

Characterizing and Avoiding Negative Transfer

Zirui Wang, Zihang Dai, Barnabás Póczos, Jaime Carbonell
 Carnegie Mellon University

{ziruiw, dzihang, bapocz, jgc}@cs.cmu.edu

Abstract

When labeled data is scarce for a specific target task, transfer learning often offers an effective solution by utilizing data from a related source task. However, when transferring knowledge from a less related source, it may inversely hurt the target performance, a phenomenon known as negative transfer. Despite its pervasiveness, negative transfer is usually described in an informal manner, lacking rigorous definition, careful analysis, or systematic treatment. This paper proposes a formal definition of negative transfer and analyzes three important aspects thereof. Stemming from this analysis, a novel technique is proposed to circumvent negative transfer by filtering out unrelated source data. Based on adversarial networks, the technique is highly generic and can be applied to a wide range of transfer learning algorithms. The proposed approach is evaluated on six state-of-the-art deep transfer methods via experiments on four benchmark datasets with varying levels of difficulty. Empirically, the proposed method consistently improves the performance of all baseline methods and largely avoids negative transfer, even when the source data is degenerate.

1. Introduction

The development of deep neural networks (DNNs) has improved the state-of-the-art performance on a wide range of machine learning problems and applications. However, DNNs often require a large amount of labeled data to train well-generalized models and as more classical methods, DNNs rely on the assumption that training data and test data are drawn from the same underlying distribution. In some cases, collecting large volumes of labeled training data is expensive or even prohibitive. Transfer learning [20] addresses this challenge of data scarcity by utilizing previously-labeled data from one or more source tasks. The hope is that this source domain is related to the target domain and thus transferring knowledge from the source can improve the performance within the target domain. This powerful paradigm has been studied under various settings [35] and has been proven effective in a wide range of appli-

cations [39, 16, 17].

However, the success of transfer learning is not always guaranteed. If the source and target domains are not sufficiently similar, transferring from such weakly related source may hinder the performance in the target, a phenomenon known as negative transfer. The notion of negative transfer has been well recognized within the transfer learning community [20, 35]. An early paper [24] has conducted empirical study on a simple binary classification problem to demonstrate the existence of negative transfer. Some more recent work [7, 10, 3] has also observed similar negative impact while performing transfer learning on more complex tasks under different settings.

Despite these empirical observations, little research work has been published to analyze or predict negative transfer, and the following questions still remain open: First, while the notion being quite intuitive, it is not clear how negative transfer should be defined exactly. For example, how should we measure it at test time? What type of baseline should we compare with? Second, it is also unknown what factors cause negative transfer, and how to exploit them to determine that negative transfer may occur. Although the divergence between the source and target domain is certainly crucial, we do know how large it must be for negative transfer to occur, nor if it is the only factor. Third and most importantly, given limited or no labeled target data, how to detect and/or avoid negative transfer.

In this work, we take a step towards addressing these questions. We first derive a formal definition of negative transfer that is general and tractable in practice. Here tractable means we can explicitly measure its effect given the testing data. This definition further reveals three underlying factors of negative transfer that give us insights on when it could occur. Motivated by these theoretical observations, we develop a novel and highly generic technique based on adversarial networks to combat negative transfer. In our approach, a discriminator estimating both marginal and joint distributions is used as a gate to filter potentially harmful source data by reducing the bias between source and target risks, which corresponds to the idea of importance reweighting [5, 38]. Our experiments

involving eight transfer learning methods and four benchmark datasets reveal the three factors of negative transfer. In addition, we apply our method to six state-of-the-art deep methods and compare their performance, demonstrating that our approach substantially improves the performance of all base methods under potential negative transfer conditions by largely avoiding negative transfer.

2. Related Work

Transfer learning [20, 36] uses knowledge learned in the source domain to assist training in the target domain. Early methods exploit conventional statistical techniques such as instance weighting [14] and feature mapping [19, 32]. Compared to these earlier approaches, deep transfer networks achieve better results in discovering domain invariant factors [37]. Some deep methods [16, 27] transfer via distribution (mis)match measurements such as Maximum Mean Discrepancy (MMD) [14]. More recent work [9, 29, 3, 26] exploit generative adversarial networks (GANs) [12] and add a subnetwork as a domain discriminator. These methods achieve state-of-the-art on computer vision tasks [26] and some natural language processing tasks [17]. However, none of these techniques are specifically designed to tackle the problem of negative transfer.

Negative transfer Early work that noted negative transfer [24] was targeted at simple classifiers such as hierarchical Naive Bayes. Later, similar negative effects have also been observed in various settings including multi-source transfer learning [7], imbalanced distributions [10] and partial transfer learning [3]. While the importance of detecting and avoiding negative transfer has raised increasing attention [35], the literature lacks in-depth analysis.

3. Rethink Negative Transfer

Notation. We will use $P_S(X, Y)$ and $P_T(X, Y)$, respectively, to denote the joint distribution in the source and the target domain, where X is the input random variable and Y the output. Following the convention, we assume having access to labeled source set $\mathcal{S} = \{(x_s^i, y_s^i)\}_{i=1}^{n_s}$ sampled from the source joint $P_S(X, Y)$, a labeled target set $\mathcal{T}_l = \{(x_t^j, y_t^j)\}_{j=1}^{n_l}$ drawn from the target joint $P_T(X, Y)$, and an unlabeled target set $\mathcal{T}_u = \{x_u^k\}_{k=1}^{n_u}$ from the target marginal $P_T(X)$. For convenience, we define $\mathcal{T} = (\mathcal{T}_l, \mathcal{T}_u)$.

Transfer Learning. Under the notation, transfer learning aims at designing an algorithm A , which takes both the source and target domain data \mathcal{S}, \mathcal{T} as input, and outputs a better hypothesis (model) $h = A(\mathcal{S}, \mathcal{T})$, compared to only using the target-domain data. For model comparison, we will adapt the standard expected risk, which is defined as

$$R_{P_T}(h) := \mathbb{E}_{x, y \sim P_T} [\ell(h(x), y)], \quad (1)$$

with ℓ being the specific task loss. To make the setting meaningful, it is often assumed that $n_s \gg n_l$.

