Active Gradual Domain Adaptation: Dataset and Approach

Shiji Zhou*, Lianzhe Wang*, Shanghang Zhang, *Member, IEEE,* Zhi Wang, *Member, IEEE,* and Wenwu Zhu, *Fellow, IEEE*

***Abstract*—Adapting deep neural networks to the changing environments is critical in practical utility, especially for online web applications, where the data distribution changes gradually due to the evolving environments. For instance, the web photo of the cellphone changes gradually over the years due to the appearance change. In this paper, we deal with such a problem via active gradual domain adaptation, where the learner continually and actively selects the most informative labels from the target to enhance the label efficiency and utilizes both labeled and unlabeled samples to improve the model adaptation under gradual domain drift. We propose the active gradual self-training (AGST) algorithm with the novel designs of active pseudolabeling and gradual semi-supervised domain adaptation. Specifically, AGST pseudolabels the samples with high confidence, and selects the most informative labels from the unconfident samples based on both uncertainty and diversity, and then gradually self-trains itself by confident pseudolabels and active queried informative labels. To study the gradual domain shift problem in the web data and verify the proposed AGST algorithm, we create a new dataset – Evolving-Image-Search (EVIS), which is collected from the web search engine and covers the time range of 12 years. Since the appearance of the products evolves over these years, such dataset naturally contains gradual domain drift. We extensively evaluate AGST on the synthetic dataset, real-world dataset, and EVIS dataset. AGST achieves up to 62% accuracy improvement (absolute value) against unsupervised gradual self-training with only 5% additional labels, and 19% accuracy improvement against directly applying CLUE, which demonstrates the effectiveness of the designs of active pseudolabel and gradual semi-supervised domain adaptation.**

***Index Terms*—Gradual Domain Drift, Gradual Domain Adaptation, Active Domain Adaptation, Web Noise Data.**

I. INTRODUCTION

DEEP neural network works remarkably well in many real-world scenarios when the models are trained with large amounts of labeled data and tested on the same data distribution [11], [33], [2], [15]. However, when the application environment keeps changing, the trained model may fail to adapt to the gradually changing domains, leading to a serve performance decay [16]. This may be solved by collecting enough labeled training data to cover all the possible distributions that occur at test time. However, it often brings prohibitively high labeling costs. This especially happens in web application scenarios. For instance, the appearance of the communication devices in the web images vary over time, as shown in Figure 2. This kind of variation could lead to the performance drop of the deep learning models initially trained based on previous data. This effect is shown in Figure 3. In the meanwhile, because of the scale of web data, annotating all the samples is expensive and impractical. Consequently, it calls for machine learning systems that can adapt to the changing environment with only limited labels, which challenges both the adaptation ability and annotation efficiency under environmental change.

Even though there are active researches on domain adaptation, conventional domain adaptation methods are severely challenged by the gradual domain drift, i.e., gradually changing data distribution caused by the evolution of environments. Unsupervised domain adaptation (UDA) [7], [30], [35] aims to improve the generalization of a pre-trained model trained by a labeled source domain to a new and fixed unlabeled target domain (Figure 1 left bottom). However, UDA is insufficient to deal with gradual domain drift, as shown in previous work [16], where they further consider unsupervised gradual domain adaptation (UGDA) to address the problem of gradual domain drift (Figure 1 middle), however, suffer from exponential error growth due to the gradual domain changes. In addition, Saito et al. [20] have shown that UDA may be insufficient to bridge a severe domain drift entirely. Since the accumulated domain drift may be large if the time step is long, it is impossible to maintain a good performance in such severe domain drift without any additional labels. Therefore, querying additional labels is necessary for successfully adapting to a changing domain.

With the help of active learning (AL) [6], [36], [1] that improves the label efficiency, active domain adaptation (ADA) [18], [23] approaches can further enhance the adaptation to a fixed domain by active querying additional informative labels (Figure 1 right bottom). However, since ADA can only adapt from the source domain to a fixed target domain, directly applying ADA to the gradual domain drift problem often leads to bad performances. In addition, it also is difficult to embed ADA in gradual domain adaptation, since

Equal contributions: Shiji Zhou is with Tsinghua-Berkeley Shenzhen Institute, Shenzhen, China. E-mail: zsj17@mails.tsinghua.edu.cn. Lianzhe Wang is with Tsinghua-Berkeley Shenzhen Institute, Shenzhen, China. E-mail: wanglz20@mails.tsinghua.edu.cn

Corresponding authors: Shanghang Zhang is with UC Berkeley, California, USA. E-mail: shz@eecs.berkeley.edu. Zhi Wang is with Tsinghua Shenzhen International Graduate School, Shenzhen, China. E-mail: wangzhi@sz.tsinghua.edu.cn. Wenwu Zhu is with Tsinghua University, Beijing, China. E-mail: wwzhu@tsinghua.edu.cn

This work is supported by the National Key Research and Development Program of China No. 2020AAA0106300, National Natural Science Foundation of China No. 62050110 and No. 61872215, and the Shenzhen Science and Technology Program of Grant No. RCYX20200714114523079. We would like to thank Kuaishou for sponsoring the research.

EVIS dataset is released at https://github.com/LianzheWang/EVIS

Code is available at https://github.com/LianzheWang/Active-Gradual-Domain-Adaptation-Dataset-and-Approach

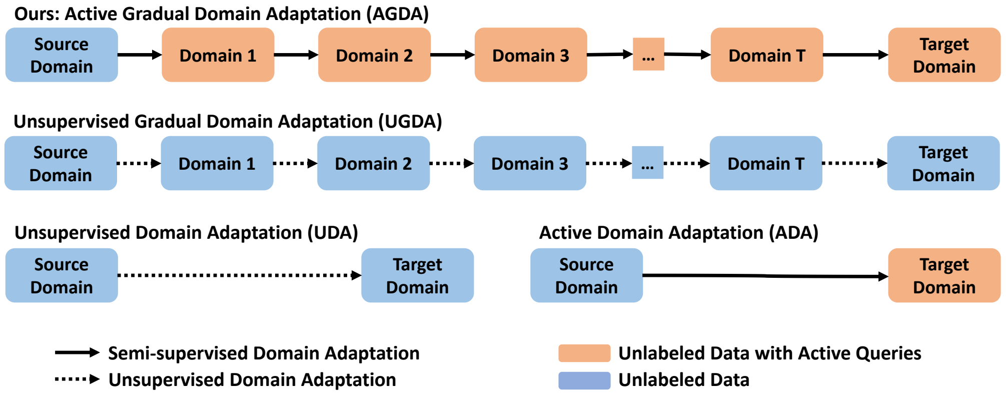

Ours: Active Gradual Domain Adaptation (AGDA)
Source Domain
Domain 1
Domain 2
Domain 3
...
Domain T
Target Domain
Unsupervised Gradual Domain Adaptation (UGDA)
Source Domain
Domain 1
Domain 2
Domain 3
...
Domain T
Target Domain
Unsupervised Domain Adaptation (UDA)
Source Domain
Target Domain
Active Domain Adaptation (ADA)
Source Domain
Target Domain
Semi-supervised Domain Adaptation
Unsupervised Domain Adaptation
Unlabeled Data with Active Queries
Unlabeled Data

Fig. 1: Comparison with related works. our proposed AGDA utilize both active queries and intermediate data to enhance the performance under gradual domain drift. In contrast, UGDA only uses intermediate data, ADA only uses active queries, and UDA does not consider both of them.

ADA needs to do a semi-supervised domain adaptation after having the labels, and the semi-supervised loss depends on the consistency between the source and target domain, which is impossible since if the time step is long, the accumulated bias caused by gradual domain drift would be too large.

