Loss-based Sequential Learning for Active Domain Adaptation

Kyeongtak Han
Inha University
han00127@inha.edu

Youngeun Kim
Yale University
youngeun.kim@yale.edu

Dongyoon Han
NAVER AI Lab
dongyoon.han@navercorp.com

Sungeun Hong
Inha University
csehong@inha.ac.kr

Abstract

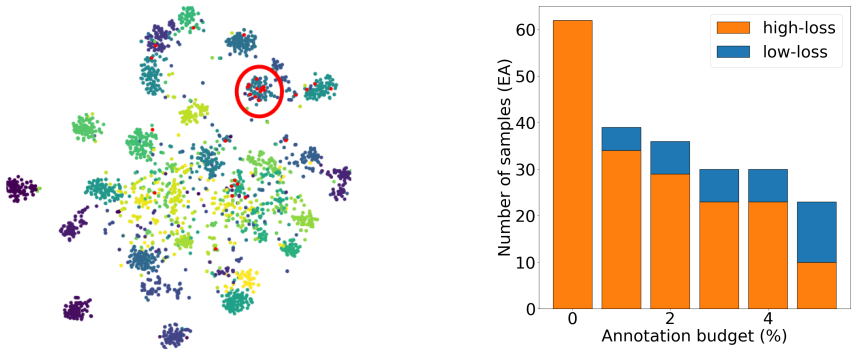
Active domain adaptation (ADA) studies have mainly addressed query selection while following existing domain adaptation strategies. However, we argue that it is critical to consider not only query selection criteria but also domain adaptation strategies designed for ADA scenarios. This paper introduces sequential learning considering both domain type (source/target) or labelness (labeled/unlabeled). We first train our model only on labeled target samples obtained by loss-based query selection. When loss-based query selection is applied under domain shift, unuseful high-loss samples gradually increase, and the labeled-sample diversity becomes low. To solve these, we fully utilize pseudo labels of the unlabeled target domain by leveraging loss prediction. We further encourage pseudo labels to have low self-entropy and diverse class distributions. Our model significantly outperforms previous methods as well as baseline models in various benchmark datasets.

1 Introduction

Although unsupervised domain adaptation (UDA) has shown promising results across various fields [1, 2], the performance gap between UDA and its supervised setting, *i.e.*, usage of target labels, is significant. The non-negligible performance gap is the main obstacle to applying UDA to real-world problems. Recently, semi-supervised domain adaptation has been actively investigated to alleviate such challenges. In the semi-supervised settings, target samples to be labeled are randomly selected [3] or sampled at an even rate across all classes [4, 5]. However, random labeling is less efficient than sampling based on specific criteria when limited annotation budget. Additionally, applying uniform sampling across classes is impractical because the target samples' labels are presumably unknown. For this reason, active domain adaptation (ADA) that aims to annotate the most informative target samples under domain shift automatically has emerged [6, 7].

One of the main challenges in the ADA task is how to select the most informative target samples under domain shift with limited annotation budgets. Early works [6, 8] use uncertainty [9] or diversity [10] as sample selection criteria. However, uncertainty estimation on the target domain is usually miscalibrated, resulting in sampling outliers or redundant instances as indicated by [11]. Furthermore, domain similarity used as diversity cannot guarantee discriminative feature space under domain shift [7]. Therefore, more recent work [11] jointly uses uncertainty and diversity as query selection criteria. Rangwani *et al.*[12] adopt submodular subset selection, which is a combination of uncertainty, diversity, and representativeness score of samples. Ma *et al.*[13] primarily utilizes uncertainty and diversity score. The state-of-the-art TQS [7] adopts query by committee scheme [14] with uncertainty and domain similarity, which utilizes consensus between multiple predictions.

Preprint. Under review.



(a) Low diversity (b) Misprediction rate

Figure 1: Limitations of loss-based query selection in active domain adaptation. (a) Labeled samples with high-loss (indicated in red) show low diversity in the target domain. Detailed description can be found in Section 4.4. (b) Mispredicted high-loss samples gradually decrease while the low-loss samples increase in unlabeled target domain, which indicates unuseful samples are labeled.

Table 1: Comparison of the proposed method on various benchmark datasets, in which Δ represents the difference in accuracy (%) between 'Learning loss + DA' and 'Ours'.

	Baseline	Learning loss [15] + DA	Ours	Δ
Office-31	80.0	83.6	92.2	+8.6
Office-Home	58.3	68.1	75.6	+7.5
VisDA	44.7	85.8	86.8	+1.0

Critically, most previous ADA methods have focused primarily on query selection, but have not investigated how to deal with the labeled source, labeled target, and unlabeled target domains after query selection. Several methods [6, 7] perform ADA by integrating the newly labeled target domain and the existing labeled source domain into a single data pool. Prabhu *et al.*[11] do not directly integrate the labeled source and labeled target domains, but treat them as one dataset in the training process. We argue that it is critical to consider not only how to choose informative unlabeled target samples, but also how to treat labeled target samples, considering domain shift.