Negative Transfer. The notion of negative transfer lacks a rigorous definition. A widely accepted description of negative transfer [20, 35] is stated as “*transferring knowledge from the source can have a negative impact on the target learner*”. While intuitive, this description conceals many critical factors underlying negative transfer, among which we stress the following three points:

1. *Negative transfer should be defined w.r.t. the algorithm.* Specifically, the informal description above does not specify what the negative impact is compared with. For example, it will be misleading to only compare with the best possible algorithm only using the target data, i.e., defining negative transfer as

$$R_{P_T}(A(\mathcal{S}, \mathcal{T})) > \min_{A'} R_{P_T}(A'(\emptyset, \mathcal{T})), \quad (2)$$

because the increase in risk may not come from using the source-domain data, but the difference in algorithms. Therefore, to study negative transfer, one should focus on a specific algorithm at a time and compare its performance with and without the source-domain data. Hence, we define the *negative transfer condition* (NTC)¹ for any algorithm A as

$$R_{P_T}(A(\mathcal{S}, \mathcal{T})) > R_{P_T}(A(\emptyset, \mathcal{T})). \quad (3)$$

For convenience, we also define the *negative transfer gap* (NTG) as a quantifiable measure of negative transfer:

$$R_{P_T}(A(\mathcal{S}, \mathcal{T})) - R_{P_T}(A(\emptyset, \mathcal{T})), \quad (4)$$

and we say that negative transfer occurs if the negative transfer gap is positive and vice versa.

2. *Divergence between the joint distributions is the root to negative transfer.* As negative transfer is algorithm specific, it is natural to ask the question that whether there exists a transfer learning algorithm that can always improve the expected risk compared to its target-domain only baseline. It turned out this depends on the divergence between $P_S(X, Y)$ and $P_T(X, Y)$ [11]. As an extreme example, assume $P_S(X) = P_T(X)$ and $P_S(Y | x)$ is uniform for any x . In the case, there is no meaningful knowledge in $P_S(X, Y)$ at all. Hence, exploiting $\mathcal{S} \sim P_S(X, Y)$ will almost surely harm the estimation of $P_T(Y | X)$, unless $P_T(Y | X)$ is uniform.

In practice, we usually deal with the case where there exists some “systematic similarity” between $P_S(X, Y)$ and $P_T(X, Y)$. Then, an ideal transfer would figure out and take advantage of the similar part, leading to improved performance. However, if an algorithm fails to discard the divergent part and instead rely on it, one can expect

¹More discussion in the supplementary.

negative transfer to happen. Thus, regardless of the algorithm choice, the distribution shift is the actual root to negative transfer.

3. *Negative transfer largely depends on the size of the labeled target data.* While the previous discussion focuses on the distribution level, an overlooked factor of negative transfer is the size of the labeled target data, which has a mixed effect.

On one hand, for the same algorithm and distribution divergence, NTC depends on how well the algorithm can do using target data alone, i.e. the RHS of Eq.(3). In zero-shot transfer learning² [8, 21] where there is no labeled target data ($n_l = 0$), only using unlabeled target data would result in a weak random model and thus NTC is unlikely to be satisfied. When labeled target data is available [24, 29, 17], a better target-only baseline can be obtained using semi-supervised learning methods and so negative transfer is *relatively* more likely to occur. At the other end of the spectrum, if there is an abundance of labeled target data, then transferring from a even slightly different source domain could hurt the generalization. Thus, this shows that negative transfer is *relative*.

On the other hand, the amount of labeled target data has a direct effect on the feasibility and reliability of discovering shared regularity between the joint distributions. As discussed above, the key component of a transfer learning algorithm is to discover the similarity between the source joint $P_S(X, Y)$ and the target joint $P_T(X, Y)$. When labeled target data is not available ($n_l = 0$), one has to resort to the similarity between the marginals $P_S(X)$ and $P_T(X)$, which though has a theoretical limitation [2]. In contrast, if one has a considerable number of samples $(x_l, y_l) \sim P_T(X, Y)$ and $(x_s, y_s) \sim P_S(X, Y)$, the problem would be manageable. Therefore, an ideal transfer learning algorithm may be able to utilize labeled target data to mitigate the negative impact of unrelated source information.

With these points in mind, we next turn to the problem of how to avoid negative transfer in a systematic way.

4. Proposed Method

As discussed in Section 3, the key to achieving successful transfer and avoiding negative effects is to discover and exploit shared underlying structures between $P_S(X, Y)$ and $P_T(X, Y)$. In practice, there are many possible regularities one may take advantage of. To motivate our proposed method, we first review an important line of work and show how the observation in section 3 helps us to identify the limitation.

²It is often referred to as unsupervised domain adaptation in literature.

4.1. Domain Adversarial Network

As a notable example, a recent line of work [16, 8, 30] has successfully utilized a domain-invariant feature space assumption to achieve knowledge transfer. Specifically, it is assumed that there exists a feature space that is both shared by both source and target domains and discriminative enough for predicting the output. By learning a feature extractor F that can map both the source and target input to the same feature space, classifier learned on the source data can transfer to the target domain.

To find such a feature extractor, a representative solution is the Domain Adversarial Neural Network (DANN) [9], which exploits a generative adversarial network (GAN) framework to train the feature extractor F such that the feature distributions $P(F(X_S))$ and $P(F(X_T))$ cannot be distinguished by the discriminator D . Based on the shared feature space, a simple classifier C is trained on both source and target data. Formally, the objective can be written as:

$$\operatorname{argmin}_{F,C} \operatorname{argmax}_D \mathcal{L}_{\text{CLF}}(F, C) - \mu \mathcal{L}_{\text{ADV}}(F, D), \quad (5)$$

$$\begin{aligned} \mathcal{L}_{\text{CLF}}(F, C) = & \mathbb{E}_{x_l, y_l \sim \mathcal{T}_L} [\ell_{\text{CLF}}(C(F(x_l)), y_l)] \\ & + \mathbb{E}_{x_s, y_s \sim S} [\ell_{\text{CLF}}(C(F(x_s)), y_s)], \end{aligned} \quad (6)$$

$$\begin{aligned} \mathcal{L}_{\text{ADV}}(F, D) = & \mathbb{E}_{x_u \sim P_T(X)} [\log D(F(x_u))] \\ & + \mathbb{E}_{x_s \sim P_S(X)} [\log(1 - D(F(x_s)))]. \end{aligned} \quad (7)$$

Intuitively, \mathcal{L}_{CLF} is the supervised classification loss on both the target and source labeled data, \mathcal{L}_{ADV} is the standard GAN loss treating $F(x_u)$ and $F(x_s)$ as the true and fake features respectively, and μ is a hyper-parameter balancing the two terms. For more details and theoretical analysis, we refer readers to the original work [8].