To further improve the adaptation efficiency of gradual domain adaptation, it is natural to study the problem of active gradual domain adaptation, where the model adapts to a gradually changing domain with only limited labels. To the best of our knowledge, no previous work has considered active domain adaptation under gradual domain drift. We define this as the Active Gradual Domain Adaptation problem (Figure 1 upper) Such problem challenges the designs of both the active query strategy, and gradual semi-supervised domain with small batches of data. To solve this, in this paper, we propose the Active Gradual Self-Training (AGST) algorithm. In each time t, AGST first pseudolabels the instances with high confidence. Then, we design a querying strategy to actively select the informative labels from the unconfident instances based on both uncertainty and diversity, where we define the uncertainty by entropy and confidence, and achieve the diversity by cluster-based active strategy [36]. After that, AGST runs semi-supervised iterations by confident pseudolabels, active queries, and data features. To eschew the chaos of noise under small batches, we add normalization to constrain the adaptation from the last model.

In our experiment, we first introduce the construction of our new web dataset – Evolving-Image-Search (EVIS) dataset, and show that natural gradual domain drift and environment noise exist. We then verify the proposed algorithm and baselines on three datasets: Rotating MNIST is a synthetic dataset without environment noise; Portraits is a real-world data labeled by human without environment noise; EVIS is a real-world data automatically collected from the web with environmental noise. The experiment results show that AGST significantly outperforms UGDA, where AGST gains over 60% accuracy increase than UGDA with only 5% active labels in Rotating MNIST, over 15% in portraits with only 2% labels in Portraits, and over 40% with 10% labels in EVIS. In contrast, UGDA only achieves a very marginal accuracy improvement than the source model, and even much worse in the EVIS dataset with environmental noise. As compared with direct applying active DA – CLUE [18], AGST achieves 19%, 15%, and 7% improvement in these three datasets, respectively. The ablation study verifies that our designs of active pseudolabel and gradual semi-supervised domain adaptation are effective.

Our contributions. Our contributions are summarized as following:

- We collect a new Evolving-Image-Search (EVIS) dataset. EVIS consists of real-world web images that appeared in Google image search results, with individually recorded searching keyword labels and uploading times. The collection process of EVIS is purely done automatically without artificial selection and annotation. We further show that this dataset has a natural gradual domain drift caused by the evolution of web images with time, and noise caused by the search randomness, which makes the trained model fail gradually.
- We formulate an active gradual domain adaptation problem, where the models need to adapt to an evolving domain with only limited labels.
- To address the challenge of limited labels, we propose an active pseudolabel strategy. It pseudolabels the confident instances and make active query from the unconfident ones by both diversity and uncertainty, where the uncertainty is defined by combining the confidence and the entropy.
- To deal with small batch and noise data, we design a gradual semi-supervised domain adaptation iteration, which regularizes the adaptation step for not forgetting the last model.
- We conduct experiments on a synthetic dataset, a real-world dataset, and a web dataset. The experiment results show the advantage of both the designs of active pseudolabeling and gradual semi-supervised domain adaptation.

The rest of the paper is organized as follows. We survey related work in Section II. We provide the system model

Fig. 2: Illustration of the prototype change for the electronic phone. We collect the figures from the google search engine from the year of 2009 to 2020. As shown, the prototype is different in each year, which makes the previous classification model fail.

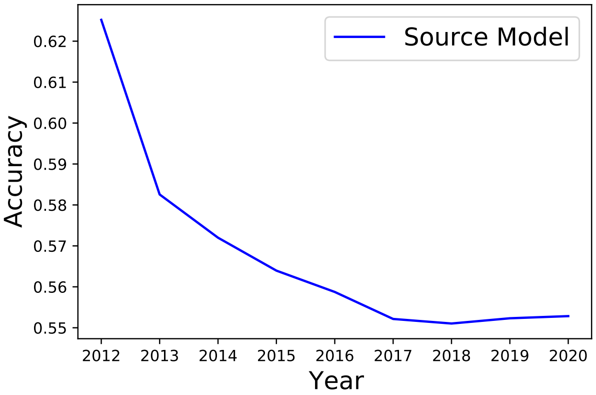


Fig. 3: Illustration of performance decay of deep learning model caused by the web images' evolution. Here we train a ResNet18 model on an initial source dataset to classify web images. The source dataset consists of Google searching result images for different electronic devices and vehicles categories from 2009 to 2011. We then test the model for the same classification task on searching result images from 2012 to 2020. This figure shows the accuracy curve w.o.t the time. The figure shows that the source model suffers from a continuous performance drop with time. It shows that the web data is undergoing gradual domain drift.

and the proposed method in Section III. We provide the construction of the new web dataset and evaluate our design in Section V. We conclude the paper in Section VI.

II. Related Works

The topic of this paper sits well in the following four bodies of literature: unsupervised domain adaptation, gradual domain adaptation, active domain adaptation, and domain adaptation dataset. Our results contribute to all these areas and hopefully will inspire more interplay among the related communities.

Unsupervised domain adaptation (UDA). Unsupervised domain adaptation is a typical method to enhance the generality of the model trained with source data, by utilizing the unlabeled samples from the target domain [35]. The key challenge for domain adaptation is that the source and target domains may be very different[22], [34], which is typical in the modern high-dimensional regime. Importance weighting-based methods [21], [14], [24] assume the domains are close, with theoretical guarantees depending on the expected density ratios between the source and target. However, in practice, even if the domains are similar, the density ratio often scales exponentially in the dimension. These methods perform poorly in high-dimensional scenarios. These methods assume that $P(Y|X)$ is the same for the source and target, while we study the case with a continually changing domain. Recent proposed methods aim to learn domain invariant representations [27], [9], [26]. However, these methods require several additional heuristics [13], and are shown to fail to deal with gradual domain drift [16]

Gradual Domain Adaptation (Gradual DA). Essentially gradual DA assumes the domain shifts gradually over time and tries to continually adapt the source domain to multiple target domains at each time. Different methods are proposed to address this challenge, such as adversarial loss [3], generative adversarial networks [31], linear transform [4], optimal transport [17], and indexed domain adaptation [29]. However, these approaches need to learn from the whole data from beginning to end. Because of the scale of real-world application, it is impractical for a machine learning system to remember all the data in history. For instance, the web application receives a large scale of data in each data, it is impossible to collect and train all the data together. Kumar et al. [16] provides a theoretical guarantee for unsupervised gradual self-training under gradual domain drift, which uses the last model to pseudolabel the current instances and then self adapts itself by these pseudolables. However, unsupervised DA (UDA) may suffer from severe performance drops in gradually changing domain without additional labels [20], since the accumulated drift could be too large to efficiently apply UDA. In our experiment, we show that gradual unsupervised self-training performs nearly the same as the initial model when the time step is long (c.f. Section V).

Active Domain Adaptation (Active DA). Active DA is first proposed by Rai et al. [19] with application to sentiment classification from text data, where they embed an online uncertainty based sample strategy in domain adaptation. Chattopadhyay et al. [5] propose a method that performs transfer and active learning simultaneously by solving a single convex optimization problem. Recently, active adversarial domain adaptation (AADA) [23] is proposed to solve the active DA problem in the context of deep learning, where AADA selects samples based on the uncertainty measured by entropy and targetness measured by the domain discriminator. Prabhu et al. [18] propose ADA-CLUE that queries labels based on uncertainty and diversity, then adopts a semi-supervised DA to transfer the domain knowledge to the target. However, previous works depend on the consistency between the source domain and target domain, which is impossible under gradual domain drift, leading to the ineffectiveness of directly applying to continual adaptation.

Domain Adaptation Dataset. There are several commonly used real-world datasets in the recent representative works of DA [26] [25] [32]. The Office-Home data [28] set is the most commonly used, which consists of images of objects in different office/home scenarios. Some works also run DA experiments on datasets transformed from common datasets like ImageNet and Cifar. However, non of the above datasets has a gradual changing property. In terms of research works about gradual DA, synthetic datasets with gradual domain drift like rotating-Mnist [29] and rotating-Gaussian [16] are often adopted. Few real-world datasets are currently being commonly used, one of which being representative is Portraits [10]. There is currently no web-image-based dataset designed for gradual DA.

