Transferable Adversarial Training: A General Approach to Adapting Deep Classifiers

Hong Liu [1 2] **Mingsheng Long** [1 3] **Jianmin Wang** [1 3] **Michael I. Jordan** [4]

Abstract

Domain adaptation enables knowledge transfer from a labeled source domain to an unlabeled target domain. A mainstream approach is adversarial feature adaptation, which learns domain-invariant representations through aligning the feature distributions of both domains. However, a theoretical prerequisite of domain adaptation is the *adaptability* measured by the expected risk of an ideal joint hypothesis over the source and target domains. In this respect, adversarial feature adaptation may potentially deteriorate the adaptability, since it distorts the original feature distributions when suppressing domain-specific variations. To this end, we propose *Transferable Adversarial Training* (TAT) to enable the adaptation of deep classifiers. The approach generates transferable examples to fill in the gap between the source and target domains, and adversarially trains the deep classifiers to make consistent predictions over the transferable examples. Without learning domain-invariant representations at the expense of distorting the feature distributions, the adaptability in the theoretical learning bound is algorithmically guaranteed. A series of experiments validate that our approach advances the state of the arts on a variety of domain adaptation tasks in vision and NLP, including object recognition, learning from synthetic to real data, and sentiment classification.

1. Introduction

Transferring knowledge from a source domain with sufficient supervision to an unlabeled target domain is advantageous, since manual annotation for a new machine learning task is often prohibitive. However, deep neural networks are sensitive to cross-domain distribution shift. The same network can make spurious prediction on a target domain dissimilar to the source domain (Quionero-Candela et al., 2009). A notable example is that models trained on labeled synthetic data which come in abundance, may fail when generalizing to real-world unlabeled data (Liu et al., 2017; Hoffman et al., 2018).

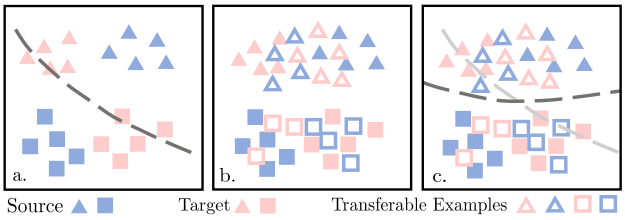


Figure 1. An overview of our approach. (a) A model trained only on the source domain is not adaptive to the target domain. (b) Our approach generates transferable examples to fill in the gap between domains. (c) The decision boundary is adapted to the target data through training with transferable examples. *Best viewed in color.*

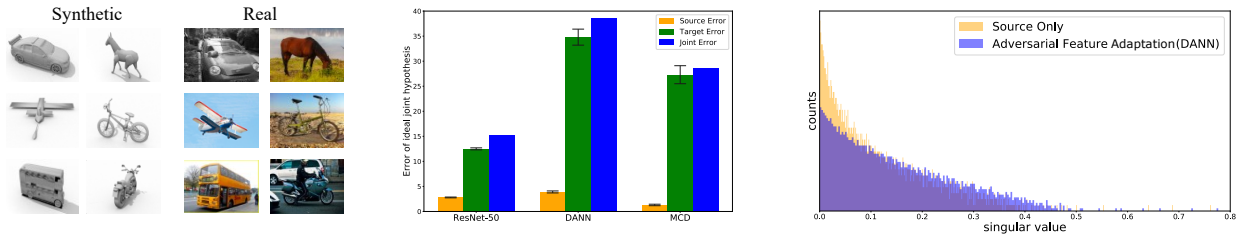
Domain adaptation aims at learning an accurate classifier for such a scenario. Recent advances in deep neural networks have enhanced the transferability of feature representations (Yosinski et al., 2014) and the disentanglement of explanatory factors behind data (Bengio et al., 2013). Therefore, a reasonable approach to domain adaptation is harnessing the power of deep neural networks to extract domain-invariant feature representations. One possible way is to minimize some measure of distance between the source and target feature distributions such as maximum mean discrepancy (Long et al., 2015). On par with distance minimizing methods, adversarial domain adaptation incorporates adversarial learning as a two-player game similar to GANs (Goodfellow et al., 2014). In this paradigm, the base network is divided into a feature extractor and a classifier. A domain discriminator is induced to discriminate the source domain from the target domain, while the feature extractor learns domain-invariant representations to fool the domain discriminator (Ganin et al., 2016; Tzeng et al., 2017; Long et al., 2018).

However, this class of techniques face critical restrictions. Based on the domain adaptation theory (Ben-David et al., 2010), an essential prerequisite for domain adaptation is the

[1]School of Software [2]Department of Electronic Engineering [3]BNRist, Research Center for Big Data, Tsinghua University, Beijing, China [4]University of California, Berkeley, USA.

Hong Liu <h-l17@mails.tsinghua.edu.cn>. Correspondence to: Mingsheng Long <mingsheng@tsinghua.edu.cn>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.



(a) VisDA-2017 (b) Test Error (c) Singular Values of Weight Matrices

Figure 2. Motivation of our approach. (a) Learning from synthetic to real data on VisDA-2017 dataset (Peng et al., 2017). (b) The error of ideal joint hypothesis. (c) The distributions of the singular values of the deep classifier's adapted weight matrices. *Best viewed in color.*

adaptability of feature representations between the source and target domains. Given feature representations, the adaptability can be explicitly quantified by the error of an ideal joint hypothesis on the source and target domains. When the adaptability is poor, we can never expect to learn a classifier with lower target error by minimizing source error. Adversarial feature adaptation is risky in this regard, since transforming the feature representations to be domain-invariant may inevitably distort the original feature distributions and enlarge the error of the ideal joint hypothesis. A preliminary empirical investigation of the adaptability on the challenging task of learning from synthetic to real is shown in Figure 2.

Another disadvantage of such techniques is the need of learning new representations in deep neural networks. When there is no well-established deep architectures for learning disentangled and transferable representations as in many important tasks of interest, e.g. sentiment polarity classification, email spam filtering and click-through rate prediction, the adversarial feature adaptation methods may perform unsatisfactorily, run very slowly, or even break down. General approaches to domain adaptation for a variety of real-world tasks should take such scenarios into consideration.

