Category-Aware Active Domain Adaptation

Wenxiao Xiao [1 *] **Jiuxiang Gu** [2] **Hongfu Liu** [1]

Abstract

Active domain adaptation has shown promising results in enhancing unsupervised domain adaptation (DA), by actively selecting and annotating a small amount of unlabeled samples from the target domain. Despite its effectiveness in boosting overall performance, the gain usually concentrates on the categories that are readily improvable, while challenging categories that demand the utmost attention are often overlooked by existing models. To alleviate this discrepancy, we propose a novel category-aware active DA method that aims to boost the adaptation for the individual category without adversely affecting others. Specifically, our approach identifies the unlabeled data that are most important for the recognition of the targeted category. Our method assesses the impact of each unlabeled sample on the recognition loss of the target data via the influence function, which allows us to directly evaluate the sample importance, without relying on indirect measurements used by existing methods. Comprehensive experiments and in-depth explorations demonstrate the efficacy of our method on category-aware active DA over three datasets.

1. Introduction

Recent research efforts have been dedicated to active domain adaptation (Su et al., 2020; Wang et al., 2024; Li et al., 2023; Wang et al., 2023b; Rangwani et al., 2021; Hwang et al., 2022), in order to bridge the significant performance gap (Chen et al., 2018; Tsai et al., 2018) between unsupervised domain adaptation (DA) and their supervised counterpart. The active DA framework actively queries ground truth labels for a small set of important unlabeled target data, providing extra supervision for the DA models to facilitate domain transfer. Most existing active DA works (Huang et al., 2023; Xie et al., 2022b; 2023) evaluate the importance of each unlabeled sample by various indirect measurements, such as diversity (Su et al., 2020), uncertainty (Prabhu et al., 2021), or domainness (Fu et al., 2021). The target samples with the highest importance estimation are annotated and subsequently incorporated into the labeled source data for re-training the adaptation model.

Nevertheless, the overall advancements brought by these methods also come with a hidden cost for certain data categories. Specifically, to achieve the maximum overall enhancement with a limited number of new annotations, existing active domain adaptation models (Prabhu et al., 2021; Liu et al., 2021; Rangwani et al., 2021; Hwang et al., 2022; Xie et al., 2023) predominantly focus on easier categories. That is, the categories that are relatively simple to improve are frequently queried by the model, even though some of them already have satisfactory performance. On the contrary, the most challenging categories seldom receive any new annotations, obtaining little improvement or even suffering impairment after active learning. The negligence of critical categories could raise significant concerns or vulnerability in real-life applications.

Contributions. To avoid these risks, we consider category-aware active domain adaptation, which seeks a strategy that facilitates adaptation for individual categories, particularly those deemed challenging. Our major contributions are summarized as follows:

- We consider a new research question, category-aware active domain adaptation, which targets the performance enhancement specific for each individual category, avoiding the potential risks caused by the neglect of critical categories. To the best of our knowledge, this is the first attempt to address the performance discrepancy among categories in active domain adaptation.
- We employ the influence function (Cook & Weisberg, 1980) and extend it into the active learning context to estimate the usefulness of target data. The influence function enables a direct evaluation of each sample's impact on the category-specific classifier, unlike other indirect measurements aiming for overall performance.

[*] Work done during an internship at Adobe Research. [1]Computer Science Department, Brandeis University, Waltham, Massachusetts, USA [2]Adobe Research, USA. Correspondence to: Wenxiao Xiao <wenxiaoxiao@brandeis.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

- We demonstrate the effectiveness of our influence-based method on targeted categories with comprehensive experiments, as well as provide answers to crucial underlying questions in category-aware active domain adaptation with various in-depth explorations.

2. Related Works

Active learning for domain adaptation. Observing the significant performance gap between unsupervised DA methods and their supervised counterparts (Chen et al., 2018; Tsai et al., 2018), Su et al. (2020) propose active domain adaptation, which utilizes importance sampling to select and annotate target samples, to facilitate knowledge transfer. Their original work measures the importance of each target sample with criteria including diversity and uncertainty, trying to find the samples that are less similar to the source domain and least confident for the predicting model. Later, CLUE (Prabhu et al., 2021) integrates both diversity and uncertainty into an uncertainty-weighted clustering framework. On this route, TQS (Fu et al., 2021) employs a "transferable committee" consisting of multiple classifiers, which calculates another criterion "transferable domainess" in addition to uncertainty and diversity, to mitigate the domain gap. S^{3}vaada (Rangwani et al., 2021) introduces a set-based criterion that extends the concepts of uncertainty and diversity to subsets of unlabeled data, aiming to identify the most informative subset. DUC (Xie et al., 2023) further explores the uncertainty miscalibration in DA problems, and employs a Dirichlet-based evidential model to select uncertain and informative samples. Observing the varied domain discrepancy within existing datasets, DiaNA (Huang et al., 2023) partitions the target data partition based on the domainness and uncertainty, then handles domain gaps with a "divide-and-adapt" approach. Observing the label distribution issue in active DA, LAMDA (Hwang et al., 2022) selects target data which best preserve the target data distribution. Inspired by the energy-based models (EBMs) (LeCun et al., 2006), EADA (Xie et al., 2022a) selects a highly informative subset of unlabeled target data under domain shift via exploiting "free energy biases" between the two domains. Recently, several works (Wang et al., 2023a; Li et al., 2022; Kothandaraman et al., 2023) extend active DA to a source-data-free setting. Despite the promising overall performance, these active DA methods often achieve such improvements at the expense of certain overlooked, hard-to-improve categories as we discussed in Section 1. Therefore, we propose the category-aware active domain adaptation to boost the recognition for each targeted category. Unlike the above methods, our method does not rely on indirect measurements such as uncertainty or diversity. Instead, our model directly estimates each target sample's impact on the recognition task with the influence function.

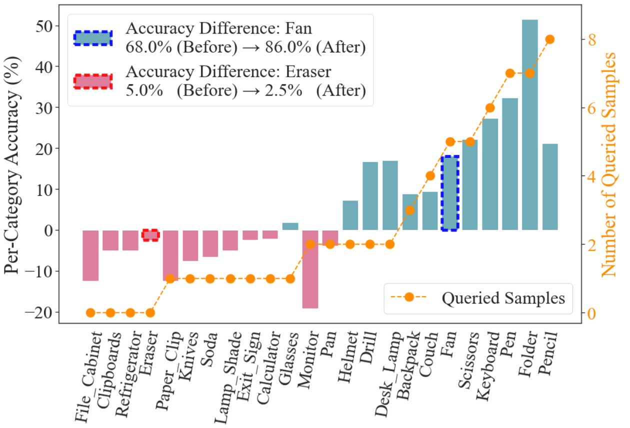


Figure 1: Each bar represents the per-category accuracy differences before and after active learning of CLUE (Prabhu et al., 2021) method for task Ar→Cl on *Office-Home* dataset (Venkateswara et al., 2017). The categories with red bars suffer performance loss after active learning, while the categories with blue bars benefit from active learning. The orange dots are the number of samples queried by CLUE from each category. We highlight two individual categories Fan and Eraser, with blue and red edges, respectively.

