Transferable Query Selection for Active Domain Adaptation

Bo Fu*, Zhangjie Cao*, Jianmin Wang, Mingsheng Long (✉)
School of Software, BNRist, Tsinghua University, China
{microhhh9,caozhangjie14}@gmail.com, {jimwang,mingsheng}@tsinghua.edu.cn

Abstract

Unsupervised domain adaptation (UDA) enables transferring knowledge from a related source domain to a fully unlabeled target domain. Despite the significant advances in UDA, the performance gap remains quite large between UDA and supervised learning with fully labeled target data. Active domain adaptation (ADA) mitigates the gap under minimal annotation cost by selecting a small quota of target samples to annotate and incorporating them into training. Due to the domain shift, the query selection criteria of prior active learning methods may be ineffective to select the most informative target samples for annotation. In this paper, we propose Transferable Query Selection (TQS), which selects the most informative samples under domain shift by an ensemble of three new criteria: transferable committee, transferable uncertainty, and transferable domainness. We further develop a randomized selection algorithm to enhance the diversity of the selected samples. Experiments show that TQS remarkably outperforms previous UDA and ADA methods on several domain adaptation datasets. Deeper analyses demonstrate that TQS can select the most informative target samples under the domain shift.

1. Introduction

Wide attention has been paid to Unsupervised Domain Adaptation (UDA) [11, 42], which adapts a model learned in the labeled source domain to the target domain with only unlabeled data. However, UDA still falls far in accuracy behind its supervised learning counterpart [37, 3]. As shown in [19], the "market value" of target labeled data is much more pronounced than that of source labeled data, and even a few target labeled data can improve Domain Adaptation (DA) models significantly. Thus, a promising DA paradigm is to informatively annotate a small quota of target data that maximally benefits the DA model. This learning paradigm is known as **Active Domain Adaptation (ADA)** [27].

Pool-based active learning [33] adopts a *query selection* strategy to select the most useful data for training from a large unlabeled data pool, which can be incorporated into DA to maximize the revenue of the small labeling budget. Previous active learning methods mainly apply three query selection strategies: *committee*, *uncertainty*, and *representativeness* [33], which respectively query the samples with maximal disagreement by a classifier committee [4, 34], of the highest uncertainty [15, 12], and more distinctive from labeled samples [13, 35]. While successful in general for single-domain active learning, these traditional criteria are *not transferable*. As we explained later, these criteria fail to select informative target samples under the *domain shift*.

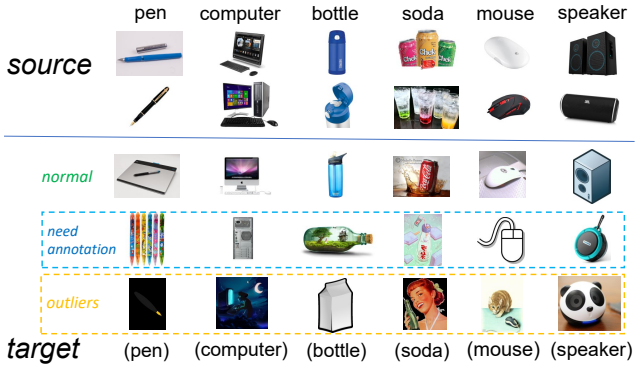


Figure 1. Examples of three kinds of target data. The *normal* row is similar to source images and can be classified correctly by UDA methods. The *need annotation* row is quite different from source images and cannot be recognized by UDA, which needs manual annotation. The *outliers* row shows outlier samples like black images or noisy objects, which are not informative for classification.

The selection criterion under domain shift is the major challenge of Active Domain Adaptation (ADA). In this paper, we propose **Transferable Query Selection (TQS)** by *transferable committee*, *transferable uncertainty*, and *transferable domainness*. The 'transferable' here means that the criteria are specially designed to mitigate the domain gap. To build a transferable committee, we enforce multiple classifiers to only pass target low-density areas by training with adversarial samples. Based on the multi-classifier architecture of the transferable committee, we compute the trans-

*Equal contribution

Table 1. Comparison of prior Active Learning (AL) and ADA methods in three dimensions: committee, uncertainty, and representativeness. **The main difference is that the existing criteria are *non-transferable* while our TQS is designed as a *transferable* criterion for ADA.**

Method	Non-Transferable			Transferable			Outliers
	Committee	Uncertainty	Representativeness	Committee	Uncertainty	Representativeness	
QBC [34]	✓	✗	✗	✗	✗	✗	✗
UCN [15]	✗	✓	✗	✗	✗	✗	✗
Cluster [43]	✗	✗	✓	✗	✗	✗	✗
ADMA [13]	✗	✓	✓	✗	✗	✗	✗
AADA [35]	✗	✓	✓	✗	✗	✗	✗
TQS (ours)	✗	✗	✗	✓	✓	✓	✓

ferable uncertainty from the *ensemble* of multiple classifiers to reduce the variance of the uncertainty. We further adopt the margin function to assess uncertainty, which is more discriminative and stable. Transferable domainness simultaneously decides whether a target sample is an outlier and measures its *distinctiveness* if it is not an outlier. The three transferable criteria are complementary to form a unified criterion that selects the most informative samples. We also propose a randomized selection mechanism to increase sample diversity. In summary:

- We propose **Transferable Query Selection(TQS)**, a novel query selection criterion for active domain adaptation, which is an integrated ensemble of transferable committee, transferable uncertainty, and transferable domainness. We demonstrate that the three criteria are complementary to enable informative query selection.
- We further design a randomized selection mechanism to increase sample diversity and prevent different selected samples from providing redundant information.
- Experimental results on several DA benchmarks show that the proposed TQS criterion selects the most informative target samples and achieves higher target domain accuracy than unsupervised domain adaptation, active learning, and active domain adaptation methods.