Now, notice that the DANN objective implicitly makes the following assumption: For any $x_s \in \mathcal{X}_s$, there exists a $x_t \in \mathcal{X}_t$ such that

$$P_S(Y|x_s) = P_T(Y|x_t) = P(Y|F(x_s)) = P(Y|F(x_t)).$$

In other words, it is assumed that every single source sample can provide meaningful knowledge for transfer learning. However, as we have discussed in Section 3, some source samples may not be able to provide any knowledge at all. Consider the case where there is a source input $x_s \in \mathcal{X}_s$ such that $P_S(Y|x_s) \neq P_T(Y|x_t)$ for any x_t . Since $P(F(X_s)) = P(F(X_t))$ as a result of the GAN objective, there exists a $x' \in \mathcal{X}_t$ such that $F(x') = F(x_s)$ and hence $P(Y|F(x')) = P(Y|F(x_s))$. Then, if $P(Y|F(x_s))$ is trained on the source data to match $P_S(Y|x_s)$, it follows

$$P(Y|F(x')) = P(Y|F(x_s)) = P(Y|x_s) \neq P(Y|x').$$

As a result, relying on such “unrelated” source samples can hurt the performance, leading to negative transfer. Motivated by this limitation, we next present a simple yet ef-

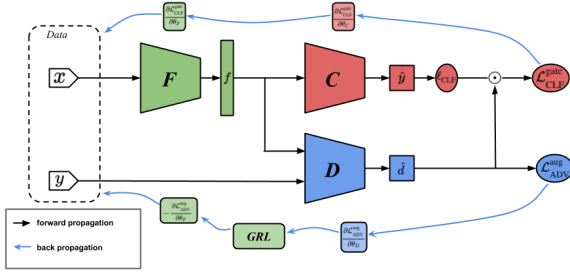


Figure 1. The architecture of proposed discriminator gate, where f is the extracted feature layer, \hat{y} and ℓ_{CLF} are predicted class label and its loss, \hat{d} is the predicted domain label, $\mathcal{L}_{\text{CLF}}^{\text{gate}}$ is the classification loss, $\mathcal{L}_{\text{ADV}}^{\text{aug}}$ is the adversarial learning loss; GRL stands for Gradient Reversal Layer and \odot is the Hadamard product.

fective method to deal with harmful source samples in a systematic way.

4.2. Discriminator Gate

The limitation of DANN comes from the unnecessary assumption that all source samples are equally useful. To eliminate the weakness, a natural idea is to reweight each source sample in some proper manner. To derive an appropriate weight, notice that the standard supervised learning objective can be rewritten as

$$\begin{aligned} \mathcal{L}_{\text{SUP}} &= \mathbb{E}_{x,y \sim P_T(X,Y)} [\ell_{\text{CLF}}(C(F(x)), y)] \\ &= \mathbb{E}_{x,y \sim P_S(X,Y)} \left[\frac{P_T(x,y)}{P_S(x,y)} \ell_{\text{CLF}}(C(F(x)), y) \right] \end{aligned} \quad (8)$$

where the density ratio $\frac{P_T(x,y)}{P_S(x,y)}$ naturally acts as an importance weight [5, 38] for the source data. Hence, the problem reduces to the classic problem of density ratio estimation.

Here, we exploit a GAN discriminator to perform the density ratio estimation [31]. Specifically, the discriminator takes both x and the paired y as input, and try to classify whether the pair is from the source domain (fake) or the target domain (true). At any point, the optimal discriminator is given by $D(x, y) = \frac{P_T(x,y)}{P_T(x,y) + P_S(x,y)}$, which implies

$$\frac{P_T(x, y)}{P_S(x, y)} = \frac{D(x, y)}{1 - D(x, y)}.$$

In our implementation, to save model parameters, we reuse the feature extractor to obtain the feature of x and instantiate $D(x, y)$ as $D(F(x), y)$. With the weight ratio, we modify the classification objective (6) in DANN as

$$\begin{aligned} \mathcal{L}_{\text{CLF}}^{\text{gate}}(C, F) &= \mathbb{E}_{x_l, y_l \sim \mathcal{T}_L} [\ell_{\text{CLF}}(C(F(x_l)), y_l)] \\ &\quad + \lambda \mathbb{E}_{x_s, y_s \sim \mathcal{S}} [\omega(x_s, y_s) \ell_{\text{CLF}}(C(F(x_s)), y_s)], \end{aligned} \quad (9)$$

$$\omega(x_s, y_s) = \text{SG} \left(\frac{D(x_s, y_s)}{1 - D(x_s, y_s)} \right)$$

where $\text{SG}(\cdot)$ denotes stop gradient and λ is another hyper-parameter introduce to scale the density ratio. As the density ratio acts like a gating function, we will refer to mechanism as discriminator gate.

On the other hand, we also augment the adversarial learning objective (7) by incorporating terms for matching the joint distributions:

$$\begin{aligned} \mathcal{L}_{\text{ADV}}^{\text{aug}}(F, D) &= \mathbb{E}_{x_u \sim P_T(X)} [\log D(F(x_u), \text{nil})] \\ &\quad + \mathbb{E}_{x_s \sim P_S(X)} [\log(1 - D(F(x_s), \text{nil}))] \\ &\quad + \mathbb{E}_{x_l, y_l \sim \mathcal{T}_L} [\log D(F(x_l), y_l)] \\ &\quad + \mathbb{E}_{x_s, y_s \sim \mathcal{S}} [\log(1 - D(F(x_s), y_s))], \end{aligned} \quad (10)$$

where nil denotes a dummy label which does not provide any label information and it is included to enable the discriminator D being used as both a marginal discriminator and a joint discriminator. As a benefit, the joint discriminator can utilize unlabeled target data since labeled data could be scarce. Similarly, under this objective, the feature network F will receive gradient from both the marginal discriminator and the joint discriminator. Theoretically speaking, the joint matching objective subsumes the the marginal matching objective, as matched joint distribution implied matched marginals. However, in practice, the labeled target data \mathcal{T}_L is usually limited, making the joint matching objective itself insufficient. This particular design choice echos our discussion about how the size of labeled target data can influence our algorithm design in Section 3.