Influence function. Influence function (Cook & Weisberg, 1980) calculates how a model's output changes in response to an infinitesimal perturbation of a training data point. Recently, the research community has dedicated significant efforts (Giordano et al., 2019; Koh et al., 2019; Ting & Brochu, 2018) on influence function. Although the accuracy of influence estimates is limited on certain non-convex, deeper networks (Basu et al., 2020), many researchers successfully integrate the influence function in various applications. For instance, Koh et al. (2019) generalizes the influence function to a group of data to understand deep convolutional networks. Later works (Han et al., 2020; Chen et al., 2023) extend the interpretative ability of the influence function to natural language processing and graph convolution networks. Besides model interpretation, the influence function is also used for tasks including poisoning attack (Fang et al., 2020), causal inference (Alaa & Van Der Schaar, 2019), and model fairness (Li & Liu, 2022). It is worth noting that ISAL (Liu et al., 2021) applies the influence function to active learning by estimating the influence of each unlabeled sample with a pseudo label. To mitigate the domain gap in DA, our method estimated the influence of the target data based on the influence of labeled samples, instead of relying on pseudo labels assigned by a domain-specific model.

3. Motivation

The existing active domain adaptation (DA) methods (Su et al., 2020; Fu et al., 2021; Prabhu et al., 2021; Xie et al., 2022b; 2023) have demonstrated potential in elevating overall accuracy. However, this enhancement comes at a con-

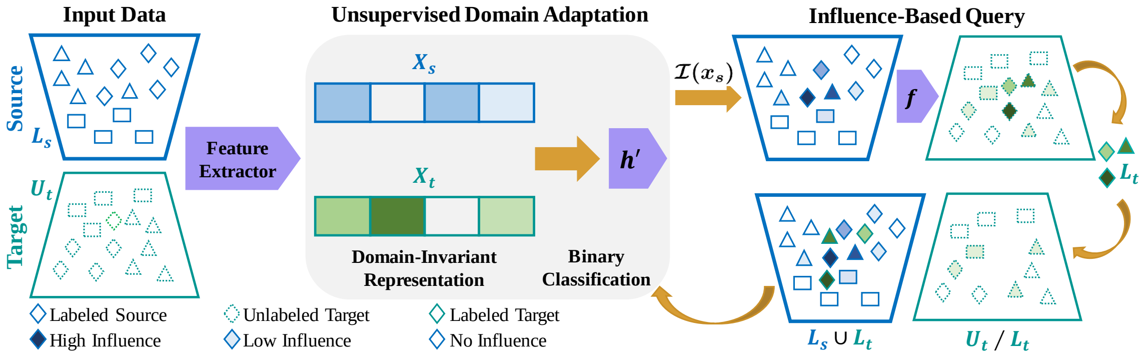


Figure 2: Our proposed category-aware active domain adaptation method operates within a specific framework, focusing on individual categories (diamond shapes). In each query iteration, our method assesses the impact of each sample's latent representation on the loss function of the classifier for the binarized data (diamond / not diamond). To achieve this, we employ a binary logistic regression model h' to classify the latent representations extracted by the base DA model. The impact of each labeled source sample is estimated with the influence function. The impact of unlabeled target samples is then predicted using a separate regression model f, which is trained using the representations and influence values derived from the labeled samples. With the influence estimations, the model selects unlabeled samples with the highest impact (depicted with darker fill shades) and incorporates them into the labeled source domain for the next training iteration.

cealed cost: a decline in performance for certain individual categories. To attain a higher overall result, current models tend to prioritize certain categories that are relatively easy to improve. Meanwhile, the truly challenging categories, which require the utmost attention, are left in the shadows. Imagine a situation where an active learning recognition algorithm is integrated into an autonomous driving system, leading to a substantial increase in average accuracy across all categories within a driving environment. However, to achieve this overall improvement, the algorithm primarily concentrates on querying easy categories, like inanimate objects, while neglecting certain challenging yet critical categories, such as humans and moving vehicles. This discrepancy in active learning may impair the performance on the unattended categories, potentially compromising model robustness and posing significant risks to vital objects.

This notable disparity among data categories is also present in current active DA models. As illustrated in Figure 1, querying extra target data results in different performance changes in multiple categories for CLUE (Prabhu et al., 2021). In categories such as *Fan*, the CLUE model queried five target samples and boosted their accuracy, although they already performed relatively well without active learning. However, the ill-performed *Eraser* category was entirely overlooked, resulting in a further decline in its performance, almost reaching 0%. The negligence of such categories hinders the reliability of active DA models' real-life applications and potentially raises significant concerns for current active DA methods. This potential limitation demands the active DA models redirect their focus from the low-hanging fruit to the harder-to-reach branches in the active domain adaptation. Consequently, we are driven to find a category-aware query strategy in active DA. This approach aims to identify target samples that enhance the recognition of specific categories, particularly those challenging ones that have been disregarded by current active DA methods.

4. Method

4.1. Preliminaries and Problem Definition

Influence function. Consider a convex model with parameters θ trained on a labeled training dataset $(X, Y) = (x_1, x_2, ..., x_N)$, the empirical risk minimization (ERM) over certain loss function $\ell(\cdot, \cdot)$ is represented as $\hat{\theta} = \arg\min_{\theta \in \Theta} \frac{1}{N} \sum_{x,y} \ell(x, y) + \frac{\lambda}{2} \|\theta\|_2^2$. If one training sample (x_i, y_i) is down-weighted by infinitesimal ϵ, the new ERM is given by $\hat{\theta}_{x_i;-\epsilon} = \arg\min_{\theta \in \Theta} \frac{1}{N} \sum_{x,y} \ell(x, y) - \epsilon \ell(x_i, y_i) + \frac{\lambda}{2} \|\theta\|_2^2$. The influence function estimates the actual change $\hat{\theta}_{x_i;-\epsilon} - \hat{\theta}$ as:

$$\mathcal{I}_{x_i;-\epsilon} = \lim_{-\epsilon \to 0} \hat{\theta}_{x_i;-\epsilon} - \hat{\theta} = -\mathbf{H}_{\hat{\theta}}^{-1} \nabla_{\hat{\theta}} \ell(x_i, y_i), \quad (1)$$

where $\mathbf{H}_{\hat{\theta}} = \frac{1}{N} \sum_{i=1}^{N} \nabla_{\hat{\theta}}^2 \ell(x_i, y_i) + \lambda \mathbf{I}$ is the positive definite Hessian matrix for $\hat{\theta}$.