In this paper, we propose sequential ADA learning, which takes into account domain type (source/target) or labelness (labeled/unlabeled). As a query selection criterion, we exploit [15] that actively selects unlabeled samples with high loss predicted by an auxiliary module. However, high-loss samples selected by loss prediction often do not help model training and cannot handle the diversity issue, as shown in Fig. 1. We analyzed the model's misprediction rate by dividing unlabeled samples into two groups (high or low) according to their expected loss values. Surprisingly, as the annotation budget increases, the number of falsely predicted high-loss samples decreases; while the number of low-loss samples increases. Furthermore, loss-based query selection does not account for sample diversity, which is exacerbated under domain shift.

To solve the aforementioned issues, we introduce a sequential adaptation strategy by learning with a small number of labelness samples and then using a large number of unlabeled samples aggressively. We first train the model using only the labeled target samples given by the oracle. Labeled target samples contribute to increasing discriminative target representation but occupy only a small fraction of the target domain. Therefore, training a model based on a small number of labeled target samples is difficult to reflect the overall target distribution. To address this, we fully utilize a number of unlabeled target domains via pseudo labeling. We further encourage pseudo labels to have low self-entropy and diverse class distributions to increase the reliability of the target pseudo labels.

Our contributions are as follows. (i) We propose a novel loss-based ADA learning that sequentially utilizes a small number of ground-truth labels by oracle and numerous pseudo labels obtained by elaborated self-learning. (ii) We analyze the limitations of loss-based query selection that have not been used in active domain adaptation, and present a simple but effective solution that outperforms the baseline models as shown in Table 1. (iii) The proposed method is superior to the existing ADA methods by a large gap in various datasets.

2 Related Work

2.1 Active learning

Over the past decades, a number of selection criteria for active learning (AL) have been suggested and these can be divided into three categories: uncertainty, committee, and diversity. Uncertainty-based approaches using confidence [16], entropy [8], or best-vs-second-best [17] have been widely used because of its simplicity and high-computational efficiency. However, most uncertainty-based approaches use task-specific uncertainty measures. Query-by-committee [14] that leverages consensus among an ensemble of multiple classifiers can handle this issue, allowing a wide range of applications. Diversity-based selection [18, 10] is also task-agnostic because selecting the samples that represent the overall distribution of the unlabeled data pool does not depend on particular tasks.

Overall, most of the existing approaches require task-specific architectures or are computationally inefficient, especially for the recent deep networks. Motivated by the fact that deep networks are trained by minimizing loss regardless of the number of tasks, task types, and model complexity, [15] propose a new task-agnostic approach. Considering the model's loss value as uncertainty, they attach a loss prediction module to the main network and trained this module to estimate the loss of unlabeled samples. Critically, since all the criteria previously used in active learning are designed considering a single domain, they are not non-transferable. Therefore, they could not select the most informative sample for annotation under domain shift.

2.2 Domain adaptation

Unsupervised domain adaptation (UDA) has been investigated a lot as it can transfer knowledge of a labeled source domain to an unlabeled target domain [19]. Even though accessing all label information of the target domain requires expensive labeling costs, labeling only a few samples and applying them to the UDA process is cost-effective. For this reason, semi-supervised domain adaptation (SSDA) methods have been proposed across various areas [4, 5, 3]. Active domain adaptation (ADA) [20] is similar to SSDA [4, 5] in that both can access few labeled target data. While SSDA approaches label target samples randomly or according to predetermined rules, ADA models automatically select target samples to be labeled.

Previous methods mainly utilize uncertainty and diversity as sample selection criteria. Su *et al*. [6] use diversity from importance weights as well as uncertainty via entropy for query selection. Prabhu *et al*. [11] identifies target samples that are uncertain and representative in feature space and the training scheme is based on the mini-max entropy (MME) [4]. Rangwani *et al*. [12] combine uncertainty, diversity and representativeness of samples via submodular subset and train the labeled source, labeled target and unlabeled target samples with improved VADA [21]. Query by committee scheme was also used for ADA and training only labeled source and labeled target, which has shown state-of-the-art performance [7]. Critically, constructing multiple classifiers and their respective data pipelines requires high computational cost. Instead of focusing on query selection criteria, we propose a new sequential process that leverages loss-based query selection.

3 Method

3.1 Overall framework

In active domain adaptation (ADA) scenarios, we are given labeled source domain $\mathcal{X}_s = \{(x_s^i, y_s^i)\}_{i=1}^{N_s}$ and unlabeled target domain $\mathcal{X}_{t_u} = \{(x_{t_u}^i)\}_{i=1}^{N_{t_u}}$ where N_s and N_{t_u} refer to the total number of the labeled source and unlabeled target domains, respectively. As model training progresses, a part of the target domain is labeled by oracle according to query selection and denoted as $\mathcal{X}_t = \{(x_t^i, y_t^i)\}_{i=1}^{N_t}$. Once the labeled target domain $\mathcal{X}_t$ is obtained, the model is trained in a semi-supervised manner with the existing $\mathcal{X}_s$ and remaining $\mathcal{X}_{t_u} \leftarrow \mathcal{X}_{t_u} \setminus \mathcal{X}_t$. Such query selection and semi-supervised domain adaptation are repeated until the annotation budget B is reached.

The main goal of our framework is to alleviate domain discrepancies between the fully labeled source domain and the partially-labeled target domain by leveraging query selection. Fig. 2 illustrates the outline of the proposed ADA framework. The main difference between the proposed method and individual-related work is the learning schemes rather than the designs of each module or architecture.