In this paper, we address the aforementioned challenges by proposing a general approach to adapting deep classifiers across domains, without performing the adversarial feature adaptation to learn domain-invariant representations. Recent advances in adversarial training (Goodfellow et al., 2015) reveal that minimizing the error under adversarial perturbations within a small Wasserstein distance to the source domain can potentially bound the error on the target domain (Lee & Raginsky, 2018). Additionally, adversarial training can push the decision boundary away from data points. For deep neural networks, it is an effective method of regularizing the model to mitigate overfitting (Miyato et al., 2018). Based on these findings, we adapt deep classifiers across domains by presenting a general approach, *Transferable Adversarial Training* (TAT). Without changing the feature representations, the approach generates *transferable examples* as adversaries to both the category classifier and the domain discriminator. Through adversarial training with these transferable examples, the category classifier can be adapted from the source to the target with guaranteed adaptability. The generation of these transferable examples and the adversarial training of both classifiers are formulated into a two-player minimax game, which can be solved in linear-time through back-propagation. Extensive experiments on vision and NLP tasks testify that our model exceeds state of the art methods on domain adaptation benchmark datasets.

2. Hidden Limitations of Adversarial Feature Adaptation

Existing adversarial feature adaptation methods are based on the domain adaptation theory (Ben-David et al., 2010).

Theorem 1. *(Ben-David et al., 2010) Let $\mathcal{H}$ be the hypothesis space and ϵ_s, ϵ_t be the generalization error of a classifier $C \in \mathcal{H}$ on the source domain X_s and the target domain X_t, respectively. Then for any classifier $C \in \mathcal{H}$,*

$$\epsilon_t(C) \leq \epsilon_s(C) + d_{\mathcal{H}\Delta\mathcal{H}}(X_s, X_t) + \lambda, \tag{1}$$

where $d_{\mathcal{H}\Delta\mathcal{H}}$ is the $\mathcal{H}\Delta\mathcal{H}$-distance between X_s and X_t,

$$d_{\mathcal{H}\Delta\mathcal{H}} \triangleq \sup_{h,h' \in \mathcal{H}} \left| \mathbb{E}_{x \sim X_s} \left[h(x) \neq h'(x) \right] - \mathbb{E}_{x \sim X_t} \left[h(x) \neq h'(x) \right] \right| \tag{2}$$

and λ is the error of an ideal joint hypothesis h^* defined as $h^* = \arg\min_{h \in \mathcal{H}} \epsilon_s(h) + \epsilon_t(h)$, such that

$$\lambda = \epsilon_s(h^*) + \epsilon_t(h^*). \tag{3}$$

The hypothesis-induced $\mathcal{H}\Delta\mathcal{H}$-distance measures the divergence between the source and target feature distributions. In adversarial feature adaptation, the feature extractor learns domain-invariant feature representations to minimize $\mathcal{H}\Delta\mathcal{H}$-distance, while the classifier is simultaneously trained on the source labeled data to minimize the source error. However, the adaptability quantified by λ is often overlooked, which should be made sufficiently small to guarantee the feasibility of domain adaptation. Adversarial feature adaptation methods align the feature distributions across domains under the assumption that λ remains small, yet they inevitably

distort the original feature representations. In consequence, there is no guarantee that λ will remain under control. We substantiate this claim with the following experimentation.

Error of ideal joint hypothesis. As aforementioned, the adaptability λ of domain adaptation is quantified by the error of the ideal joint hypothesis h^* as $\lambda = \epsilon_s(h^*) + \epsilon_t(h^*)$. To compute λ, we train a new classifier over the feature representations learned by existing methods: **ResNet-50** (He et al., 2016) pre-trained on ImageNet (Russakovsky et al., 2015), Domain Adversarial Neural Network (**DANN**) (Ganin et al., 2016), and Maximum Classifier Discrepancy (**MCD**) (Saito et al., 2018). The ideal joint hypothesis h^* is found by training on both source labeled data and target labeled data. Note that in this case, the ground truth labels of target data are only used to reason about the adaptability. The error of the ideal joint hypothesis on the source domain, the target domain, and their sum λ are shown in Figure 2(b).

It is somewhat unexpected that the adaptability λ, as quantified by the error of the ideal joint hypothesis h^*, worsens substantially in the adversarial feature adaptation methods DANN and MCD, compared to the non-adaptation method ResNet-50. We reasonably postulate that this undesirable effect is caused by the distortion of feature distributions in the process of adversarial representation learning, which is generally performed in adversarial feature adaptation methods. Diminishing domain-specific variations inevitably breaks the discriminative structures of the original representations.

Singular values of weight matrices. We further justify that the feature distributions are distorted in adversarial feature adaptation such that the discriminative structures of the feature representations are substantially deteriorated. To this end, we compute the singular values of the weight matrices of the layers corresponding to the adapted feature representations learned by **ResNet-50** (He et al., 2016) pre-trained on ImageNet (Russakovsky et al., 2015) and Domain Adversarial Neural Network (**DANN**) (Ganin & Lempitsky, 2015). The distributions of singular values are shown in Figure 2(c). The singular values of weight matrix from DANN (Ganin et al., 2016) have higher variations than those from ResNet-50. Further, the singular value distribution of DANN matrix is more heavy-tailed, indicating a worse-conditioned and more-distorted feature representation (Bjorck et al., 2018).

Motivation of this work. The above findings reveal that existing adversarial feature learning generally deteriorates the adaptability λ and makes adaptation models vulnerable. Thus, we have to rethink the current paradigm and propose alternatives to the mainstream adversarial feature adaptation approaches. A natural solution is to fix the feature representations and instead adapt the deep classifiers, which apparently guarantees the adaptability λ. This is possible by extending the adversarial training approaches (Goodfellow et al., 2015) to domain adaptation.

3. Transferable Adversarial Training

Consider the problem of unsupervised domain adaptation, with n_s i.i.d. observations $\{\mathbf{x}_s^{(i)}, \mathbf{y}_s^{(i)}\}_{i=1}^{n_s}$ from a source domain of distribution $P(\mathbf{x}_s, \mathbf{y}_s)$, and n_t i.i.d. observations $\{\mathbf{x}_t^{(i)}\}_{i=1}^{n_t}$ from a target domain of distribution $Q(\mathbf{x}_t, \mathbf{y}_t)$. Note that the i.i.d. assumption is violated across domains as $P \neq Q$. Our goal is to adapt a deep category classifier $\mathbf{y} = C(\mathbf{f})$ under the feature representation $\mathbf{f} = F(\mathbf{x})$, which guarantees lower generalization error on the target domain.

