Pattern Recognition 180 (2026) 114299

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

GALA: A GlobAl-LocAl cluster approach for Multi-Source Active Domain Adaptation

Juepeng Zheng [a,b], Yibin Wen [a], Peifeng Zhang [a], Zurong Mai [a], Qingmei Li [c,*], Yang Zhang [a], Haohuan Fu [c,b]

[a] School of Artificial Intelligence, Sun Yat-Sen University, Zhuhai, 519000, China
[b] National Supercomputing Center in Shenzhen, Shenzhen, 515055, China
[c] Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, 518057, China

ARTICLE INFO

Keywords:
Active learning
Domain adaptation
Multisource domain adaptation
Global-local cluster strategy

ABSTRACT

Domain Adaptation (DA) provides an effective way to tackle target-domain tasks by leveraging knowledge learned from source domains. Recent studies have extended this paradigm to Multi-Source Domain Adaptation (MSDA), which exploits multiple source domains carrying richer and more diverse transferable information. However, a substantial performance gap still remains between adaptation-based methods and fully supervised learning. In this paper, we explore a more practical and challenging setting, named Multi-Source Active Domain Adaptation (MS-ADA), to further enhance target-domain performance by selectively acquiring annotations from the target domain. The key difficulty of MS-ADA lies in designing selection criteria that can jointly handle inter-class diversity and multi-source domain variation. To address these challenges, we propose a simple yet effective GlobAl-LocAl strategy (GALA), which combines a global k-means clustering step for target-domain samples with a cluster-wise local selection criterion. The global step encourages diverse and uncertain target sample selection, while the local step further considers the relationship between target samples and multiple source domains to improve transferability. Our proposed GALA is plug-and-play and can be seamlessly integrated into existing DA frameworks without introducing any additional trainable parameters. Extensive experiments on three standard DA benchmarks, including Digit-Five, Office-Home, and DomainNet, demonstrate that GALA consistently outperforms prior active learning and active DA methods, achieving performance comparable to the fully supervised target-domain reference while using only 1% of the target annotations.

1. Introduction

Despite the remarkable breakthroughs of deep learning across a wide spectrum of applications, it remains a challenge to achieve a generalized deep learning model which can generalize well on unseen samples, especially given the vast diversity of real-world data and problems. A significant drop in accuracy occurs due to various differences between the training data and new target dataset, for aspects such as statistical distribution, dimension, context, etc. As a result, we usually need to further tune this model for a new specific target distribution. Yet, with the explosion of data acquisition, additional annotating and tuning soon becomes infeasible. Domain adaptation (DA) offers significant benefits in this scenario by adapting the pre-trained models towards new domains with different distributions and properties from the original domain [1–3]. Moreover, DA provides great potential for efficiency improvement by reducing the need for retraining on new data from scratch, which has been well adopted in various fields. However, current DA strategies mainly focus on just a single source domain and single target domain, *i.e.*, Single-Source DA (SSDA) (see Fig. 1(a)).

Recently, Multi-Source Domain Adaptation (MSDA) (see Fig. 1(c)) attracts more and more attention since we usually have multiple source domains in practice [4]. In MSDA, labeled samples derive from multiple distinct source distributions and models are trained to generalize to a target domain, where ground truth labels are entirely absent. For instance, images in autonomous driving may be captured across diverse weather conditions, or remote sensing data may originate from different generations of sensors or satellites. A variety of approaches have been proposed aiming at different application scenarios for MSDA, such as text classification [5], semantic segmentation [6], facial expression recognition [7], etc. Some popular strategies, such as adversarial

* Corresponding author.
E-mail address: qingmeili@sz.tsinghua.edu.cn (Q. Li).

https://doi.org/10.1016/j.patcog.2026.114299
Received 12 February 2026; Received in revised form 4 June 2026; Accepted 16 June 2026
Available online 22 June 2026

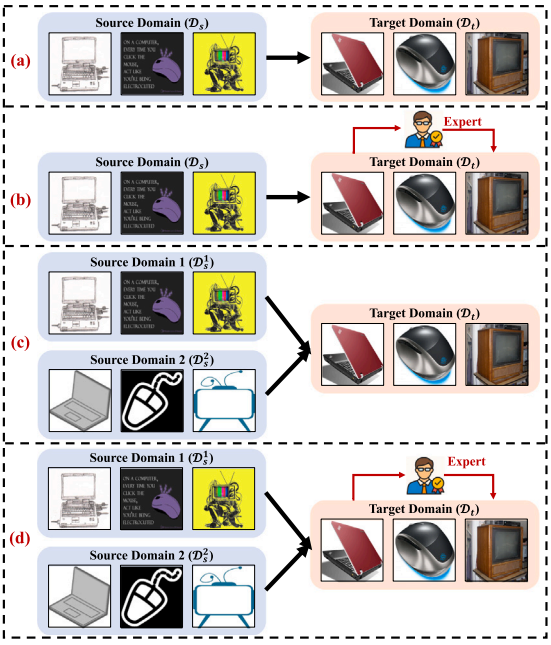

Fig. 1. Illustration of different domain adaptation and active domain adaptation settings. (a) **Single-Source Domain Adaptation (SSDA)** adapts a model from one labeled source domain to an unlabeled target domain. (b) **Single-Source Active Domain Adaptation (SS-ADA)** further queries a small number of informative target samples for annotation to facilitate adaptation from one source domain. (c) **Multi-Source Domain Adaptation (MSDA)** leverages multiple labeled source domains with different styles or distributions to adapt to the target domain. (d) **Multi-Source Active Domain Adaptation (MS-ADA)** combines multi-source adaptation with active target annotation, aiming to select representative and informative target samples under source-domain variations.

learning [8] and source distilling [9], have been introduced to enhance target-domain accuracy by leveraging annotated source data. In all these cases, we wish to learn from all source domains.

Although abovementioned MSDA approaches could improve the performance of MSDA setting, there still exists considerable accuracy gap between the adaptation and the Fully-Supervised (FS) target-domain reference (with 14.4% and 2.0% for **DomainNet** and **Digit-Five**, respectively) (See Fig. 2). Therefore, we try to further close the gap between MSDA and FS using active learning for the target domain. Nowadays, Active Domain Adaptation (ADA) approaches have emerged, aiming at selecting unlabeled instances that are uncertain to the model or reflect the statistical features of target domain. Most of them focus on Single-Source ADA (SS-ADA) (see Fig. 1(b)) using density weighted entropy [10], hard examples mining [11] or exploiting free energy biases [12], etc.

Different from SS-ADA or other active learning settings, we focus on another more practical setting named Multi-Source ADA (MS-ADA) (see Fig. 1(d)), where we could annotate a small subset of data from the target domain to overcome the problems of domain shifts according to the multiple source domains. Although Zhang et al. [13] firstly propose Detective method for MS-ADA by integrating uncertainty calculation strategy with a dynamic DA framework, they have not considered the specific challenges in MS-ADA. In fact, there are two major challenges in MS-ADA: ***Challenge 1***: the performance imbalance among different classes originates from multiple source domains, so that the query strategy algorithm must guarantee inter-class diversity from the perspectives of target dataset; ***Challenge 2***: facing multiple source domains with style and texture variations, the model has to consider a new metric to enhance the transferability of model by conducting suitable target sample selection. Due to these two challenges, existing ADA methods that directly count on utilizing the uncertainty [10] or free energy [12] between two specific domains become infeasible.