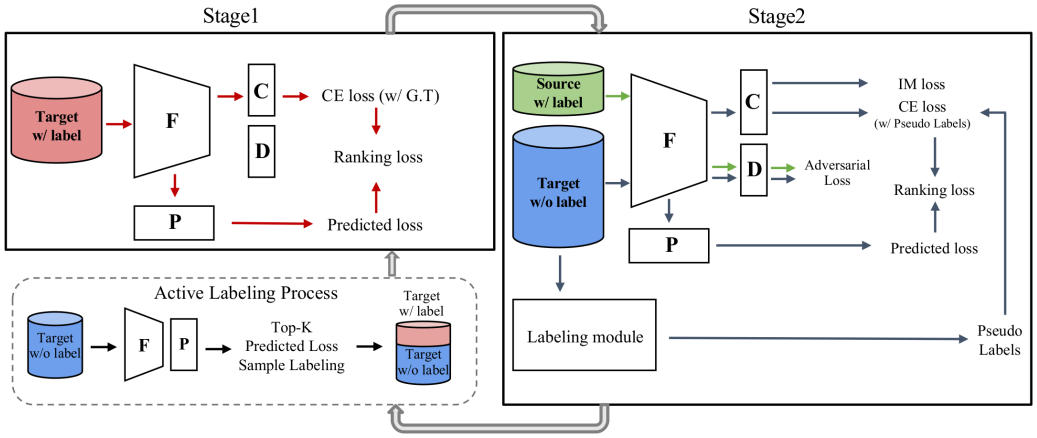


Figure 2: Outline of the proposed ADA framework considering domain type (source/target) and labelness (labeled/unlabeled).

Unlike most existing ADA methods [7, 22], we separate model training for labeled source domain and labeled target domain as in [23, 24]. Concretely, we pre-train a feature extractor F and a classifier C with the labeled source domain $\mathcal{X}_s$. We then freeze the classifier and alternately perform domain adaptation and query selection within the annotation budget B to optimize each sub-network.

As a first step, we train the model mainly on labeled target samples $\mathcal{X}_t$ obtained by oracle, similar to existing ADA approaches. We exploit loss-based query selection in our framework. However, if the model is trained only on labeled target samples by loss-based query selection, the model could be overfitted by a small number of labeled samples with low diversity. To increase generalization power in the target domain, we fully utilize a majority of unlabeled target samples by leveraging pseudo labeling and information maximization loss [25, 26]. Simultaneously, we train an adversarial domain discriminator to reduce domain discrepancy. After that, the samples with the top-K high-loss predicted by the auxiliary loss predictor P in the unlabeled target domain pool $\mathcal{X}_{t_u}$ are labeled as described in Algorithm 1.

Algorithm 1: Query selection

Input: unlabeled target domain: $\mathcal{X}_{t_u}$, non-trainable modules : (F, P), annotation budget: B
Output: set of queries $Q(X)$
initialize a predicted loss set L = []
initialize an active sample set Q = []
for $i = 1$ *to* N_{t_u} **do**
  $\hat{l}^i = P(F(x^i_{t_u}))$
  $\mathrm{L} \leftarrow \mathrm{L} \cup \hat{l}^i$
end
for $i = 1$ *to* B **do**
  $\mathrm{Q} \leftarrow \mathrm{Q} \cup \mathrm{argmax}_{x^i_{t_u}}(\mathrm{L})$
  $\mathrm{L} \leftarrow \mathrm{L} \setminus \hat{l}^i$
end

3.2 Model training with labeled target domain

An intuition of loss-based query selection is to actively label high-loss samples because samples with high loss have a high probability of being incorrectly recognized by the model. To this end, we train the auxiliary loss predictor P using labeled target samples $\mathcal{X}_t$ obtained through query selection previously. Specifically, once labeled target domain $\mathcal{X}_t = \{(x^i_t, y^i_t)\}^{N_t}_{i=1}$ is obtained by oracle, we train the feature extractor F using conventional cross-entropy loss $\mathcal{L}_{ce}$. The obtained actual loss value $\mathcal{L}_{ce}$ is then used as a ground-truth label for the auxiliary loss predictor P. The scale of the real loss $\mathcal{L}_{ce}$ steadily decreases as the model training progresses, so we use margin ranking loss as follows:

$$L_{loss} = \mathbb{E}_{x_t \in \mathcal{X}_t} max(0, -\mathbb{1}(l_n, l_m) \cdot (\hat{l}_n - \hat{l}_m) + \Delta)$$
$$s.t.\ \mathbb{1}(l_n, l_m) = \begin{cases} +1, & \text{if } l_n > l_m \\ -1, & \text{otherwise,} \end{cases} \tag{1}$$

where the n and m refer to the ranking pair index in the training mini-batch and Δ is a pre-defined margin. Unlike the conventional loss-based query selection for active learning [15], the gradient from our loss predictor P flows to the feature extractor F, so they are learned jointly. This approach is effective for generating pseudo labels of unlabeled target samples by making our feature extractor F learn loss under domain shift.