To sum up, no previous work have considered adopting active learning that queries limited labels to further enhance the effectiveness, this work takes the first step. That is requiring far fewer labels than full supervision, e.g., 2% and 5% in our experiments (Section V), while maintaining a good performance. In this paper, we address this problem by designing a novel algorithm AGST with active pseudolabeling and gradual semi-supervised learning. Compared to gradual DA, we allow the adaptation model to queried additional labels to eschew the ineffectiveness of UDA, and propose an efficient sample strategy to enhance the label efficiency. Compared to active DA, we study the problem under gradual domain drift, and design a confidence-based pseudolabeling and a gradual semi-supervised DA suitable for such scenarios. Also, by making the EVIS dataset, we become the first to propose a web image-based dataset for gradual domain adaptation.

III. Method

We address the problem of active gradual domain adaptation (AGDA), where the goal is to continually adapt a model trained on a source domain to a gradually changing target domain, with the option to query a budget of labels from the target domain. In this section, we present a novel algorithm – Active Gradual Self-Training (AGST Algorithm 1) for AGDA, as shown in Figure 4, which performs consistently well under gradual domain drift. We will first introduce the system model and present the two novel designs: active pseudolabel and gradual semi-supervised domain adaptation.

A. System Model

In AGDA, the learning algorithm has access to a set of labeled data from the source domain $(X_{\mathcal{S}}, Y_{\mathcal{S}})$, and unlabeled data from the target domain $X_{\mathcal{T}}^t$ at time t, where the target domain evolves with the time. In each time t, the leaner is allowed to query labels from the target with a budget B, which is small related to the amount of unlabeled data. The active queries are denoted as $(X_{\mathcal{LT}}^t, Y_{\mathcal{LT}}^t) \subset (X_{\mathcal{T}}^t, Y_{\mathcal{T}}^t)$, where $Y_{\mathcal{T}}^t$ is the target labels in hindsight. In time t, the task is to gradually adapt the previous neural network $f_{t-1} \circ \phi_{t-1} : X \to Y$ to the changing target domain, and get a better model $f_t \circ \phi_t$ with a good performance, where $\phi_t : X \to Z$ is the feature extractor and $f_t : Z \to Y$ is the classifier. Denote the instances $x_{\mathcal{S}} \in X_{\mathcal{S}}, x_{\mathcal{T}} \in X_{\mathcal{T}}^t$, and labels $y_{\mathcal{S}} \in Y_{\mathcal{S}}, y_{\mathcal{T}} \in Y_{\mathcal{T}}^t$ with categorical variables $y \in \{1, 2, \ldots, C\}$. Denote the whole time horizon is T. Since the target and intermediate domains are the training data, we actually know the time horizon.

Algorithm 1 Active Gradual Self-Training (AGST)

```
Input: Confidence threshold value $\alpha$, pseudolabeled data loss weight $\lambda_{\mathcal{PT}}$, active queried data loss weight $\lambda_{\mathcal{LT}}$, regularization weight $\lambda_{\mathcal{R}}$, and entropy weight $\lambda_{\mathcal{H}}$.
Initial: Learn from source data $X_{\mathcal{S}}, Y_{\mathcal{S}}$, and get initial feature extractor $\phi_0$ and classifier $f_0$.
for $t = 1, \ldots, T$ do
    Received unlabeled data $X_{\mathcal{T}}^t$.
    Active Pseudolabel:
    Compute the confidence for each instance $\rho(x; f, \phi)(Eq.1)$ in $X_{\mathcal{T}}^t$.
    Give the pseudolabels for instanstance with high confidence $\rho(x) > \alpha$ by the initial model, and get pseudolabeled data $X_{\mathcal{PT}}^t, Y_{\mathcal{PT}}^t$.
    Active Query:
    Compute the uncertainty $\mathcal{S}(x; f, \phi)$ (Eq. 2) in $X_{\mathcal{T}}^t \backslash X_{\mathcal{PT}}^t$.
    Run weighted KMeans++ for $X_{\mathcal{T}}^t \setminus X_{\mathcal{PT}}^t$ with weight of $\mathcal{S}(x; f, \phi)$, and get the centroids.
    Query labels for the nearest neighbors to centroids, and get active queried data $X_{\mathcal{LT}}^t, Y_{\mathcal{LT}}^t$.
    Gradual Semi-supervised Domain Adaptation:
    Joint update the feature extractor $\phi$ and classifier $f$ by
        $f_t = \arg\min_{f} L_t(f, \phi) + R_t^f(f) - \lambda_{\mathcal{H}} \sum_{x \in X_T^t} \mathcal{H}(x; f, \phi),$
        $\phi_t = \arg\min_{\phi} L_t(f, \phi) + R_t^{\phi}(\phi) + \lambda_{\mathcal{H}} \sum_{x \in X_T^t} \mathcal{H}(x; f, \phi).$
    Return: Prediction $f_t(\phi_t(X_{\mathcal{T}}^t))$.
end for
```

TABLE I: Notations

t	time slot
T	time horizon
$x_{\mathcal{S}} \in X_{\mathcal{S}}$	data sample from the source domain
$y_{\mathcal{S}} \in Y_{\mathcal{S}}$	data label from the source domain
$x_{\mathcal{T}}^t \in X_{\mathcal{T}}^t$	data sample from the target domain in time t
$Y_{\mathcal{T}}^t$	data label from the target domain in time t
$X_{\mathcal{PT}}^t$	data sample of the pseudolabeled data in time t
$Y_{\mathcal{PT}}^t$	pseudolabel of the pseudolabeled data in time t
$X_{\mathcal{LT}}^t$	data sample of the active queries in time t
$Y_{\mathcal{LT}}^t$	data label of the active queries in time t
ϕ_t	feature extractor of the deep learning model in time t
f_t	classifier of the deep learning model in time t
C	number of classes
B	query budget

B. Active Pseudolabel

The goal of the gradual active query is to identify the most informative samples from the target domain. To this end, we design a novel query algorithm, which pseudolabels the instances with high confidence and active query from the unconfident ones with high uncertainty and diversity.

Confidence-based pseudolabeling. Self-training has been proposed to solve the domain adaptation problem. It first pseudolabels the target data using the initial model trained with source data and then retrain the model with such labeled data. However, a considerable part of instances from the target data are unconfident with the initial model, if the domain drift is

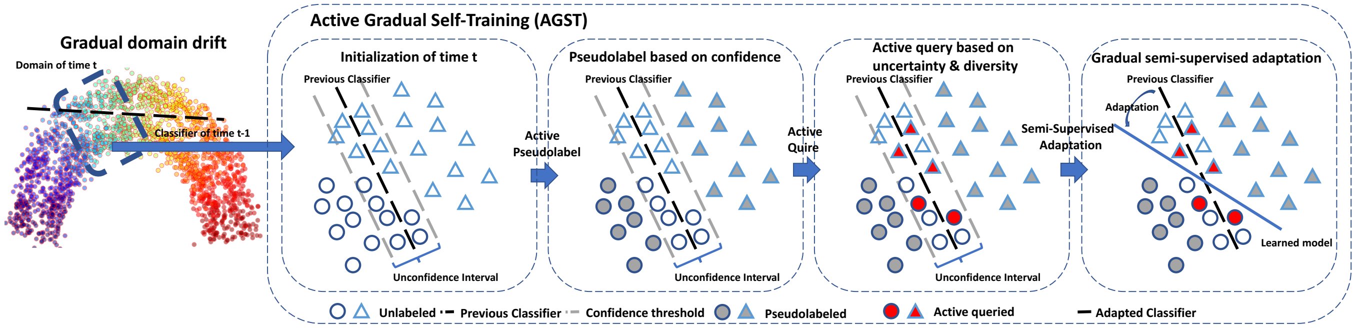


Fig. 4: Illustration of the AGST algorithm. The gradual domain drift is shown in the left figure, where the color represents the domain time index. Our method is presented in the right framework, where the triangle and circle represent data points of two different classes. In each time t, the adaptor receives unlabeled data and calculates the unconfident interval based on the initial classifier. It then pseudolabels the confident instances and active queries from unconfident instances by uncertainty and diversity. Finally, it adapts itself from the initial classifier by confident pseudolabels, active queries, and data features.

severe. Therefore, pseudolabeling unconfident instances would bring significant label noise, leading to a severe performance drop. A more wise choice is to pseudolabel only the confident instances, which is mostly correct, and then leave the rest data to the active agent to query the most informative ones. Specifically, we measure the confidence of pseudolable by

$$\rho(x; f, \phi) = \max_c p(Y = c \mid x; f, \phi), \tag{1}$$

where $p(Y = c|x; f, \phi)$ is the soft prediction, i.e. the softmax layer output of class c, and pseudolabel instances with $\rho(x; f, \phi) > \alpha$, where α is the thresholding value. We denote this pseudolabeled data as $(X^t_{\mathcal{PT}}, Y^t_{\mathcal{PT}})$. After that, we collect a set of labeled data with confident pseudolabels.