For the validation loss $\mathcal{L}_v = \ell(V; \hat{\theta})$, V being the validation set, the change after removing a training sample can further be estimated using $\nabla_{\hat{\theta}} \mathcal{L}_v \mathbf{H}_{\hat{\theta}}^{-1} \ell(x_i, y_i)$, as demonstrated in the work by Koh & Liang (2017). Therefore, we can estimate the impact of a specific training data point (x_i, y_i) on the validation loss, by calculating the difference of $\mathcal{L}_v$ before and after removing (x_i, y_i) from training as:

$$\mathcal{I}_{\mathcal{L}_v}(x_i) = -\nabla_{\hat{\theta}} \mathcal{L}_v \mathbf{H}_{\hat{\theta}}^{-1} \nabla_{\hat{\theta}} \ell(x_i, y_i), \quad (2)$$

Active domain adaptation. Active domain adaptation incorporates active learning with the unsupervised domain adaptation, where a backbone DA model $\mathcal{B}$ trained on a labeled source domain L_s needs to be adapted to perform well on a different unlabeled target domain U_t. Both domains contain the same set of categories C. The active learning process consists of K iterations, where the pre-trained DA model $\mathcal{B}$ actively queries and annotates a small number of unlabeled target samples L_t in each iteration. These annotated samples are added into the source domain, and the model $\mathcal{B}$ is then retrained with the new source $L_s = L_s \cup L_t$ and target domains $U_t = U_t \setminus L_t$. Ideally, the model can achieve better overall performance on the target domain.

Category-aware active domain adaptation. As discussed in Section 3, category-aware active DA focuses on finding a query strategy that can enhance the performance for a specific category $c \in C$. Therefore, the model trains a category-specific classifier h^c to recognize the samples in category c, with the domain-invariant representations extracted by the base DA model $\mathcal{B}$. Based on h^c, the active learning algorithm selects b unlabeled samples that are most beneficial for the targeted category c, and then queries the ground truth for the selected samples. Finally, the category-specific classifier is re-trained after acquiring the annotated target samples L_t^c in each iteration. For simplicity, unless stated otherwise, we exclude c from the notations throughout the rest of our discussion, as category-aware learning is equally applicable to every specific category.

4.2. Framework Overview

Category-aware active DA aims to identify the advantageous samples for the targeted category c. This necessitates the active learning module to assess the impact of annotating each unlabeled sample for individual categories. Contrarily, the indirect criteria, such as uncertainty or domainness, utilized by current active DA methods (Fu et al., 2021; Prabhu et al., 2021; Xie et al., 2023) can only estimate the overall effect of each sample on the entire dataset, neglecting the discrepancy among categories. To address this challenge, our method directly estimates the impact of each data point on predicting the specific target categories using the influence function (Koh et al., 2019). In predicting the utility of the unlabeled target samples, we extend the influence function into the active learning context, enabling it to operate without accessing the ground truth labels. This straightforward influence-based query strategy acquires annotations for unlabeled samples to enhance specific categories, thereby facilitating category-aware DA through active learning.

Figure 2 illustrates the active learning process in one training iteration of our proposed method. After an unsupervised base DA model projects both source and target data into a shared representation space, our active learning algorithm selects the target data for annotation with the help of the influence function. Technically, the category-specific impact estimation module calculates the impact of each source sample on a binary surrogate classifier h', which is trained to distinguish samples of the targeted category c from the rest of the data. The influence values of the source data are then used to train an influence predictor f, enabling the target influence approximation module to assess the usefulness of the target data without labels. In the final selection step, the query set is randomly sampled from a candidate set $\mathcal{C}$ consisting of the target samples with higher influence estimations. Lastly, the queried samples are annotated with ground truth labels and incorporated into the labeled source data to re-train the classifier in the subsequent iteration.

Algorithm 1 Category-Aware Active DA for Category c

```
Input: Binarized labeled source data representations $\{(x_s, y_s)\} \in L_s$, unlabeled target data representations $\{(x_t)\} \in U_t$, per-round budget $b$ and total rounds $R$.
Output: Classification model $h^c$ for category $c$.
Train a surrogate classification model $h'$ with $L_s$.
for $i = 1$ to $R$ do
    for $(x_s, y) \in L_s$ do
        Calculate the source influences $\mathcal{I}(x_s)$ by Eq. (4).
    end for
    Train the influence estimator $f$ with Eq. (5).
    for $x_t \in U_t$ do
        Calculate the target influences $\mathcal{I}(x_t) = f(x_t)$.
    end for
    Choose the $5 \times b$ samples with highest influence as $\mathcal{C}$.
    Sample and Annotate $b$ data points from $\mathcal{C}$ as $L_t$.
    Retrain the surrogate model $h'$ by Eq. (6).
    Update both domains $L_s = L_s \cup L_t$ and $U_t = U_t \setminus L_t$.
end for
Set classification model $h^c = h'$.
Return $h^c$.
```

4.3. Sample Selection

In the k-th iteration of the active learning, we have an unlabeled target dataset U_t^k and a labeled source dataset L_s^k. Notice that L_s^k contains the original source dataset L_s^0 and the annotated target data $\{L_t^i\}_{i=1}^{k-1}$ from all previous k-1 iterations. To select the most informative samples for each individual category c, we first binarize all available labels in L_s^k into the targeted category c ("One") and the non-targeted categories *not* c ("Non-One"). This allows us to focus on differentiating category c from other the other categories in U_t^k. As active DA queries the target samples in iterations, the following section will focus on the query strategy in one iteration and omit iteration notation k.

Initial iteration. In the first iteration, we do not have any annotated target samples for influence estimation. To avoid the cold start, we selected a diverse query set based on the pseudo labels $h'(x_t)$ of all target samples. Specifically, we randomly select $\frac{b}{2}$ target samples with $h'(x_t) = 0$ and $\frac{b}{2}$ target samples with $h'(x_t) = 1$ in this initial iteration.

Category-specific impact estimation. In the following iterations, we use a logistic regression model h' to distinguish the targeted category c from the rest categories. Here h' serves a convex surrogate for the non-convex classifiers utilized by the current active DA models, satisfying the convex requirement for the influence function, Mathematically, it can be expressed as:

$$\mathcal{L}_b = \frac{1}{n_s} \Sigma_{(x_s,y)\in L_s} \ell(h'(x_s), y), \tag{3}$$

where L_s is the source samples with binary labels and ℓ is the binary cross entropy loss.