3.3 Sequential adaptation with unlabeled target domain

After training the model with labeled target samples $\mathcal{X}_t$, which occupy a very small number of the target domain, we fully utilize the majority of unlabeled target samples for regularization. Recall that the query selection of the proposed ADA framework is based on the loss predictor. We train the main network (*i.e.*, F and C) and the auxiliary loss predictor P using unlabeled target samples. To train the loss predictor P, we need ground-truth labels of samples. Unfortunately, there is no label in the unlabeled target domain $\mathcal{X}_{t_u}$. To solve this, we utilize target pseudo labels obtained as follows:

$$\tilde{y}_t = \underset{c}{argmax}(\sigma(C(F(x_{t_u})))), \tag{2}$$

where $\sigma(\cdot)$ is the Softmax function, c is a class index, and $\tilde{y}_t$ refers to the target pseudo labels from the model inference. Note that pseudo labels are updated at pre-determined intervals γ instead of every step, and they are used to train the feature extractor F in a self-training manner. We then use the pseudo-cross-entropy loss to train our loss predictor P using margin ranking loss as in Eq.1.

A natural question may arise about the reliability of pseudo labels for model training. To increase the reliability of the target pseudo labels, we use information maximization considering self-entropy and class diversity, which can be formulated as follows:

$$\begin{aligned} L_{im} = &- \mathbb{E}_{x_{t_u} \in \mathcal{X}_{t_u}} \sum_{c=1}^{C} [\mathbb{1}_{[c=\hat{y}^t]} log(\sigma(C(F(x_{t_u}))))] \\ &+ \xi D_{KL}(\hat{y}_t, \frac{1}{C}\mathbf{1}_C) - log(C). \end{aligned} \tag{3}$$

The first term represents the self-entropy, which encourages our model to assign disparate one-hot encodings to the feature representations of $\mathcal{X}_{t_u}$. The second term is used to avoid situations where the target pseudo labels $\tilde{y}_t$ are assigned to only a small number of classes, *i.e.*, low diversity. In the above equation, σ denotes Softmax function and $\mathbb{1}_{[\cdot]}$ is indicator function. C is the number of classes and $\mathbf{1}_C$ is a vector with all elements equal to 1 and the same size as the number of classes. Importantly, $\hat{y}_t = \mathbb{E}_{x_t \in \mathcal{X}_{t_u}}[\sigma(C(F(x_{t_u})))]$ is the mean output probability of the whole unlabeled target domain. D_{KL} denotes Kullback–Leibler divergence and ξ stands for controlling variable between two-loss terms. As a result, our model produces progressively more reliable target pseudo labels $\tilde{y}_t$ with our information maximization loss for diverse classes.

To alleviate domain discrepancy between the labeled source domain $\mathcal{X}_s$ and unlabeled target domain $\mathcal{X}_{t_u}$, we exploit a conventional mini-max game between the feature extractor F and the domain discriminator D. The domain discriminator D attempts to classify the domain label of the given samples while the feature extractor F tries to deceive the domain discriminator D as follows:

$$\begin{aligned} L_{dis} = &- \mathbb{E}_{x_s \in \mathcal{X}_s}[logD(F(x_s))] \\ &- \mathbb{E}_{x_{t_u} \in \mathcal{X}_{t_u}}[log(1 - D(F(x_{t_u})))]. \end{aligned} \tag{4}$$

$$L_{adv} = -L_{dis}. \tag{5}$$

The total loss consists of the losses mentioned above as:

$$L_{total} = L_{loss} + L_{adv} + L_{im} + L_{dis}. \tag{6}$$

As a result, we apply a constraint to generate reliable target pseudo labels while alleviating domain discrepancy. Importantly, our loss predictor actively utilizes unlabeled target samples occupying the majority of target domains and the whole process is described in Algorithm 2.

Algorithm 2: Sequential adaptation procedure

Input: labeled source domain: $\mathcal{X}_s$, unlabeled target domain: $\mathcal{X}_{t_u}$, trainable modules: (F, P, D), non-trainable module : C, total iterations N, $\theta_{(F,P,D)}$: parameters of each module, γ: intervals for pseudo labeling.

for $n = 1$ ***to*** N **do**

  if $n\%\ \gamma$ == *0* **then**

    get pseudo labels via Eq.2

  end

  $\hat{d}_i \leftarrow D(F(x_t, x_s))$

  $\hat{l}_i \leftarrow P(F(x_t))$

  $l_{G.T} \leftarrow \mathcal{L}_{ce}$ with target pseudo labels

  $L_{loss} \leftarrow Rankingloss(l_{G.T}, \hat{l}_i)$ via Eq.1

  $L_{im} \leftarrow L_{ent}$ + L_{class_div} via Eq.3

  $L_{dis} \leftarrow -log(\hat{d}_i) - log(1 - \hat{d}_i)$ via Eq.4

  $L_{adv} \leftarrow -L_{dis}$ via Eq.5

  optimize three modules F, P, and D in turn:

$$\theta_F \leftarrow \theta_F - \lambda \nabla_{\theta_F}(\mathcal{L}_{im} + \mathcal{L}_{loss} + \mathcal{L}_{adv})$$