Herein, we propose a straightforward yet high-performing method, named GlobAl-LocAl strategy (GALA), to better tackle the aforementioned two challenges in MS-ADA scenarios. This work delivers three key contributions:

- We point out two main difficulties in MS-ADA: the inter-class diversity as well as the multiple source domain variation, and propose a plug-and-play approach to any other domain adaptation approaches without adding any trainable parameters.
- We propose a straightforward yet high-performing GlobAl-LocAl (GALA) strategy to overcome abovementioned two challenges. GALA consists of a global k-means step for target domain data and a cluster-wise local selection with a new criterion that includes diversity and transferability.
- We conduct comprehensive experiments on three common DA benchmarks (**Digit-Five**, **Office-31** and **DomainNet**), and empirical results show that GALA outperforms SOTA ADA methods with considerable gains and achieves competitive accuracy against the fully supervised target-domain reference with only 1% annotation.

2. Related works

2.1. Domain adaptation

The primary objective of Domain Adaptation (DA) is to minimize domain shift, i.e., the statistical discrepancy between source and target domains. Most existing DA methods focus on Single-Source DA (SSDA) (see Fig. 1(a)) and have been widely applied to classification [14–17], semantic segmentation [18,19], and object detection [20]. To reduce domain discrepancies, existing methods mainly rely on adversarial learning inspired by GANs [21,22], or statistical distribution alignment techniques such as MMD [23] and JMMD [24]. More recently, refined strategies such as dual semantic correlation alignment [20] have been proposed to address more complex domain shifts in object detection.

Beyond the standard closed-set setting, recent DA studies have explored more realistic scenarios, including few-shot unsupervised domain adaptation with limited labeled source samples [25], blended-target domain adaptation with multiple latent target domains [26], and source-free universal domain adaptation with category-space discrepancies [27]. Specifically, ADCC [25] exploits collaborative clustering and an intermediate mix-up domain for smoother source–target alignment; EDEAN [26] uses free-energy bias, pseudo-label fusion, and energy-based reweighting to handle blended target domains; and EGRS [27] reactivates source-model knowledge with adversarial perturbation, prototype refinement, and pseudo-label clustering. Nevertheless, SSDA methods may suffer from limited transferable knowledge when only one source domain is available. Therefore, this paper focuses on Multi-Source DA (MSDA), where multiple source domains can provide richer and more diverse knowledge for target-domain adaptation.

2.2. MSDA

Compared to the extensive studies on SSDA, MSDA (see Fig. 1(c)) remains less explored. Existing MSDA methods mainly address source–target and inter-source discrepancies through adversarial alignment, moment matching, dynamic transfer, and class-aware alignment. For example, MDAN [8] and DCTN [28] adopt adversarial learning to extract domain-invariant features, while M^{3}SDA [29] reduces discrepancies by matching high-order moments. DRT [30] tailors model parameters to individual samples to alleviate domain conflicts, and PFDA [31] designs multiple alignment losses to preserve both class discrimination and domain-specific cluster structures. SIG [32] further introduces class-aware conditional alignment to handle target shifts caused

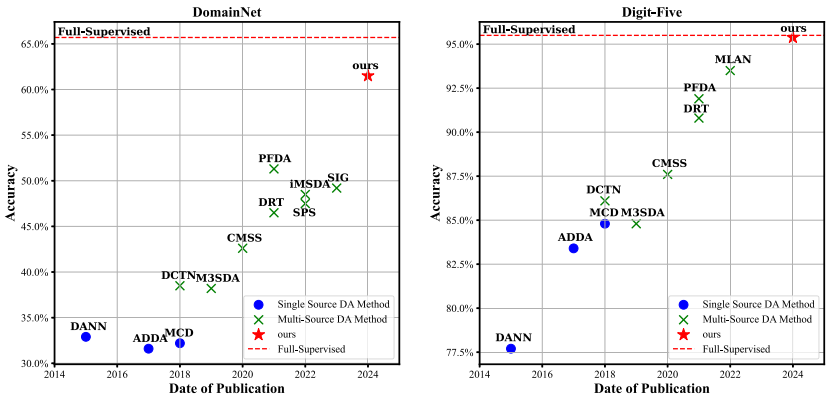

Fig. 2. Accuracy comparison of representative DA methods on the **DomainNet** and **Digit-Five** datasets. The blue circles denote single-source DA methods, the green crosses denote multi-source DA methods, and the red stars denote our proposed GALA. The red dashed line indicates the fully supervised target-domain reference, which is obtained by directly training the same backbone with labeled target-domain data. Although existing MSDA methods generally improve over SSDA methods, a clear gap still remains between adaptation-based methods and this supervised target-domain reference (with 14.4% and 2.0% on DomainNet and Digit-Five, respectively). By actively annotating only 1% of the target data, GALA consistently improves the adaptation performance and narrows this gap.

by domain-varying label distributions. Recent studies have extended MSDA to more practical scenarios. C-AMDN [33] explores diffusion-based MSDA with causal-guided adaptation, and DFM-Trans [34] studies multi-source-free domain adaptation by dynamically fusing multiple source models without accessing source data. Beyond the standard single-target MSDA setting, MAN [35] investigates multi-source and multi-target domain adaptation, while SAUE [36] studies multi-source blended-target domain adaptation by combining target-style augmentation and uncertainty estimation.

However, as shown in Fig. 2, although these MSDA methods generally outperform SSDA methods, a considerable accuracy gap still remains between MSDA methods and the fully supervised target-domain reference, with gaps of 14.4% and 2.0% on DomainNet and Digit-Five, respectively.

2.3. Active domain adaptation

Conventional active learning aims to select informative samples for annotation based on uncertainty [37] or diversity [38]. Active Domain Adaptation (ADA) further extends this idea to target-domain sample selection under distribution shifts. Representative methods include AADA [10], which balances entropy with adversarial target probabilities; CLUE [39], which combines entropy with k-means-based diversity sampling; SDM [11], which leverages a margin-based loss for active selection; and the prototype-guided method of Peng et al. [37], which alleviates classifier bias with perturbation-based uncertainty. ADA has also been explored in medical image analysis, including dual-reference source-free ADA for nasopharyngeal carcinoma tumor segmentation [40], VLM-driven ADA for fair cross-domain medical image segmentation [41], and source-free ADA for medical video polyp segmentation [42], demonstrating its generality in privacy-sensitive and annotation-scarce scenarios.