2. Related Work

Domain Adaptation (DA). One of the DA paradigms is Unsupervised Domain Adaptation (UDA) where no labeled data are available in the target domain. Early UDA methods minimize the marginal distribution distance [39, 20], while adversarial learning based methods [38, 10, 31, 21] generally achieve stronger performance. However, UDA still performs worse than its supervised learning counterpart [37, 3]. Labeling small target data is a practical trade-off of fully supervised learning and unsupervised domain adaptation.

Our work is also related to Semi-Supervised Domain Adaptation (SSDA) [30, 6, 22, 23] and Few-Shot Domain Adaptation (FSDA) [24] that allow a few target labeled data. SSDA and FSDA assume that a few labeled data are passively given beforehand, while ADA actively selects the most informative samples to annotate. Due to their orthogonality to ADA, SSDA and FSDA can be readily applied to the samples actively selected and annotated by ADA to further boost the performance with the limited labeling budget.

Active Learning (AL). Active learning methods select which sample to annotate instead of providing labeled data beforehand [33]. They can be mainly categorized into query by committee, by uncertainty, and by representativeness. Query by committee methods use Gibbs training [34, 9], random sampling [4] or Dropout [8] to generate diverse classifiers and measure their disagreement. Early uncertainty methods are based on SVM [32, 36], confidence [18] or margin [1, 7]. Recently, methods based on deep networks estimate uncertainty by confidence [41], entropy minimization [13], best-vs-second-best [15] and mutual information [17]. Representativeness methods [43, 5, 25] mainly pre-cluster unlabeled samples, in which [13] achieves state-of-the-art performance through the distinctiveness of target samples on multi-layer feature maps. However, all the criteria used in previous AL works are designed for the *single-domain situation*. When exposed to the domain shift, such *non-transferable* criteria may not select the most informative samples for annotation.

Active Domain Adaptation (ADA). ADA was firstly addressed by a two-phase training [27] or confidence-based metric [29], but they only fall into shallow learning regimes. Deep ADA method [35] adopts entropy as uncertainty and domain similarity as representativeness. However, the *uncalibrated* entropy is unreliable across domains and the domain similarity is *indiscriminative* under domain shift [44], hence both are *not transferable*. [14] applies deep ADA to driving but directly uses classic query selection criteria.

With the above discussion, we conclude that all previous active learning and active domain adaptation methods use non-transferable criteria, which means that the criteria suffer from the domain gap and become inaccurate when applied to the target domain. Also, prior criteria fail to detect non-informative outliers. Instead, as discussed below, the three criteria in TQS are specially designed to mitigate the domain gap. We summarize the difference between TQS and prior methods in Table 1.

3. Approach

In Active Domain Adaptation (ADA), we have a labeled source domain $\mathcal{D}_s = \{(x_s, y_s)\}$ and an unlabeled target domain $\mathcal{D}_t = \{x_t\}$ drawn from different distributions. We follow the standard domain adaptation setting [10] where the source and target domains share identical label space. We further have a labeling budget B in the target domain, which is the maximum number of samples we can annotate by human experts. The goal of this work is to design a transferable criterion for ADA that selects and annotates the B most informative samples to maximize the target domain accuracy.

3.1. Transferable Committee (TQS$_c$)

The original query by committee methods adopt multiple classifiers and seek samples maximally disagreed by these classifiers [34]. However, they need to train the classifiers on labeled data that are from the source domain in the ADA setting. So the classifiers may pass high-density areas of the target domain due to the domain shift. Then different committees may select extremely different sets of target samples, *e.g.*, committee C_1, C_2 and committee C_3, C_4 select entirely different samples in Figure 2. So high randomness exists in sample selection, which may lead the original committee criterion to select non-informative target samples.

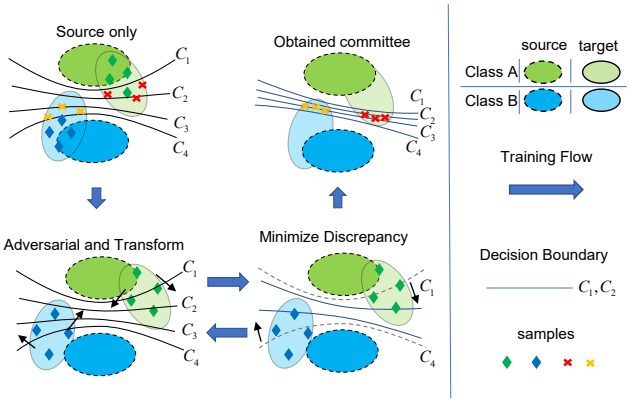


Figure 2. The top left figure shows that source classifiers may pass target high-density areas. Two different committees C_1, C_2 and C_3, C_4 disagree on red samples and orange samples respectively, so they select different sets of target samples. The two bottom figures show that we enforce the predictions of original and adversarial samples to be consistent, which gradually pushes the classifiers to low-density areas. In the top right figure, after adversarial training, different committees C_1, C_2 and C_3, C_4 select a similar set of target samples, which are more likely to be informative.

To mitigate the randomness caused by different committees in query selection, we enforce that the classifiers should not pass through target high-density areas. To this end, we slightly perturb each target sample and make the predictions of the perturbed samples stay the same. The number of perturbations is exponential to the dimension of the data, which is inefficient to sample. So we select the perturbation that mostly changes the prediction, known as adversarial samples. We denote the M classifiers by $C_1, C_2, \ldots, C_M$ and the feature extractor as F. The loss $\mathcal{L}_{\text{com}}$ for generating the proposed *transferable committee* can be defined as

$$\mathcal{L}_{\text{com}} = \sum_{m=1}^{M} \mathop{\mathbb{E}}_{x_t \sim \mathcal{D}_t} \left[\max_{|x_t' - x_t| \leq \epsilon} \left| C_m(F(x_t)) - C_m(F(x_t')) \right| \right] \tag{1}$$

where ϵ is the maximum perturbation allowed and the perturbed sample x_t' is learned to maximize the prediction discrepancy from x_t. With the above loss, we can increase the consistency between different committees, making it more probable to select the most informative samples. Enforcing consistent prediction between adversarial samples can improve uncertainty estimation, which further enhances the transferability of the uncertainty criterion. For each target sample x, we compute $Q_c(x)$, the *transferable committee criterion* (TQS$_c$) by the disagreement of M classifiers:

$$Q_c(x) = \sqrt{\frac{\sum_{m=1}^{M} \left\| C_m(F(x)) - \frac{1}{M} \sum_{m'=1}^{M} C_{m'}(F(x)) \right\|}{M}} \tag{2}$$

We use standard deviation to measure the disagreement of predictions, which lets Q_c stay stably in the range $[0, \frac{\sqrt{2}}{2}]$. We normalize it into range $[0, 1]$ by further multiplying $\sqrt{2}$.