In this paper, we present a general approach to adapting deep classifiers across domains with guaranteed adaptability λ. The approach, *Transferable Adversarial Training* (TAT), constitutes two alternating steps: adversarial generation of transferable examples and adversarial training with transferable examples, both without distorting feature distributions.

3.1. Adversarial Generation of Transferable Examples

Existing adversarial feature adaptation methods diminish domain-specific variations by learning domain-invariant representations. Denote by $\mathbf{f} = F(\mathbf{x})$ the feature extractor and by $\mathbf{d} = D(\mathbf{f})$ the domain discriminator. D and F form a two-player minimax game: D is trained to distinguish the source from the target while F is trained simultaneously to confuse D. However, such a procedure may deteriorate the adaptability λ. To guarantee adaptability, we propose to fix the feature representations and generate *transferable examples* to bridge domain gap. Concretely, we still train the domain discriminator D to distinguish the source domain from the target domain through the following loss function:

$$\begin{aligned}\ell_d\left(\theta_D, \mathbf{f}\right) = &-\frac{1}{n_s}\sum_{i=1}^{n_s}\log[D(\mathbf{f}_s^{(i)})] \\ &-\frac{1}{n_t}\sum_{i=1}^{n_t}\log[1-D(\mathbf{f}_t^{(i)})].\end{aligned} \tag{4}$$

The deep category classifier C is also trained to perform well on the source domain through the cross-entropy loss:

$$\ell_c(\theta_C, \mathbf{f}) = \frac{1}{n_s}\sum_{i=1}^{n_s}\ell_{ce}(C(\mathbf{f}_s^{(i)}), \mathbf{y}_s^{(i)}). \tag{5}$$

Different from existing adversarial feature adaptation methods, we diminish the distributional variations by filling the gap between the source and target domains with transferable examples generated in a new adversarial training paradigm. To enable adversarial generation of transferable examples, we compute the gradients of the above loss functions ℓ_d and ℓ_c w.r.t. each example in terms of learned features $\mathbf{f}_s$ and $\mathbf{f}_t$. Note that we opt not to compute the gradients w.r.t. the raw images $\mathbf{x}_s$ and $\mathbf{x}_t$ as in standard adversarial training (Goodfellow et al., 2015). Further, the feature extractor F

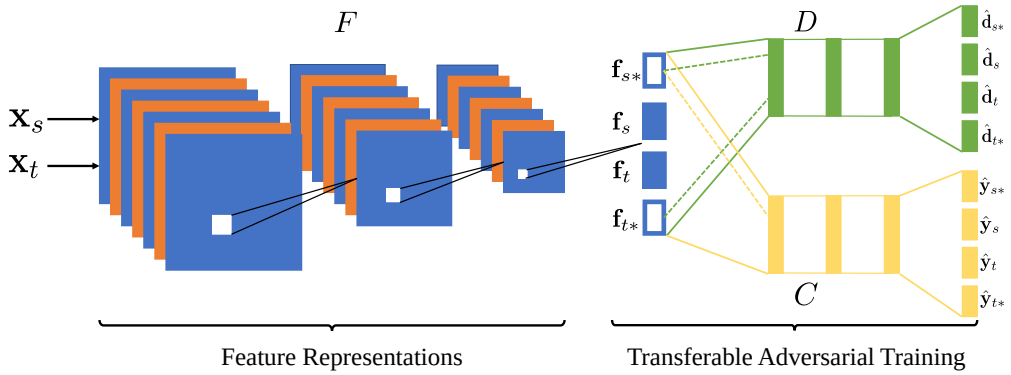


Figure 3. Transferable Adversarial Training (TAT) for adapting deep classifiers. The feature extractor F yields representations $\mathbf{f}_s$ and $\mathbf{f}_t$ of the source and target data, which are *fixed* in the training process to guarantee adaptability λ. The dashed lines indicate the adversarial generation of transferable examples $\mathbf{f}_{s*}$ and $\mathbf{f}_{t*}$ through maximizing the errors of the category classifier C and domain discriminator D. We adversarially train the classifiers with transferable examples: C to minimize the source error and D to distinguish source from target.

is not updated using these gradients. As analyzed in the previous section, this guarantees good adaptability λ.

More formally, we propose to generate the transferable examples taking the philosophy of adversarial training. First, the transferable examples should effectively confuse the domain discriminator D, such that they can fill in the gap and bridge the source and target domains. Second, the transferable examples should be able to deceive the category classifier C, such that they can push the decision boundary away from data points. Hence, the transferable examples are generated adversarially through a joint loss of ℓ_c and ℓ_d:

$$\begin{aligned}\mathbf{f}_{t^{k+1}} \leftarrow \mathbf{f}_{t^k} &+ \beta \nabla_{\mathbf{f}_{t^k}} \ell_d(\theta_D, \mathbf{f}_{t^k}) \\ &- \gamma \nabla_{\mathbf{f}_{t^k}} \ell_2(\mathbf{f}_{t^k}, \mathbf{f}_{t^0}),\end{aligned} \tag{6}$$

$$\begin{aligned}\mathbf{f}_{s^{k+1}} \leftarrow \mathbf{f}_{s^k} &+ \beta \nabla_{\mathbf{f}_{s^k}} \ell_d(\theta_D, \mathbf{f}_{s^k}) \\ &- \gamma \nabla_{\mathbf{f}_{s^k}} \ell_2(\mathbf{f}_{s^k}, \mathbf{f}_{s^0}) \\ &+ \beta \nabla_{\mathbf{f}_{s^k}} \ell_c(\theta_C, \mathbf{f}_{s^k}),\end{aligned} \tag{7}$$

where K is the number of iterations for generating each transferable example, and $k = 0, 1, \cdots, K-1$ is the current iteration. Note that $\mathbf{f}_{t^0} = \mathbf{f}_t$, $\mathbf{f}_{s^0} = \mathbf{f}_s$, $\mathbf{f}_{t*} = \mathbf{f}_{t^K}$, $\mathbf{f}_{s*} = \mathbf{f}_{s^K}$. To generate examples that are sufficiently transferable, we need a sufficient number of iterations, typically $K = 10$. In addition, to avoid divergence of the generated examples, we control the ℓ_2-distance between the generated examples and the original examples with hyper-parameters γ and β.