Uncertainty and diversity active querying. Unsupervised DA is insufficient to completely bridge a servere domain drift [18]. Since the gradual domain drift is often severe when the time is long [16], unsupervised domain adaptation may perform poorly in the end, leading to the necessary of additional labels. In this paper, we design a querying strategy that actively selects the labels based on both uncertainty and diversity to enhance the label efficiency. The uncertainty is measured by both entropy and confidence, defined as

$$\mathcal{S}(x; f, \phi) = 1 - \rho(x; f, \phi) + \mathcal{H}(x; f, \phi), \tag{2}$$

where the predictive entropy is defined as

$$\mathcal{H}(x; f, \phi) = -\sum_{c=1}^{C} p(Y = c|x; f, \phi) \log p(Y = c|x; f, \phi).$$

Such uncertainty is better than solely using confidence or entropy for multiclass classification, because entropy suffers low discriminability for highly uncertain and extremely sharp predictions, and confidence only consider the class with highest prediction value while ignoring other information [8]. For instance, if $p(Y|x) = (1/2, 1/2, 0, \ldots, 0)$, the entropy is $\log(2)$, which is tiny as compared with the max value $\log(C)$ if C is large; And if $p(Y|x) = (1/2, 1/4, 1/4, 0, \ldots, 0)$, the confidence is the same as the previous one, which ignore the rest changes.

In addition, querying labels only based on uncertainty leads to budget waste since similar instances share similar uncertainty, leading to the repetition of similar instances. We further consider the diversity of queried instances based on the feature space distribution of uncertainty samples. Specifically, we use uncertainty-weighted KMeans to create B/T clusters and query labels for nearest neighbors to the cluster centroids. Recall that B is the whole budget for how many labels the algorithm can query, then the budget is B/T for each training round. Since we query the label for each cluster, the number of clusters equals the number of queries. The intuition behind this is that the uncertainty-weighted KMeans optimizes the following objective

$$\min_{\mu_1, \ldots, \mu_B} \sum_{k=1}^{B} \mathcal{S}(x; f, \phi) \|\phi(x) - \mu_k\|,$$

where the center μ_k tends to approach the point with high uncertainty $\mathcal{S}(x; f, \phi)$. Since the centroids are naturally diverse, the nearest neighbors are diverse and uncertain.

C. Gradual Semi-supervised Domain Adaptation

We next introduce our design for gradual semi-supervised domain adaptation (SSDA) after having the confident pseudolabels and the queried labels.

Active-supervised objective. Our supervised objective loss consists of two parts: loss of pseudolabeled data, loss of active queried data, as following

$$\begin{aligned} L_t(f, \phi) = {} & \lambda_{\mathcal{PT}} \sum_{(x,y) \in (X_{\mathcal{PT}}, Y_{\mathcal{PT}})} l_{ce}(f_t \circ \phi_t(x), y) \\ & + \lambda_{\mathcal{LT}} \frac{|X_{\mathcal{T}}| - |X_{\mathcal{PT}}|}{|X_{\mathcal{LT}}|} \sum_{(x,y) \in (X_{\mathcal{LT}}, Y_{\mathcal{LT}})} l_{ce}(f_t \circ \phi_t(x), y), \end{aligned}$$

where l_{ce} denotes the cross-entropy, and $\lambda_{\mathcal{PT}}, \lambda_{\mathcal{LT}}$ are scalar weights. We here use different parameter, because the importance of pseudolabeled instances and active queried instances are different. In addition, we add weight $(|X_{\mathcal{T}}| - |X_{\mathcal{PT}}|)/|X_{\mathcal{LT}}|$, since the active queries represent all the unlabeled data $X_{\mathcal{T}} \setminus X_{\mathcal{PT}}$.

Gradual regularization. Since the batch size may be small each time, it is too aggressive to "forget" the previous model and retrain a new one from only limited samples, which could lead to large performance decay in the noise setting. We thus regularize the update of the model by adding the following regularization

$$R_t^f(f) = \lambda_{\mathcal{R}} \|f - f_{t-1}\|_1, R_t^\phi(\phi) = \lambda_{\mathcal{R}} \|\phi - \phi_{t-1}\|_1,$$

where $\lambda_{\mathcal{R}}$ is the regularization weight. Since the domain is gradually changing, the domain bias is small between each consequent time. Within this regularization, the model tends to find the solution near the previous one, i.e., gradually updating the model, aligned with the gradual domain drift. We here use the l_1 norm to enforce the model to sparsely update, since the gradual drift in feature space is often sparse, e.g., in most of the years, the evaluation of phone focuses its sub-modular such as screens or frames, leading to the sparse update of the model.

Minimax entropy. The minimax entropy (MME) [20] is a typical method to enhance the domain alignment for semi-supervised domain adaptation. In MME, the classifier f_t tends to maximize the entropy to increase the model discriminability, and the feature extractor ϕ_t tends to minimize the entropy for increasing the representation ability. MME gives us an approach to utilize the rest of unlabeled data, i.e. instances that are not pseudolabeled or active queried, to enhance the model discriminability and the representation ability.

Gradual SSDA iteration. Based on the previous designs, we are now ready to propose our Gradual SSDA iteration, as following

$$f_t = \arg\min_f L_t(f, \phi) + R_t^f(f) - \lambda_{\mathcal{H}} \sum_{x \in X_T^t} \mathcal{H}(x; f, \phi),$$
$$\phi_t = \arg\min_\phi L_t(f, \phi) + R_t^\phi(\phi) + \lambda_{\mathcal{H}} \sum_{x \in X_T^t} \mathcal{H}(x; f, \phi),$$

where $\lambda_{\mathcal{H}}$ denotes the weight of the entropy. Here we use an alternative iteration for the classifier f_t and the feature extractor ϕ_t by the framework of MME. By gradual SSDA iteration, the model is trained with unlabeled, active labeled, and pseudolabeled data. This optimization finds the optimal solution that has good discriminability and representation ability with gradual update.

IV. Dataset

Started a new section for dataset introduction here. Added two subsection titles. To evaluate the model performance on web applications with gradual domain drift, we use an automatic approach to construct a new Evolving-Image-Search (EVIS) dataset. In this section, we introduce the construction of the new web dataset and analyze its properties.

A. Construction of Evolving-Image-Search (EVIS) Dataset

There are massive image data resources on the Internet. In different years, the web images are undergoing a gradual domain shift, e.g., the “mobile phone” related images on the Internet have undergone drastic and continuous changes in recent years. Therefore, web images have great potential to be adopted in gradual domain shift adaptation-related research work. However, there is currently no existing automatic approach to construct a web image dataset for gradual domain shift learning researches. The construction work of the EVIS dataset could fill this gap to some extent.

EVIS consists of real-world web images that appeared in Google image search results, with individually recorded searching keyword labels and uploading times. The collection process of EVIS is purely done automatically without any manual selection or annotation.