The impact of each source sample on the target data can be then estimated with the influence function. As we do not have labels for all target samples, we use all queried target samples from previous iterations, *i.e.*, $V = L_t$, to calculate the validation loss $\mathcal{L}_v$ in Eq. (2). The influence of a source sample $(x_s, y) \in L_s$ can be calculated as:

$$\mathcal{I}_{\mathcal{L}_v}(x_s) = -\nabla_{\hat{\theta}} \mathcal{L}_v \mathbf{H}_{\hat{\theta}}^{-1} \nabla_{\hat{\theta}} \ell(h'(x_s), y), \tag{4}$$

where $\hat{\theta}$ is the optimized parameter of h'. We will omit the $\mathcal{L}_v$ and notate the influence estimation for a data point x as $\mathcal{I}(x)$ in the following discussion, as the validation loss remains unchanged in the same iteration.

Target influence approximation. The influence calculation with Eq. (4) requires a ground truth label, which is not available for an unlabeled target data $x_t \in U_t$. To address this limitation, we employ a regression model f to predict the influence values of each target sample without accessing their labels. This influence predictor f is trained on the labeled source data and their corresponding influence estimations with the following Mean Squared Error loss:

$$\mathcal{L}_f = \frac{1}{n_s} \Sigma_{x_s \in L_s} \ell_{\text{mse}}(f(x_s), \mathcal{I}(x_s)), \tag{5}$$

where $\mathcal{I}(x_s)$ is the influence estimation of source data x_s.

This influence predictor f approximately calculates the influence value directly from the training samples. Therefore, for each unlabeled target $x_t \in U_t$, we can estimate its influence value as $\mathcal{I}(x_t) = f(x_t)$.

Final selection. The queried samples are selected from the unlabeled target dataset based on the influence estimation. We employ a random sampling strategy to increase the diversity of the queried samples. The queried target samples are then annotated with the binary labels for re-training the model. We summarized our category-aware active DA in Algorithm 1. As we discussed earlier, we omit iteration number k in all notations.

The time complexity will be $\mathcal{O}(nd)$ for the surrogate model h', where n is the number of samples, and d is the model dimension, which is relatively small for logistic regression. Similarly, the influence estimator f will also take $\mathcal{O}(nd)$. For the influence function, the explicit calculation of Hessian matrix will take $\mathcal{O}(nd^2)$ time complexity and its inverse will take $\mathcal{O}(d^3)$. To accelerate the this calculation, we apply conjugate gradients and stochastic estimations of Hessian-vector products, which reduce the time complexity to $\mathcal{O}(nd)$. Hence, the overall time complex will be $\mathcal{O}(nd)$.

4.4. Model Retraining

After obtaining the labels of the queried samples, we update the classifier h' by training with the newly annotated data L_t^k in each iteration. To preserve the knowledge learned in the previous iterations, we also include all labeled data L_s^k in the training with a smaller weight than L_t^k. Mathematically, the re-training is supervised by the following loss term:

$$\begin{aligned}\mathcal{L}_h =& \frac{n_t}{n_s + n_t} \Sigma_{(x_s,y)\in L_s^k} \ell(h'(x_s), y) \\ &+ \frac{n_s}{n_s + n_t} \Sigma_{(x_t,y)\in L_t^k} \ell(h'(x_t), y),\end{aligned} \tag{6}$$

where n_s is the size of L_s^k and n_t is the size of L_t^k.

5. Experiments

We illustrate the performance of our method in this section. We first introduce the experimental setup, and then report the algorithmic performances in category-aware domain adaptation. Finally, we provide various in-depth analyses to answer crucial questions in category-aware active DA.

5.1. Experimental Setup

Datasets. We choose three popular DA benchmark datasets, *Office-Home* (Venkateswara et al., 2017), *DomainNet-126* (Peng et al., 2019) and *VisDa-2017* (Peng et al., 2017) in our experiments. We conduct experimenters on all 12 adaptation tasks on *Office-Home*, 6 tasks adapting source domains Cl and Sk to the rest of target domains on *DomainNet-126* (Peng et al., 2019) and the challenging S→R task on *VisDa-2017* (Peng et al., 2017). The detailed descriptions of the datasets can be found in Appendix A.

Baseline methods. We select four active DA baseline methods in our compassion, including Random Sampling, CLUE (Prabhu et al., 2021), DUC (Xie et al., 2023), ISAL (Liu et al., 2021). We also add an "Always One" method, which chooses the unlabeled samples that are most likely to belong to the "One" category, i.e., the target category that is expected to improve, according to the classifier.

Table 1: Average per-category predicting accuracy (%) for Category-Aware Active Domain Adaptation on *Office-Home*. *Avg.* column represents the average accuracy across all 6 tasks in the same table.

Office-Home	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Avg.
DANN (Ganin et al., 2016)	36.99	51.83	58.81	36.54	48.59	59.17	48.65
Random	47.19	55.41	70.14	50.92	65.08	62.77	58.58
CLUE (Prabhu et al., 2021)	48.14	59.34	74.34	54.23	78.74	66.92	63.62
DUC (Xie et al., 2023)	51.23	61.81	75.31	56.41	79.82	70.02	65.76
ISAL (Liu et al., 2021)	64.84	79.23	78.66	62.93	81.24	75.62	73.75
Always One	66.27	85.85	81.93	**70.91**	82.15	77.94	77.51
Ours	**69.17**	**88.42**	**86.25**	68.89	**84.47**	**82.74**	**81.01**
Office-Home	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg.
DANN (Ganin et al., 2016)	32.83	32.98	59.26	52.28	68.93	45.36	48.61
Random	48.41	45.77	72.11	61.92	77.61	55.78	60.26
CLUE (Prabhu et al., 2021)	53.17	59.83	75.13	65.17	80.42	61.14	65.81
DUC (Xie et al., 2023)	59.23	60.73	77.34	67.35	79.68	65.02	68.22
ISAL (Liu et al., 2021)	64.17	65.02	83.51	71.42	80.41	71.51	72.67
Always One	67.44	63.22	83.42	78.16	87.83	67.65	74.62
Ours	**69.75**	**67.85**	**88.31**	**79.61**	**92.11**	**74.22**	**78.64**

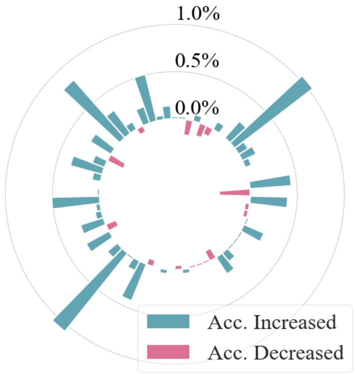


Figure 3: The performance changes for all test samples that belong to the "Non-One" categories after category-aware active learning for adaptation task Cl→Pr on *Office-Home* dataset.