$$\theta_P \leftarrow \theta_P - \lambda \nabla_{\theta_P}(\mathcal{L}_{loss})$$

$$\theta_D \leftarrow \theta_D - \lambda \nabla_{\theta_D}(\mathcal{L}_{dis})$$

end

4 Experiments

4.1 Datasets and implementations

We perform experiments on Digits, Office-31, Office-Home, and VisDA. **Digits** mainly consists of two subsets, SVHN (S) [27] and MNIST (M) [28]. SVHN consists of 73,257 RGB images and MNIST consists of 60,000 grayscale images. **Office-31** [29] consists of 4,652 images with 31 categories collected from three different domains: Amazon (A), Webcam (W), and DSLR (D). **Office-Home** [30] consists of 15,588 images collected from four domains with 65 categories: Artistic (A), Clipart (C), Product (P), and Real-World (R). **VisDA** [31] (2017 Ver.) is a large-scale Sim-to-Real dataset consisting of 280,000 images with 12 categories. For a fair comparison with existing methods, we follow the official UDA protocol in all datasets and employ the same architecture for the feature extractor. The baseline models for active learning follow the setting of [7], and we cite the reported results of previous studies if the experimental protocol is the same as ours.

4.2 Comparison results

We extensively compare our method with previous ADA methods and various baseline models on four public datasets. As shown in Fig. 3, our method rapidly increases the performance with very few annotation budgets compared to BADGE and ADA-CLUE on Digits. AADA has high initial performance, but the growth rate is not steep compared to the proposed method. From Table 2, we can see that the proposed method outperforms various baseline models and the state-of-the-art method (*i.e.*, TQS) in Office-31 at 5% annotation budget. Especially, in $A \Rightarrow D$ and $A \Rightarrow W$ scenarios, our method achieves more than 95% accuracy only using 5% of annotation budget. Given 10% annotation budget in Office-31, the performance of our method is comparable to TQS as

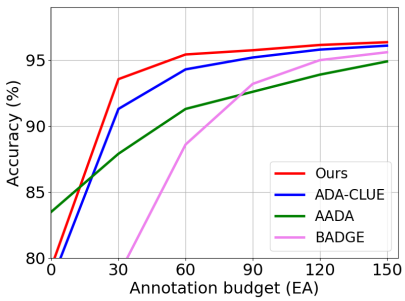


Figure 3: Comparison results on Digits (S $\Rightarrow$ M) with respect to annotation budget.

Table 2: Classification accuracy (%) on Office-31 with 5% annotation budget

Method	Office-31						
	A⇒D	A⇒W	D⇒A	D⇒W	W⇒A	W⇒D	Avg
ResNet (source only) [32]	81.5	75.0	63.1	95.2	65.7	99.4	80.0
RAN (Random Sampling)	87.1	84.1	75.5	98.1	75.8	99.6	86.7
UCN [17]	89.8	87.9	78.2	99.0	78.6	100.0	88.9
QBC [33]	89.7	87.3	77.1	98.6	78.1	99.6	88.4
Cluster [18]	88.1	86.0	76.2	98.3	77.4	99.6	87.6
Learning loss [15]	80.9	84.0	69.5	98.0	69.4	99.8	83.6
ADMA [8]	90.0	88.3	79.2	100.0	79.1	100.0	89.4
AADA [6]	89.2	87.3	78.2	99.5	78.7	100.0	88.8
ADA-CLUE [11]	92.0	87.3	79.0	99.2	79.6	99.8	89.5
S3VAAD [12]	93.0	93.7	75.9	99.4	78.2	100	90.0
TQS [7]	92.8	92.2	**80.6**	**100.0**	80.4	**100.0**	91.1
Ours	**96.6**	**96.8**	79.9	99.8	**81.7**	99.8	**92.2**

Table 3: Classification accuracy (%) on Office-Home and VisDA with 5% annotation budget

Method	VisDA	Office-Home												
		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 (source only) [32]	44.7	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
RAN (Random Sampling)	78.1	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
UCN [17]	81.3	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
QBC [33]	80.5	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
Cluster [18]	79.8	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
Learning loss [15]	85.8	58.2	74.2	77.4	62.6	72.8	73.4	62.1	56.6	79.6	70.7	55.1	75.9	68.1
ADMA [8]	81.4	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
AADA [6]	80.8	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
ADA-CLUE [11]	85.2	63.6	79.3	80.9	68.8	77.5	76.7	66.3	57.9	81.4	75.6	60.8	86.3	72.5
S3VAAD [12]	77.7	57.3	73.9	76.6	60.3	76.5	71.1	57.6	56.0	78.7	71.4	63.1	83.3	68.8
TQS [7]	83.1	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
Ours	**86.8**	**63.7**	**83.9**	**82.5**	**69.7**	**82.7**	**81.4**	**70.3**	**61.2**	**84.6**	**77.4**	**63.4**	85.9	**75.6**

shown in Table 4. Also, our proposed method shows state-of-the-art performance in Office-Home and VisDA with 5% and 10% annotation budgets significantly as shown in Table 3 and Table 5. Note that the performance of S3VAAD was partially reported on the various dataset and therefore not compared all in our experiment. The performance of the ADA-CLUE was not reported in 10% annotation budget due to the reproducibility issue.