However, most existing ADA methods focus on the Single-Source ADA (SS-ADA) setting (see Fig. 1(b)). Although methods such as BADGE [43] and CLUE [39] consider both uncertainty and diversity, they do not explicitly model discrepancies among multiple source domains. D^{3}GU [44] considers Multi-Target ADA, while Detective [13] is the first MS-ADA method, but it relies on additional trainable modules and is not a general plug-and-play strategy for arbitrary MSDA methods. In contrast, GALA decomposes MS-ADA sample selection into a global diversity-aware step and a local transferability-aware step, enabling the selected target samples to be both diverse and informative with respect to multiple heterogeneous source distributions. Moreover, GALA introduces no extra trainable parameters and can be seamlessly integrated into both SSDA and MSDA frameworks.

3. Methodology

3.1. Problem setting and notations

In contrast to SSDA, MSDA is established on multiple source domains $\mathcal{D}_S = \{D_s^k\}_{k=1}^K = \{(\mathbf{x}_s^i, \mathbf{y}_s^i)\}_{i=1}^{n_S}$, where K is the number of source domains and $n_S = \sum_{k=1}^K n_s^k$ denotes the total quantity of images in the multi-source domain. $\mathbf{y}_s^i$ is the corresponding label of $\mathbf{x}_s^i$. In the MS-ADA setting, we utilize the labeled multiple source domains and the unlabeled target domain $\mathcal{D}_t = \left\{(\mathbf{x}_t^j)\right\}_{j=1}^{n_t}$, where n_t denotes the quantity of images in the target domain. Then, the model selects and annotates images from $\mathcal{D}_t$ at multiple stages to build the selected target set $\mathcal{D}_{st} = \left\{(\mathbf{x}_t^j, \mathbf{y}_t^j)\right\}_{j=1}^N$ for training, in which $N = R \times B$ is the total number of selected target images (R is the total active rounds and B is the quantity of queries for each round). Hence, if we directly adopt existing SS-ADA methods and regard all multiple source domains as a single source dataset straightforwardly, the training objective and sample selection strategy would tend to learn representations invariant to the combined source set $\mathcal{D}_S$ rather than K source domains $\{D_s^k\}_{k=1}^K$.

As shown in Fig. 3, the proposed GlobAl-LocAl (GALA) method can be seamlessly embedded into different DA approaches without introducing additional trainable parameters. GALA consists of three components: the global step, the local step, and the underlying DA approach. The global step addresses ***Challenge 1*** by encouraging inter-class diversity in target sample selection, while the local step tackles ***Challenge 2*** by considering the style and texture variations brought by multiple source domains.

During the rth active learning round, GALA queries a batch of B unlabeled target samples by jointly considering diversity and transferability. After annotation, these selected samples are combined with the labeled multi-source data to update the model and progressively refine the target-domain decision boundary. This process is repeated over R active learning rounds. The overall workflow is illustrated in Fig. 3 and summarized in Algorithm 1.

3.2. Global step

Specifically, we utilize k-means to achieve the global step on the hypothetical gradient embedding for the target data D_t. Traditional active learning methodologies frequently incorporate gradient embeddings [43]. It is intuitive that the larger the gradient embedding is, the more uncertain the corresponding sample is. We assume that z is the embedding vector immediately preceding the final layer W associated with $x_t^j \in D_t^{r-1}$, where D_t^{r-1} denotes the unlabeled target dataset before

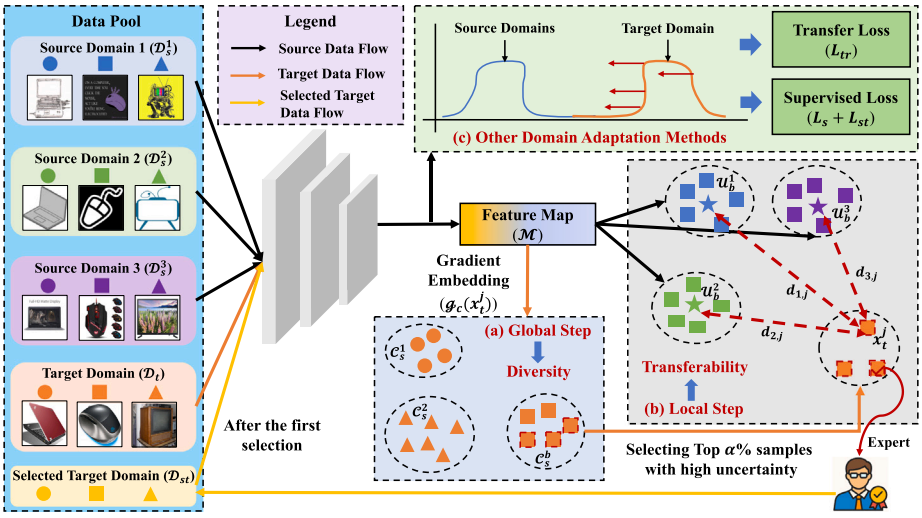

Fig. 3. The framework for our proposed GlobAl-LocAl strategy (GALA). (a) **Global step**: We select the target queries (D_{st}) to contain both diversity in the gradient embedding space ($g_c\left(x_t^j\right)$) and uncertainty from the perspective of the training model (addressing ***Challenge 1***). (b) **Local step**: We select the highest value with high uncertainty and large domain gap for each cluster (C_s^b) of v according to Eq. (6) (addressing ***Challenge 2***). (c) **Other DA methods**: Our proposed GALA could be easily plug-and-play to any other DA approaches without any trainable parameters. Pseudo algorithm for GALA is provided in Algorithm 1.

the annotating round of r. As we do not have access to the ground-truth annotations for the target domain, we employ their pseudo-labels generated from the model. Here, we introduce the gradient derived from the negative cross-entropy objective function relative to the last layer of encoder, which is induced by a pseudo label. Note that we only consider the gradients associated with the predicted label (i.e., $g_{\hat{y}}\left(x_t^j\right)$) to reduce computational overhead.

$$g_c\left(x_t^j\right) = -\frac{\partial}{\partial W_c}\mathcal{L}_{CE}\left(x_t^j, \hat{y}; \Theta^r\right) = \mathcal{M}\cdot\left(\mathbb{1}_{[\hat{y}=c]} - p_c\right), \tag{1}$$

where W_c denotes the weights connected to the cth logit neuron, Θ^r denotes the model parameters at round r, p_c is the predicted probability of class c, and $\hat{y}$ is the pseudo-label defined as $\hat{y} = \arg\max_{c\in[C]} p_c$. After that, we compute B cluster centers $(C_s^1, \ldots, C_s^B)$ within the gradient embedding space adopting the standard k-means algorithm by minimizing:

$$J = \sum_{j=1}^{n_t}\sum_{b=1}^{B} w_{j,b}\|g_{\hat{y}}\left(x_t^j\right) - C_s^b\|^2, \tag{2}$$