3.2. Transferable Uncertainty (TQS$_u$)

Existing active learning criteria measure the uncertainty, *e.g.* entropy, based on a single classifier. But we find that the uncertainty estimated by a single classifier suffers from large variance and it leads to unreliable queries. As shown in Figure 3, there are different classifiers well separating two source classes. But due to the domain gap, a target sample can lie on the decision boundary for one classifier (meaning that it has high uncertainty) while being far from the decision boundary for another classifier. Thus, different source classifiers have different uncertainties on the target samples, causing a large variance if using a single classifier for the uncertainty estimation. When using the ensemble of multiple classifiers, only samples nearest to all decision boundaries have high uncertainty. These samples are extremely ambiguous for the source classifiers and annotating them is valuable to reduce the ambiguity. Thus, we compute the *transferable uncertainty* based on the multiple classifiers of the transferable committee in TQS$_c$. The cross-entropy loss for the K-way classification of M classifiers corresponds to

$$\mathcal{L}_{\text{class}} = \sum_{m=1}^{M} \mathop{\mathbb{E}}_{(x_s, y_s) \sim \mathcal{D}_s} \left[-\sum_{k=1}^{K} \mathbb{1}_{[k=y_s]} \log \left(C_m^k(F(x_s)) \right) \right] \tag{3}$$

The M classifiers are initialized with different random initializations to ensure diversity. Unlike most previous works that use entropy to estimate uncertainty, we use the *margin* of the highest and the second-highest probability in the predicted class distribution $\hat{y}$. Note that with more classes, the maximum value of entropy increases while the *margin* stays in the range $[0, 1]$. Thus, if normalizing entropy into $[0, 1]$, the margin is more sensitive than entropy when only a small portion of classes have high probabilities, which is common for uncertain samples. Based on the multiple classifiers and the margin function, we define the *transferable uncertainty* criterion $Q_u(x)$ for each target sample x as

$$Q_u(x) = \sum_{m=1}^{M} \frac{\left[1 - \left(\max_i \hat{y}_m^i - \max_{j|j\neq \text{argmax}_k \hat{y}_m^k} \hat{y}_m^j\right)\right]}{M} \tag{4}$$

where $\hat{y}_m = C_m(F(x))$ is the class distribution predicted by classifier C_m and $\hat{y}_m^i$ is probability of the ith class for x. We use minus margin since smaller margin means higher uncertainty. The transferable uncertainty only assigns high uncertainty to most ambiguous samples that are disagreed by multiple classifiers, which reduces the variance of uncertainty estimation for the target data under the domain shift. It reuses the multiple classifiers of our transferable committee and introduces no more computational and storage costs.

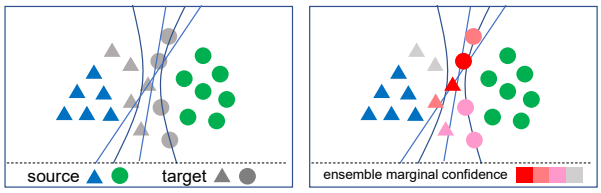
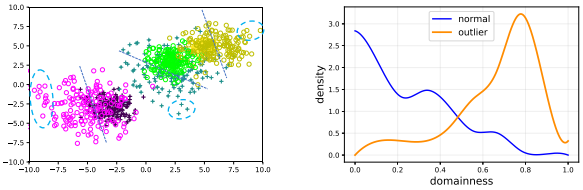


Figure 3. The left shows two classes of source samples (blue and green) and unlabeled target samples (gray). Four source classifiers are derived. The right shows the overall uncertainty of target samples estimated by four classifiers, where redder color means larger uncertainty. Only samples nearest to all decision boundaries have high uncertainty, so multiple classifiers of our transferable committee can select the most uncertain samples under domain shift.

3.3. Transferable Domainness (TQS$_d$)

Recent active learning or ADA methods [13, 35] consider samples with higher distinctiveness from the source to capture the unique part in the target domain. However, as shown in Figure 4(a), we find that outliers exist in the target domain, which are useless or even harmful for target classification. When the domain shift is on, both normal target samples disjoint from the source domain and target outliers are far from the source domain. Thus, prior distinctiveness measures cannot discriminate target-specific samples from outliers and may select outliers to waste the labeling budget.

In light of this finding, we design *transferable domainness*, which assesses how private each sample to the target



(a) Outlier illustration (b) Density of D's output

Figure 4. (a) Plot of source (○) and target (+) samples. Dotted lines indicate the best linear classifier to discriminate source from target. The dotted ellipses highlight the outliers. (b) The density of domain discriminator D's output for normal samples and outliers. The output of D is continuously distributed in the range $[0, 1]$ instead of concentrated on 0 and 1 since source and target samples are not perfectly separable and we use a one-layer classifier. The similar observation is also found in [44].

domain with consideration of outliers. We employ a domain discriminator D trained to classify whether a sample is from the source domain or target domain, with the following loss:

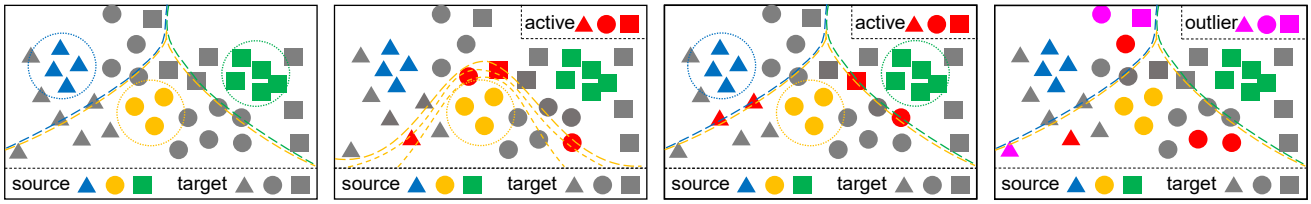
$$\begin{aligned}\mathcal{L}_{\text{dom}} = &- \mathbb{E}_{(x_s,y_s)\sim\mathcal{D}_s}\left[\log\left(1 - D(F(x_s))\right)\right] \\ &- \mathbb{E}_{x_t\sim\mathcal{D}_t}\left[\log\left(D(F(x_t))\right)\right]\end{aligned} \tag{5}$$

with label 0 for source and label 1 for target. Unlike the existing ADA method [35] which trains D and F adversarially, we do not back-propagate $\mathcal{L}_{\text{dom}}$ to F as it may spoil the quality of F [44]. D's output reflects the probability of a sample belonging to the target domain. We plot in Figure 4(b) the density of D's output for normal target samples and outliers on Office-Home dataset [40]. The outliers here are detected by outlier detection methods [2] with target labels (just for this showcase) and verified manually. We can observe that outliers have clearly higher D outputs.

With the above analysis, we can conclude that the extremely low D's output means that the sample is close to the source domains, and the extremely high D's output means that the sample is likely to be an outlier. So both samples with extremely low or high D's output should not be selected. Based on this conclusion, to derive domainness from the D's output, we need a function with a bell shape and with constrained values. So we use the Gaussian density function to define the *transferable domainness* $Q_d(x)$ for each target sample x as

$$Q_d(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(D(F(x)) - \mu)^2}{2\sigma^2}\right) \tag{6}$$

where $0 \leq \mu \leq 1$. Q_d based on Gaussian density is in the range $[0, 1]$ and has a nice property for measuring domainness. For samples with D's output lower than μ, increasing D's output increases the domainness, because higher D's output means that the sample is closer to the target domain and is more likely to represent the target unique part. For samples with D's output higher than μ, increasing D's



(a) source decision boundary (b) samples selected by committee (c) samples selected by uncertainty (d) samples selected by domainness

Figure 5. The three criteria are complementary. (a) Source samples and source decision boundary. (b) Samples selected by the transferable committee, which are disagreed by classifiers; (c) Samples selected by transferable uncertainty, which are close to all decision boundaries; (d) Samples selected by transferable domainness, where red samples are selected to cover target domain while pink outliers are removed.

output decreases the domainness, because we regard target samples far from the source domain (with $D(F(x))$ close to 1) as *outliers* and assign them with low domainness. We can control μ to decide a soft separation between inliers and outliers. We can tune σ to control the difference between the largest domainness ($D(F(x)) = \mu$) and the smallest one ($D(F(x)) = 0$ or 1), which controls the extent we aim to filter outliers. In the experiments, we find that a fixed set of μ and σ work well for all the datasets.

3.4. Transferable Criterion for Query Selection

We integrate transferable committee TQS$_c$, transferable uncertainty TQS$_u$, and transferable domainness TQS$_d$ into a unified *transferable query selection* (TQS) criterion $Q(x)$:

$$Q(x) = Q_c(x) + Q_u(x) + Q_d(x). \tag{7}$$

We do not impose trade-offs between the three criteria since the three criteria are all in range $[0, 1]$ and empirical results show that adding trade-offs does not introduce much higher accuracy. Our three criteria are *complementary* to address query selection in ADA. (1) Samples with high Q_u and Q_c can also be outliers as the prediction on outliers is random, while Q_d can help eliminate these outliers. (2) Q_d can only select target unique samples but some of them are already classified correctly by source classifiers, while Q_u and Q_c can prioritize those more informative samples. (3) Q_u and Q_c share multiple diverse classifiers in an ensemble, where Q_u provides $\mathcal{L}_{\text{class}}$ to train diverse classifiers required by Q_c while Q_c provides $\mathcal{L}_{\text{com}}$ to improve the prediction required by Q_u. As showcased in Figure 5, samples selected by each criterion have a large disjoint part.

The architecture of the proposed ADA approach based on TQS is shown in Figure 6. Instead of using $C_1, ..., C_M$, we use another classifier C purely trained with classification loss for prediction to remove the influence of Equation (1):

$$\begin{aligned}\mathcal{L} = &- \mathop{\mathbb{E}}_{(x_s,y_s)\sim\mathcal{D}_s}\left[\sum_{k=1}^{K}\mathbb{1}_{[k=y_s]}\log\left(C^k(F(x_s))\right)\right] \\ &- \mathop{\mathbb{E}}_{(x_t,y_t)\sim\mathcal{D}_t}\left[\mathbb{1}_{[I_{Q(x_t)}]}\sum_{k=1}^{K}\mathbb{1}_{[k=y_t]}\log\left(C^k(F(x_t))\right)\right]\end{aligned} \tag{8}$$

$I_{Q(x)}$ is an *indicator*, which is true when the target sample x is selected by the criterion $Q(x)$ and false otherwise. y_t is only annotated for samples with I_Q being true, which are added to the training set for learning classifier C. Similarly, we use these data to train the M classifiers $C_1, \ldots, C_M$. So the loss in Equation (3) can be iteratively extended as

$$\begin{aligned}\mathcal{L}'_{\text{class}} = &- \sum_{m=1}^{M}\mathop{\mathbb{E}}_{(x_s,y_s)\sim\mathcal{D}_s}\left[\sum_{k=1}^{K}\mathbb{1}_{[k=y_s]}\log\left(C_m^k(F(x_s))\right)\right] \\ &- \sum_{m=1}^{M}\mathop{\mathbb{E}}_{(x_t,y_t)\sim\mathcal{D}_t}\left[\mathbb{1}_{[I_{Q(x_t)}]}\sum_{k=1}^{K}\mathbb{1}_{[k=y_t]}\log\left(C_m^k(F(x_t))\right)\right]\end{aligned} \tag{9}$$

The overall optimization problem can be defined as follows:

$$\begin{aligned}&\min_{F,C,C_1,\ldots,C_M}\ \mathcal{L} + \mathcal{L}'_{\text{class}} \\ &\min_{C_1,\ldots,C_M,D}\ \mathcal{L}_{\text{com}} + \mathcal{L}_{\text{dom}}\end{aligned} \tag{10}$$

Note that we do not back-propagate $\mathcal{L}_{\text{com}}$ and $\mathcal{L}_{\text{dom}}$ to F as they may spoil the quality of features for final classification.