4.3 Ablation study and analysis

To validate each module of the proposed method, we evaluate diverse experimental settings on Office-31 as shown in Fig. 4. Firstly, we set the experiment as loss-based active learning (AL) schemes, which start with Cold-Start (*i.e.*, random selection at the initial step) and train the model with only labeled data. The second ablation is corresponding to Cold-Start with S_1 and S_2 without utilizing pseudo label and IM loss training. By comparing Ablation1 and Ablation2, we found that sequential learning is effective for active domain adaptation. The result of Ablation3 shows the performance degradation when training the model only with selected high-loss samples. Ablation4 supports the effectiveness of using pseudo labels from the majority of unlabeled target samples. The last ablation shows that

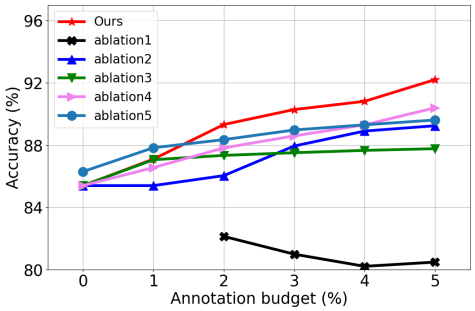


Figure 4: Ablation study on Office-31 with 5% annotation budget. The setting of each ablation set is presented in Table 6.

Table 4: Classification accuracy (%) on Office-31 with 10% annotation budget

Method	Office-31						
	A⇒D	A⇒W	D⇒A	D⇒W	W⇒A	W⇒D	Avg
ADMA [8]	94.0	93.4	84.4	100.0	84.6	100.0	92.7
AADA [6]	93.5	93.1	83.2	99.7	84.2	100.0	92.3
ADA-CLUE [6]	93.5	93.1	83.2	99.7	84.2	100.0	92.3
S3VAAD [12]	**98.0**	95.6	81.0	99.4	81.6	100.0	92.6
TQS [7]	96.4	96.4	**86.4**	**100.0**	**87.1**	100.0	**94.4**
Ours	97.8	**97.9**	85.0	99.8	85.3	**100.0**	94.3

Table 5: Classification accuracy (%) on Office-Home and VisDA with 10% annotation budget

Method	VisDA	Office-Home												
		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
ADMA [8]	84.8	66.5	85.4	82.8	63.8	80.9	76.3	67.7	61.6	80.9	74.3	66.8	89.7	74.7
AADA [6]	84.6	65.8	84.5	82.2	64.1	80.6	76.1	67.6	62.6	80.1	73.7	66.1	88.6	74.3
S3VAAD [12]	81.1	64.6	81.4	80.6	62.6	82.8	76.2	61.7	62.2	81.9	73.0	65.3	87.1	73.3
TQS [7]	87.2	68.0	87.7	85.7	67.0	83.0	78.7	69.3	64.5	83.9	77.8	68.9	**90.6**	77.1
Ours	**90.2**	**70.7**	**87.9**	**86.9**	**74.3**	**87.4**	**85.4**	**74.5**	**69.2**	**87.4**	**81.4**	**70.2**	90.4	**80.5**

the proposed sequential order has the best performance between S_1 to S_2 and S_2 to S_1 settings. Note that S_1 conducts domain adaptation and pseudo labeling process like algorithm 2 and S_2 instead conducts selected sampling training in Ablation4. Overall, we empirically demonstrate the effectiveness of our sequential learning and reliable pseudo labels by information loss.

4.4 Visualization of sample diversity

We claim that loss-based query selection cannot handle the sample diversity issues in domain shift scenarios. To support this, we visualize selected target samples in the target domain according to random selection and loss-based query selection. From Fig. 5, we can see that the samples selected by loss-based query selection are clustered in some regions rather than spread out in the target domain compared to the randomly selected samples. Although sample diversity is low when loss-based query selection is used alone, diversity can be increased by using the proposed model regularization, which leads to improved performance.

4.5 Parameter sensitivity analysis

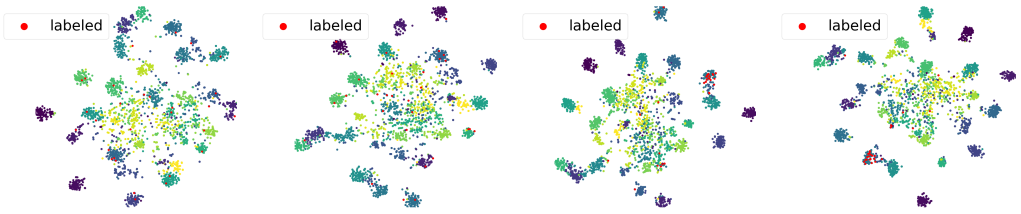
We present sensitivity analysis on hyperparameters including batch size, balance parameter ξ for information maximization loss, and intermediate dimension for the loss predictor. Fig. 6 implies that the batch size or intermediate dimension for the loss predictor is not sensitive, and low self-entropy is more important in information maximization loss. Furthermore, we analyze pseudo-label-update interval parameter γ. From Table 7, we can confirm that updating the pseudo label too often slows down the learning time, but does not improve the performance.

Table 6: Ablation study setting (RAN : random selection, L : loss-based selection, S_1 : selected sample training, S_2-L : sequential adaptation w/ [G.T label or pseudo label], S_2-IM : sequential adaptation w/ IM loss).