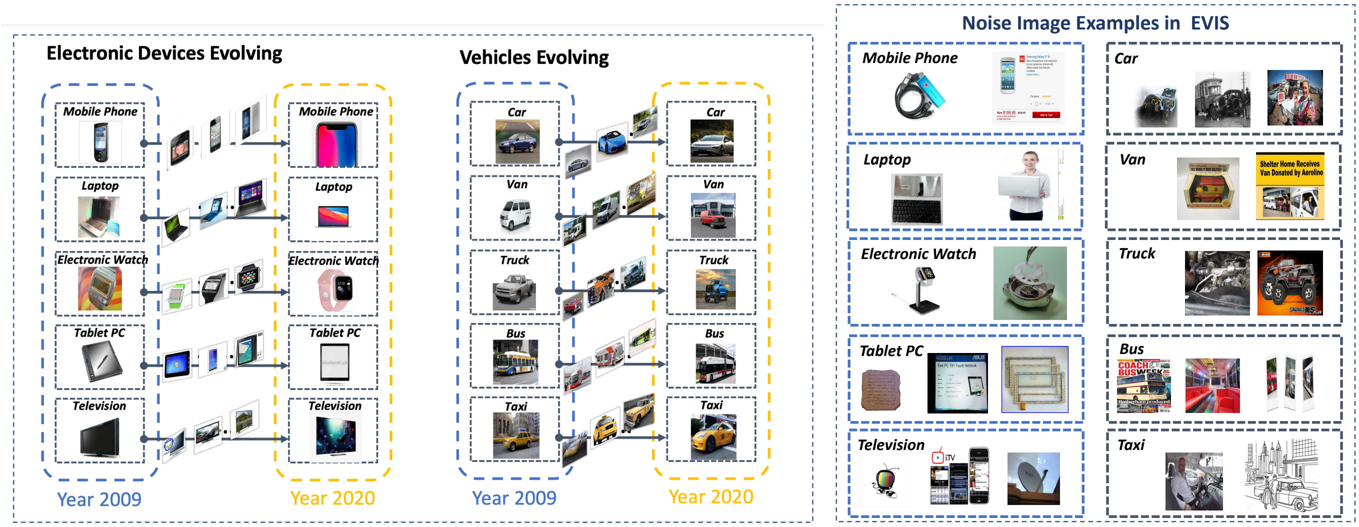
We select 10 objects with strong changes in this era (5 types of electronic products: mobile phone, laptop, tablet PC, television, electronic watch. 5 types of vehicles: car, van, truck, bus, taxi), use their names as keywords to search and collect on the Google image search engine. In particular, we will restrict the upload time range of the images in the search results through the API, and the length of each search interval is set to one month. For each month, we perform one set of searches, and each set of searches includes one search for each of the above keywords. To reduce the overlap and similarity of results in different searches, we prefix the search keyword with the word “new”. We also filter out search results that are too small (less than 200×200 pixels) or too large (greater than 1500×1500 pixels). The time range for the data collection is from 2009 to 2020, as shown in Figure 5a.

A crawler program is developed by us to automatically complete the above searching process and download the first 40 downloadable images for each keyword searching result. The search keywords are recorded as labels each time. All downloaded images are resized to 256×256 pixels and saved in JPG format. In this way, the EVIS dataset has a total of 12×12×10×40 (years, months, categories, downloads per search) = 57600 pictures, as shown in Table II.

B. Properties of EVIS Dataset

There could be a certain amount of noise in the dataset purely collected from the search engine, such as unrelated images that appear in the search results, as shown in Figure 5b. We deliberately retain this part of the noise to simulate the noise environment that the model needs to deal with in real application scenarios and retain our dataset constructing method’s fully automatic collecting property. We tested the dataset and found that deep learning model properly trained on EVIS could reach over 80% accuracy on testset while only learning limited samples. It’s a decent result for such a classification task for 10 categories, showing that the EVIS dataset has the quality of deep model learning research.

Through experiments, we are able to prove the gradual domain shift characteristics of the data in the EVIS. We first define the data in EVIS from 2009 to 2011 as source data and train a source model on it. We select ResNet18 as the model for the experiment and train it on EVIS data from 2009 to 2011. The training data consists of 7200 images. (Random selected 20 images per category per month are used for training. The total amount of training samples is 3(year)×12(month)×10(category)×20 = 7200.) After training the source model, we test the model on the EVIS data from 2012 to 2020 and evaluate the performance by each year. We found that the prediction accuracy on the test data shows a smooth downward trend year by year, as shown in Figure 3.



(a) Illustration of gradual changes

(b) Illustration of label noise

Fig. 5: Illustration of Evolving Image Search (EVIS) dataset. (a) For each category, the dataset covers the records of the Google image search results taking the category name as a keyword from 2009 to 2020. The images in the dataset show significant uniform gradual evolving characteristics with respect to time, i.e., with gradual domain drift property. (b) There is a certain amount of noisy images inside the EVIS dataset. Possible origins of the noises include irrelevant results generated by search engines and unusual forms of the searched objects in the images.

The results of this experiment help to show that the data domain distribution in the EVIS dataset gradually shifts over time.

Furthermore, we study the degree of evolution over the last decade for the 10 selected objects. We first again use the source model trained on data before 2011 to predict data from 2012 to 2020. Then, for each year, we record the per-category accuracy. Finally, for each category, we make a linear regression for the accuracy record from 2012 to 2020, take the regression slope value as the approximation of the accuracy decreasing rate (the average amount of accuracy decrease per year for each specific category). The result is shown in Figure 6.

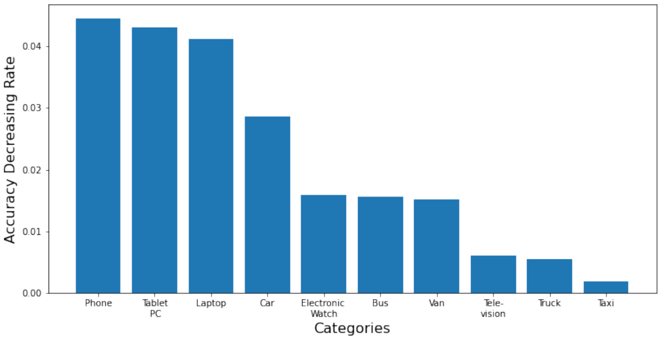


Fig. 6: Result of the measurement of the degree of evolving over the last decade for the 10 selected objects. Accuracy decreasing rate is the approximated average amount of accuracy decrease per year for each category (this approximation is obtained by linear regression). Here we could observe that the object with the highest degree of evolution is the phone, followed by tablet computers and laptops, all of which have a decline rate larger than 0.04. The car also has a relatively high degree of evolution and is the vehicle whose decline rate is closest to the above three electronic products. The three least evolved objects are television, truck, and taxi. Taxi is holding a minimal decreasing rate which is close to 0.

TABLE II: Statistic of the web data.

Years	Image Size	Data Size	Categories
12	256×256	57,600	10

V. Experiment

In this section, we evaluate the performance of the proposed method.

A. Experiment Setup

1) dataset: We evaluate the proposed method with three datasets: a synthetic dataset, a real-world dataset, and a new web dataset collected from a web search engine.

- **Rotating MNIST.** We randomly select and shuffle 35,000 images from the original MNIST dataset and use the first 2,000 images with no Rotating as the source dataset. The next 26,000 images are gradually rotated from 0° to 90° counterclockwise each time to be the target dataset with gradual domain drift. We set the time interval size by 2,000. Then the angle rotates $90°/13$ each time.
- **Portraits.** It is a realistic dataset, which contains 37,921 photos of high school seniors labeled by gender across a century. As shown in previous works [10], [16], since

the sex ratio and dress-up are evolving with years, this real dataset suffers from a natural gradual domain shift, including covariate shift and label shift. We downsample all the images to 32x32 pixels and do no other preprocessing. We take the first 2000 images as the source domain to learn the initial model. We use the next 30000 images as target data with a gradually changing domain.