in which $w_{j,b}$ serves as an indicator function. If $g_{\hat{y}}\left(x_t^j\right)$ is assigned to C_s^b, $w_{j,b} = 1$. Otherwise $w_{j,b} = 0$. Actually, Eq. (1) indicates that the gradient embedding ($g_c\left(x_t^j\right)$) essentially represents a scaled version of the feature map $\mathcal{M}$, especially with the scale of uncertainty (or confidence). In other words, if a target sample exhibits uncertainty when predicting (low value of $p_{\hat{y}}$, which also means low confidence), its corresponding gradient vector becomes larger. To this end, Eq. (2) could be considered as a weighted variant of k-means applied to the feature map through utilizing the function of softmax response. After that, we divide the target samples into B clusters according to their diversity and uncertainty. Also, we retain the top α% most uncertain candidates calculated by Eq. (1) for each cluster C_s^b. It is worth noting that the global step does not merely provide cluster assignments. After clustering the target samples into B clusters, we further retain the top α% most uncertain samples within each cluster as the candidate subset $D_{t,b}^{r-1}$. This candidate subset serves as the input pool for the subsequent local step. In other words, the local step does not search over all target samples, but only performs fine-grained selection within the uncertainty-filtered candidates of each cluster. This design allows the global step to first guarantee coarse inter-cluster diversity and remove low-uncertainty samples, while the local step further considers multi-source domain discrepancy for final query selection. Therefore, the global step of GALA enables the selected target set (D_{st}) to balance feature-space diversity with model-based uncertainty, which effectively addresses ***Challenge 1***.

3.3. Local step

After the global step, we divide the target samples into B clusters according to their diversity and uncertainty. For each cluster, we obtain an uncertainty-filtered candidate subset $D_{t,b}^{r-1}$ from the global step. Given this candidate subset, the local step further selects one target sample from $D_{t,b}^{r-1}$ instead of searching over the whole target pool. Specifically, we assign all source samples from multiple source domains to these B clusters using their feature embeddings ($\mathcal{M}$). In this way, source samples and target candidates are organized under the same cluster structure, enabling source-domain information to be compared within each cluster. After that, we calculate the distance ($d_{k,j}, k = 1, \ldots, K$) between each target sample ($x_t^j \in D_{t,b}^{r-1}$) and these K centroids from different source domains in the feature embedding space. To make the distance calculation explicit, we denote the feature map extracted from the last convolutional block as $F(x) \in \mathbb{R}^{C_f\times H_f\times W_f}$. For ResNet-50 and ResNet-101, $C_f = 2048$; under the standard 224×224 input resolution, the spatial size before global average pooling is 7×7. We apply global average pooling to obtain a channel-wise feature vector $h(x) \in \mathbb{R}^{C_f}$. Therefore, for a target sample x_t^j, its statistics μ_t^j and σ_t^j are computed as the mean and standard deviation over the channel dimension of $h(x_t^j)$. These statistics characterize the channel-wise feature distribution of an individual target sample, rather than the statistics over a mini-batch or over multiple target samples. For the bth cluster and the kth source domain, let S_b^k denote the set of source samples from domain k assigned to cluster b. The corresponding source-domain centroid is computed by directly averaging their channel-wise feature vectors:

$$\mathcal{U}_b^k = \frac{1}{|S_b^k|}\sum_{x_s^{i,k}\in S_b^k} h(x_s^{i,k}). \tag{3}$$

The statistics μ_s^k and σ_s^k in Eq. (4) are then computed as the mean and standard deviation over the channel dimension of the centroid vector $\mathcal{U}_b^k$. Thus, Eq. (4) measures the discrepancy between the normalized channel-wise statistics of each target candidate and those of

the source-domain centroid in the same cluster.

$$d_{k,j} = \left| \frac{\mu_s^k}{\sqrt{\left(\sigma_s^k\right)^2 + \epsilon}} - \frac{\mu_t^j}{\sqrt{\left(\sigma_t^j\right)^2 + \epsilon}} \right|, \tag{4}$$

where ϵ is a small number to prevent us from numerical problems in case of zero variance. μ_s^k and σ_s^k are the mean and the standard deviation for the centroid of kth source domain ($\mathcal{U}_b^k$). μ_t^j and $\sigma_t^{(j)}$ are the mean and the standard deviation for the centroid of jth target unlabeled data. Therefore, we adopt the max value from different j ($d_j = \min d_{k,j}, k = 1, 2, \ldots, K$) to describe the deviation of the target sample and the multiple source domains. In Section 5.4 we will compare the performance for different ways including the average and the minimum.

For each target candidate x_t^j in the bth cluster, we compute its distance to each source-domain centroid $\mathcal{U}_b^k$, where $k \in \{1, \ldots, K\}$. This produces a set of source–target distances $\{d_{k,j}\}_{k=1}^K$. We define the domain gap of x_t^j as its distance to the nearest source-domain centroid:

$$d_j = \min_{k \in \{1,\ldots,K\}} d_{k,j}. \tag{5}$$

The minimum is taken over source domains. A large value of d_j means that x_t^j is far from even its closest source domain, indicating poor transferability from the available source domains.

To this end, we design a new metric v that can be formulated as Eq. (6) for the target data x_t^j. $v\left(x_t^j\right)$ considers both the domain gap ($d_j / \max d$) generated from Eq. (4) using the feature embedding space and the uncertainty value ($\| g_c\left(x_t^j\right) \|_2$) generated from Eq. (1) using gradient embedding space. Notably, we normalize the distance (d_j) so that it could be $0 \sim 1$ and calculate the L2-normalization of $g_c\left(x_t^j\right)$ as the uncertainty.

$$v\left(x_t^j\right) = \| g_c\left(x_t^j\right) \|_2 \times \frac{d_j}{\max d}. \tag{6}$$

Here, $\| g_c(x_t^j) \|_2$ measures model uncertainty, while $d_j / \max d$ denotes the normalized source–target discrepancy. Therefore we select the target sample with the highest value of $v\left(x_t^j\right)$ for each cluster (C_s^b) that was introduced in the global step. Formally, the selected target sample in the bth cluster at round r is given by:

$$x_{t,b}^r = \arg \max_{x_t^j \in \mathcal{D}_{t,b}^{r-1}} v(x_t^j). \tag{7}$$

Therefore, the "minimum" is used to compute the nearest-source distance for each target candidate, while the "maximum" is used to select the final target sample among all candidates in the same cluster. This way could effectively pay more attention on those samples that are not very transferable during the sample selection process, which could better address ***Challenge 2***. Finally, we will select B instance at the round r for annotating. Algorithm 1 introduces the comprehensive framework of our GALA.

3.4. Embedding GALA into other DA methods

We integrate GALA into several representative DA frameworks to verify its generality and effectiveness, including Domain Adversarial Neural Network (DANN) [15], M^{3}SDA [29], Subspace Identification Guarantee (SIG) [32], and Cycle Self-Refinement (CSR) [45]. We also embed GALA into a baseline CNN model to evaluate its behavior under a standard non-adversarial DA setting. For all variants, we keep the original supervised source loss (L_s) and transfer loss (L_{tr}) unchanged. When selected target samples are annotated through active querying, an additional supervised target loss (L_{st}) is introduced to further improve target-domain discrimination. Since GALA only serves as a sample query strategy and introduces no additional trainable parameters, it can be seamlessly incorporated into different DA architectures while maintaining a lightweight and plug-and-play design.