	Query selection	S_1	S_2-L	S_2-IM
Ablation1	RAN + L	X	G.T	X
Ablation2	RAN + L	O	G.T	X
Ablation3	L	O	G.T	X
Ablation4	L	X	P.L	O
Ablation5	L	swap-S_2	swap-S_1	X
Ours	L	O	O	O

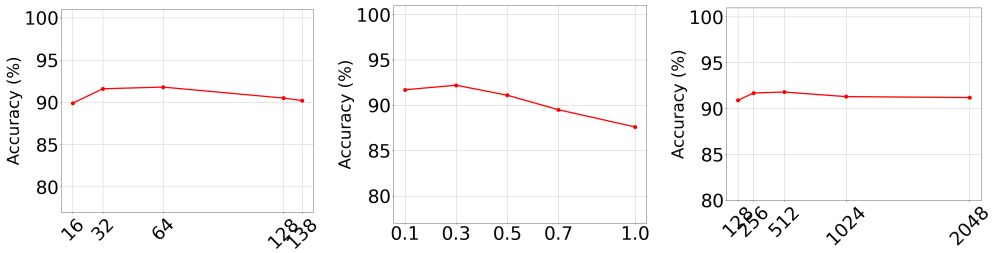
Table 7: Accuracy (%) change with respect to annotation budget $B\%$ and pseudo label update intervals γ in Office-31.

	$\gamma = 2$	$\gamma = 10$	$\gamma = 20$	$\gamma = 30$
$B = 0$	87.6	**88.6**	87.7	87.7
$B = 1$	91.0	91.8	**91.8**	91.0
$B = 2$	93.4	92.2	**94.8**	91.8
$B = 3$	94.6	93.6	**95.6**	93.2
$B = 4$	95.0	94.2	**95.6**	93.4
$B = 5$	95.0	95.0	**96.0**	94.2



(a) Random selection (b) Random selection (c) learning loss [15] (d) learning loss [15]

Figure 5: t-SNE visualization on Office-31 (D⇒A). (a) and (b) visualize the result of random query selection while (c) and (d) is the visualized results using plain loss-based query selection. The red dots represent labeled target samples by query selection, and the colors indicate each class.



(a) (b) (c)

Figure 6: Parameter sensitivity analysis on Office-31. (a) batch size, (b) balancing parameter ξ, (c) loss predictor's dimension.

4.6 Pseudo-label reliability

We actively use target pseudo labels as the ground truth label for loss prediction and adaptation. To increase the reliability of pseudo labels, we encourage pseudo labels to have low self-entropy and diverse class distributions. As shown in Fig. 7, our proposed method gradually improves the reliability of the pseudo label for the unlabeled target domain across all the adaptation scenarios in Office-Home. Surprisingly, half of the total 12 adaptation scenarios show an accuracy of over 80% when the annotation budget is 5%. The result of increasing pseudo-label accuracy as the annotation budget increases demonstrates the effectiveness of the proposed ADA framework including information loss with sample diversity.

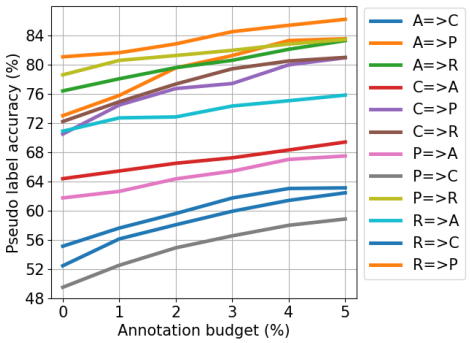


Figure 7: Pseudo labels accuracy (%) in Office-Home with 5% of annotation budget.

5 Conclusion

While previous active domain adaptation methods focus on sample selection criteria, *e.g.*, uncertainty, diversity, and committee, we exploit loss-based query selection and propose model regularization schemes. The main difference between the proposed method and individual-related work is the sequential learning scheme considering domain type (source/target) and labelness (labeled/unlabeled). We first train our main network including an auxiliary loss predictor with a small number of ground-truth labels by oracle. We then fully utilize numerous pseudo labels where the reliability is improved by information maximization loss. We extensively show the limitations of applying only loss-based query selection to active domain adaptation and extensively present analysis for our method. Our model achieves state-of-the-art performance in various active domain adaptation scenarios.

References

[1] Wang, M., Deng, W.: Deep visual domain adaptation: A survey. Neurocomputing **312** (2018) 135–153

[2] Zhao, S., Yue, X., Zhang, S., Li, B., Zhao, H., Wu, B., Krishna, R., Gonzalez, J.E., Sangiovanni-Vincentelli, A.L., Seshia, S.A., et al.: A review of single-source deep unsupervised visual domain adaptation. IEEE Trans. on Neural Networks and Learning Systems (TNNLS) (2020)

[3] Li, B., Wang, Y., Zhang, S., Li, D., Keutzer, K., Darrell, T., Zhao, H.: Learning invariant representations and risks for semi-supervised domain adaptation. In: Proc. of Computer Vision and Pattern Recognition (CVPR). (2021) 1104–1113

[4] Saito, K., Kim, D., Sclaroff, S., Darrell, T., Saenko, K.: Semi-supervised domain adaptation via minimax entropy. In: Proc. of Int'l Conf. on Computer Vision (ICCV). (2019) 8050–8058

[5] Jiang, P., Wu, A., Han, Y., Shao, Y., Qi, M., Li, B.: Bidirectional adversarial training for semi-supervised domain adaptation. In: Proc. of Int'l Joint Conf. on Artificial Intelligence. (2020) 934–940