For the learning process, we first train F, C and $C_m|_{m=1}^{M}$ on source data. Then as previous active learning methods [13], we use the labeling budget B in a gradual manner. We repeat the following steps until using up the labeling budget. We compute the criterion Q based on the current network parameters and select b (a portion of the total budget B) target samples to annotate and train with all the selected samples by Equation (10) until convergence.

To further ensure that the selected samples are representative of the target domain, we design a *randomize selection* (RS) algorithm to select more diverse target samples. In each selection step, instead of selecting b samples with the highest Q, we first select a set of $b' > b$ candidates with the highest Q. Then we sample points from the candidates where the probability of each point is proportional to its Q value. RS can increase the diversity of the selected samples while still selects more informative samples.

Complexity. In terms of space cost, TQS only uses M more classifiers and one more domain discriminator, which are one-layer networks and cheap compared to the backbone. The computation of transferable criteria Q_c, Q_u and Q_d only needs a forward pass of the whole network, which

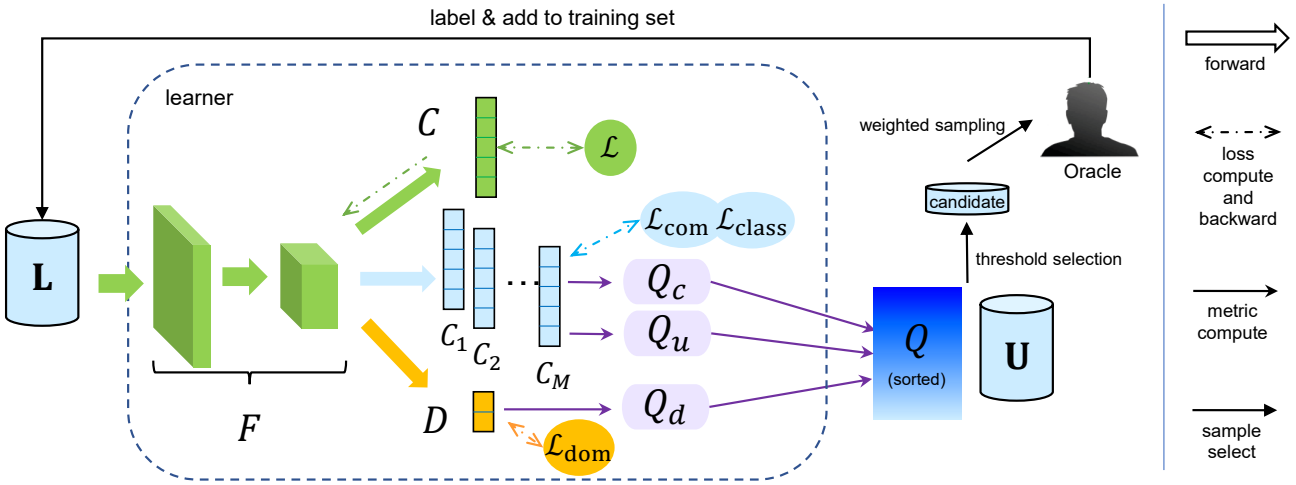


Figure 6. The architecture of TQS. The training repeats many steps of model training and query selection until the labeling budget B runs out. Each step uses labeled data to train $F, C, C_1, ..., C_M, D$ as Equation (10). Three criteria Q_c, Q_u and Q_d are computed on unlabeled target samples to form the transferable selection criterion Q as Equation (7). A set of samples with the highest Q values are selected from the pool as candidates. Then we sample b points from the candidates with the probability proportional to its Q value. The oracle annotates the ground-truth label for the b selected samples, which are removed from the unlabeled set and added to labeled data.

costs comparable time to previous active learning methods. Compare to unsupervised domain adaptation (UDA), we have multiple stages of data selection, and in each stage, we initialize with the converged model of the previous stage. So the model converges faster in later stages. Also, UDA needs to compute an extra domain alignment loss while TQS only needs to compute the classification loss. Therefore, the time complexity for TQS and UDA methods is comparable. In our experiments with the Office-31 dataset [28], TQS needs about 20 epochs to converge (25 minutes), while a typical UDA method [21] needs about 40 epochs (110 minutes).

4. Experiments

We conduct experiments on three widely-used domain adaptation datasets: Office-31 [28], Office-Home [40] and VisDA-2017 [26]. We compare the proposed TQS approach with Source-Only (ResNet), Random (RAN, which randomly selects target examples to annotate), and active learning methods including query by uncertainty (UCN) [15], pre-cluster (Cluster) [25], query by committee (QBC) [4], Active Adversarial Domain Adaptation (AADA) [35] and Active Deep Model Adaptation (ADMA) [13], which is the state-of-the-art ADA method. We also compare TQS with state-of-the-art UDA methods CDAN [21], AFN [42] and CAN [16]. Note that the data transformations and multiple classifiers are only used to learn query selection criteria in TQS. For a fair comparison, we employ a single classifier and the same image pre-processing for all methods when doing classification. We explain details on the datasets and the experimentation in Section 2.1 of the supplementary materials.

The code is available at https://github.com/thuml/Transferable-Query-Selection.

4.1. Results

The classification results of Office-31, Office-Home, and VisDA-2017 are shown in Tables 2–3. The labeling budget of all active methods is 5% of target samples. The variance in the 'Avg' column is the mean of the variances of all tasks.

