun and Saenko (2016) matches the second-order statistical moment (covariance) in shared feature space, while Li et al. (2017); Maria Carlucci et al. (2017); Mancini et al. (2019) match the first-order (mean) and the second-order (variance) statistical moments in the Batch Normalization layer (Ioffe and Szegedy 2015). Tzeng et al. (2014); Long et al. (2015, 2017) align the cross-domain means in Reproducing Kernel Hilbert Space (RKHS) based on Maximum Mean Discrepancy (MMD) (Gretton et al. 2012) and its variants. Motivated by GANs (Goodfellow et al. 2014), the adversarial confusing approaches learn domain-invariant feature representations by training the feature extractor and the domain discriminator in an adversarial way. Ganin et al. (2016) leverages Gradient Reverse Layer (GRL) to perform one-step optimization of adversarial min-max problem while Tzeng et al. (2017) learns asymmetric feature extractors for source domain and target domain. Long et al. (2018) builds a conditional domain discriminator based on Conditional-GANs (Mirza and Osindero 2014). Later, Pei et al. (2018); Saito et al. (2018a,b) propose various improvements for pursuing more fine-grained alignment. Most recently, Hoffman et al. (2018); Chen et al. (2020a) adopt Cycle-GANs (Zhu et al. 2017) to achieve impressive performances.

How to balance and consolidate the enhancement of discriminability and transferability in feature learning is getting more attention from researchers. Chen et al. (2019b) attempt to penalize the transferability for enhancing discriminability heuristically. They consider learning transferability too much may harm the discriminability. Wang et al. (2019a); Kurmi, Kumar, and Namboodiri (2019); Liu et al. (2019) try to let the transfer module asymmetrically guide the discriminative module, while Liang et al. (2021) designs a target-classification-mimicking loss to leverage the target predictions. Moreover, Wei et al. (2021) model the two optimizations in a meta-learning scheme.

This paper further argues that the guidance between the transfer module and the discriminative module can be reciprocal through cooperative samples. The proposed CALE model further extends the transferability and discriminability that the existing feature learning approaches can reach.

6 Conclusions

Discriminability and transferability are two goals of feature learning for domain adaptation, as we aim to find the transferable features from the source domain that are helpful for discriminating in the target domain. We claim that the enhancement of discriminability and transferability should be considered jointly instead of separately. We propose Cooperative and Adversarial LEarning (CALE) for achieving the two goals simultaneously by letting the transfer module and the discriminative module assist each other. Besides, We demonstrate the benefits of using one module to guide another and suggest that there may be different ways to unify the learning of discriminability and transferability, which could be explored in future research.

Acknowledgements

This research was supported by NSFC (62076121, 61921 006).

References

Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; and Vaughan, J. W. 2010. A theory of learning from different domains. *Machine Learning*, 79(1): 151–175.

Ben-David, S.; Blitzer, J.; Crammer, K.; and Pereira, F. 2006. Analysis of Representations for Domain Adaptation. In *Advances in Neural Information Processing Systems*.

Bickel, S.; Brückner, M.; and Scheffer, T. 2007. Discriminative learning for differing training and test distributions. In *Proceedings of the International Conference on Machine Learning*, 81–88.

Blitzer, J.; McDonald, R.; and Pereira, F. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 120–128.

Bousmalis, K.; Trigeorgis, G.; Silberman, N.; Krishnan, D.; and Erhan, D. 2016. Domain separation networks. In *Advances in Neural Information Processing Systems*, 343–351.

Cai, T.; Gao, R.; Lee, J.; and Lei, Q. 2021. A theory of label propagation for subpopulation shift. In *Proceedings of the International Conference on Machine Learning*, 1170–1182. PMLR.

Chen, C.; Xie, W.; Wen, Y.; Huang, Y.; and Ding, X. 2020a. Multiple-source domain adaptation with generative adversarial nets. *Knowledge-Based Systems*, 199: 105962.

Chen, X.; Wang, S.; Fu, B.; Long, M.; and Wang, J. 2019a. Catastrophic forgetting meets negative transfer: Batch spectral shrinkage for safe transfer learning. In *Advances in Neural Information Processing Systems*, volume 32.