[6] Su, J.C., Tsai, Y.H., Sohn, K., Liu, B., Maji, S., Chandraker, M.: Active adversarial domain adaptation. In: Proc. of Winter Conf. on Applications of Computer Vision (WACV). (2020) 739–748

[7] Fu, B., Cao, Z., Wang, J., Long, M.: Transferable query selection for active domain adaptation. In: Proc. of Computer Vision and Pattern Recognition (CVPR). (2021) 7272–7281

[8] Huang, S.J., Zhao, J.W., Liu, Z.Y.: Cost-effective training of deep cnns with active model adaptation. In: Proc. of Int'l Conf. on Knowledge Discovery and Data Mining. (KDD). (2018) 1580–1588

[9] Joshi, A.J., Porikli, F., Papanikolopoulos, N.: Multi-class active learning for image classification. In: Proc. of Computer Vision and Pattern Recognition (CVPR), IEEE (2009) 2372–2379

[10] Sener, O., Savarese, S.: Active learning for convolutional neural networks: A core-set approach. In: Proc. of Int'l Conf. on Learning Representation (ICLR). (2018)

[11] Prabhu, V., Chandrasekaran, A., Saenko, K., Hoffman, J.: Active domain adaptation via clustering uncertainty-weighted embeddings. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). (October 2021) 8505–8514

[12] Rangwani, H., Jain, A., Aithal, S.K., Babu, R.V.: S3vaada: Submodular subset selection for virtual adversarial active domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). (October 2021) 7516–7525

[13] Ma, X., Gao, J., Xu, C.: Active universal domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2021) 8968–8977

[14] Beluch, W.H., Genewein, T., Nürnberger, A., Köhler, J.M.: The power of ensembles for active learning in image classification. In: Proc. of Computer Vision and Pattern Recognition (CVPR). (2018) 9368–9377

[15] Yoo, D., Kweon, I.S.: Learning loss for active learning. In: Proc. of Computer Vision and Pattern Recognition (CVPR). (2019) 93–102

[16] Wang, K., Zhang, D., Li, Y., Zhang, R., Lin, L.: Cost-effective active learning for deep image classification. IEEE Trans. on Circuits and Systems for Video Technology. (TCSVT) **27**(12) (2016) 2591–2600

[17] Joshi, A.J., Porikli, F., Papanikolopoulos, N.P.: Scalable active learning for multiclass image classification. IEEE Trans. on Pattern Anal. Mach. Intell. (TPAMI) **34**(11) (2012) 2259–2273

[18] Nguyen, H.T., Smeulders, A.: Active learning using pre-clustering. In: Proc. of Int'l Conf. on Machine Learning (ICML). (2004) 79

[19] Gong, B., Shi, Y., Sha, F., Grauman, K.: Geodesic flow kernel for unsupervised domain adaptation. In: Proc. of Computer Vision and Pattern Recognition (CVPR), IEEE (2012) 2066–2073

[20] Chattopadhyay, R., Fan, W., Davidson, I., Panchanathan, S., Ye, J.: Joint transfer and batch-mode active learning. In: Proc. of Int'l Conf. on Machine Learning (ICML), PMLR (2013) 253–261

[21] Shu, R., Bui, H.H., Narui, H., Ermon, S.: A dirt-t approach to unsupervised domain adaptation. arXiv preprint arXiv:1802.08735 (2018)

[22] Rangwani, H., Jain, A., Aithal, S.K., Babu, R.V.: S3vaada: Submodular subset selection for virtual adversarial active domain adaptation. In: Proc. of Int'l Conf. on Computer Vision (ICCV). (2021) 7516–7525

[23] Liang, J., Hu, D., Feng, J.: Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In: Proc. of Int'l Conf. on Machine Learning (ICML), PMLR (2020) 6028–6039

[24] Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: Proc. of Computer Vision and Pattern Recognition (CVPR). (2017) 7167–7176

[25] Krause, A., Perona, P., Gomes, R.: Discriminative clustering by regularized information maximization. Proc. of Neural Information Processing Systems (NeurIPS) **23** (2010)

[26] Shi, Y., Sha, F.: Information-theoretical learning of discriminative clusters for unsupervised domain adaptation. Proc. of Int'l Conf. on Machine Learning (ICML) (2012)

[27] Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading digits in natural images with unsupervised feature learning. Proc. of Neural Information Processing Systems Workshops (NeurIPSW) (2011)

[28] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11) (1998) 2278–2324

[29] Saenko, K., Kulis, B., Fritz, M., Darrell, T.: Adapting visual category models to new domains. In: Proc. of European Conf. on Computer Vision (ECCV), Springer (2010) 213–226

[30] Venkateswara, H., Eusebio, J., Chakraborty, S., Panchanathan, S.: Deep hashing network for unsupervised domain adaptation. In: Proc. of Computer Vision and Pattern Recognition (CVPR). (2017) 5018–5027

[31] Peng, X., Usman, B., Kaushik, N., Hoffman, J., Wang, D., Saenko, K.: Visda: The visual domain adaptation challenge. arXiv preprint arXiv:1710.06924 (2017)

[32] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. of Computer Vision and Pattern Recognition (CVPR). (2016) 770–778

[33] Dagan, I., Engelson, S.P.: Committee-based sampling for training probabilistic classifiers. In: Machine Learning Proceedings 1995. Elsevier (1995) 150–157