- **EVIS.** EVIS is our newly constructed real-world dataset, consisting of images from web searching results. The images are annotated with uploading time, uniformly distributed from the year 2009 to the year 2020. All the images are downsampled to 64×64 pixels in the experiment, normalized by overall mean and variance. We take the images from 2009 to 2011 as the source domain to train the initial model. The images from 20012 to 2019 are taken as target data with a gradually changing domain.

2) Baselines: As we are the first to study the active gradual domain adaptation problem, most of the existing domain adaptation methods are not suitable to compare in such a setting. We therefore compare with the following baselines to verify the effectiveness of our design.

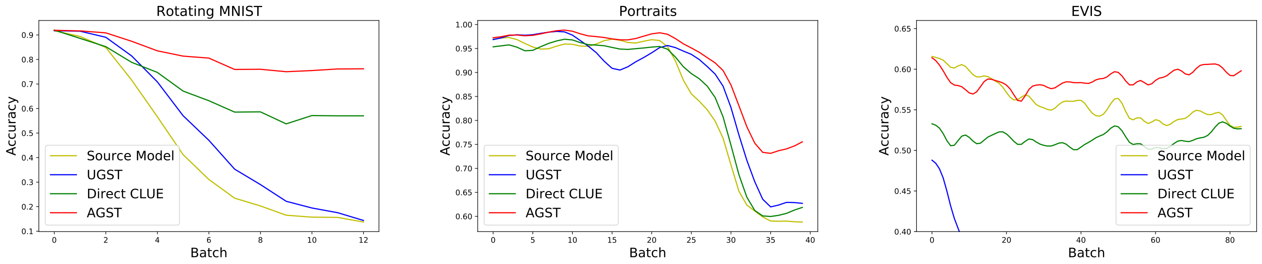
- **Source model.** We compare with the source model, i.e. the model trained by the source data, to verify the necessity of domain adaptation in the changing environments.
- **Unsupervised Gradual Self-Training.** We compare with the unsupervised gradual self-training (UGST) [16], to verify the effectiveness of the proposed method.
- **Direct CLUE.** We compare with the direct CLUE, i.e. directly applying Clustering Uncertainty-weighted Embeddings (CLUE) from the source to the target without gradual domain adaptation on the intermediate domain.
- **AGST w/o pseudolabel.** We compare with the AGST w/o pseudolabel, i.e. AGST without pseudolabeling the confident instances, to verify the design of pseudolabeling.
- **AGST w/o active query.** We compare with the AGST w/o active query, i.e. AGST using only random querying, to verify the design of our uncertainty & diversity-based active query strategy. This baseline could be regarded as a semi-supervised method because to random query labels is equivalent to passively obtaining labels.
- **AGST w/o regularization.** We compare with the AGST w/o pseudolabel, i.e., AGST without the regularization of the distance to the last model, to verify the effectiveness of regularization in gradual semi-supervised domain adaptation.
- **AGST w/o uncertainty metric.** We compare with the AGST w/o uncertainty metric, i.e., AGST makes the original K-Means clustering for sample features in the active query phase. By comparing AGST with this baseline, we could verify the design of uncertainty metric and uncertainty-weighted K-Means in AGST.
- **Baseline query by confidence.** This baseline method is the same as AGST except for the active query part. It queries the least confident k samples at each active query phase. This baseline is designed to verify the effectiveness of our representative-diversity-based query strategy.

3) Implementation Detail: Models for each dataset are as follows.

- For Portraits and Rotating MNIST, we design the same neural network feature extractor with 3 conv layers. For each layer, we use a filter size of 5×5, stride of 2×2, 32 output channels, and relu activation. After the final convolution layer, we add a dropout layer with the probability of 0.5 and a batchnorm layer after dropout. The extracted features are then flattened and fed into fully connected layers with 2 and 10 outputs for Portraits and Rotating MNIST. Each of the output neurons is matched with a specific prediction class.
- For EVIS, we adopt the ResNet18 [12] as the backbone model for the classification task. Non-pretrained weights are used for the model initialization. Common pre-training scenarios, such as ImageNet, have covered web images after 2012. The model initialized in this way would be equivalent to having made a certain degree of adaptation to the target domain in advance, violating our setting.

Parameter settings for each dataset as following.

- For Rotating MNIST, the batch size is set to be 2000. We set confidence threshold value $\alpha = 0.1$,, pseudo-labeled data loss weight $\lambda_{\mathcal{PT}} = 1$, active queried data loss weight $\lambda_{\mathcal{LT}} = 5$, regularization weight $\lambda_{\mathcal{R}} = 0.01$, and entropy weight $\lambda_{\mathcal{H}} = 0.01$. Initial model is trained on source data for 200 epochs, then take 20 epochs of learning for each of the batches. The model optimizer used is the Adam optimizer, with a learning rate of 0.003. 100 active queries are made for each batch, i.e. $B = 100$, and the query rate is 5%.
- For Portraits, the batch size is set to be 500. We set confidence threshold value $\alpha = 0.6$,, pseudo-labeled data loss weight $\lambda_{\mathcal{PT}} = 1$, active queried data loss weight $\lambda_{\mathcal{LT}} = 1$, regularization weight $\lambda_{\mathcal{R}} = 0.08$, and entropy weight $\lambda_{\mathcal{H}} = 0.01$. Initial model is trained on source data for 200 epochs, then takes 20 epochs of learning for each of the batches. The model optimizer used is the Adam optimizer, with a learning rate of 0.002. 10 active queries are made for each batch, i.e., $B = 10$, and the query rate is 2%.
- For EVIS, the batch size is set to be 400 (i.e. one batch for one month). The input images are randomly cropped by size 60×60 and randomly horizontally flipped by the probability of 0.5 to achieve data augmentation. We set confidence threshold value $\alpha = 0.5$,, pseudo-labeled data loss weight $\lambda_{\mathcal{PT}} = 1$, active queried data loss weight $\lambda_{\mathcal{LT}} = 7.5$, regularization weight $\lambda_{\mathcal{R}} = 0.025$, and entropy weight $\lambda_{\mathcal{H}} = 0.0025$. Initial model is trained on source data for 60 epochs, then takes 2 epochs of learning for each batch. The model optimizer used is the Adam optimizer, with a learning rate of 0.000125 for extractor and classifier. 40 active queries are made for each batch, i.e. $B = 40$, and the query rate is 10%.



(a) Rotating MNIST (b) Portraits (c) EVIS

Fig. 7: Classification accuracy v.s. batch of AGST and baselines. (a) The result of the rotating MNIST, which is a synthesis dataset without environment noise. (b) The result of the portraits, which is a real-world dataset labeled by a human without environmental noise. (c) The result of the EVIS, which is a real-world dataset automatically collected from the web with environment noise.

TABLE III: Classification accuracies on the final target domain for AGST and baseline models with 90% confidence intervals for the mean over 5 runs.

	Rotating MNIST		**Portraits**		**EVIS**	
	Accuracy	Labels	Accuracy	Labels	Accuracy	Labels
Source Model	13.75±3.20%	0%	59.08±1.32%	0%	52.65±2.18%	0%
UGDA	14.38±3.30%	0%	62.60±1.02%	0%	17.70±3.73%	0%
Direct CLUE	57.03±6.87%	5%	62.12±8.10%	2%	53.25±4.91%	10%
AGST	**76.20**±2.41%	5%	**77.60**±1.01%	2%	**60.35**±3.10%	10%

B. Experiment Results

Our experiment results are illustrated in Figure 7 and Table III, from which we make the following observations:

AGST works consistently well. AGST works consistently well on synthetic data is artificially selected and does not suffer great noise, portrait data that is real-world and artificially selected without labeling noise, and our EVIS data that is automatically sampled from web search engine and suffers significant noise. Specifically, as shown in Table III, AGST achieves over 60% accuracy increase than UGDA with only 5% active labels and 19% better than direct CLUE in Rotating MNIST, over 15% in portraits with only 2% labels and 15% better than direct CLUE in Portraits, around 8% with 10% labels and 7% better than direct CLUE in EVIS. From the Figure 7, we observe that AGST also consistently outperforms other benchmarks during the adaptation process. These results show the effectiveness of our proposed algorithm.