Chen, X.; Wang, S.; Long, M.; and Wang, J. 2019b. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In *Proceedings of the International Conference on Machine Learning*, 1081–1090. PMLR.

Chen, X.-H.; Jiang, S.; Xu, F.; Zhang, Z.; and Yu, Y. 2021. Cross-modal Domain Adaptation for Cost-Efficient Visual Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 34.

Chen, Y.; Wei, C.; Kumar, A.; and Ma, T. 2020b. Self-training avoids using spurious features under domain shift. *Advances in Neural Information Processing Systems*, 33: 21061–21071.

Chen, Z.; Xiao, R.; Li, C.; Ye, G.; Sun, H.; and Deng, H. 2020c. ESAM: Discriminative Domain Adaptation with Non-Displayed Items to Improve Long-Tail Performance. In *International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 579–588.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference Computer Vision Pattern Recognition*, 248–255. IEEE.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1): 1–35.

Gao, J.; Fan, W.; Jiang, J.; and Han, J. 2008. Knowledge transfer via multiple model local structure mapping. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 283–291.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27.

Goodfellow, I.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Gretton, A.; Borgwardt, K. M.; Rasch, M. J.; Schölkopf, B.; and Smola, A. 2012. A kernel two-sample test. *The Journal of Machine Learning Research*, 13: 723–773.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference Computer Vision Pattern Recognition*, 770–778. IEEE.

Hoffman, J.; Tzeng, E.; Park, T.; Zhu, J.-Y.; Isola, P.; Saenko, K.; Efros, A.; and Darrell, T. 2018. Cycada: Cycle-consistent adversarial domain adaptation. In *Proceedings of the International Conference on Machine Learning*, 1989–1998. PMLR.

Ioffe, S.; and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the International Conference on Machine Learning*, 448–456. PMLR.

Jiang, J.; Chen, B.; Fu, B.; and Long, M. 2020. Transfer-Learning-library. https://github.com/thuml/Transfer-Learning-Library. Accessed: 2022-08-15.

Jin, Y.; Wang, X.; Long, M.; and Wang, J. 2019. Minimum Class Confusion for Versatile Domain Adaptation. In *Proceedings of the European Conference on Computer Vision*, 464–480.

Kanagawa, H.; Kobayashi, H.; Shimizu, N.; Tagami, Y.; and Suzuki, T. 2019. Cross-domain recommendation via deep domain adaptation. In *European Conference on Information Retrieval (ECIR)*, 20–29. Springer.

Kontar, R.; Raskutti, G.; and Zhou, S. 2020. Minimizing negative transfer of knowledge in multivariate gaussian processes: A scalable and regularized approach. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 43(10): 3508–3522.

Kumar, A.; Ma, T.; and Liang, P. 2020. Understanding self-training for gradual domain adaptation. In *Proceedings of the International Conference on Machine Learning*, 5468–5479. PMLR.

Kurmi, V. K.; Kumar, S.; and Namboodiri, V. P. 2019. Attending to discriminative certainty for domain adaptation. In *Proceedings of the IEEE Conference Computer Vision Pattern Recognition*, 491–500. IEEE.

Lekhtman, E.; Ziser, Y.; and Reichart, R. 2021. DILBERT: Customized Pre-Training for Domain Adaptation with Category Shift, with an Application to Aspect Extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 219–230.

Li, X.; and Caragea, D. 2020. Domain adaptation with reconstruction for disaster tweet classification. In *International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 1561–1564.

Li, Y.; Wang, N.; Shi, J.; Liu, J.; and Hou, X. 2017. Revisiting Batch Normalization For Practical Domain Adaptation. In *Proceedings of the International Conference on Learning Representations*. OpenReview.net.

Liang, J.; Gong, K.; Li, S.; Liu, C. H.; Li, H.; Liu, D.; and Wang, G. 2021. Pareto domain adaptation. *Advances in Neural Information Processing Systems*, 34: 12917–12929.

Liu, H.; Long, M.; Wang, J.; and Jordan, M. I. 2019. Transferable adversarial training: A general approach to adapting deep classifiers. In *Proceedings of the International Conference on Machine Learning*, 4013–4022. PMLR.