Combining the gated classification objective (9) and the augmented adversarial learning objective (10), we arrive at our proposed approach to transfer learning

$$\underset{F, C}{\text{argmin}} \underset{D}{\text{argmax}} \mathcal{L}_{\text{CLF}}^{\text{gate}}(F, C) - \mu \mathcal{L}_{\text{ADV}}^{\text{aug}}(F, D). \quad (11)$$

The overall architecture is illustrated in Figure 1. Finally, although the presentation of the proposed method is based on DANN, our method is highly general and can be applied directly to other adversarial transfer learning methods. In fact, we can even extend non-adversarial methods to achieve similar goals. In our experiments, we adapt six deep methods [16, 27, 8, 30, 4, 26] of three different categories to demonstrate the effectiveness of our method.

5. Experiments

We conduct extensive experiments on four benchmark datasets to (1) analyze negative transfer and its three underlying aspects, and (2) evaluate our proposed discriminator gate on six state-of-the-art methods.

5.1. Datasets

We use four standard datasets with different levels of difficulties: (1) small domain shift: Digits dataset, (2) moderate domain shift: Office-31 dataset, and (3) large domain shift: Office-Home and VisDA datasets.

Digits contains three standard digit classification datasets: MNIST, USPS, SVHN. Each dataset contains large amount of images belonging to 10 classes (0-9). This dataset is relatively easy due to its simple data distribution and therefore we only consider a harder case: **SVHN**→**MNIST**. Specifically, SVHN [18] contains 73K images cropped from house numbers in Google Street View images while MNIST [15] consists of 70K handwritten digits captured under constrained conditions.

Office-31 [25] is the most widely used dataset for visual transfer learning. It contains 4,652 images of 31 categories from three domains: Amazon(**A**) which contains images from amazon.com, Webcam(**W**) and DSLR(**D**) which consist of images taken by web camera and SLR camera. We evaluate all methods across three tasks: **W**→**D**, **A**→**D**, and **D**→**A**. We select these three settings because the other three possible cases yield similar results.

Office-Home [33] is a more challenging dataset that consists of about 15,500 images of 65 categories that crawled through several search engines and online image directories. In particular, it contains four domains: Artistic images(**Ar**), Clip Art(**Cl**), Product images(**Pr**) and Real-World images(**Rw**). We want to test on more interesting and practical transfer learning tasks involving adaptation from synthetic to real-world and thus we consider three transfer tasks: **Ar**→**Rw**, **Cl**→**Rw**, and **Pr**→**Rw**. In addition, we choose to use the first 25 categories in alphabetic order to make our results more comparable to previous studies [4].

VisDA [22] is another challenging synthetic to real dataset. We use the training set as the synthetic source and the testing set as the real-world target (**Synthetic**→**Real**). Specifically, the training set contains 152K synthetic images generated by rendering 3D models and the testing set contains 72K real images from crops of Youtube Bounding Box dataset [23], both contain 12 categories.

5.2. Experimental Setup

To better study negative transfer effect and evaluate our approach, we need to control the three factors discussed in Section 3, namely *algorithm factor*, *divergence factor* and *target factor*. In our experiments, we adopt the following mechanism to control each of them.

Divergence factor: Since existing benchmark datasets usually contain domains that are similar to each other, we need to alter their distributions to better observe negative transfer effect. In our experiments, we introduce two perturbation rates ϵ_x and ϵ_y to respectively control the marginal divergence and the conditional divergence between two domains. Specifically, for each source domain data we independently draw a Bernoulli variable of probability ϵ_x , and if it returns one, we add a series of random noises to the input image such as random rotation, random salt&pepper noise, random flipping, etc (examples shown Figure 2). Ac-

cording to studies in [28, 1], such perturbation is enough to cause misclassification for neural networks and therefore is sufficient for our purpose. In addition, we draw a second independent Bernoulli variable of probability ϵ_y and assign a randomly picked label if it returns one.



Figure 2. Example images before & after perturbation

Target factor: Similar to previous works, we use all labeled source data for training. For the target data, we first split 50% as training set and the rest 50% for testing. In addition, we use all of target training data as unlabeled target data and use $L\%$ percent of them as labeled target data. A symmetric study of source data can be found in [34].

Algorithm factor: To provide a more comprehensive study of negative transfer, we evaluate the performance of eight transfer learning methods of five categories: **TCA** [19], **KMM** [14], **DAN** [16], **DCORAL** [27], **DANN** a.k.a RevGrad [8], **ADDA** [29], **PADA** [4], **GTA** [26]. Specifically, (1) TCA is a conventional method based on MMD-regularized PCA, (2) KMM is a conventional sample reweighting method, (3) DAN and DCORAL are non-adversarial deep methods which use a distribution measurement as an extra loss, (4) DANN, ADDA and PADA use adversarial learning and directly train a discriminator, (5) GTA is a GAN based method that includes a generator to generate actual images in addition to the discriminator. We mainly follow the default settings and training procedures for model selection as explained in their respective papers. However, for fair comparison, we use the same feature extractor and classifier architecture for all deep methods. In particular, we use a modified LeNet as detailed in [26] for the Digits dataset. For other datasets, we fine-tune from the ResNet-50 [13] pretrained on ImageNet with an added 256-dimension bottleneck layer between the *res5c* and *fc* layers. To compare the performance of our proposed approach, we adapt a gated version for each of the six deep methods (e.g **DANN_{gate}** is the gated DANN). Specifically, we extend DANN, ADDA and PADA straightforwardly as described in Section 4.2. For GTA, we extend the discriminator to take in class labels and output domain label predictions as gates. For DAN and DCORAL, we add an extra discriminator network to be used as gates but the general network is not trained adversarially. For hyper-parameters, we set $\lambda = 1$ and μ progressively increased from 0 to 1 in all our experiments. For each transfer task, we compare the average classification accuracy over five random repeats. To

Table 1. Classification accuracy (%) of DANN and DANN_{gate} on tasks W→D and A→D. Perturbation rates are set equal, i.e. $\epsilon = \epsilon_x = \epsilon_y$. NTG₁ and NTG₂ are negative transfer gaps for DANN and DANN_{gate}. Δ is the performance gain of DANN_{gate} compared to DANN.