Direct apply active DA is insufficient. We observe that Direct CLUE can significantly improve the performance of unsupervised methods, e.g. up to 42% improvement as compared with UGDA in Rotating MNIST, 3% in Portraits and 35% in EVIS. However, it is significantly worse than AGST that utilizes the intermediate data, demonstrating the necessity of gradual domain adaptation.

Unsupervised model drops with time. As illustrated in Figure 7, we observe that the accuracy of the source model keeps decreasing with time, which implies the gradual domain drift in these datasets. From Figure 7a and Figure 7b, the unsupervised approach can indeed slightly help the adaptation in datasets without noise, however, is still aligned with the decreasing trend, leading to only marginal enhancement on the final target domain, e.g., 0.6% in Rotating MINIST, 3.6% in portraits, and no enhancement in EVIS. These results support our claim that the unsupervised methods can not deal with gradual domain drift.

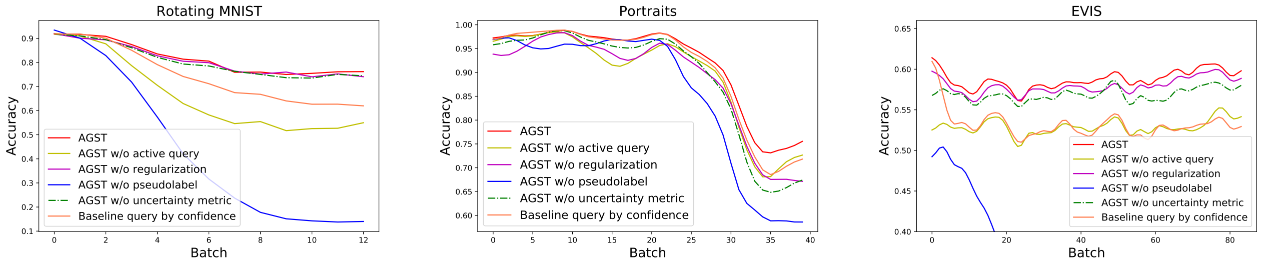
Unsupervised method fails with noise. As illustrated in Figure 7c and Table III EVIS, we observe that the UGST drops severe and suffers only 17.70% accuracy in the final target domain, while even the source model has 52.65% accuracy. Note that we do not show the whole line of UGST, since the performance of UGST is too low, showing the whole figure makes the range of the y axis too large to distinguish the comparison with other baselines. This implies that the unsupervised approach may harm the result in the noise setting. In contrast, with the help of active queries, AGST can maintain a good performance with 10% labels, which verify the necessity of active querying in the gradually changing domain.

In summary, the experiment results show that the previous unsupervised approach fails in dealing with gradual domain drift, especially in noise setting such as web application environments, and the direct active DA method is insufficient to get a good result. In contrast, our method achieves a significant performance gain by active gradual domain adaptation.

C. Ablation Study

Our experiment results of the ablation study are illustrated in Figure 8 and Table IV, from which we make the following observations:

Pseudolabeling is crucial. As illustrated in Table IV, we observe that AGST significantly outperforms AGST w/o pseudolabel, which only learns from active queried labels without the pseudolables. Specifically, the accuracy of AGST w/o



(a) Rotating MNIST (b) Portrait (c) EVIS

Fig. 8: Classification accuracy v.s. batch of AGST and baselines for ablation studies on (a) MNIST, (b) Portraits, and (c) EVIS dataset. Note that we have three key designs on AGST, i.e. active query, pseudolabel learning and gradual regularization. To verify the efficacy of each component, we compare with five baselines: AGST w/o active query, AGST w/o pseudolabel, AGST w/o regularization, AGST w/o uncertainty metric, and baseline query by confidence.

TABLE IV: Classification accuracies on the final target domain for AGST and ablation study baseline models with 90% confidence intervals for the mean over 5 runs.

	Rotating MNIST		**Portraits**		**EVIS**	
	Accuracy	Labels	Accuracy	Labels	Accuracy	Labels
AGST	**76.20**±2.41%	5%	**77.60**±1.01%	2%	**60.35**±3.10%	10%
AGST w/o pseudolabel	14.00±1.28%	5%	59.08±1.01%	2%	20.12±3.60%	10%
AGST w/o active query	54.99±4.83%	5%	73.44±9.91%	2%	55.15±0.81%	10%
AGST w/o regularization	74.08±4.21%	5%	67.20±6.79%	2%	59.85±0.12%	10%
AGST w/o uncertainty metric	74.46±3.68%	5%	67.80±7.75%	2%	58.55±0.81%	10%
Baseline query by confidence	61.99±4.63%	5%	72.68±5.63%	2%	52.95±2.05%	10%

pseudolabel is only 14% in Rotating MINIST, 59% on Portraits, and 20% in EVIS, and has a significant bias with AGST. As shown in Figure 8, the performance of AGST w/o pseudolabel drops rapidly over time even with the same queried labels as AGST, while AGST with pseudolabels performs consistently well. These results demonstrate that pseudolabeling the confident instances is crucial to improve the model generalization ability under gradual domain drift. The reason is that the pseudolabel augments the labeled data and helps the gradual semi-supervised domain adaptation.

Active query strategy is efficient. As illustrated in Table IV, we observe that AGST significantly outperforms AGST w/o active query, which uses a random query strategy with the same querying ratio with AGST. Specifically, the accuracy of AGST w/o active query is 22% worse than AGST on Rotating MINIST, 4% worse on Portraits, and 5% worse on EVIS. As shown in Figure 8, the performance of AGST w/o active query is consistently worse than AGST with active query strategy during the process. We conduct experiments on different datasets with different percentages of queried labels, e.g., 2%, 5% and 10%. The experimental results show that our proposals consistently show advantages. These results show that active query is effective in the gradual domain adaptation problem. Moreover, by observing the result in Table IV and Figure 8, we could find that AGST also consistently outperforms the baseline query by confidence for all the experiments. These results demonstrate that our active query strategy, based on both uncertainty and diversity, is more efficient than the confidence-based query strategy under gradual domain drift. The reason is that our active strategy selects the most informative samples by measuring the uncertainty and diversity, which largely improves the efficiency of queried samples. Finally, AGST consistently outperforms AGST w/o uncertainty metric in the experiments for all three datasets, supporting the effectiveness of AGST's uncertainty evaluation metric and uncertainty-weighted K-Means clustering for the active query.

Gradual regularization helps to improve robustness. As illustrated in Table IV, we observe that AGST significantly outperforms AGST w/o regularization, which does not constraint the update from the last model by adding a regularization. Specifically, the accuracy of AGST w/o regularization is 2% worse than AGST on Rotating MINIST, 10% worse on Portraits, and 0.5% worse on EVIS, where adding the regularization leads to robust performance. As shown in Figure 8, the performance of AGST w/o regularization is consistently worse than AGST with active regularization during the process. The reason is that the regularization helps to eschew the noise under a small batch.

In summary of the ablation results, we verify that all the designs, consisting of pseudolabeling, active query, and gradual regularization, effectively help the model adapt to the gradually changing target domains. All the novel designs contribute to the performance of AGST.