Liu, H.; Wang, J.; and Long, M. 2021. Cycle Self-Training for Domain Adaptation. In *Advances in Neural Information Processing Systems*.

Liu, Y.; Zhang, W.; and Wang, J. 2021. Source-free domain adaptation for semantic segmentation. In *Proceedings of the IEEE Conference Computer Vision Pattern Recognition*, 1215–1224. IEEE.

Long, M.; Cao, Y.; Wang, J.; and Jordan, M. I. 2015. Learning transferable features with deep adaptation networks. In *Proceedings of the International Conference on Machine Learning*, 97–105. PMLR.

Long, M.; Cao, Z.; Wang, J.; and Jordan, M. I. 2018. Conditional Adversarial Domain Adaptation. In *Advances in Neural Information Processing Systems*, 1647–1657.

Long, M.; Zhu, H.; Wang, J.; and Jordan, M. I. 2017. Deep transfer learning with joint adaptation networks. In *Proceedings of the International Conference on Machine Learning*, 2208–2217. PMLR.

Mancini, M.; Porzi, L.; Bulo, S. R.; Caputo, B.; and Ricci, E. 2019. Inferring latent domains for unsupervised deep domain adaptation. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 43(2): 485–498.

Maria Carlucci, F.; Porzi, L.; Caputo, B.; Ricci, E.; and Rota Bulo, S. 2017. Autodial: Automatic domain alignment layers. In *Proceedings of the IEEE International Conference on Computer Vision*, 5077–5085. IEEE.

Mihalkova, L.; Huynh, T.; and Mooney, R. J. 2007. Mapping and revising markov logic networks for transfer learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 608–614. IEEE.

Mirza, M.; and Osindero, S. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Miyato, T.; Maeda, S.-i.; Koyama, M.; and Ishii, S. 2018. Virtual Adversarial Training : A Regularization Method for Supervised and. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 41(8): 1979–1993.

Pan, S. J.; and Yang, Q. 2010. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10): 1345–1359.

Pei, Z.; Cao, Z.; Long, M.; and Wang, J. 2018. Multi-Adversarial Domain Adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 3934–3941. IEEE.

Rao, K.; Harris, C.; Irpan, A.; Levine, S.; Ibarz, J.; and Khansari, M. 2020. Rl-cyclegan: Reinforcement learning aware simulation-to-real. In *Proceedings of the IEEE Conference Computer Vision Pattern Recognition*, 11157–11166. IEEE.

Saito, K.; Ushiku, Y.; Harada, T.; and Saenko, K. 2018a. Adversarial dropout regularization. In *Proceedings of the International Conference on Learning Representations*. OpenReview.net.

Saito, K.; Watanabe, K.; Ushiku, Y.; and Harada, T. 2018b. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference Computer Vision Pattern Recognition*, 3723–3732. IEEE.

Shu, R.; Bui, H. H.; Narui, H.; and Ermon, S. 2018. A DIRT-T Approach to Unsupervised Domain Adaptation.

Sohn, K.; Berthelot, D.; Li, C. L.; Zhang, Z.; Carlini, N.; Cubuk, E. D.; Kurakin, A.; Zhang, H.; and Raffel, C. 2020. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*.

Sugiyama, M.; Nakajima, S.; Kashima, H.; Buenau, P.; and Kawanabe, M. 2007. Direct importance estimation with model selection and its application to covariate shift adaptation. *Advances in Neural Information Processing Systems*, 20.

Sun, B.; and Saenko, K. 2016. Deep coral: Correlation alignment for deep domain adaptation. In *Proceedings of the European Conference on Computer Vision*, 443–450. Springer.

Tan, C.; Sun, F.; Kong, T.; Zhang, W.; Yang, C.; and Liu, C. 2018. A survey on deep transfer learning. In *Proceedings of the IEEE International Conference on Artificial Neural Networks*, 270–279. Springer.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *Proceedings of the International Conference on Machine Learning*, 10347–10357. PMLR.

Tzeng, E.; Hoffman, J.; Saenko, K.; and Darrell, T. 2017. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference Computer Vision Pattern Recognition*, 7167–7176. IEEE.