Algorithm 1 MS-ADA framework with GALA algorithm

Input: Initial parameter Θ. Multiple labeled source dataset $\{D_s^k\}_{k=1}^K$ Unlabeled target dataset $\mathcal{D}_t^0$. Selected labeled target dataset $\mathcal{D}_{st}^0$. Labeling Budget B. Active learning round R.
Output: Well-trained parameter Θ^R.

Train a model from the scratch through any DA method using $\{D_s^k\}_{k=1}^K$ and $\mathcal{D}_t^0$.
#Global Step
for $r = 1 : R$ **do**
  For each $x_t^j \in \mathcal{D}_t^{r-1}$, calculate the gradient embedding $g_{\hat{y}}\left(x_t^j\right)$ by Eq. (1).
  Cluster $\mathcal{D}_t^{r-1}$ into B clusters ($C_s^1, \ldots, C_s^B$) by Eq. (2).
  Select $\alpha\%$ for $\mathcal{D}_t^{r-1}$ in each C_s^b cluster with high normalization of $g_{\hat{y}}\left(x_t^j\right)$: $\mathcal{D}_{t,b}^{r-1}$.
  Assign each $x_s^i \in \{D_s^i\}_{i=1}^{n_S}$ into these B clusters according to their feature embedding space ($\mathcal{M}_s^i$).
  Selected target sample at round r: $\mathcal{X}_t^r = \emptyset$.
  #Local Step
  for $b = 1 : B$ **do**
    Cluster $\{D_s^i\}_{i=1}^k$ that belong to C_s^b according to the domain label with k centroids ($\mathcal{U}_b^1, \ldots, \mathcal{U}_b^K$).
    For each $x_t^j \in \mathcal{D}_{t,b}^{r-1}$, calculate the distance between x_t^j and the centroid $\mathcal{U}_b^k$ by Eq. (4).
    Calculate the metric $v\left(x_t^j\right)$ according to Eq. (6).
    Select the sample $x_{t,b}^r$ with highest value of v among $\mathcal{D}_{t,b}^{r-1}$ in each C_s^b.
    $\mathcal{X}_t^r = \mathcal{X}_t^r \cup x_{t,b}^r$.
  end for
  $\mathcal{D}_t^r = \mathcal{D}_t^{r-1} / \mathcal{X}_t^r$
  $\mathcal{D}_{st}^r = \mathcal{D}_{st}^{r-1} \cup \mathcal{X}_t^r$
end for
Continue training a model from Θ^R through any DA method using $\{D_s^k\}_{k=1}^K$, $\mathcal{D}_t^R$ and $\mathcal{D}_{st}^R$.
return $\Theta^{R*} = \Theta^R$.

It is worth noting that GALA serves only as a query strategy and does not change the optimization pipeline of the underlying DA method. For all DA methods, including DANN, M^{3}SDA, SIG, and CSR, we adopt the same active training protocol. At the rth active learning round, GALA selects a batch of target samples X_t^r for annotation. After receiving their labels, these samples are moved from the unlabeled target pool $\mathcal{D}_t^{r-1}$ to the selected labeled target set $\mathcal{D}_{st}^r$. The model is then continuously trained with the same DA framework, rather than through a separate fine-tuning stage. For all these variants, we maintain the same supervised loss for the source samples (L_s) as well as the transfer loss (L_{tr}) as used in their original implementations, ensuring a fair comparison. When target-domain annotations become available through the active querying process, we further introduce a supervised loss term on the selected target samples (L_{st}), which guides the model toward more discriminative and domain-invariant representations. Specifically, the original supervised loss on the labeled source domains, denoted as L_s, and the original transfer loss of each DA method, denoted as L_{tr}, are kept unchanged. The selected target samples are still regarded as target-domain samples for computing the DA loss, while their annotations further enable a supervised cross-entropy loss on the target domain, denoted as L_{st}. Therefore, the overall training objective can be written as:

$$L = L_s + \lambda_{tr} L_{tr} + \lambda_{st} L_{st}, \tag{8}$$

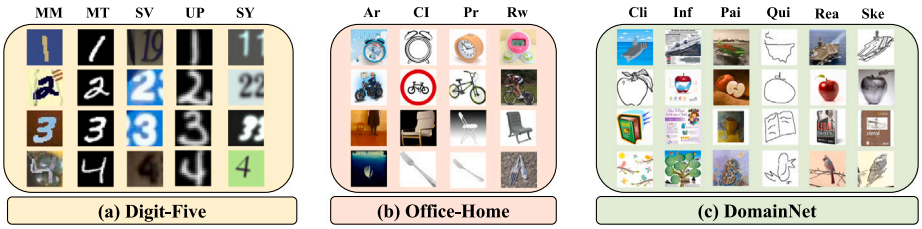

Fig. 4. Examples for each dataset: (a) Digit-Five: Images from MNIST-M (MM), MNIST (MT), SVHN (SV), USPS (UP), and Synthetic Digits (SY). (b) Office-Home: Sample images from Artistic (Ar), Clip-Art (Cl), Product (Pr), and Real-World (Rw). (c) DomainNet: Visuals from Clipart (Cli), Infograph (Inf), Painting (Pai), Quickdraw (Qui), Real (Rea), and Sketch (Ske).

where L_{tr} follows the original design of the corresponding DA method, and L_{st} is computed only on the selected labeled target samples. We use the same training schedule, active query rounds, and annotation budget for all methods to ensure a fair comparison.

It is worth emphasizing that GALA can be seamlessly incorporated into a wide range of DA frameworks without introducing any additional trainable parameters. Despite its lightweight nature, it consistently improves target-domain classification accuracy across different architectures and adaptation paradigms, demonstrating its strong generalizability and efficiency compared to existing other ADA approaches.

4. Experiments

4.1. Datasets

We conduct comprehensive evaluations on three standard benchmarks of domain adaptation: **Digit-Five**, **Office-Home** and **DomainNet**, with dataset samples visualized in Fig. 4. **Digit-Five** [28] consists of 25,000 training images and 9000 testing images in each domain: MNIST-M (**MM**), MNIST (**MT**), SVHN (**SV**), USPS (**UP**) and Synthetic Digits (**SY**). **Office-Home** [46] serves as a more demanding dataset, which collects 15,500 images in total from up to 65 categories of everyday objects. There are four significantly different domain: Artistic (**Ar**), Clip-Art (**Cl**), Product (**Pr**), and Real-World (**Rw**) images. **DomainNet** [29] stands out as an extremely large-scale and difficult DA benchmark, which involves six diverse domains: Clipart (**Cli**), Infograph (**Inf**), Painting (**Pai**), Quickdraw (**Qui**), Real (**Rea**) and Sketch (**Ske**). In total, this benchmark contains nearly 600,000 images across training and testing splits, and covers 345 distinct object classes.