	W→D					A→D					$L_{\%}$
	$\epsilon=0.0$	$\epsilon=0.3$	$\epsilon=0.7$	$\epsilon=0.9$	Avg	$\epsilon=0.0$	$\epsilon=0.3$	$\epsilon=0.7$	$\epsilon=0.9$	Avg	
DANN	99.1±0.8	83.2±1.4	47.2±2.7	32.2±3.5	65.4	76.2±1.5	40.9±1.1	21.3±2.7	12.9±3.7	37.8	0%
NTG ₁	-96.5	-80.3	-44.1	-28.3	-62.3	-73.7	-37.3	-17.2	-9.7	-34.5	
DANN _{gate}	98.9±0.6	83.3±2.1	48.4±2.5	32.1±3.1	65.7	76.0±1.2	41.0±1.6	21.5±3.1	13.2±2.4	37.9	
NTG ₂	-96.3	-80.4	-45.3	-28.2	-62.6	-73.5	-37.4	-17.4	-10.0	-34.6	
Δ	↓0.2	↑0.1	↑1.2	↓0.1	↑0.3	↓0.2	↑0.1	↑0.2	↑0.3	↑0.1	
DANN	99.5±0.4	86.8±2.8	73.1±3.3	48.8±4.3	77.0	78.6±2.7	54.8±3.1	49.6±2.1	32.3±2.6	53.8	10%
NTG ₁	-48.7	-37.8	-23.6	1.6	-27.1	-28.4	-4.4	1.2	18.4	-3.3	
DANN _{gate}	99.2±0.3	85.4±2.6	79.4±2.9	50.4±3.2	78.6	85.1±1.7	60.2±2.1	58.3±2.0	49.1±2.5	63.2	
NTG ₂	-48.4	-36.4	-29.9	0.0	-28.7	-34.9	-9.8	-7.5	1.6	-12.7	
Δ	↓0.3	↓1.4	↑6.3	↑1.6	↑1.6	↑6.5	↑5.4	↑8.7	↑16.8	↑9.4	
DANN	99.6±0.2	89.7±1.6	78.4±2.5	70.5±4.3	84.6	80.2±2.0	73.3±2.2	70.2±3.3	51.3±4.3	68.8	30%
NTG ₁	-18.5	-10.3	1.8	8.2	-4.7	-1.5	6.5	8.9	28.4	10.6	
DANN _{gate}	100.0±0.1	90.4±1.8	82.0±1.8	79.9±3.8	88.1	89.0±1.5	82.6±1.0	81.3±2.1	80.6±1.8	83.4	
NTG ₂	-18.9	-11.0	-1.8	-1.2	-8.2	-10.3	-2.8	-2.2	-0.9	-4.1	
Δ	↑0.4	↑0.7	↑3.6	↑9.4	↑2.6	↑8.8	↑9.3	↑11.1	↑29.3	↑14.6	
DANN	100.0±0.0	92.2±1.7	85.8±2.3	78.2±4.8	89.1	84.5±1.9	77.6±3.8	70.6±4.9	65.4±6.3	74.5	50%
NTG ₁	-11.7	-3.2	3.8	10.4	-0.2	4.6	12.1	18.8	23.2	14.7	
DANN _{gate}	100.0±0.0	93.3±1.7	91.2±1.5	89.5±3.4	92.5	93.2±1.3	91.4±1.2	90.2±2.0	89.8±1.9	91.2	
NTG ₂	-11.7	-4.3	-1.6	-0.9	-4.6	-4.1	-1.7	-0.8	-1.2	-2.0	
Δ	→0.0	↑1.1	↑5.4	↑11.3	↑4.5	↑8.7	↑13.8	↑19.6	↑24.4	↑16.7	

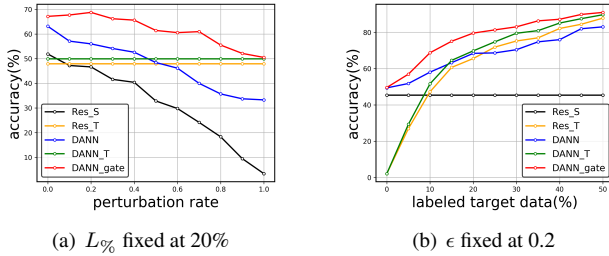


Figure 3. Incremental performance on task Pr→Rw. Res_S and Res_T are ResNet-50 baselines trained using only source data and only target data. Perturbation rates are set equal, i.e. $\epsilon = \epsilon_x = \epsilon_y$.

test whether negative transfer occurs, we measure the negative transfer gap (NTG) as the gap between the accuracy of target-only baseline and that of the original method. For instance, for DANN, the target-only baseline is DANN_T which treats labeled target data as “source” data and uses unlabeled data as usual. A positive NTG indicates the occurrence of negative transfer and vice versa.

5.3. Results and Analysis

5.3.1 Study of Negative Transfer

To reveal the three dependent factors, we study the effect of negative transfer under different methods with varying perturbation rates (ϵ_x, ϵ_y) and target labeled data ($L_{\%}$).

Divergence factor. The performance of DANN under different settings of ϵ and $L_{\%}$ on two tasks of Office-31 are shown in Table 1. We observe an increasing negative transfer gap as we increase the perturbation rate in all cases. In some cases such as $L_{\%} = 10\%$, we can even observe a

change in the sign of NTG. For a more fine-grained study, we investigate a wider spectrum of distribution divergence by gradually increasing ϵ from 0.0 to 1.0 in Figure 3(a). Although DANN is better than DANN_T when ϵ is small, its performance degrades quickly as ϵ increases and drops below DANN_T, indicating the occurrence of negative transfer. On the other hand, by fixing $\epsilon_y = 0$ and using two domains **W** and **D** that are known to be particularly similar, we study negative transfer under the assumption of covariate shift in Table 3, and observe that negative transfer does *not* occur even with high ϵ_x and descent $L_{\%}$. These experimental results confirms that the distribution divergence is an important factor of negative transfer.

Table 3. Classification accuracy (%) under the Covariate Shift assumption on task W→D. ϵ_y is fixed at 0. Negative transfer gap is shown in brackets.

Method	$\epsilon_x=0.7$ $L_{\%}=10\%$	$\epsilon_x=1.0$ $L_{\%}=30\%$
DAN	81.2(-29.3)	85.8(-6.2)
DANN	83.0(-30.8)	86.1(-6.5)
GTA	85.5(-33.5)	88.1(-8.0)

Target factor. Fixing a specific ϵ , we observe that the negative transfer gap increases as $L_{\%}$ increases in Table 1. In the extreme case of unsupervised adaptation ($L_{\%} = 0\%$), NTG stays negative even if two domains are far apart ($\epsilon = 0.9$). In Figure 3(b), we fix $\epsilon = 0.2$ and plot the performance curve as $L_{\%}$ increases. We can see that while both DANN and DANN_T perform better with more labeled target data, DANN is affected by the divergence factor and outperformed by DANN_T when $L_{\%}$ becomes larger. This observation shows that negative transfer is relative and it depends on target labeled data.