Specifically, it queries the unlabeled samples with the highest soft-max output for the "One" category of h'. For our influence-based method, we provide the implementation details and codes of our method in Appendix B.

5.2. Algorithmic Performance

To evince the efficiency of our method in improving the performance of the targeted individual categories, we compare the per-category accuracy of our model and the baselines. For individual category c, we train the model for five active iterations and calculate the recognition accuracy for the test target data of category c. We conduct the category-aware active DA experiments for all categories for each transfer task, and report the average testing per-category accuracy across all targeted categories for the same task. We provide the detailed experimental protocol in Appendix C.

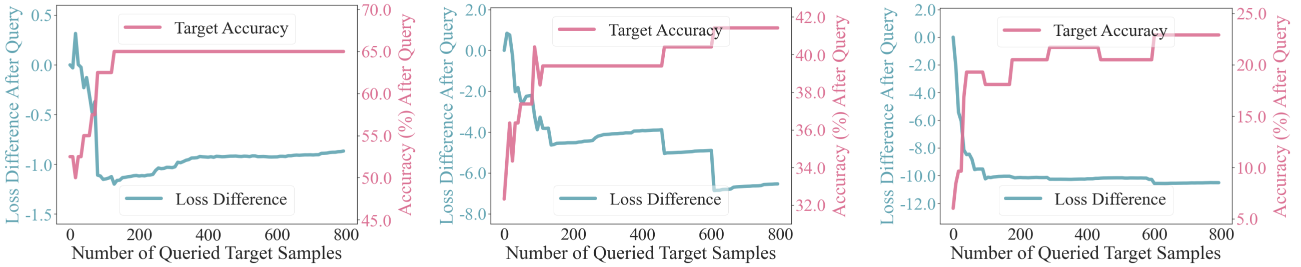
As shown in Table 1, our method attains the best average per-category accuracy on 11 out of 12 tasks in *Office-Home* (Venkateswara et al., 2017), achieving an average of 79.32% across 12 tasks. The significant performance margin validates the effectiveness of our influence-based method over CLUE (Prabhu et al., 2021) and DUC (Xie et al., 2023) in the category-aware settings. Our method also outperforms the ISAL (Liu et al., 2021) by 6.11% (ISAL obtains an average of 73.21% in 12 tasks), which also incorporates the influence function. This demonstrates that our influence approximation module provides more accurate impact estimations for the target domain, compared with the ISAL's influence calculation based on pseudo labels. The "Always One" method, which we consider to be a straightforward baseline suitable for the category-aware setting, falls behind our proposed method by 3.51%. Therefore we believe that both "One" and "Non-One" samples contribute to the prediction of the targeted category for our influenced-based method. We also notice that this straightforward "Always One" method attains the highest accuracy among the baselines, which illustrates the necessity for a method specifically tailored for category-aware active DA.

Table 2: Average per-category predicting accuracy (%) for Category-Aware Active Domain Adaptation on *DomainNet-126* and *VisDA-2017*. *Avg.* column represents the average accuracy across all 3 tasks of *DomainNet-126* in this table.

Method	*DomainNet-126*				*VisDA-2017*
	cl→pt	cl→rl	cl→sk	Avg.	S→R
DANN 2016	34.18	42.15	36.12	37.48	47.84
Random	36.25	44.75	40.74	40.58	54.19
CLUE 2021	38.06	45.34	42.28	46.47	58.54
DUC 2023	37.53	48.91	39.84	41.89	60.12
ISAL 2021	45.05	56.94	51.31	51.10	62.31
Always One	47.13	57.63	52.84	52.53	63.25
Ours	**52.23**	**64.21**	**56.34**	**57.59**	**66.68**

In addition to boosting the performance of the targeted category, we also want to avoid significantly sacrificing the accuracy of other categories, as we discussed in Section 3. To that end, we calculated the accuracy change for test samples not belonging to the target category c for adaptation task Cl→Pr on *Office-Home* (Venkateswara et al., 2017), and plot the results after training for each targeted category in Figure 3. As demonstrated in the figure, the category-aware active learning for most targeted categories does not impair other categories. Specifically, only 15 out of 65 categories suffer performance loss on the non-targeted categories, and the largest drop is only -0.33%, which validates that our method can achieve category-aware improvement without hurting the recognition of other categories.

We also conduct experiments for *DomainNet-126* (Peng



(a) Category "Calendar"

(b) Category "Helmet"

(c) Category "Fork"

Figure 4: Classification loss on the validation set and predicting accuracy on the target domain in three different categories in the Ar→Cl task on *Office-Home* (Venkateswara et al., 2017) dataset. For each category, our category-aware active DA method queries different numbers of unlabeled target samples, and calculates the validation loss and the test predicting accuracy after re-training the model with newly added annotations.

et al., 2019) and *VisDa-2017* (Peng et al., 2017) datasets to illustrate the generalizability of our method to large-scale datasets in Table 2. Due to limited space, we only include results for three tasks for *DomainNet-126* and report the full results in Appendix D. Although the large data volume increases the difficulty of selecting target samples, our method still outperforms all active domain adaptation baselines. Overall, the improvements are less prominent for all methods on these two datasets, due to the limited query budget in contrast to the substantial data size. Despite the challenging datasets, our method still obtains results that are at least 5.06% and 3.43% higher than the baselines on *DomainNet-126* and *VisDa-2017*, respectively. Similar to *Office-Home*, the "Always One" method exceeds other active learning methods that do not emphasize the enhancement of individual categories, further validating our motivation for category-aware active domain adaptation.

5.3. In-depth Exploration

In our in-depth exploration, we would like to answer the following questions for our category-aware active DA method:

1. As our method aims at better performance for individual categories, can we estimate the upper limit of the recognition accuracy of the target category?
2. In our pursuit of this per-category upper limit, is it beneficial to annotate some "Non-One" data, *i.e.*, samples out of the targeted category?
3. Given we have chosen the important target data after answering Questions 1 and 2, can we apply strategies utilized by previous active learning methods to further boost the performance?
4. After achieving category-specific improvement, we observed a slight performance drop for certain "Non-One" categories. How can we compensate for such loss within our category-aware active DA framework?

Upper limit for active learning. In the active learning tasks, there is a natural question: *How much we can improve the performance, and how many samples do we need to reach that upper limit?* As our method selects the query set by estimating each sample's impact on the loss function of validation data, we conjure that the validation loss could indicate the upper limit of the performance gain after annotating more target samples. To that end, we conduct experiments to further explore the association between validation loss, predicting accuracy, and the number of queried samples.