4.2. Training details

Our methods are implemented in PyTorch [47]. We use LeNet [48] for **Digit-Five**, and ImageNet-pretrained ResNet-50/ResNet-101 [49, 50] for **Office-Home** and **DomainNet**, respectively. Modules initialized from scratch use a learning rate ten times larger than that of the pretrained backbone. All models are optimized by mini-batch SGD with momentum 0.95 and progressive training. The active learning process contains five stages at epochs $\{10, 12, 14, 16, 18\}$, with the annotation budget evenly distributed across rounds. Following Section 4.5, we set $\alpha = 60\%$ and use an active ratio of 1% in all experiments and ablation studies. Since **Office-Home** has no test images, the fully supervised target-domain reference is reported only for **Digit-Five** and **DomainNet**.

For fair comparison, all active selection methods follow the same MS-ADA protocol. Since most ADA baselines are designed for single-source ADA, we adapt them to the multi-source setting by applying their selection criteria on the same MSDA backbone. Each method selects the same number of target samples from the same unlabeled target pool after training on all labeled source domains, and the selected samples are then annotated and used for subsequent training under the same optimization schedule. All methods share the same source/target splits, active ratio, query rounds, warm-up epochs, total training epochs, optimizer, batch size, learning rate schedule, and data augmentation.

State-of-the-art. LeNet [48] and ResNet [49] are used as baseline backbones without domain adaptation or active learning. To evaluate the generality of GALA, we integrate it into representative DA/MSDA methods, including DANN [15], M^{3}SDA [29], Subspace Identification Guarantee (SIG) [32], and Cycle Self-Refinement (CSR) [45]. Specifically, DANN is a classical adversarial DA baseline, M^{3}SDA reduces multi-source discrepancies by moment matching, SIG separates domain-agnostic and domain-specific features, and CSR iteratively exploits transferable information from individual sources.

We further compare GALA with three AL methods, including Random, BADGE [43], and APFWA [51]; five SS-ADA methods, including AADA [10], CLUE [39], SDM [11], DUC [52], and D^{3}GU [44]; and one MS-ADA method, Detective [13]. These methods cover random sampling, uncertainty-diversity-based active learning, adversarial active adaptation, clustering-based selection, margin-based selection, calibration-based selection, multi-target ADA, and MS-ADA. Since D^{3}GU is originally designed for multi-target domains, we set its domain number to 1 in our experiments. Detective is evaluated on Digit-Five, Office-Home, and DomainNet as the closest existing MS-ADA baseline. The active ratio is set to 1% in all experiments and ablation studies.

4.3. Results

Experiments on Digit-Five. As listed in Table 1, our approach GALA overpasses all other active learning strategies in different domain adaptation scenarios by at least 1.2% improvement on average. Among all methods, SDM achieves the second-highest accuracy across all tasks, performing the strongest as a comparison approach. However, our method still demonstrates superior performance over SDM especially in challenging tasks. For example, with M^{3}SDA method, the accuracy of GALA is 1.5% and 1.2% higher than that of SDM on tasks **MM** and **SV**, respectively. GALA approach achieves the state-of-the-art in the MS-ADA setting with M^{3}SDA method and is very close to Fully-Supervised. It is remarkable that GALA achieves higher accuracy on transfer tasks of **MT**, **UP** and **SY** compared to Full-Supervised.

Experiments on Office-Home. The accuracy on **Office-Home** are summarized in Table 2. It is evident that our GALA achieves the highest average accuracy (76.8%, 76.7% and 80.1%) compared to other active learning strategies in these three methods. Notably, when benchmarked against **Detective** which is dedicated to the MS-ADA scenario, GALA still shows clear advantages on this dataset. When datasets become more complex, the effectiveness of most active learning strategies begins to diminish. For example, SDM, which performs second best on the **Digit-Five** dataset, does not perform as well as AADA in several scenarios. However, our method GALA is competent at maintaining the efficiency. Especially, GALA promotes the average accuracy by at least 2.2% for all methods. In the most challenging task **Cl**, GALA is the only strategy with more than 60% accuracy and surpasses 2.1% over the current leading active learning strategy CLUE in backbone setting. GALA outperforms all existing active learning strategies with M^{3}SDA

Table 1
Overall accuracy (%) on **Digit-Five** for state-of-the-art methods. We report the mean and standard deviation over three repeated runs. In all experiments, LeNet serves as the underlying architecture and we set the active ratio as 1%.