Table 2. Classification accuracies (%) on **Office-31** with 5% target samples as the labeling budget for active learning methods.

Method	Office-31						
	A→D	A→W	D→A	D→W	W→A	W→D	Avg
ResNet	81.5	75.0	63.1	95.2	65.7	99.4	80.0±0.1
RAN	87.1	84.1	75.5	98.1	75.8	99.6	86.7±0.5
UCN	89.8	87.9	78.2	99	78.6	100.0	88.9±0.3
QBC	89.7	87.3	77.1	98.6	78.1	99.6	88.4±0.2
Cluster	88.1	86.0	76.2	98.3	77.4	99.6	87.6±0.1
AADA	89.2	87.3	78.2	99.5	78.7	100.0	88.8±0.3
ADMA	90.0	88.3	79.2	100.0	79.1	100.0	89.4±0.3
TQS	**92.8**	**92.2**	**80.6**	**100.0**	**80.4**	**100.0**	**91.1±0.3**

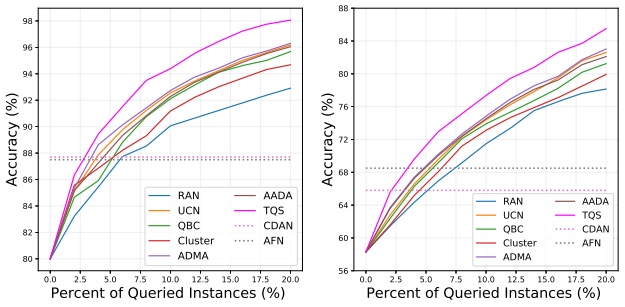
Randomly selecting samples can still achieve higher performance than ResNet, which implies that ADA is a promising solution for domain adaptation. All different kinds of active learning methods outperform random selection (RAN). Though not tailored to ADA, previous active learning criteria still bias to highly informative samples. TQS consistently outperforms these methods, because TQS consists of transferable committee, uncertainty, and domainness specially designed for ADA. Note that TQS also outperforms ADMA, which emphasizes the importance of reducing the variance of uncertainty and detecting outliers.

Table 3. Classification accuracies (%) on **Office-Home** and **VisDA-2017** with 5% target samples as the labeling budget for active methods.

Method	Office-Home													VisDA-2017
	A→C	A→P	A→R	C→A	C→P	C→R	P→A	P→C	P→R	R→A	R→C	R→P	Avg	
ResNet	42.1	66.3	73.3	50.7	59.0	62.6	51.9	37.9	71.2	65.2	42.6	76.6	58.3±0.1	44.7±0.1
RAN	52.5	74.3	77.4	56.3	69.7	68.9	57.7	50.9	75.8	70.0	54.6	81.3	65.8±0.5	78.1±0.6
UCN	56.3	78.6	79.3	58.1	74.0	70.9	59.5	52.6	77.2	71.2	56.4	84.5	68.2±0.3	81.3±0.4
QBC	56.9	78.0	78.4	58.5	73.3	69.6	60.2	53.3	76.1	70.3	57.1	83.1	67.9±0.2	80.5±0.3
Cluster	56.0	76.8	78.1	58.4	72.6	69.2	58.4	51.2	75.4	70.1	56.4	82.4	67.1±0.2	79.8±0.2
AADA	56.6	78.1	79.0	58.5	73.7	71.0	60.1	53.1	77.0	70.6	57.0	84.5	68.3±0.2	80.8±0.4
ADMA	57.2	79.0	79.4	58.2	74.0	71.1	60.2	52.2	77.6	71.0	57.5	85.4	68.6±0.3	81.4±0.4
TQS	**58.6**	**81.1**	**81.5**	**61.1**	**76.1**	**73.3**	**61.2**	**54.7**	**79.7**	**73.4**	**58.9**	**86.1**	**70.5±0.3**	**83.1±0.4**

4.2. Analyses

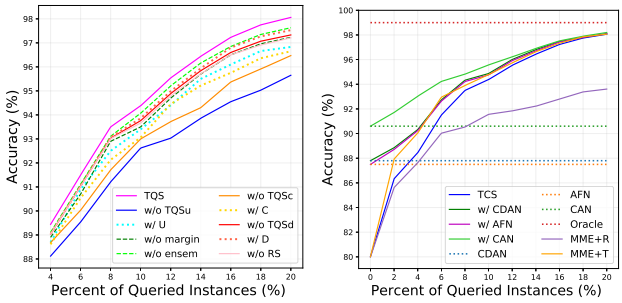
Varying Labeling Budget. In this experiment, we want to show two claims by varying the labeling budget: (1) TQS consistently outperforms other active learning or active domain adaptation methods with varying labeling budgets; (2) TQS outperforms UDA methods only with a small labeling budget. For each budget, we conduct experiments on all 6 tasks of Office-31 and all 12 tasks of Office-Home and compute the average accuracy on each dataset. The results are shown in Figures 7(a)–7(b). We can observe that TQS consistently outperforms previous active learning and active domain adaptation methods with various labeling budgets. In particular, with only about 3% labeling budget, TQS performs comparably with UDA methods while outperforms UDA with a larger budget, indicating that TQS only needs little labeling burden but boosts the domain adaptation performance significantly.



(a) Office (b) Office-Home

Figure 7. Accuracies with varying percent of labeling budget averaged on *all* tasks of Office and Office-Home datasets respectively.