We choose three different categories, namely, "Calendar," "Helmet," and "Fork" in the Ar→Cl task on *Office-Home* dataset, and plot their change of validation loss and predicting accuracy after increasing the number of queried samples. For one individual category, we use our influence-based method to select 10 more samples for the target category in each iteration, up to 800 target samples. As demonstrated in Figure 4, the target accuracy shows a strong negative association with the change of validation loss in all three categories for task Ar→Cl, and the per-category prediction accuracy reaches the upper limit when the validation loss drops down to the lowest value. These results substantiate our above conjunction that validation loss change can be used to assess the upper accuracy limit and query size for our active learning method, and help the user to leverage between performance gain and query budget.

Effect of "Non-One" samples. The experimental results in Section 5.2 indicate that the "Non-One" samples for other categories might also help recognize the target category, as we notice that the "Always One" method, which only selects the unlabeled samples that are most likely to belong to the target category, performs worse than our influenced-based method. Hence, we aim to delve deeper into the effect of the samples labeled as "Non-One" in this section.

Firstly, we check how many "One" samples are queried in task Pr→Ar on *Office-Home* (Venkateswara et al., 2017) for different categories with various performance changes after active learning. In Figure 5(a), we plot the ratio of "One" samples and the per-category accuracy differences after

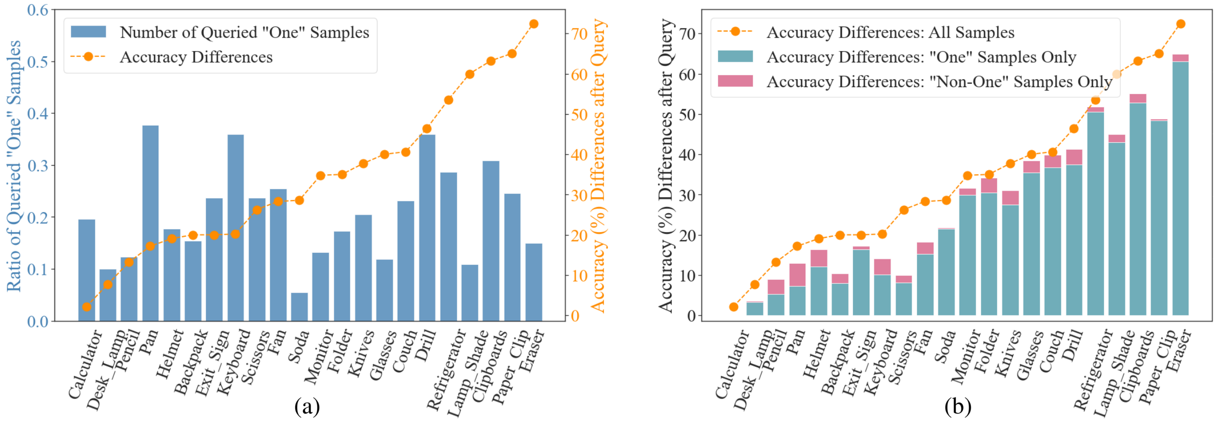


Figure 5: Per-category accuracy differences for the same set of categories in task Pr→Ar on *Office-Home* (Venkateswara et al., 2017) after active learning are plotted with orange lines in both figures. (a) Blue bars represent the ratio of "One" samples queried by our category-aware active learning method; (b) The green bars and red bars indicate the accuracy differences when re-training the model with only "One" and "Non-One" samples from the queried data, respectively.

querying 5% data from the target domain. As shown in the figure, a significant performance gain does not necessarily require a large number of annotated target samples from the targeted category. For example, both "Lamp_Shade" and "Eraser" classes attain more than 50% improvement after actively learning, while only less than 20% of the queried samples are from the targeted class. Conversely, categories like "Pan" and "keyboard" only get a moderate boost despite being trained with more samples of the targeted classes.

We also conduct experiments isolating the queried samples, *i.e.*, we only annotate the selected "One" or "Non-One" samples and re-train the classifier separately. In Figure 5(b), we again plot the per-category accuracy differences after active learning with the orange line, along with the accuracy differences when only re-training with newly annotated "One" or "Non-One" exclusively, using the red and green bars. As demonstrated in the figure, even though re-training the model with the "One" samples can achieve considerable improvements, the "Non-One" samples alone also benefit each individual category. Besides, the performance gain of all samples is larger than the gains of only including "One" and "Non-One" samples combined, evincing that both kinds of queried samples help recognize the targeted categories.

Diversity sampling and data re-weighting. Inspired by previous active DA works like (Su et al., 2020; Fu et al., 2021), we add a sampling step to obtain a query set with higher diversity, and increase the weight of the queried data during the re-training. In this section, we conduct an ablation study for each module in our method for adaptation task Rw→Ar on *Office-Home* dataset. For each method, we run the experiments 3 times, then report the average results and standard deviations in Table 3.

Table 3: Accuracy with different model components for domain adaptation task Rw→Ar on *Office-Home* dataset

Model	Infl.	Re-Wt.	Samp.	Acc. (%)
DANN	×	×	×	52.28 ± 3.63
Infl. Only	√	×	×	72.74 ± 5.41
Infl. + Re-Wt.	√	√	×	76.52 ± 6.37
Infl. + Samp.	√	×	√	74.36 ± 3.18
Full Model	√	√	√	79.15 ± 2.92

Our influence-based active learning strategy (Infl. Only) without re-weighting and sampling boosts the performance by 20.46% over DANN (Ganin et al., 2016) with the extra target annotations. Adding data re-weighting (Re-Wt.) further improves the accuracy by 2.74%, indicating the benefit of emphasizing the newly annotated samples. Diversity sampling (Samp.) alone only brings 1.64% performance gain, but reduces the standard deviation over 3 random runs. Compared with the above results, the final accuracy of the full model demonstrated the efficiency of integrating both diversity sampling and data re-weighting. Together, these two modules increase the performance by 6.41% and reduce the deviation compared with the influence-only model.