Algorithm factor. In Table 2, we compare the results

Table 2. Classification accuracy (%) of state-of-the-art methods on four benchmark datasets with negative transfer gap shown in brackets. Perturbation rates are fixed at $\epsilon_x = \epsilon_y = 0.7$. Target labeled ratio is set at $L\% = 10\%$ and we further enforce each task to use at most 3 labeled target samples per class.

Method	Digits	Office-31			Office-Home			VisDA	Avg
	SVHN→MNIST	W→D	A→D	D→A	Ar→Rw	Cl→Rw	Pr→Rw	Synthetic→Real	
TCA[19]	58.7(18.2)	54.2(-4.2)	11.4(20.5)	13.1(18.4)	-	-	-	-	34.4(13.2)
KMM[14]	70.9(6.0)	58.7(-8.5)	18.5(13.4)	17.7(13.8)	-	-	-	-	41.5(6.2)
DAN[16]	78.5(-4.4)	76.3(-19.5)	55.0(-1.3)	39.2(4.9)	43.2(3.8)	30.2(5.8)	47.2(4.0)	28.4(7.2)	49.8(0.1)
DAN _{gate}	82.2(-8.1)	78.7(-21.9)	60.4(-6.7)	43.9(0.2)	46.8(0.2)	38.0(-2.0)	50.4(0.8)	36.2(-0.6)	54.6(-4.7)
Δ DAN	$\uparrow 3.7$	$\uparrow 2.4$	$\uparrow 5.4$	$\uparrow 4.7$	$\uparrow 3.6$	$\uparrow 7.8$	$\uparrow 3.2$	$\uparrow 7.8$	$\uparrow 4.8$
DCORAL[27]	75.2(-1.2)	75.7(-18.9)	53.8(-0.4)	37.4(5.0)	44.0(3.7)	32.4(4.1)	48.0(2.2)	30.5(5.7)	49.6(0.0)
DCORAL _{gate}	81.0(-7.0)	78.2(-21.4)	59.0(-5.6)	43.2(-0.8)	48.5(-0.8)	40.0(-3.5)	51.6(-1.4)	35.8(0.4)	54.7(-5.1)
Δ DCORAL	$\uparrow 5.8$	$\uparrow 2.5$	$\uparrow 5.2$	$\uparrow 5.8$	$\uparrow 4.5$	$\uparrow 7.6$	$\uparrow 3.6$	$\uparrow 5.3$	$\uparrow 5.1$
DANN[8]	68.3(7.7)	75.0(-19.2)	51.0(2.3)	38.2(5.6)	42.8(4.2)	28.5(7.7)	42.0(10.0)	29.9(6.0)	47.0(3.0)
DANN _{gate}	78.1(-2.1)	80.2(-24.4)	61.8(-8.5)	48.3(-4.5)	51.2(-4.2)	43.8(-7.6)	55.2(-3.2)	40.5(-4.6)	57.4(-7.4)
Δ DANN	$\uparrow 9.8$	$\uparrow 5.2$	$\uparrow 10.8$	$\uparrow 10.1$	$\uparrow 9.4$	$\uparrow 14.7$	$\uparrow 13.2$	$\uparrow 10.6$	$\uparrow 10.4$
ADDA[30]	63.2(12.2)	74.5(-18.1)	49.9(2.2)	38.3(5.1)	41.4(6.0)	25.2(13.5)	43.2(7.2)	28.0(7.3)	45.5(4.4)
ADDA _{gate}	79.4(-4.0)	82.9(-26.5)	64.2(-12.1)	47.7(-4.3)	52.2(-4.8)	48.0(-9.3)	58.2(-7.8)	43.0(-7.7)	59.5(-9.6)
Δ ADDA	$\uparrow 16.2$	$\uparrow 8.4$	$\uparrow 14.3$	$\uparrow 9.4$	$\uparrow 10.8$	$\uparrow 22.8$	$\uparrow 15.0$	$\uparrow 15.0$	$\uparrow 14.0$
PADA[4]	69.7(6.5)	75.5(-19.0)	50.2(1.9)	38.7(5.1)	43.2(3.8)	30.1(5.5)	43.4(6.6)	32.2(5.5)	47.9(2.0)
PADA _{gate}	81.8(-5.6)	81.6(-25.1)	62.1(-10.0)	44.8(-1.0)	52.8(-5.8)	45.2(-9.6)	54.5(-4.5)	41.4(-5.7)	58.0(-8.1)
Δ PADA	$\uparrow 12.1$	$\uparrow 5.9$	$\uparrow 11.9$	$\uparrow 6.1$	$\uparrow 9.6$	$\uparrow 15.1$	$\uparrow 11.1$	$\uparrow 11.2$	$\uparrow 10.1$
GTA[26]	81.2(-6.8)	78.9(-20.5)	58.4(-7.2)	42.2(2.8)	48.2(1.0)	33.1(5.1)	50.2(-0.1)	31.2(4.2)	52.9(-2.7)
GTA _{gate}	83.3(-8.9)	85.8(-27.4)	66.7(-15.5)	48.5(-3.5)	55.0(-5.8)	44.9(-6.7)	58.0(-7.7)	43.8(-8.4)	60.8(-10.6)
Δ GTA	$\uparrow 2.1$	$\uparrow 6.9$	$\uparrow 8.3$	$\uparrow 6.3$	$\uparrow 6.8$	$\uparrow 11.8$	$\uparrow 7.8$	$\uparrow 12.6$	$\uparrow 7.9$
Δ Avg	$\uparrow 8.3$	$\uparrow 5.2$	$\uparrow 8.1$	$\uparrow 7.1$	$\uparrow 7.5$	$\uparrow 13.3$	$\uparrow 8.9$	$\uparrow 10.4$	

of all methods under a more practically interesting scenario of moderately different distributions and limited amount of labeled target data. We observe that some methods are more vulnerable to negative transfer than the other even using the same training data. For conventional methods, instance-reweighting method KMM achieves smaller NTG compared to feature selection method TCA, possibly because KMM can assign small weights to source instances with dissimilar input features. For deep methods, we find GTA to be the most robust method against negative transfer since it takes both label information and random noises as inputs to the generator network. More interestingly, we observe that methods based on distribution measurement such as MMD (e.g. DAN) achieve smaller NTG than methods based on adversarial networks (e.g. DANN), even though the latter tends to perform better when distributions are similar. This is consistent with findings in previous works [3] and one possible explanation is that adversarial network's better capability of matching source and target domains leads to more severe negative transfer. Similarly, ADDA has better matching power by using two separate feature extractors, but it results in larger NTG compared to DANN.