Ablation Study. We go deeper into the efficacy of each of our contributions. For the three criteria: TQS_c, TQS_u and TQS_d, we design two sets of variants: removing each criterion from TQS (w/o TQS_c, w/o TQS_u, and w/o TQS_d) and replacing the transferable criterion with its non-transferable version. The non-transferable versions are corresponding previous state-of-the-art criteria: For committee, we remove the min-max optimization (w/ c); For uncertainty, we use the entropy on a single classifier (w/ u); For domainness, we replace TQS_d with the domain similarity in [35] (w/ d). For the transferable uncertainty, we further analyze the efficacy of the ensemble of multiple classifiers and the margin function by only using one classifier to estimate Q_u (w/o ensemble) and replacing the margin function with the well-known entropy function (w/o margin) in Equation (4) respectively. For these two variants, the computation of the other two criteria: TQS_c and TQS_d, are not influenced. We verify the efficacy of the random selection algorithm by comparing it with removing RS (w/o RS). The performance is evaluated as the average accuracy of all 6 tasks on the Office-31 dataset.



(a) Ablation Study of TQS (b) Combine TQS with UDA/SSDA

Figure 8. (a) Accuracy of TQS and its variants averaged on 6 tasks of Office-31. (b) Combining TQS with UDA/SSDA methods.

In Figure 8(a), we can observe that TQS outperforms w/o TQS_c, w/o TQS_u and w/o TQS_d, indicating that TQS_c, TQS_u and TQS_d are all important to selecting informative samples. TQS outperforms w/ c, w/ u and w/ d, indicating that our transferable committee, transferable uncertainty and transferable domainness are more suitable to the cross-domain setting than the original committee, uncertainty and representativeness designed for traditional active learning. TQS outperforms w/o ensemble, indicating the efficacy of ensemble to reduce variance in the uncertainty estimation. TQS outperforms w/o margin, which proves that the margin function is a more stable and discriminative uncertainty measurement. TQS outperforms w/o RS, which verifies that

the RS mechanism can select more diverse target samples to increase the information gain of the selected samples.

Combination with UDA/SSDA Methods. We have analyzed that UDA methods are orthogonal to our work. We will further show that the selected samples by TQS are more beneficial for UDA methods, meaning that if given a fixed labeling budget, annotating the samples selected by TQS can improve the performance of UDA methods more than random selection. We compare the UDA methods, the combination of UDA methods with samples selected by TQS (CDAN/AFN/CAN+TQS), and the supervised oracles. We combine TQS with UDA methods by applying distribution matching losses to labeled source data and unlabeled target data, and applying the classification loss to labeled source data and labeled target data. We show the performance averaged on all 6 tasks of the Office-31 dataset.

As shown in Figure 8(b), there is a large gap between UDA and oracle. With the labeling budget increased from 0% to 15%, TQS with UDA methods achieves much better accuracy than both TQS and UDA methods and is close to the oracle, indicating that TQS is complementary to UDA methods. Furthermore, with enough budget of more than 15%, TQS can achieve similar performance with and without UDA methods, showing that even a modest labeling budget can *eliminate* the effect of unsupervised domain adaptation. We show the comparison of UDA methods with randomly selected samples in the supplementary materials.

As we stated in the related work, SSDA methods are orthogonal to our work and TQS can be naturally combined with SSDA methods by using TQS selected samples as labeled samples for SSDA. We further show that using the selected samples by TQS is much more beneficial for SSDA methods than randomly selected samples, which means that combining TQS with SSDA further boosts the performance of SSDA. Similar to UDA, in Figure 8(b), we compare SSDA+TQS with SSDA+RAN in terms of average accuracy on all 6 tasks of Office-31, where we use the state-of-the-art SSDA work MME [30]. We can observe that MME+TQS outperforms MME+RAN, showing that TQS is orthogonal to SSDA methods and can improve them to higher accuracy.

4.3. Analysis of the Trade-offs between Criteria

We argue that the trade-offs are not needed between the three criteria in Equation (7). We empirically demonstrate that directly adding the three criteria without trade-offs achieves the optimal performance. We employ two parameters λ_1 and λ_2 to compute the criterion as

$$Q' = Q_c + \lambda_1 Q_u + \lambda_2 Q_d. \tag{11}$$

We show the accuracy on the A→D task of Office-31 with varying λ_1 and λ_2. We select the task A→D because the target domain D is small and the performance is more unstable than other tasks. If we demonstrate the stability of the performance with respect to λ_1 and λ_2 on A→D, we can easily generalize the conclusion to other tasks.

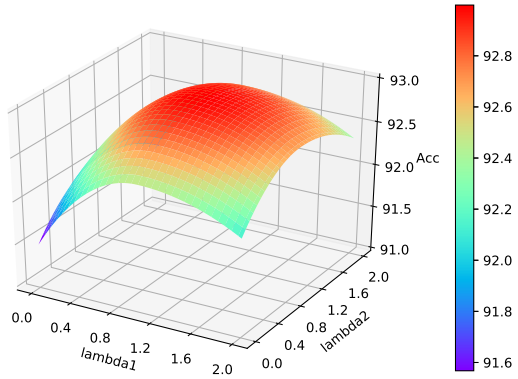


Figure 9. The smoothed classification accuracy on Office-31 A→D task (5% budget) with varying trade-offs.

From Figure 9, we can observe that the performance changes very slightly within a small range of λ_1 and λ_2 around 1. Furthermore, the performance of $\lambda_1 = 1$ and $\lambda_2 = 1$ is also almost the best. These observations demonstrate TQS is not sensitive to trade-offs in a reasonable range. Adding the three criteria without any trade-off does not drop the performance much and can be widely used across different tasks. One may ask whether the stable performance is due to that Q_u and Q_d actually do not influence the performance. We observe that when the parameter λ_1 or λ_2 is close to 0, the performance drops much, which demonstrates that both Q_u and Q_d are required for the method.

5. Conclusion

This paper presents a new Transferable Query Selection (TQS) approach to active domain adaptation, consisting of transferable uncertainty, transferable domainness and transferable committee that are complementary to each other for selecting informative target samples under domain shift. The random selection algorithm further increases the diversity of selected samples. A large volume of experimental results on three benchmarks show that TQS is an effective approach for active domain adaptation. Deep analyses indicate that TQS can be used in various labeling budgets and collaborate with unsupervised and semi-supervised domain adaptation methods to further boost performance.

Acknowledgments

This work was supported by the AI project granted by China's Ministry of Industry and Information Technology, NSFC grants (62022050, 62021002, 61772299, 71690231), and Beijing Nova Program (Z201100006820041).