3.2. Adversarial Training with Transferable Examples

We aim at enabling the robustness of the category classifier C against domain distribution shift. To reach this goal, a reasonable way is to train the classifier to make accurate predictions for the transferable examples $\mathbf{f}_{s*}$ in the source domain. Furthermore, we require the classifier to make consistent predictions for the transferable examples $\mathbf{f}_{t*}$ and their original counterparts $\mathbf{f}_t$ in the target domain. As analyzed empirically by Miyato et al. (2018), adversarial training improves local smoothness of the output distribution. Taking similar explanation, training the category classifier with transferable examples can be interpreted as improving the robustness of the classifier's prior distribution against both adversarial perturbations and domain variations. Note that we have access to labels of the source domain but the target labels are absent. Thus, the loss function for adversarial training of the category classifier C is formulated as follows,

$$\begin{aligned}\ell_{c,adv}(\theta_C, \mathbf{f}_*) = &\frac{1}{n_s}\sum_{i=1}^{n_s} \ell_{ce}(C(\mathbf{f}_{s*}^{(i)}), \mathbf{y}_{s*}^{(i)}) \\ &+ \frac{1}{n_t}\sum_{i=1}^{n_t} \Big| C((\mathbf{f}_{t*}^{(i)})) - C((\mathbf{f}_t^{(i)})) \Big|.\end{aligned} \tag{8}$$

Similar to training the category classifier, we also train the domain discriminator with generated transferable examples. This is important to stabilize the adversarial training process, otherwise the generated transferable examples will diverge. Another key perspective is to leverage these transferable examples to bridge the domain discrepancy. Simply fooling the domain discriminator on original data cannot guarantee that generated examples are transferable from one domain to the other. Hence, we propose to adversarially train the domain discriminator to further distinguish transferable examples from the source and target, using the following loss

$$\begin{aligned}\ell_{d,adv}(\theta_D, \mathbf{f}_*) = &-\frac{1}{n_s}\sum_{i=1}^{n_s} \log[D(\mathbf{f}_{s*}^{(i)})] \\ &-\frac{1}{n_t}\sum_{i=1}^{n_t} \log[1 - D(\mathbf{f}_{t*}^{(i)})].\end{aligned} \tag{9}$$

Finally, we enable transferable adversarial training (TAT) by generating transferable examples and training classifiers on them. We jointly minimize error (4) and error (9) with

Algorithm 1 Transferable Adversarial Training (TAT)

Input: original features $\{\mathbf{f}_s^{(i)}, \mathbf{y}_s^{(i)}\}_{i=1}^{n_s}$ and $\{\mathbf{f}_t^{(i)}\}_{i=1}^{n_t}$.
Output: learned model parameters $\theta = (\theta_C, \theta_D)$.
Initialize $\theta = (\theta_C, \theta_D)$ randomly.
for `iter` $= 1$ **to** `MaxIter` **do**
  Sample a mini-batch of $\{(\mathbf{f}_s, \mathbf{y}_s)\}$ and $\{\mathbf{f}_t\}$ uniformly from the training dataset in terms of original features.
  for $k = 0$ **to** $K-1$ **do**
    $\mathbf{f}_{s^{k+1}} \leftarrow \mathbf{f}_{s^k} + \beta \nabla_{\mathbf{f}_{s^k}} \ell_d(\theta_D, \mathbf{f}_{s^k}) - \gamma \nabla_{\mathbf{f}_{s^k}} \ell_2(\mathbf{f}_{s^k}, \mathbf{f}_{s^0}) + \beta \nabla_{\mathbf{f}_{s^k}} \ell_c(\theta_C, \mathbf{f}_{s^k})$
    $\mathbf{f}_{t^{k+1}} \leftarrow \mathbf{f}_{t^k} + \beta \nabla_{\mathbf{f}_{t^k}} \ell_d(\theta_D, \mathbf{f}_{t^k}) - \gamma \nabla_{\mathbf{f}_{t^k}} \ell_2(\mathbf{f}_{t^k}, \mathbf{f}_{t^0})$
  end for
  $\theta_C \leftarrow \theta_C - \alpha \nabla_{\theta_C} [\ell_c(\theta_C, \mathbf{f}) + \ell_{c,adv}(\theta_C, \mathbf{f}_*))]$
  $\theta_D \leftarrow \theta_D - \alpha \nabla_{\theta_D} [\ell_d(\theta_D, \mathbf{f}) + \ell_{d,adv}(\theta_D, \mathbf{f}_*))]$
end for

respect to D, and error (5) and error (8) with respect to C. This leads to the optimization problem for the TAT approach:

$$\min_{\theta_D, \theta_C} \ell_d(\theta_D, \mathbf{f}) + \ell_c(\theta_C, \mathbf{f}) + \ell_{d,adv}(\theta_D, \mathbf{f}_*) + \ell_{c,adv}(\theta_C, \mathbf{f}_*). \tag{10}$$

We summarize the detailed training procedure in Algorithm 1. TAT runs over the feature-level examples $\mathbf{f}$ and propagates only through the deep classifier C (usually of no more than three layers), which is very computationally efficient (an order of magnitude faster than feature adaptation methods).

4. Theoretical Understanding

In this section, we give a theoretical understanding of the proposed approach, making use of the domain adaptation theory (Ben-David et al., 2010) and the adversarial training theory (Sinha et al., 2018). While unifying both theories turns out to be nontrivial, we will leave it as our future work.