5.3.2 Evaluation of Discriminator Gate

We compare our gated models with their respective state-of-the-art methods on the benchmarks in Table 2. Even using limited amount of labeled target data, our proposed method consistently improves the performance for all deep methods on all tasks. More importantly, our method can largely eliminate the negative impact of less related source data and

Table 4. Ablation Study on task A→D. DANN_{gate}-only applies only the discriminator gate while DANN_{label}-only only uses label information without the gate. DANN_{joint} is a variant of DANN_{gate} where the feature network only matches the joint distribution (last two lines of Eq.10), DANN_{marginal} only matches the marginal distribution, and DANN_{none} matches none of them. DANN_{oracle} excludes perturbed source data via human oracle.

Method	Setting ($\epsilon, L\%$)				Avg
	0.7, 30%	0.7, 10%	0.3, 30%	0.3, 10%	
DANN	70.4	49.4	72.5	54.3	61.7
DANN _T	79.5	50.7	80.3	50.1	65.2
DANN _{oracle}	81.6	58.5	89.1	85.4	78.7
DANN _{gate} -only	76.3	53.8	78.0	55.7	66.0
DANN _{label} -only	74.4	52.5	77.5	55.0	64.9
DANN _{joint}	82.3	57.6	83.1	59.4	70.6
DANN _{marginal}	80.6	56.5	81.5	58.6	69.3
DANN _{none}	79.6	52.4	79.7	57.5	67.3
DANN _{gate}	82.5	58.7	82.7	60.7	71.2

avoid negative transfer (e.g. DANN_{gate} achieves negative average NTG while DANN gets positive NTG). Specifically, our method achieves larger accuracy gains on harder tasks such as synthetic to real-world tasks in Office-Home and VisDA. This is mainly because source domains in these tasks tend to contain more unrelated samples. This finding is also consistent with results in Table 1 and Figure 3(a) where we can observe larger performance gains as perturbation rates increase. In the extreme case where the source domain is degenerate ($\epsilon = 1.0$ in Figure 3(a)), the gated model achieves comparable results to those of DANN_T. On the other hand, the results of DANN and DANN_{gate} are similar when source domain is closely related to the target ($\epsilon = 0.0$ on task W→D in Table 1). This indicates that the discriminator gate can control the trade-off between maxi-

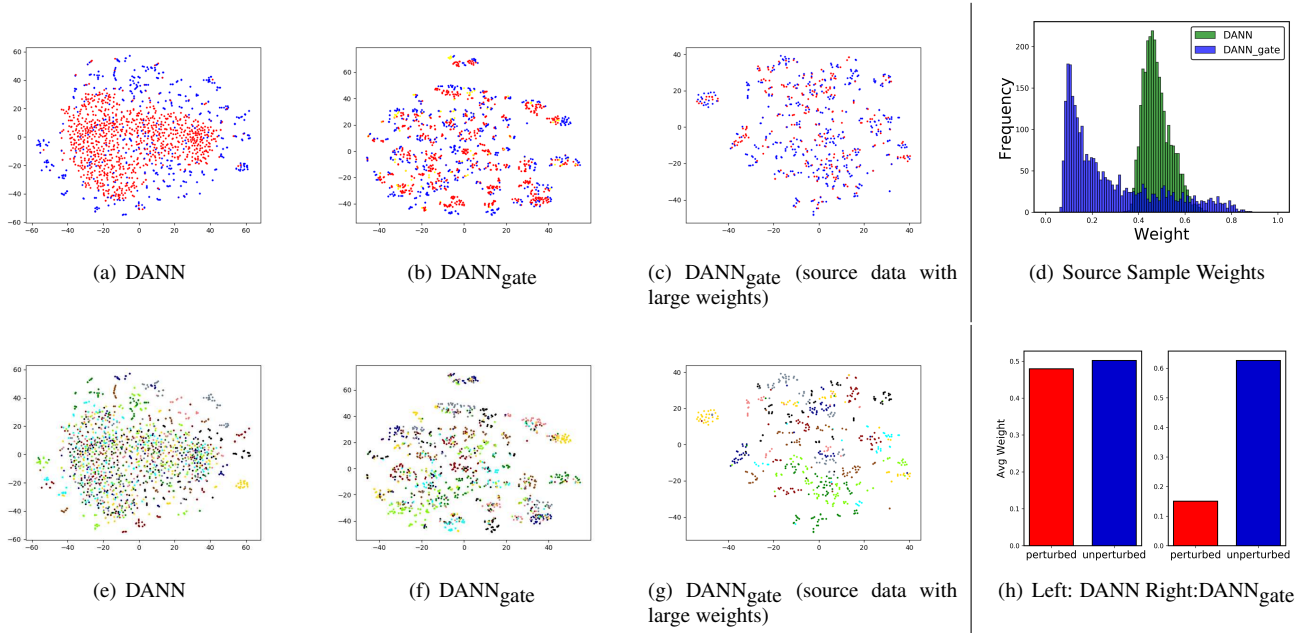


Figure 4. Visualization on A→W, with $\epsilon = 0.7$, $L\% = 30\%$. **Left:** The t-SNE visualization. First row shows domain info with red for source samples (yellow for weights > 0.4) and blue for target samples. Second row shows corresponding class info. **Right:** Top shows the histogram of discriminator weights for source samples. Bottom shows average weights for perturbed and unperturbed samples.

mal transfer and alleviating negative impact.

Ablation Study. We report the results of ablation study in Table 4 and analyze the effects of several components in our method subject to different settings of transfer tasks. First, both $\text{DANN}_{\text{gate-only}}$ and $\text{DANN}_{\text{label-only}}$ perform better than DANN but worse than $\text{DANN}_{\text{gate}}$, showing that the discriminator gate and estimating joint distributions can both improve performance but their combination yields full performance benefit. Second, $\text{DANN}_{\text{joint}}$ obtains higher accuracy results than $\text{DANN}_{\text{marginal}}$ and $\text{DANN}_{\text{none}}$ since matching joint distributions is the key to avoid negative transfer when both marginal and conditional distributions shift. However, while $\text{DANN}_{\text{joint}}$ achieves comparable results as $\text{DANN}_{\text{gate}}$ when $L\% = 30\%$, it performs worse than $\text{DANN}_{\text{gate}}$ when $L\% = 10\%$. This shows that utilizing unlabeled target data to match marginal distributions can be beneficial when labeled target data is scarce. Lastly, it is inspiring to see $\text{DANN}_{\text{gate}}$ outperforms $\text{DANN}_{\text{oracle}}$ when perturbation rate is high. This is because less unperturbed source data are used for $\text{DANN}_{\text{oracle}}$ but $\text{DANN}_{\text{gate}}$ can utilize perturbed source data that contain related information. This further shows the effectiveness of our approach.