Remedy for "Non-One" categories. In Section 5.2, we noticed a slight accuracy drop for "Non-One" samples after active training for 15 targeted categories on adaptation task Cl→Pr. To rectify this minor detriment, we employ a remedy modification to our query strategy to alleviate the negative impact on "Non-One" samples. Technically, we partition "One" and "Non-One" samples in the validation set V into V_1 and V_0 respectively, where V contains all the annotated target data queried by the model. Subsequently, we compute the influence of each source point x_s on V_1 and V_0 separately using Eq. (4). Following this, we proceed to train individual target influence predictors f_1 and f_0 to estimate the influence of an unlabeled sample x_t on "One" and "Non-One" data. Finally, we exclude the samples with negative influence on "Non-One" data, *i.e.*, $f_0(x_t) < 0$, and

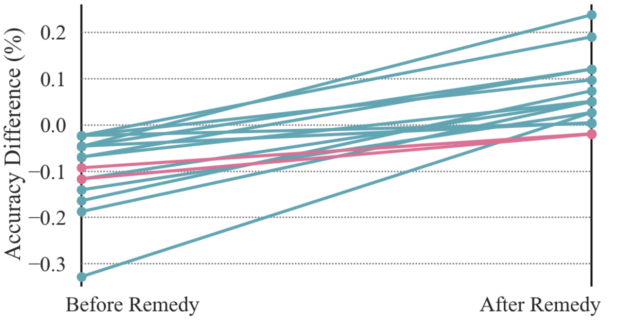


Figure 6: Performance on our remedy on the 15 categories with performance drop in Figure 3 through two individual target influence predictors on the "One" and "Non-One."

query the samples with highest influence $f_1(x_t)$ from the rest of the unlabeled data. As shown in Figure 6, the remedy modification successfully eliminates the performance drop for 13 categories after active learning and reduces the adverse impact on the other 2 categories to almost 0%. It's noteworthy that even after applying the remedy, our model continues to achieve comparable improvements in the 15 targeted categories, just as it did before the remedy.

6. Conclusion

To sum up, we proposed a category-aware framework to enhance the adaptation for individual categories in active DA, mitigating the potential risks associated with overlooking critical categories. To the best of our knowledge, this is the initial effort to tackle the performance discrepancy among categories in active DA. To achieve category-aware enhancement, our method initially trains a binary classifier dedicated to recognizing a specific category. The category-aware active learning module subsequently utilizes the influence function to directly estimate the importance of each unlabeled sample for the category-specific classifier. Samples with the highest influence estimations are selected for annotation, and these annotations are then utilized to supervise the retraining of the classifier, thereby facilitating adaptation for the specific category. The efficacy of our method was manifested by experiments on three benchmark datasets, and we also conducted extensive in-depth explorations to answer some critical questions in category-aware active DA.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. We expand the well-defined active domain adaptation task to the category-aware setting, which tackles performance discrepancy among data categories brought by active learning. Addressing the disparities overlooked by current active domain adaptation methods is essential, as it helps prevent biases in knowledge transfer and enhances model robustness, avoiding significant risks to vital entities. Beyond the impact mentioned above, there are also other potential societal consequences of our work, none of which we feel must be specifically highlighted here.

Limitations

One potential limitation of our method is that the binarization might cause an imbalance in training data. We acknowledge the data imbalance issues in the binary classification, but this is not the primary focus of this work. Therefore, we did not focus on the data imbalance, which has been extensively studied in various previous works (Elhassan et al., 2016; Dai et al., 2022; Gong & Kim, 2017; Popel et al., 2018; Ahmed et al., 2019) dedicated to this issue.

References

Ahmed, S., Rayhan, F., Mahbub, A., Rafsan Jani, M., Shatabda, S., and Farid, D. M. Liuboost: locality informed under-boosting for imbalanced data classification. In *Proceedings of Emerging Technologies in Data Mining and Information Security*, 2019.

Alaa, A. and Van Der Schaar, M. Validating causal inference models via influence functions. In *Proceedings of the International Conference on Machine Learning*, 2019.

Basu, S., Pope, P., and Feizi, S. Influence functions in deep learning are fragile. In *Proceedings of the International Conference on Learning Representations*, 2020.

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., and Varoquaux, G. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013.

Chen, Y., Li, W., Sakaridis, C., Dai, D., and Van Gool, L. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

Chen, Z., Li, P., Liu, H., and Hong, P. Characterizing the influence of graph elements. In *Proceedings of the International Conference on Learning Representations*, 2023.

Cook, R. D. and Weisberg, S. Characterizations of an empirical influence function for detecting influential cases in regression. *Technometrics*, 1980.

Dai, Q., Liu, J.-w., and Liu, Y. Multi-granularity relabeled under-sampling algorithm for imbalanced data. *Applied Soft Computing*, 2022.

Elhassan, T., M, A., F, A.-M., and Shoukri, M. Classification of imbalance data using tomek link(t-link) combined with random under-sampling (rus) as a data reduction method. *Global Journal of Technology and Optimization*, 2016.

Fang, M., Gong, N. Z., and Liu, J. Influence function based data poisoning attacks to top-n recommender systems. In *Proceedings of The Web Conference*, 2020.

Fu, B., Cao, Z., Wang, J., and Long, M. Transferable query selection for active domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

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

Giordano, R., Stephenson, W., Liu, R., Jordan, M., and Broderick, T. A swiss army infinitesimal jackknife. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2019.

Gong, J. and Kim, H. Rhsboost: Improving classification performance in imbalance data. *Computational Statistics & Data Analysis*, 2017.

Han, X., Wallace, B. C., and Tsvetkov, Y. Explaining black box predictions and unveiling data artifacts through influence functions. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2020.

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

Huang, D., Li, J., Chen, W., Huang, J., Chai, Z., and Li, G. Divide and adapt: Active domain adaptation via customized learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Hwang, S., Lee, S., Kim, S., Ok, J., and Kwak, S. Combating label distribution shift for active domain adaptation. In *Proceedings of the European Conference on Computer Vision*, 2022.

Koh, P. W. and Liang, P. Understanding black-box predictions via influence functions. In *Proceedings of the International Conference on Machine Learning*, 2017.

Koh, P. W. W., Ang, K.-S., Teo, H., and Liang, P. S. On the accuracy of influence functions for measuring group effects. *Advances in Neural Information Processing Systems*, 2019.

Kothandaraman, D., Shekhar, S., Sancheti, A., Ghuhan, M., Shukla, T., and Manocha, D. Salad: Source-free active label-agnostic domain adaptation for classification, segmentation and detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023.

LeCun, Y., Chopra, S., Hadsell, R., Ranzato, A., and Huang, F. J. A tutorial on energy-based learning. *Predicting Structured Data*, 2006.

Li, P. and Liu, H. Achieving fairness at no utility cost via data reweighing with influence. In *Proceedings of the International Conference on Machine Learning*, 2022.

Li, S., Zhang, R., Gong, K., Xie, M., Ma, W., and Gao, G. Source-free active domain adaptation via augmentation-based sample query and progressive model adaptation. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

Li, X., Du, Z., Li, J., Zhu, L., and Lu, K. Source-free active domain adaptation via energy-based locality preserving transfer. In *Proceedings of the ACM International Conference on Multimedia*, 2022.