4.1. Domain Adaptation Theory

Recall the domain adaptation theory in Theorem 1,

$$\epsilon_t(C) \le \epsilon_s(C) + d_{\mathcal{H}\Delta\mathcal{H}}(X_s, X_t) + \lambda. \tag{11}$$

We have analyzed that the adversarial feature adaptation methods weaken the adaptability λ. TAT fixes the feature extractor F throughout the training procedure to keep λ unchanged, which is a complementary improvement to the previous methods. Further, by training the category classifier on the source domain, TAT minimizes error $\epsilon_s(C)$. Most importantly, our approach generates the transferable examples towards the opposite domains by confusing the domain discriminator. And training domain discriminator D against these transferable examples will explicitly bound the $\mathcal{H}\Delta\mathcal{H}$-distance, as justified in Ganin & Lempitsky (2015). Since the transferable examples will fill in the gap across domains (intuitively demonstrated in Figures 1 and 4), the $d_{\mathcal{H}\Delta\mathcal{H}}(X_s, X_t)$ term is further bounded in our approach. In summary, our approach conforms well with the domain adaptation theory, further yielding guaranteed adaptability.

4.2. Adversarial Training Theory

The proposed approach generates transferable examples, which are essentially the adversarial examples against both category classifier C and domain discriminator D. This is in line with the adversarial training theory (Sinha et al., 2018).

Theorem 2. *(Sinha et al., 2018) Assume $|\ell(\theta, x)| \le M_\ell$ for all models $\theta \in \Theta$ and examples $x \in \mathcal{X}$. Then, for a fixed $t > 0$ and numerical constants $b_1, b_2 > 0$, with probability at least $1 - e^{-t}$, simultaneously for all $\theta \in \Theta$, $\rho \ge 0$, $\gamma \ge 0$,*

$$\sup_{Q: W(P,Q) \le \rho} \mathbb{E}_Q[\ell(\theta, X)] \le \gamma\rho + \mathbb{E}_{\hat{P}_n}[\phi_\gamma(\theta, X)] + O(1/\sqrt{n}, M_\ell, b_1, b_2), \tag{12}$$

where $W(P, Q)$ is the Wasserstein distance between P and Q, $\hat{P}_n$ is the empirical distribution of P, $O(\cdot)$ is a complexity function, and $\phi_\gamma(\theta, X)$ is the robust surrogate loss,

$$\phi_\gamma(\theta, X) \triangleq \sup_{x \in \mathcal{X}} \ell(\theta, x) - \gamma W(x_0, x). \tag{13}$$

Our approach trains the category classifier C with the worst-case distributions within distance ρ away from the source domain, which guarantees good performance if the target domain is in the range of distance ρ. By introducing hyper-parameters β and γ in Eq. (6)–(7), our approach executes the adversarial training procedure guaranteed by the theory.

5. Experiments

We evaluate TAT on five domain adaptation datasets. Codes and datasets are made available at `github.com/thuml/Transferable-Adversarial-Training`.

5.1. Experimental Setup

Office-31 (Saenko et al., 2010) is a standard dataset for visual domain adaptation. It contains 4,652 images across 31 categories from three domains: Amazon (**A**), Webcam (**W**), and DSLR (**D**). From this dataset, we construct 6 transfer tasks: A→W, D→W, W→D, A→D, D→A, and W→A.

ImageCLEF-DA (Long et al., 2017) is a dataset organized by selecting the 12 common classes shared by three public datasets (domains): Caltech-256 (**C**), ImageNet ILSVRC 2012 (**I**), and Pascal VOC 2012 (**P**). We evaluate all methods on 6 transfer tasks: I→P, P→I, I→C, C→I, C→ P, and P→C.

Table 1. Classification accuracies (%) on Office-31 for unsupervised domain adaptation with ResNet-50.

METHOD	A→W	D→W	W→D	A→D	D→A	W→A	AVG.
RESNET-50 (HE ET AL., 2016)	68.4±0.2	96.7±0.1	99.3±0.1	68.9±0.2	62.5±0.3	60.7±0.3	76.1
DAN (LONG ET AL., 2015)	80.5±0.4	97.1±0.2	99.6±0.1	78.6±0.2	63.6±0.3	62.8±0.2	80.4
DANN (GANIN ET AL., 2016)	82.6±0.4	96.9±0.2	99.3±0.2	81.5±0.4	68.4±0.5	67.5±0.5	82.7
ADDA (TZENG ET AL., 2017)	86.2±0.5	96.2±0.3	98.4±0.3	77.8±0.3	69.5±0.4	68.9±0.5	82.9
VADA (SHU ET AL., 2018)	86.5±0.5	98.2±0.4	99.7±0.2	86.7±0.4	70.1±0.4	70.5±0.4	85.4
GTA (SANKARANARAYANAN ET AL., 2018)	89.5±0.5	97.9±0.3	99.7±0.2	87.7±0.5	72.8±0.3	71.4±0.4	86.5
MCD (SAITO ET AL., 2018)	88.6±0.2	98.5±0.1	**100.0**±.0	92.2±0.2	69.5±0.1	69.7±0.3	86.5
CDAN (LONG ET AL., 2018)	**93.1**±0.1	98.6±0.1	**100.0**±.0	92.9±0.2	71.0±0.3	69.3±0.3	87.5
TAT	92.5±0.3	**99.3**±0.1	**100.0**±.0	**93.2**±0.2	**73.1**±0.3	**72.1**±0.3	**88.4**

Table 2. Classification accuracies (%) on Image-CLEF for unsupervised domain adaptation with ResNet-50.

METHOD	I→P	P→I	I→C	C→I	C→P	P→C	AVG.
RESNET-50 (HE ET AL., 2016)	74.8±0.3	83.9±0.1	91.5±0.3	78.0±0.2	65.5±0.3	91.2±0.3	80.7
DAN (LONG ET AL., 2015)	74.5±0.4	82.2±0.2	92.8±0.2	86.3±0.4	69.2±0.4	89.8±0.4	82.5
DANN (GANIN ET AL., 2016)	75.0±0.3	86.0±0.3	96.2±0.4	87.0±0.5	74.3±0.5	91.5±0.6	85.0
CDAN (LONG ET AL., 2018)	76.7±0.3	90.6±0.3	97.0±0.4	90.5±0.4	74.5±0.3	93.5±0.4	87.1
TAT	**78.8**±0.2	**92.0**±0.2	**97.5**±0.3	**92.0**±0.3	**78.2**±0.4	**94.7**±0.4	**88.9**

Office-Home (Venkateswara et al., 2017) is a new dataset more difficult than Office-31. It has 15,500 images across 65 classes in office and home settings from four domains of large domain discrepancy: *Artistic* images (**Ar**), *Clip Art* (**Cl**), *Product* images (**Pr**), and *Real-World* images (**Rw**).