Method	Active learning strategy	AL or ADA	MM	MT	UP	SV	SY	Avg
LeNet [Proc. IEEE 1998]	None	–	63.4 ± 0.4	90.5 ± 0.6	88.7 ± 0.3	63.5 ± 0.8	82.4 ± 0.6	77.7 ± 0.6
	Random	AL	82.0 ± 1.1	95.8 ± 0.2	97.0 ± 0.4	72.0 ± 0.6	86.5 ± 0.4	86.7 ± 0.5
	BADGE [ICLR 2020]	AL	82.3 ± 0.9	96.2 ± 0.1	96.8 ± 0.1	74.9 ± 0.6	85.4 ± 0.3	87.1 ± 0.4
	APFWA [CVPR 2023]	AL	83.3 ± 0.7	97.1 ± 0.1	96.5 ± 0.2	74.3 ± 0.4	84.9 ± 0.2	87.2 ± 0.2
	AADA [WACV 2020]	ADA	84.5 ± 0.2	97.3 ± 0.3	97.9 ± 0.1	78.5 ± 0.2	87.0 ± 0.3	89.0 ± 0.2
	CLUE [ICCV 2021]	ADA	82.7 ± 0.3	96.7 ± 0.2	96.3 ± 0.2	74.6 ± 0.1	87.2 ± 0.5	87.5 ± 0.3
	SDM [CVPR 2022]	ADA	86.3 ± 0.5	99.1 ± 0.1	98.5 ± 0.4	83.1 ± 0.2	90.2 ± 0.1	91.4 ± 0.2
	DUC [ICLR 2023]	ADA	85.3 ± 0.1	98.6 ± 0.1	98.1 ± 0.2	79.6 ± 0.3	88.2 ± 0.5	90.0 ± 0.2
	D^3GU [WACV 2024]	ADA	86.8 ± 0.8	98.1 ± 0.1	97.2 ± 0.3	80.9 ± 0.2	89.6 ± 0.1	90.5 ± 0.3
	GALA (Ours)	ADA	**87.6 ± 0.1**	**99.3 ± 0.1**	**99.0 ± 0.4**	**85.6 ± 0.2**	**95.1 ± 0.2**	**93.3 ± 0.2**
DANN [JMLR 2016]	None	–	70.8 ± 0.8	97.9 ± 0.2	93.5 ± 0.3	67.8 ± 0.3	86.9 ± 0.6	83.4 ± 0.5
	Random	AL	78.6 ± 0.4	97.1 ± 0.1	96.1 ± 0.4	73.6 ± 0.6	84.4 ± 0.1	86.0 ± 0.3
	BADGE [ICLR 2020]	AL	80.0 ± 0.0	96.8 ± 0.2	96.9 ± 0.1	74.5 ± 0.8	86.8 ± 0.5	87.0 ± 0.3
	APFWA [CVPR 2023]	AL	79.6 ± 0.5	98.0 ± 0.1	95.6 ± 0.2	74.7 ± 0.2	85.1 ± 0.5	86.6 ± 0.2
	AADA [WACV 2020]	ADA	83.2 ± 0.1	97.9 ± 0.0	97.2 ± 0.2	78.2 ± 0.1	87.1 ± 0.3	88.7 ± 0.1
	CLUE [ICCV 2021]	ADA	80.6 ± 0.7	94.1 ± 0.4	96.6 ± 0.3	69.3 ± 0.6	85.9 ± 0.1	85.3 ± 0.4
	SDM [CVPR 2022]	ADA	89.0 ± 0.1	98.9 ± 0.3	98.7 ± 0.2	84.2 ± 0.5	93.3 ± 0.0	92.8 ± 0.2
	DUC [ICLR 2023]	ADA	86.6 ± 0.3	96.9 ± 0.2	97.0 ± 0.1	86.1 ± 0.6	87.8 ± 0.2	90.9 ± 0.3
	D^3GU [WACV 2024]	ADA	87.4 ± 0.7	96.7 ± 0.0	97.3 ± 0.3	83.5 ± 0.3	89.6 ± 0.4	90.9 ± 0.4
	GALA (Ours)	ADA	**91.9 ± 0.2**	**99.2 ± 0.0**	**99.1 ± 0.3**	**87.0 ± 0.2**	**95.3 ± 0.3**	**94.5 ± 0.2**
M^3SDA [ICCV 2019]	None	–	72.8 ± 0.3	98.4 ± 0.2	96.1 ± 0.2	81.3 ± 0.6	89.6 ± 0.3	87.6 ± 0.3
	Random	AL	80.9 ± 1.0	98.6 ± 0.2	96.5 ± 0.1	82.5 ± 0.4	89.9 ± 0.4	89.7 ± 0.4
	BADGE [ICLR 2020]	AL	82.6 ± 0.5	98.8 ± 0.2	97.5 ± 0.0	82.5 ± 0.8	92.1 ± 0.7	90.7 ± 0.4
	APFWA [CVPR 2023]	AL	83.1 ± 0.3	98.6 ± 0.1	97.3 ± 0.2	83.1 ± 0.4	91.7 ± 0.3	90.8 ± 0.1
	AADA [WACV 2020]	ADA	84.6 ± 0.1	98.6 ± 0.4	97.1 ± 0.1	82.1 ± 0.3	90.9 ± 0.3	90.7 ± 0.2
	CLUE [ICCV 2021]	ADA	82.0 ± 0.5	98.2 ± 0.4	96.5 ± 0.6	83.7 ± 0.1	94.3 ± 0.1	90.9 ± 0.3
	SDM [CVPR 2022]	ADA	91.8 ± 0.4	98.7 ± 0.2	98.9 ± 0.2	86.7 ± 0.6	97.0 ± 0.1	94.6 ± 0.3
	DUC [ICLR 2023]	ADA	88.6 ± 0.4	98.4 ± 0.2	97.4 ± 0.2	84.3 ± 0.5	94.2 ± 0.4	92.6 ± 0.3
	D^3GU [WACV 2024]	ADA	89.4 ± 0.6	98.2 ± 0.1	97.8 ± 0.3	85.6 ± 0.8	95.1 ± 0.2	93.2 ± 0.4
	GALA (Ours)	ADA	**93.3 ± 0.7**	**99.3 ± 0.3**	**99.2 ± 0.4**	**87.9 ± 0.1**	**97.2 ± 0.5**	**95.4 ± 0.4**
Detective [CVPR 2024]		ADA	80.0 ± 1.7	97.3 ± 0.5	96.8 ± 0.5	88.5 ± 0.0	99.2 ± 0.2	92.4 ± 0.4
Full-Supervised			94.7 ± 0.4	99.0 ± 0.1	99.2 ± 0.2	87.6 ± 0.2	97.0 ± 0.1	95.5 ± 0.2

method except with DANN on Pr and achieves the state-of-the-art on the MS-ADA setting.

Experiments on DomainNet. The results on **DomainNet**, the most difficult dataset to date, are listed on Table 3. Our proposed GALA improves the average accuracy by at least 0.7% (CLUE [39] vs. GALA in ResNet-101) compared to any other active learning strategy with any method listed in this table. When our transfer tasks become complex and diverse, different methods exhibit distinct advantages. Each comparative method, except for random, achieves the highest accuracy in a specific task. However, our method remains superior, achieving optimal performance in 14 out of 18 transfer scenarios with different methods. In addition, it is encouraging that GALA overpasses SDM with 3.8% and 11.1% improvement in two of the most difficult tasks **Inf** and **Qui**, respectively. Furthermore, GALA with M^3SDA is higher (with 1.9% and 1.2%) than Full-supervised in **Cli** and **Pai**.

4.4. Computational complexity

As shown in Table 4, we use Office-Home with Art as the target domain for comparison, and take Random as the reference baseline. Here, N denotes the size of the unlabeled target pool, C denotes the number of classes, d denotes the feature dimension, B denotes the query budget, and I denotes the number of k-means iterations. For GALA, α denotes the retained uncertain-sample ratio after global clustering, which is set to 60% in our experiments. Although GALA is slower than simple uncertainty-based methods such as SDM and AADA, it is more efficient than BADGE, CLUE, and D^3GU. More importantly, GALA introduces no additional trainable parameters and does not change the training pipeline. Therefore, the overall training cost remains comparable to the underlying DA backbone, with only a moderate additional cost during active sample selection.

4.5. Sensitive analysis of α

As introduced in Section 3.2, after the global step, we first select a set of α% (the only hyper-parameter in GALA.) candidates with the highest uncertainty calculated by Eq. (1) for each cluster C_b. Fig. 5(a) and (b) displays the performance under different α for **Digit-Five** and **Office-Home**. We can observe that when $\alpha = 0.6$, that means we select top 60% target samples with high uncertainty for each cluster (C_b) could achieve higher performance. This strategy filters those target samples with low uncertainty but large domain gap, which actually already have high confidence in label prediction.

5. Discussions

5.1. Analysis of distance metrics

To validate our choice of metric $\mu/\sqrt{\sigma^2+\epsilon}$, we benchmark it against two alternative distance measures. The first method is $\left|\mu_s^i - \mu_t^j\right|$ while the second is Wasserstein distance [53]: $\left\|\mu_s^i - \mu_t^j\right\|_2^2 + \left((\sigma_s^i)^2 + \left(\sigma_t^j\right)^2 - 2\sigma_s^i\sigma_t^j \right)$. Table 5 details the comparative results of these distance metrics under the same conditions. We can observe that $\mu/\sqrt{\sigma^2+\epsilon}$ achieves higher overall average accuracy (mean accuracy for four average accuracies of four DA datasets) than the other two metrics. A plausible explanation is that relying solely on mean differences maybe that the distance μ cannot capture the distributional variance σ, while the Wasserstein distance lacks a mechanism to effectively balance the relative influence of means versus variances. On the other hand, our distance calculation scheme is consistent with BN [54], representing the distance between means with variance normalization and performing the best results among these three distance ways. Moreover, empirical

Table 2
Overall accuracy (%) on **Office-Home** for state-of-the-art methods. We report the mean and standard deviation over three repeated runs. In all experiments, ResNet-50 serves as the underlying architecture and we set the active ratio as 1%.