Feature Visualization. We visualize the t-SNE embeddings [6] of the bottleneck representations in Figure 4. The first column shows that, when perturbation rate is high, DANN cannot align the two domains well and it fails to discriminate both source and target classes as different classes are mixed together. The second column illustrates the discriminator gate can improve the alignment by assigning less

weights to unrelated source data. For instance, we can see some source data from different classes mixed in the yellow cluster at the center right but they get assigned smaller weights. The third column shows the embeddings after we remove source data with small discriminator weights (< 0.4). We can observe that target data are much better clustered compared to that of DANN. These in-depth results demonstrate the efficacy of discriminator gate method.

Statistics of Instance Weights. We illustrate the discriminator output ($\frac{P_T(x_s^i, y_s^i)}{P_T(x_s^i, y_s^i) + P_S(x_s^i, y_s^i)}$) for each source data in Figure 4(d). We can observe that DANN fails to discriminate unrelated source data as all weights concentrate around 0.5 in the middle. On the other hand, $\text{DANN}_{\text{gate}}$ assigns smaller weights to a large portion of source data (since perturbation rate is high) and thus filters out unrelated information. Figure 4(h) further shows that DANN assign similar average weights for perturbed and unperturbed source data while $\text{DANN}_{\text{gate}}$ outputs much smaller values for perturbed data but higher ones for unperturbed data.

6. Conclusion

In this work, we analyze the problem of negative transfer and propose a novel discriminator gate technique to avoid it. We show that negative transfer directly relates to specific algorithms, domain divergence and target data. Experiments demonstrate these factors and the efficacy of our method. Our method consistently improves the performance of base methods and largely avoids negative transfer. Understanding negative transfer in more complex transfer tasks and settings should be addressed in a future research.

References

- [1] A. Azulay and Y. Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *arXiv preprint arXiv:1805.12177*, 2018.
- [2] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems (NIPS)*, pages 137–144, 2007.
- [3] Z. Cao, M. Long, J. Wang, and M. I. Jordan. Partial transfer learning with selective adversarial networks. 2018.
- [4] Z. Cao, L. Ma, M. Long, and J. Wang. Partial adversarial domain adaptation. 2018.
- [5] C. Cortes, Y. Mansour, and M. Mohri. Learning bounds for importance weighting. In *Advances in neural information processing systems*, pages 442–450, 2010.
- [6] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655, 2014.
- [7] L. Duan, D. Xu, and S.-F. Chang. Exploiting web images for event recognition in consumer videos: A multiple source domain adaptation approach. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1338–1345. IEEE, 2012.
- [8] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, pages 1180–1189, 2015.
- [9] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.
- [10] L. Ge, J. Gao, H. Ngo, K. Li, and A. Zhang. On handling negative transfer and imbalanced distributions in multiple source transfer learning. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 7(4):254–271, 2014.
- [11] M. Gong, K. Zhang, T. Liu, D. Tao, C. Glymour, and B. Schölkopf. Domain adaptation with conditional transferable components. In *International conference on machine learning*, pages 2839–2848, 2016.
- [12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
- [13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [14] J. Huang, A. Gretton, K. M. Borgwardt, B. Schölkopf, and A. J. Smola. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems (NIPS)*, pages 601–608, 2007.
- [15] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [16] M. Long, Y. Cao, J. Wang, and M. I. Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning*, 2015.
- [17] S. Moon and J. Carbonell. Completely heterogeneous transfer learning with attention-what and what not to transfer. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2508–2514, 2017.
- [18] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng. Reading digits in natural images with unsupervised feature learning. volume 2011, page 5, 2011.
- [19] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2011.
- [20] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.
- [21] Z. Pei, Z. Cao, M. Long, and J. Wang. Multi-adversarial domain adaptation. In *AAAI Conference on Artificial Intelligence*, 2018.
- [22] X. Peng, B. Usman, N. Kaushik, J. Hoffman, D. Wang, and K. Saenko. Visda: The visual domain adaptation challenge, 2017.
- [23] E. Real, J. Shlens, S. Mazzocchi, X. Pan, and V. Vanhoucke. Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 7464–7473. IEEE, 2017.
- [24] M. T. Rosenstein, Z. Marx, L. P. Kaelbling, and T. G. Dietterich. To transfer or not to transfer. In *NIPS 2005 workshop on transfer learning*, volume 898, pages 1–4, 2005.
- [25] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010.
- [26] S. Sankaranarayanan, Y. Balaji, C. D. Castillo, and R. Chellappa. Generate to adapt: Aligning domains using generative adversarial networks. *Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [27] B. Sun and K. Saenko. Deep coral: Correlation alignment for deep domain adaptation. pages 443–450. Springer, 2016.
- [28] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *ICLR*, 2014.
- [29] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4068–4076, 2015.
- [30] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. In *Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 4, 2017.
- [31] M. Uehara, I. Sato, M. Suzuki, K. Nakayama, and Y. Matsuo. Generative adversarial nets from a density ratio estimation perspective. *arXiv preprint arXiv:1610.02920*, 2016.
- [32] S. Uguroglu and J. Carbonell. Feature selection for transfer learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 430–442. Springer, 2011.
- [33] H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan. Deep hashing network for unsupervised domain

- adaptation. In *(IEEE) Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [34] Z. Wang and J. Carbonell. Towards more reliable transfer learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 794–810, 2018.
 - [35] K. Weiss, T. M. Khoshgoftaar, and D. Wang. A survey of transfer learning. *Journal of Big Data*, 3(1):9, 2016.
 - [36] L. Yang, S. Hanneke, and J. Carbonell. A theory of transfer learning with applications to active learning. *Machine learning*, 90(2):161–189, 2013.
 - [37] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328, 2014.
 - [38] Y.-L. Yu and C. Szepesvári. Analysis of kernel mean matching under covariate shift. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1147–1154, 2012.
 - [39] A. R. Zamir, A. Sax, W. Shen, L. Guibas, J. Malik, and S. Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3712–3722, 2018.