VisDA-2017 (Peng et al., 2017) is the largest dataset to date for visual domain adaptation, providing with two distinct domains. **Synthetic**: renderings of 3D models from different angles and with different lightning conditions; **Real**: real-world images collected from MSCOCO (Lin et al., 2014).

Multi-Domain Sentiment (Blitzer et al., 2007) was widely adopted as the benchmark for domain adaptation in sentiment classification. It comprises of product reviews from `amazon.com` in four product domains: books (**B**), dvds (**D**), electronics (**E**), and kitchen appliances (**K**). Each review is assigned with a positive or negative polarity and is represented by Bag of Words (BoW) using term frequency.

We investigate state of the art domain adaptation methods: Deep Adaptation Network (**DAN**) (Long et al., 2015), Domain Adversarial Neural Network (**DANN**) (Ganin et al., 2016), Adversarial Discriminative Domain Adaptation (**ADDA**) (Tzeng et al., 2017), Virtual Adversarial Domain Adaptation (**VADA**) (Shu et al., 2018), Generate to Adapt (**GTA**) (Sankaranarayanan et al., 2018), Maximum Classifier Discrepancy (**MCD**) (Saito et al., 2018), and Conditional Domain Adversarial Network (**CDAN**) (Long et al., 2018). For the sentiment classification tasks, we further compare with classic Marginalized Denoising Autoencoders (**mSDA**) (Chen et al., 2012) and the state of the art Transfer Denoising Autoencoders (**TDA**) (Long et al., 2016).

For image datasets, we use **ResNet-50** (He et al., 2016) pre-trained on ImageNet (Russakovsky et al., 2015) to extract original feature representations. We use Adam (Kingma & Ba, 2014) with initial learning rate $\eta_0 = 10^{-4}$. We adopt the inverse-decay strategy of DANN (Ganin et al., 2016), where the learning rate changes by $\eta_p = \frac{\eta_0}{(1+\omega p)^\phi}$, $\omega = 10$, $\phi = 0.75$, and p is the progress ranging from 0 to 1. We use reverse validation for hyperparameter selection (Zhong et al., 2010). For image datasets, $\beta = 5$ and $\gamma = 1$. For sentiment datasets, we use Bag of Word (BoW) vectors and mSDA (Chen et al., 2012) representations as input.

For the proposed TAT approach, the category classifier is a two-layer fully connected network ($2048 \times 256 \times \#\text{classes}$), and the domain discriminator consists of two fully connected layers with BatchNorm (Ioffe & Szegedy, 2015) and LeakyReLU non-linearity in the first layer. For the adversarial feature adaptation methods, we adopt gradient reversal layers (Ganin et al., 2016) for jointly training the domain discriminator with the category classifier.

5.2. Results

We report results of Office-31 based on ResNet-50 in Table 1. The proposed TAT outperforms all comparison methods on most tasks. We clearly observe that on **W→A** and **D→A** with relatively large domain shift and imbalanced domain scales, TAT exceeds all feature adaptation methods by large margins and even performs better than models incorporating complex generative architectures. This further testifies that transferable examples augment the original source domain and therefore mitigate the imbalance between domains.

On ImageCLEF-DA, TAT exceeds comparison methods in all tasks, but the boost is relatively minor, since domain scale is the same and the domain discrepancy is smaller.

When domain discrepancy is significant as on Office-Home,

Table 3. Classification accuracies (%) on Office-Home for unsupervised domain adaptation (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 (HE ET AL., 2016)	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 (LONG ET AL., 2015)	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 (GANIN ET AL., 2016)	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 (LONG ET AL., 2018)	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
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**

Table 4. Classification accuracies (%) on Multi-Domain Sentiment Dataset for unsupervised domain adaptation.

METHOD	B→D	B→E	B→K	D→B	D→E	D→K	E→B	E→D	E→K	K→B	K→D	K→E	AVG.
SVM (BOW) (CHANG & LIN, 2011)	79.9	74.8	76.9	74.3	74.8	74.6	70.5	72.6	84.7	70.7	73.6	84.2	76.0
DANN (BOW) (GANIN ET AL., 2016)	78.4	73.3	77.9	72.3	75.4	78.3	71.3	73.8	85.4	70.9	74.0	84.3	76.3
MSDA (BOW) (CHEN ET AL., 2012)	83.5	74.5	**84.6**	**83.6**	79.5	**87.1**	78.8	**80.9**	85.3	**80.2**	**82.3**	86.9	82.3
TAT (BOW)	**84.5**	**80.1**	83.6	81.9	**81.9**	84.0	**83.2**	77.9	**90.0**	75.8	77.7	**88.2**	**82.4**
DANN (MSDA) (GANIN ET AL., 2016)	82.9	80.4	84.3	82.5	80.9	84.9	77.4	78.1	88.1	71.8	78.9	85.6	81.3
TDA (MSDA) (LONG ET AL., 2016)	84.1	85.0	87.5	84.9	85.7	88.6	82.0	82.7	87.7	**81.5**	83.3	86.8	85.0
TAT (MSDA)	**86.8**	**85.9**	**88.6**	**86.4**	**86.4**	**89.4**	**83.7**	**83.5**	**90.4**	81.4	**84.7**	**89.2**	**86.3**

Table 5. Classification accuracies(%) on VisDA-2017 (ResNet-50).

METHOD	ACCURACY
RESNET-50 (HE ET AL., 2016)	40.2
DANN (GANIN ET AL., 2016)	63.7
MCD (SAITO ET AL., 2018)	69.2
GTA (SANKARANARAYANAN ET AL., 2018)	69.5
CDAN (LONG ET AL., 2018)	70.0
TAT	**71.9**

TAT still achieves strong performance across all the tasks, as shown in Table 3. By guaranteeing adaptability, it improves substantially over adversarial feature adaptation methods.