Tzeng, E.; Hoffman, J.; Zhang, N.; Saenko, K.; and Darrell, T. 2014. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *The Journal of Machine Learning Research*, 9(11).

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Wang, C.; and Mahadevan, S. 2008. Manifold alignment using procrustes analysis. In *Proceedings of the International Conference on Machine Learning*, 1120–1127. PMLR.

Wang, X.; Li, L.; Ye, W.; Long, M.; and Wang, J. 2019a. Transferable attention for domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 5345–5352. IEEE.

Wang, Z.; Dai, Z.; Póczos, B.; and Carbonell, J. 2019b. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE Conference Computer Vision Pattern Recognition*, 11293–11302. IEEE.

Wei, G.; Lan, C.; Zeng, W.; and Chen, Z. 2021. Metaalign: Coordinating domain alignment and classification for unsupervised domain adaptation. In *Proceedings of the IEEE Conference Computer Vision Pattern Recognition*, 16643–16653. IEEE.

Xie, S.; Zheng, Z.; Chen, L.; and Chen, C. 2018. Learning semantic representations for unsupervised domain adaptation. In *Proceedings of the International Conference on Machine Learning*, 5423–5432. PMLR.

Xu, T.; Chen, W.; Wang, P.; Wang, F.; Li, H.; and Jin, R. 2021. CDTrans: Cross-domain transformer for unsupervised domain adaptation. *arXiv preprint arXiv:2109.06165*.

Zhu, H.; Wang, Z.; Zhang, H.; Liu, M.; Zhao, S.; and Qin, B. 2021. Less Is More: Domain Adaptation with Lottery Ticket for Reading Comprehension. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1102–1113.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2223–2232. IEEE.

A Appendix

A.1 Comparison with State-of-the-art on Office-31

We here report the experimental results on *Office-31* in Table 4. It can be observed in the table that CALE achieves performance improvement with both ResNet and DeiT backbone.

A.2 Hyper-Parameter Sensitivity of Performance and Convergence

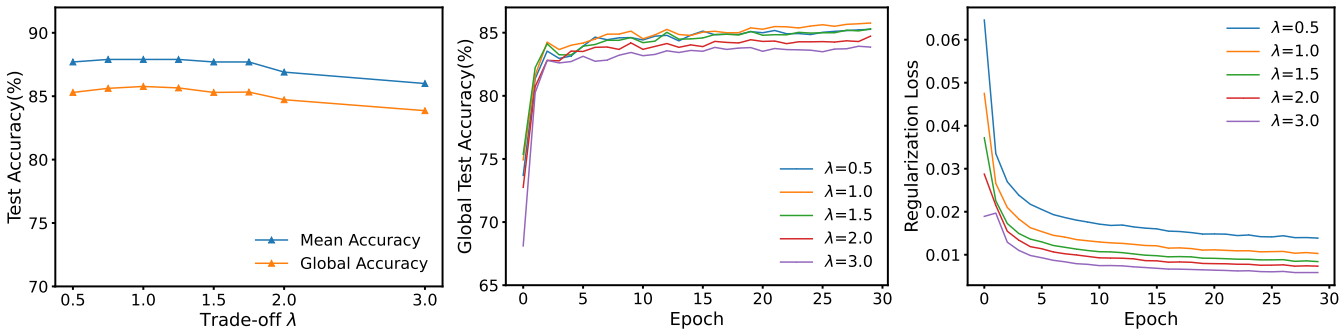
We study the sensitivity of the hyper-parameter λ in Equation 13, which determines the strength of the CALE regularization loss. The results of test accuracy on *VisDA-2017* are shown in Figure 4a, which reveal that global accuracy and class mean accuracy are stable with the various value of λ. Besides, we investigate the convergence of test global accuracy and the CALE regularization loss when setting different values of λ. The results on the convergence are shown in Figure 4b and Figure 4c.