References

[1] Maria-Florina Balcan, Andrei Broder, and Tong Zhang. Margin based active learning. In *International Conference on Computational Learning Theory*, pages 35–50. Springer, 2007. 2

[2] Raghavendra Chalapathy and Sanjay Chawla. Deep learning for anomaly detection: A survey. *arXiv preprint arXiv:1901.03407*, 2019. 4

[3] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3339–3348, 2018. 1, 2

[4] Ido Dagan and Sean P Engelson. Committee-based sampling for training probabilistic classifiers. In *Machine Learning*, pages 150–157. Elsevier, 1995. 1, 2, 6

[5] Sanjoy Dasgupta and Daniel Hsu. Hierarchical sampling for active learning. In *Proceedings of International Conference on Machine Learning*, pages 208–215. ACM, 2008. 2

[6] Jeff Donahue, Judy Hoffman, Erik Rodner, Kate Saenko, and Trevor Darrell. Semi-supervised domain adaptation with instance constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 668–675, 2013. 2

[7] Pinar Donmez, Jaime G Carbonell, and Paul N Bennett. Dual strategy active learning. In *European Conference on Machine Learning*, pages 116–127. Springer, 2007. 2

[8] Melanie Ducoffe and Frederic Precioso. Qbdc: query by dropout committee for training deep supervised architecture. *arXiv preprint arXiv:1511.06412*, 2015. 2

[9] Yoav Freund, H Sebastian Seung, Eli Shamir, and Naftali Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2-3):133–168, 1997. 2

[10] Yaroslav Ganin and Victor S. Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, pages 1180–1189, 2015. 2, 3

[11] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(1):2096–2030, 2016. 1

[12] Sheng-Jun Huang, Rong Jin, and Zhi-Hua Zhou. Active learning by querying informative and representative examples. In *Advances in Neural Information Processing Systems*, pages 892–900, 2010. 1

[13] Sheng-Jun Huang, Jia-Wei Zhao, and Zhao-Yang Liu. Cost-effective training of deep cnns with active model adaptation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1580–1588. ACM, 2018. 1, 2, 4, 5, 6

[14] Yue Huang, Zhenwei Liu, Minghui Jiang, Xian Yu, and Xinghao Ding. Cost-effective vehicle type recognition in surveillance images with deep active learning and web data. *IEEE Transactions on Intelligent Transportation Systems*, 2019. 2

[15] Ajay J Joshi, Fatih Porikli, and Nikolaos P Papanikolopoulos. Scalable active learning for multiclass image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2259–2273, 2012. 1, 2, 6

[16] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4893–4902, 2019. 6

[17] Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *arXiv preprint arXiv:1906.08158*, 2019. 2

[18] David D Lewis and Jason Catlett. Heterogeneous uncertainty sampling for supervised learning. In *Machine Learning*, pages 148–156. Elsevier, 1994. 2

[19] Mingsheng Long, Yue Cao, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Transferable representation learning with deep adaptation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. 1

[20] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, pages 97–105, 2015. 2

[21] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pages 1640–1650, 2018. 2, 6

[22] David Lopez-Paz, Jose M Hernández-lobato, and Bernhard Schölkopf. Semi-supervised domain adaptation with non-parametric copulas. In *Advances in neural information processing systems*, pages 665–673, 2012. 2

[23] Zelun Luo, Yuliang Zou, Judy Hoffman, and Li F Fei-Fei. Label efficient learning of transferable representations acrosss domains and tasks. In *Advances in Neural Information Processing Systems*, pages 165–177, 2017. 2

[24] Saeid Motiian, Quinn Jones, Seyed Iranmanesh, and Gianfranco Doretto. Few-shot adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pages 6670–6680, 2017. 2

[25] Hieu T Nguyen and Arnold Smeulders. Active learning using pre-clustering. In *Proceedings of International Conference on Machine Learning*, page 79. ACM, 2004. 2, 6

[26] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017. 6

[27] Piyush Rai, Avishek Saha, Hal Daumé III, and Suresh Venkatasubramanian. Domain adaptation meets active learning. In *Proceedings of the NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing*, pages 27–32. Association for Computational Linguistics, 2010. 1, 2

[28] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European Conference on Computer Vision*, pages 213–226, 2010. 6

[29] Avishek Saha, Piyush Rai, Hal Daumé, Suresh Venkatasubramanian, and Scott L DuVall. Active supervised domain adaptation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 97–112. Springer, 2011. 2

[30] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8050–8058, 2019. 2, 8

[31] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2018. 2

[32] Greg Schohn and David Cohn. Less is more: Active learning with support vector machines. In *International Conference on Machine Learning*, volume 2, page 6. Citeseer, 2000. 2

[33] Burr Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009. 1, 2

[34] H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings on Computational Learning Theory*, pages 287–294, 1992. 1, 2, 3

[35] Jong-Chyi Su, Yi-Hsuan Tsai, Kihyuk Sohn, Buyu Liu, Subhransu Maji, and Manmohan Chandraker. Active adversarial domain adaptation. *CoRR*, abs/1904.07848, 2019. 1, 2, 4, 6, 7

[36] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 2(Nov):45–66, 2001. 2

[37] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7472–7481, 2018. 1, 2

[38] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2962–2971, 2017. 2

[39] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014. 2

[40] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5385–5394, 2017. 4, 6

[41] Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin. Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(12):2591–2600, 2016. 2

[42] Ruijia Xu, Guanbin Li, Jihan Yang, and Liang Lin. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1426–1435, 2019. 1, 6

[43] Zhao Xu, Kai Yu, Volker Tresp, Xiaowei Xu, and Jizhi Wang. Representative sampling for text classification using support vector machines. In *European Conference on Information Retrieval*, pages 393–407. Springer, 2003. 2

[44] Jing Zhang, Zewei Ding, Wanqing Li, and Philip Ogunbona. Importance weighted adversarial nets for partial domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2018. 2, 4