Results of VisDA-2017 are shown in Table 5, where TAT outperforms both feature-level adaptation and generative pixel-level adaptation methods. Note that TAT only involves two-layer fully connected networks, much simpler than the generative methods that incorporate complex architecture tailored to the synthetic-to-real domain adaptation problem.

TAT, with the same architecture choice as the vision tasks, even outperforms strong competitors in the multi-domain sentiment classification tasks by large margins. The two-layer TAT can achieve comparable accuracy to five-layer mSDA on the 30000-dimension BoW input, and improve mSDA by 3.9% if further trained on mSDA representations. This verifies TAT's effectiveness on non-visual domain adaptation tasks. To our knowledge, TAT is the first approach that performs well in both vision and NLP adaptation scenarios.

5.3. Analysis

Toy dataset. We study the behavior of TAT on the *rotating twinning moon* dataset. We generate 1000 samples for each domain with **scikit-learn** (Pedregosa et al., 2013). Samples of the target domain are rotated 30° from the source domain. We depict the decision boundary of TAT and compare it with the model trained solely on the source domain. We also show the distributions of the transferable examples.

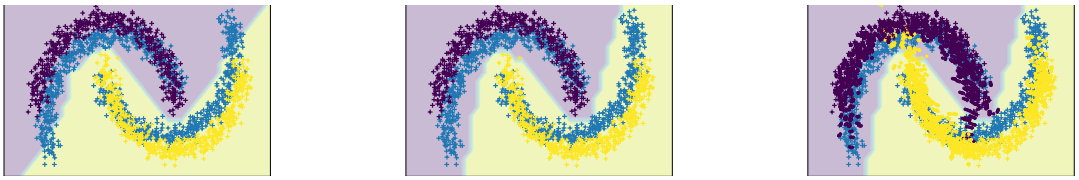
As shown in Figure 4(a), the model trained on the source domain cannot accurately classify the target examples. In contrast, TAT's decision boundary separates most target examples correctly (Figure 4(b)). In Figure 4(c), we illustrate the distribution of transferable examples. The generated examples are shown to fill in the gap between the source and target domains and thus reduce the domain discrepancy. By enforcing consistent predictions over transferable examples, we drive the decision boundary away from all examples.

Ablation study. We study the strategies for transferable examples generating and transferable adversarial training. The comparison between TAT and its variants are provided in Table 6. TAT (w\o c) and TAT (w\o d) refer to the proposed model without transferable examples generated from the gradient of category classifier C and domain discriminator D, respectively. Results indicate that adversarial training towards the category classifier and the domain discriminator are both beneficial to bridging cross-domain discrepancy.

Cross-domain A-distance. As shown in the domain adaptation theory (1), two important factors bound the generalization error: *adaptability* λ and *discrepancy* $d_{\mathcal{H}\Delta\mathcal{H}}$. Figure 2 shows that our approach yields the highest adaptability (that of ResNet-50). We further show in Table 7 the cross-domain A-distance (Ben-David et al., 2010), a proxy of $d_{\mathcal{H}\Delta\mathcal{H}}$. We compute the A-distance of TAT based on the transferable examples, which turns out to be the smallest in all methods.

6. Related Work

Domain Adaptation Domain adaptation generalizes a model under dataset shift (Pan & Yang, 2010). Moment matching minimizes the distance between feature statistics



(a) Source Only Model (b) TAT (c) Transferable Examples

Figure 4. Behaviors on the *two moon* problem. Purple and yellow "+"s indicate source samples, blue "+"s are target samples, while dots are transferable examples. (a) The source only model. (b) The decision boundary of TAT. (c) The distribution of the transferable examples.

Table 6. Classification accuracies (%) of the TAT variants on Office-31 for unsupervised domain adaptation (ResNet-50).

METHOD	A→W	D→W	W→D	A→D	D→A	W→A	AVG.
TAT (W\O C)	87.2±0.2	98.3±0.2	99.9±0.1	88.4±0.2	71.0±0.1	69.8±0.1	85.8
TAT (W\O D)	88.5±0.3	98.7±0.2	100.0±.0	90.6±0.2	71.4±0.2	72.0±0.1	86.8
TAT	**92.5**±0.3	**99.3**±0.1	**100.0**±.0	**93.2**±0.2	**73.1**±0.3	**72.1**±0.3	**88.4**

Table 7. Cross-domain A-distance of different approaches.

METHOD	D→W	W→A
RESNET-50 (HE ET AL., 2016)	1.27	1.86
DANN (GANIN ET AL., 2016)	1.23	1.44
MCD (SAITO ET AL., 2018)	1.22	1.60
TAT	**1.06**	**1.04**

of the source and target. DAN and DDC minimize the maximum mean discrepancies (Long et al., 2015; Tzeng et al., 2014). Zellinger et al. (2017) proposed the central moment discrepancy as a discrepancy measure between distributions.

The success of GANs (Goodfellow et al., 2014) inspires the adversarial feature adaptation approaches. DANN trains a domain discriminator to distinguish the source and target while the features are learned to fool the discriminator (Ganin et al., 2016). This paradigm incorporates minimax training objectives and can be interpreted as approximating the $\mathcal{H}\Delta\mathcal{H}$-distance in the domain adaptation theory (Ben-David et al., 2010). Based on adversarial feature adaptation, a line of works improve the domain discriminator or the procedure of adversarial learning. ADDA uses asymmetric feature extractors for the source and target (Tzeng et al., 2017). CDAN conditions the domain discriminator on classifier predictions (Long et al., 2018). Another way of adversarial feature adaptation generates target features to minimize the $\mathcal{H}\Delta\mathcal{H}$-distance, which is computed by the disagreement of independent classifiers (Saito et al., 2018).

On par with feature-level adaptation methods, pixel-level adaptation methods translate the source data into the target domain or vice versa by Image to Image Translation. Liu et al. (2017) proposed to learn a shared latent space with translated images. GTA generates source-like images using source features and target-like images using target features (Sankaranarayanan et al., 2018). Inspired by CycleGAN (Zhu et al., 2017), CyCADA enforces semantic consistency of the image translation to improve the pixel-level methods (Hoffman et al., 2018).