VI. Conclusion

In this paper, we study a new but practical problem: the gradual domain adaptation with limited labels, which challenges machine learning systems in many real-world scenarios, especially web applications. To address this, we establish an effective algorithm – Active Gradual Self-Training (AGST) with the key designs of the active pseudolabeling and the gradual semi-supervised domain adaptation. To verify the effectiveness of the proposed method, we first create a new dataset – Evolving-Image-Search (EVIS) collected from the web search engine without any manual selection. We conduct the experiments on synthetic, real-world, and EVIS datasets, and the results show that AGST performs consistently well. Our ablation study shows that both the active pseudolabeling and the gradual semi-supervised domain adaptation contribute to this remarkable performance. Our results take the first step towards the problem of active gradual domain adaptation, and we believe that this paper could stimulate future work on the design of algorithms with stronger adaptation and active query strategies with better efficiency.

References

[1] Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*, 2019.

[2] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[3] Adeleh Bitarafan, Mahdieh Soleymani Baghshah, and Marzieh Gheisari. Incremental evolving domain adaptation. *IEEE Transactions on Knowledge and Data Engineering*, 28(8):2128–2141, 2016.

[4] Andreea Bobu, Eric Tzeng, Judy Hoffman, and Trevor Darrell. Adapting to continuously shifting domains. 2018.

[5] Rita Chattopadhyay, Wei Fan, Ian Davidson, Sethuraman Panchanathan, and Jieping Ye. Joint transfer and batch-mode active learning. In *International conference on machine learning*, pages 253–261. PMLR, 2013.

[6] David Cohn, Les Atlas, and Richard Ladner. Improving generalization with active learning. *Machine learning*, 15(2):201–221, 1994.

[7] Gabriela Csurka. Domain adaptation for visual applications: A comprehensive survey. *arXiv preprint arXiv:1702.05374*, 2017.

[8] Bo Fu, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. Learning to detect open classes for universal domain adaptation. In *European Conference on Computer Vision*, pages 567–583. Springer, 2020.

[9] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

[10] Shiry Ginosar, Kate Rakelly, Sarah Sachs, Brian Yin, and Alexei A Efros. A century of portraits: A visual historical record of american high school yearbooks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1–7, 2015.

[11] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[13] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. PMLR, 2018.

[14] Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex Smola. Correcting sample selection bias by unlabeled data. *Advances in neural information processing systems*, 19:601–608, 2006.

[15] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7310–7311, 2017.

[16] Ananya Kumar, Tengyu Ma, and Percy Liang. Understanding self-training for gradual domain adaptation. In *International Conference on Machine Learning*, pages 5468–5479. PMLR, 2020.

[17] Guillermo Ortiz-Jimenez, Mireille El Gheche, Effrosyni Simou, Hermina Petric Maretic, and Pascal Frossard. Cdot: Continuous domain adaptation using optimal transport. *arXiv preprint arXiv:1909.11448*, 2019.

[18] Viraj Prabhu, Arjun Chandrasekaran, Kate Saenko, and Judy Hoffman. Active domain adaptation via clustering uncertainty-weighted embeddings. *arXiv preprint arXiv:2010.08666*, 2020.

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

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

[21] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

[22] Rui Shu, Hung Bui, Hirokazu Narui, and Stefano Ermon. A dirt-t approach to unsupervised domain adaptation. In *International Conference on Learning Representations*, 2018.

[23] Jong-Chyi Su, Yi-Hsuan Tsai, Kihyuk Sohn, Buyu Liu, Subhransu Maji, and Manmohan Chandraker. Active adversarial domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 739–748, 2020.

[24] Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5), 2007.

[25] Yu Sun, Eric Tzeng, Trevor Darrell, and Alexei A Efros. Unsupervised domain adaptation through self-supervision. *arXiv preprint arXiv:1909.11825*, 2019.

[26] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017.

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

[28] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017.

[29] Hao Wang, Hao He, and Dina Katabi. Continuously indexed domain adaptation. *arXiv preprint arXiv:2007.01807*, 2020.

[30] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.

[31] Markus Wulfmeier, Alex Bewley, and Ingmar Posner. Incremental adversarial domain adaptation for continually changing environments. In *2018 IEEE International conference on robotics and automation (ICRA)*, pages 4489–4495. IEEE, 2018.

[32] Jiaolong Xu, Liang Xiao, and Antonio M López. Self-supervised domain adaptation for computer vision tasks. *IEEE Access*, 7:156694–156706, 2019.

[33] Yifan Zhang, Ying Wei, Qingyao Wu, Peilin Zhao, Shuaicheng Niu, Junzhou Huang, and Mingkui Tan. Collaborative unsupervised domain adaptation for medical image diagnosis. *IEEE Transactions on Image Processing*, 29:7834–7844, 2020.

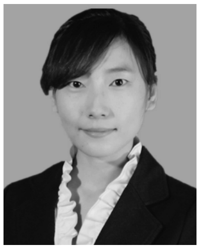
[34] Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In *International Conference on Machine Learning*, pages 7523–7532. PMLR, 2019.

[35] Sicheng Zhao, Xiangyu Yue, Shanghang Zhang, Bo Li, Han Zhao, Bichen Wu, Ravi Krishna, Joseph E Gonzalez, Alberto L Sangiovanni-Vincentelli, Sanjit A Seshia, et al. A review of single-source deep unsupervised visual domain adaptation. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[36] Fedor Zhdanov. Diverse mini-batch active learning. *arXiv preprint arXiv:1901.05954*, 2019.

Shiji Zhou is currently a Ph.D. student at Tsinghua University. He received his bachelor's degree in Chinese Academy of Science - Beihang University, Hua Luogeng Mathematics Honors Class. His research areas include non-stationary online learning, domain adaptation, and multimedia networks.

Lianzhe Wang is currently a Master student at Tsinghua University. He received his bachelor's degree from Sichuan University. His research interests focus on transfer learning, especially in the field of open-world domain adaptation.



Shanghang Zhang is currently a postdoc research fellow in the Berkeley AI Research Lab (BAIR), EECS, UC Berkeley. Her research covers multimedia intelligence and machine learning, especially sample efficient learning, as reflected in her publications on top-tier journals and conferences, including TMM, TNNLS, NeurIPS, ICLR, ACM MM, CVPR, ICCV, and AAAI. She has also been the chief author and editor of the book "Deep Reinforcement Learning: Fundamentals, Research and Applications" published by Springer Nature. She has received the Best Paper Award at AAAI 2021, "2018 Rising Stars in EECS, USA", and Qualcomm Innovation Fellowship (QInF) Finalist Award. Dr. Zhang has been the chief organizer of several workshops on ICML/NeurIPS, and the special issue on ICMR. She received her Ph.D. from Carnegie Mellon University; and her Master from Peking University.

Zhi Wang is currently an associate professor at Tsinghua Shenzhen International Graduate School. His research areas include multimedia networks, mobile cloud computing, and large-scale machine learning systems. He received the Outstanding Doctoral Dissertation Award from China Computer Federation in 2014, Best Paper Award at ACM Multimedia 2012, and Best Student Paper Award at MMM 2015. He is a recipient of the Second Prize of National Natural Science Award and the First Prize of Natural Science Award of Ministry of Education in 2017. He is an Associate Editor of IEEE TMM and Guest Editor of ACM TIST and JCST. His research has been covered by prestigious media including MIT Technology Review.

Wenwu Zhu is currently a Professor at Computer Science Department of Tsinghua University and Vice Dean of National Research Center on Information Science and Technology. Prior to his current post, he was a Senior Researcher and Research Manager at Microsoft Research Asia. He was the Chief Scientist and Director at Intel Research China from 2004 to 2008. He worked at Bell Labs New Jersey as a Member of Technical Staff during 1996-1999. He served as the Editor-in-Chief for the IEEE Transactions on Multimedia (T-MM) from January 1, 2017 to December 31, 2019. He has been serving as the chair of the steering committee for IEEE T-MM and Vice EiC for IEEE Transactions on Circuits and Systems for Video Technology (TCSVT) since January 1, 2020. His current research interests are in the areas of cross-media big data and intelligence, and multimedia edge computing. He received nine Best Paper Awards. He is an IEEE Fellow, AAAS Fellow, SPIE Fellow and a member of the European Academy of Sciences (Academia Europaea).