Liu, Z., Ding, H., Zhong, H., Li, W., Dai, J., and He, C. Influence selection for active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 2019.

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

Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., and Wang, B. Moment matching for multi-source domain adaptation. In *Proceedings of the International Conference on Computer Vision*, 2019.

Popel, M. H., Hasib, K. M., Ahsan Habib, S., and Muhammad Shah, F. A hybrid under-sampling method (husboost) to classify imbalanced data. In *Proceedings of the International Conference of Computer and Information Technology*, 2018.

Prabhu, V., Chandrasekaran, A., Saenko, K., and Hoffman, J. Active domain adaptation via clustering uncertainty-weighted embeddings. In *Proceedings of the International Conference on Computer Vision*, 2021.

Rangwani, H., Jain, A., Aithal, S. K., and Babu, R. V. S3vaada: Submodular subset selection for virtual adversarial active domain adaptation. In *Proceedings of the International Conference on Computer Vision*, 2021.

Su, J.-C., Tsai, Y.-H., Sohn, K., Liu, B., Maji, S., and Chandraker, M. Active adversarial domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020.

Ting, D. and Brochu, E. Optimal subsampling with influence functions. *Advances in Neural Information Processing Systems*, 2018.

Tsai, Y.-H., Hung, W.-C., Schulter, S., Sohn, K., Yang, M.-H., and Chandraker, M. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

Venkateswara, H., Eusebio, J., Chakraborty, S., and Panchanathan, S. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.

Wang, F., Han, Z., Zhang, Z., He, R., and Yin, Y. Mhpl: Minimum happy points learning for active source free domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023a.

Wang, F., Han, Z., Zhang, Z., He, R., and Yin, Y. Mhpl: Minimum happy points learning for active source free domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023b.

Wang, F., Han, Z., and Yin, Y. Bias: Bridging inactive and active samples for active source free domain adaptation. *Knowledge-Based Systems*, 2024.

Xie, B., Yuan, L., Li, S., Liu, C. H., Cheng, X., and Wang, G. Active learning for domain adaptation: An energy-based approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022a.

Xie, M., Li, Y., Wang, Y., Luo, Z., Gan, Z., Sun, Z., Chi, M., Wang, C., and Wang, P. Learning distinctive margin toward active domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022b.

Xie, M., Li, S., Zhang, R., and Liu, C. H. Dirichlet-based uncertainty calibration for active domain adaptation. In *Proceedings of the International Conference on Learning Representations*, 2023.

Appendix

A. Datasets

In our experiments, we choose three popular DA benchmark datasets, *Office-Home* (Venkateswara et al., 2017), *DomainNet-126* (Peng et al., 2019) and *VisDa-2017* (Peng et al., 2017). *(i) Office-Home* (Venkateswara et al., 2017) is one of the most popular DA benchmark datasets, which contains images of 65 different categories from four domains: Art (Ar), Clip Art (Cl), Product (Pr), and Real World (Rw). We include all 12 available adaptation tasks in our experiments. *(ii) DomainNet-126* is a subset of *DomainNet* (Peng et al., 2019), the current largest DA benchmark dataset, which consists of 345 categories from 6 different domains. We follow the same data protocol as (Prabhu et al., 2021) and choose 126 categories from 4 different domains: Real (Rl), Clip Art (Cl), Painting (Pt), and Sketch (Sk) since the labels for certain domains and categories are very noisy. We conduct six tasks adapting source domains Cl and Sk to the rest three target domains. *(iii) VisDa-2017* (Peng et al., 2017) is a large-scale dataset containing 12 categories from 2 domains, Synthetic (S) and Real (R). In this work, we focus on the challenging S→R task.

B. Implementation Details

We implement[1] our model using PyTorch(Paszke et al., 2019) and scikit-learn (Buitinck et al., 2013) with one NVIDIA TITAN RTX GPU. We use the ResNet-50 (He et al., 2016) pre-trained on *ImageNet* as the backbone feature extractor and train a DANN model as the base DA method following CLUE (Prabhu et al., 2021). The feature representations for both the labeled source domain and the unlabeled target domain are obtained with a fixed feature extractor pre-trained by the base DA method. For each specific category c, we convert the labels into binary, *i.e.*, $(x_s, y) \in L_s$ and train the category-specific surrogate model h' with L_s. With this convex surrogate classifier, we choose our first query set as described in Section 4.3. In the following iterations, we actively query and annotate b target samples with Algorithm 1.

C. Experimental protocol

For each task in one dataset, we conduct the category-aware active learning for each category c. We first binarize all available labels, i.e., "One" for category c and "Non-One" for all other categories. Based on the binary labels and the representations extracted by the base DA model, we initially train a category-specific logistic regression model h' as the surrogate classifier. Subsequently, we perform active learning in five iterations. In each iteration, we select b samples for the target domain as described in Algorithm 1 and label these samples L_t with ground truth. Next, we retrain the classifier h' with Eq. (6). After 5 iterations of active learning, we compute the predicting accuracy of the "One" samples for category c in the test set. For all three datasets, we train the model for 5 iterations. For *Office-Home* (Venkateswara et al., 2017), we select 1% of the target data in each iteration. For the large datasets *VisDa-2017* (Peng et al., 2017) and *DomainNet-126* (Peng et al., 2019), we select a fixed number of 100 samples in each iteration. For each transferring task, we take the average accuracy across all categories and report it in Table 1 and Table 2. We apply the same training and testing protocol in the experiments for baseline methods, and report the average per-category accuracy.

D. Experimental Results for *DomainNet-126*

Table 4: Average per-category predicting accuracy (%) for Category-aware Active Domain Adaptation on *DomainNet-126* (Peng et al., 2019). Avg. column represents the average accuracy across all 6 tasks in *DomainNet-126*.

Method	cl→pt	cl→rl	cl→sk	sk→cl	sk→pt	sk→rl	Avg.
DANN (Ganin et al., 2016)	34.18	42.15	36.12	31.82	56.59	47.24	41.35
Random	36.25	44.75	40.74	34.41	58.31	49.53	44.00
CLUE (Prabhu et al., 2021)	38.06	45.34	42.28	35.63	60.12	57.37	46.47
DUC (Xie et al., 2023)	37.53	48.91	39.84	36.12	61.35	58.26	47.02
ISAL (Liu et al., 2021)	45.05	56.94	51.31	40.34	62.27	61.64	52.92
Always One	47.13	57.63	52.84	41.62	61.61	59.15	53.33
Ours	**52.23**	**64.21**	**56.34**	**42.49**	**67.94**	**63.41**	**57.78**

[1] Our code is available at *https://github.com/wxxiaoss/Category_Aware_DA*.