Adversarial Training Szegedy et al. (2014) first discovered the intriguing weakness of deep networks to minor adversarial perturbations. Goodfellow et al. (2015) delved into adversarial examples and pointed out the advantages of adversarial training. Training with adversarial examples results in regularizing the network to mitigate over-fitting (Zheng et al., 2016). Sinha et al. (2018) derived a theory of principled adversarial training with robustness guarantees.

In addition to enhancing the robustness of deep networks, adversarial training is also promising in a variety of machine learning problems. Miyato et al. (2018) incorporated virtual adversarial training (VAT) in semi-supervised context to smooth the output distributions as a regularization of deep networks. Virtual Adversarial Domain Adaptation (VADA) improves adversarial feature adaptation with VAT and harnesses the cluster assumption (Chapelle & Zien, 2005; Shu et al., 2018). Different from our method, it generates adversarial examples against only the classifier and still performs adversarial feature adaptation. Volpi et al. (2018) explored adversarial training in domain generalization scenarios.

7. Conclusion

We present a general approach, transferable adversarial training, to adapting deep classifiers. By deceiving both category classifier and domain discriminator, the approach generates transferable examples which bridge the gap across domains. Both high adaptability and small distribution discrepancy expected by the domain adaptation theory are achieved by the approach, as justified on both vision and NLP datasets.

8. Acknowledgements

This work was supported by the National Natural Science Foundation of China (61772299, 71690231, and 61672313).

References

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Machine Learning*, 79(1-2):151–175, 2010.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(8):1798–1828, 2013.

Bjorck, N., Gomes, C. P., Selman, B., and Weinberger, K. Q. Understanding batch normalization. In *Advances in Neural Information Processing Systems 31*, pp. 7705–7716. 2018.

Blitzer, J., Dredze, M., and Pereira, F. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 440–447, 2007.

Chang, C.-C. and Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Chapelle, O. and Zien, A. Semi-supervised classification by low density separation. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics, AISTATS*, pp. 57–64, 2005.

Chen, M., Xu, Z., Weinberger, K., and Sha, F. Marginalized denoising autoencoders for domain adaptation. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pp. 767–774, 2012.

Ganin, Y. and Lempitsky, V. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Marchand, M., and Lempitsky, V. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pp. 2672–2680, 2014.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *Proceedings of International Conference on Learning Representations*, 2015.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Hoffman, J., Tzeng, E., Park, T., Zhu, J., Isola, P., Saenko, K., Efros, A. A., and Darrell, T. Cycada: Cycle-consistent adversarial domain adaptation. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 1994–2003, 2018.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 448–456, 2015.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv*, abs/1412.6980, 2014.

Lee, J. and Raginsky, M. Minimax statistical learning with wasserstein distances. In *Advances in Neural Information Processing Systems 31*, pp. 2692–2701. 2018.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pp. 740–755, 2014.

Liu, M.-Y., Breuel, T., and Kautz, J. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*, pp. 700–708. 2017.

Long, M., Cao, Y., Wang, J., and Jordan, M. I. Learning transferable features with deep adaptation networks. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pp. 97–105, 2015.

Long, M., Wang, J., Cao, Y., Sun, J., and Yu, P. S. Deep learning of transferable representation for scalable domain adaptation. *IEEE Transactions on Knowledge and Data Engineering*, 28(8):2027–2040, 2016.

Long, M., Zhu, H., Wang, J., and Jordan, M. I. Deep transfer learning with joint adaptation networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 2208–2217, 2017.

Long, M., Cao, Z., Wang, J., and Jordan, M. I. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems 31*, pp. 1640–1650. 2018.

Miyato, T., Maeda, S., Ishii, S., and Koyama, M. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

Pan, S. J. and Yang, Q. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22 (10):1345–1359, 2010.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., and Louppe, G. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(10):2825–2830, 2013.

Peng, X., Usman, B., Kaushik, N., Hoffman, J., Wang, D., and Saenko, K. Visda: The visual domain adaptation challenge. *arXiv*, abs/1710.06924, 2017.

Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. *Dataset Shift in Machine Learning*. The MIT Press, 2009.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Saenko, K., Kulis, B., Fritz, M., and Darrell, T. Adapting visual category models to new domains. In *European Conference on Computer Vision*, pp. 213–226, 2010.

Saito, K., Watanabe, K., Ushiku, Y., and Harada, T. Maximum classifier discrepancy for unsupervised domain adaptation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3723–3732, 2018.

Sankaranarayanan, S., Balaji, Y., Castillo, C. D., and Chellappa, R. Generate to adapt: Aligning domains using generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8503–8512, 2018.

Shu, R., Bui, H., Narui, H., and Ermon, S. A DIRT-t approach to unsupervised domain adaptation. In *Proceedings of International Conference on Learning Representations*, 2018.

Sinha, A., Namkoong, H., and Duchi, J. Certifiable distributional robustness with principled adversarial training. In *Proceedings of International Conference on Learning Representations*, 2018.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. In *Proceedings of International Conference on Learning Representations*, 2014.

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

Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. Adversarial discriminative domain adaptation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7167–7176, 2017.

Venkateswara, H., Eusebio, J., Chakraborty, S., and Panchanathan, S. Deep hashing network for unsupervised domain adaptation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5018–5027, 2017.

Volpi, R., Namkoong, H., Sener, O., Duchi, J. C., Murino, V., and Savarese, S. Generalizing to unseen domains via adversarial data augmentation. In *Advances in Neural Information Processing Systems 31*, pp. 5339–5349. 2018.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems 27*, pp. 3320–3328. 2014.

Zellinger, W., Grubinger, T., Lughofer, E., Natschläger, T., and Saminger-Platz, S. Central moment discrepancy (CMD) for domain-invariant representation learning. In *Proceedings of the International Conference on Learning Representations*, 2017.

Zheng, S., Song, Y., Leung, T., and Goodfellow, I. Improving the robustness of deep neural networks via stability training. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4480–4488, 2016.

Zhong, E., Fan, W., Yang, Q., Verscheure, O., and Ren, J. Cross validation framework to choose amongst models and datasets for transfer learning. In *Machine Learning and Knowledge Discovery in Databases*, pp. 547–562, 2010.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *The IEEE International Conference on Computer Vision (ICCV)*, pp. 2223–2232, 2017.