A.3 Feature Visualization

To present the adaptation process intuitively, we leverage t-SNE (Van der Maaten and Hinton 2008) to visualize the feature representations of *VisDA-2017*. The results are shown in Figure 5, where different colors mean different categories, and circles and triangles denote source data and target data, respectively. Apparently, in Figure 5a, the feature of target before adaptation is non-discriminative and non-transferable. However, in Figure 5b, the target's feature adapted by CALE is discriminative (target features that come from different categories can be separated easily) and transferable (source feature and target feature are confused).

Method	$A \to D$	$A \to W$	$D \to A$	$D \to W$	$W \to A$	$W \to D$	Avg
ResNet-50	68.9	68.4	62.5	96.7	60.7	99.3	76.1
DAN	78.6	80.5	63.6	97.1	62.8	99.6	80.4
DANN	79.7	82.0	68.2	96.9	67.4	99.1	82.2
CDAN	89.8	93.1	70.1	98.2	68.0	**100.0**	86.6
MCD*	91.4	89.2	70.0	98.2	68.5	**100.0**	86.2
MCC	**95.0**	**94.7**	73.0	98.6	73.6	**100.0**	89.2
CST*	94.9	85.0	75.6	99.2	73.3	99.9	88.0
TADA	91.6	94.3	72.9	98.7	73.0	99.8	88.4
TAT	93.2	92.5	73.1	**99.3**	72.1	**100.0**	88.4
BSP	93.0	93.3	73.6	98.2	72.6	**100.0**	88.5
CALE	92.8	91.6	**77.0**	98.6	**76.9**	**100.0**	**89.5**
DeiT-Base*	85.5	88.1	74.8	98.9	75.9	**100.0**	87.2
DAN*	89.2	91.8	77.4	98.9	75.9	**100.0**	88.9
DANN*	87.7	93.8	80.1	98.3	79.6	**100.0**	89.9
MCD*	94.7	95.1	72.9	98.6	73.8	99.8	89.2
MCC*	97.0	96.5	80.5	98.9	80.6	**100.0**	92.3
CST*	96.6	94.2	78.2	**99.2**	79.7	**100.0**	91.3
CDTrans	97.0	96.7	81.1	99.0	**81.9**	**100.0**	92.6
CALE	**97.6**	**98.0**	**82.4**	99.0	80.7	**100.0**	**93.0**

Table 4: Classification accuracy (%) on Office-31 with ResNet-50 (upper) and DeiT-Base (lower).

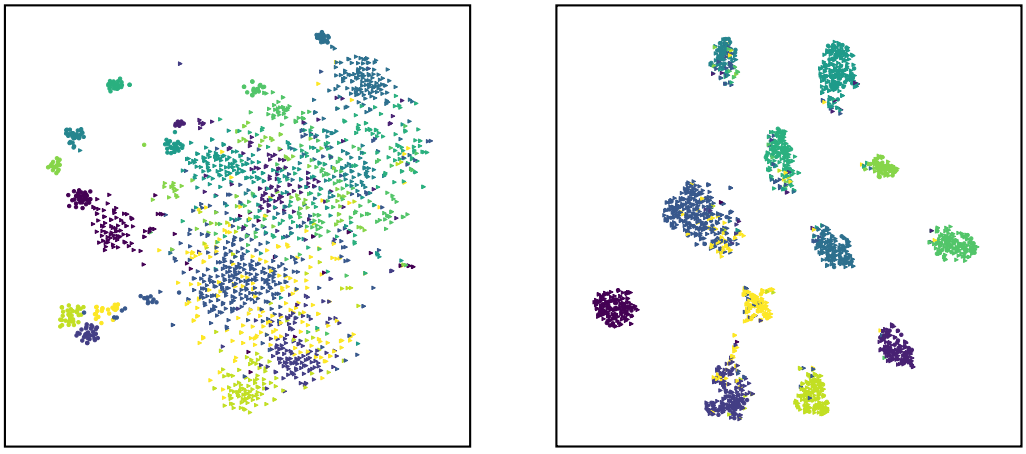


(a) Test global accuracy and class mean accuracy w.r.t. λ.

(b) Convergence of the global accuracy with various λ.

(c) Convergence of the regularization loss with various λ.

Figure 4: Hyper-parameter sensitivity of classification performance and convergence.



(a) Before adaptation.

(b) After CALE adaptation.

Figure 5: Feature visualizations.