Method	Strategy	AL/ADA	Ar	Cl	Pr	Rw	Avg
ResNet-50 [CVPR 2016]	None	–	64.6 ± 0.6	52.3 ± 0.6	77.6 ± 0.1	80.7 ± 0.4	68.8 ± 0.4
	Random	AL	71.1 ± 0.5	59.0 ± 1.0	82.5 ± 0.5	82.6 ± 0.2	73.8 ± 0.6
	BADGE [ICLR 2020]	AL	71.3 ± 0.8	59.1 ± 1.0	81.2 ± 0.2	81.7 ± 0.1	73.3 ± 0.5
	AADA [WACV 2020]	ADA	70.2 ± 0.3	58.9 ± 0.4	83.0 ± 0.7	82.7 ± 0.6	73.7 ± 0.5
	CLUE [ICCV 2021]	ADA	71.7 ± 0.4	59.9 ± 0.5	82.2 ± 0.4	81.9 ± 0.1	73.9 ± 0.3
	SDM [CVPR 2022]	ADA	72.5 ± 0.6	58.6 ± 0.5	82.9 ± 0.2	83.7 ± 0.7	74.4 ± 0.5
	D^{3}GU [WACV 2024]	ADA	70.9 ± 0.3	57.6 ± 0.7	81.3 ± 0.1	81.8 ± 0.4	72.9 ± 0.4
	GALA (Ours)	ADA	**74.9 ± 0.2**	**62.0 ± 0.2**	**84.9 ± 0.1**	**85.4 ± 0.1**	**76.8 ± 0.1**
DANN [JMLR 2016]	None	–	64.3 ± 0.6	58.0 ± 0.1	76.4 ± 0.0	78.8 ± 0.4	69.4 ± 0.3
	Random	AL	69.3 ± 0.4	60.5 ± 0.5	79.4 ± 0.9	80.7 ± 0.0	72.5 ± 0.5
	BADGE [ICLR 2020]	AL	69.0 ± 0.6	59.2 ± 0.7	79.8 ± 0.2	79.8 ± 0.1	72.0 ± 0.4
	AADA [WACV 2020]	ADA	72.1 ± 0.9	59.0 ± 0.2	**84.6 ± 0.4**	82.5 ± 0.4	74.5 ± 0.5
	CLUE [ICCV 2021]	ADA	68.8 ± 0.1	60.7 ± 0.3	79.7 ± 0.6	80.0 ± 0.5	72.3 ± 0.4
	SDM [CVPR 2022]	ADA	72.5 ± 0.1	63.0 ± 0.6	80.6 ± 0.1	81.6 ± 0.8	74.4 ± 0.4
	D^{3}GU [WACV 2024]	ADA	71.5 ± 0.5	62.4 ± 0.6	81.4 ± 0.2	80.7 ± 0.1	74.0 ± 0.3
	GALA (Ours)	ADA	**74.2 ± 0.0**	**64.5 ± 0.7**	83.6 ± 0.1	**84.3 ± 0.2**	**76.7 ± 0.2**
M^{3}SDA [ICCV 2019]	None	–	66.2 ± 0.5	58.5 ± 0.3	79.5 ± 0.2	81.3 ± 0.3	71.4 ± 0.3
	Random	AL	70.9 ± 0.4	63.1 ± 0.1	83.5 ± 0.6	85.4 ± 0.3	75.7 ± 0.4
	BADGE [ICLR 2020]	AL	72.8 ± 0.6	64.2 ± 0.6	84.9 ± 0.1	85.3 ± 0.6	76.8 ± 0.5
	AADA [WACV 2020]	ADA	74.3 ± 0.3	65.0 ± 0.2	86.1 ± 0.6	87.4 ± 0.0	78.2 ± 0.3
	CLUE [ICCV 2021]	ADA	75.8 ± 0.5	66.1 ± 0.3	85.6 ± 0.5	86.9 ± 0.1	78.6 ± 0.3
	SDM [CVPR 2022]	ADA	73.5 ± 0.3	66.2 ± 0.4	85.9 ± 0.9	86.1 ± 0.6	77.9 ± 0.6
	D^{3}GU [WACV 2024]	ADA	73.1 ± 0.4	65.5 ± 0.4	86.2 ± 0.2	85.3 ± 0.1	77.5 ± 0.3
	GALA (Ours)	ADA	**76.8 ± 0.4**	**67.9 ± 0.1**	**87.5 ± 0.3**	**88.3 ± 0.4**	**80.1 ± 0.3**
SIG [NeurIPS 2023]	None	–	76.4 ± 0.1	63.9 ± 0.4	85.4 ± 0.6	85.8 ± 0.4	77.9 ± 0.4
	Random	AL	78.2 ± 0.5	64.8 ± 0.2	86.1 ± 0.5	86.6 ± 0.8	78.9 ± 0.5
	BADGE [ICLR 2020]	AL	77.9 ± 0.7	65.7 ± 0.7	87.6 ± 0.3	87.9 ± 0.7	79.8 ± 0.6
	AADA [WACV 2020]	ADA	79.9 ± 0.7	65.9 ± 0.4	87.0 ± 0.6	87.1 ± 0.1	80.0 ± 0.5
	CLUE [ICCV 2021]	ADA	78.1 ± 0.2	67.0 ± 0.3	87.1 ± 0.8	88.4 ± 0.2	80.2 ± 0.4
	SDM [CVPR 2022]	ADA	**81.5 ± 0.6**	66.7 ± 0.7	87.9 ± 0.2	88.5 ± 0.2	81.2 ± 0.4
	D^{3}GU [WACV 2024]	ADA	79.3 ± 0.5	65.7 ± 0.2	86.6 ± 0.6	88.2 ± 0.2	80.0 ± 0.4
	GALA (Ours)	ADA	81.1 ± 0.3	**67.2 ± 0.4**	**88.2 ± 0.6**	**89.1 ± 0.3**	**81.4 ± 0.4**
CSR [AAAI 2024]	None	–	76.7 ± 0.2	71.4 ± 0.3	86.8 ± 0.3	85.5 ± 0.6	80.1 ± 0.3
	Random	AL	77.3 ± 0.6	70.6 ± 0.3	86.9 ± 0.2	86.9 ± 0.7	80.4 ± 0.5
	BADGE [ICLR 2020]	AL	78.7 ± 0.4	72.6 ± 0.3	87.4 ± 0.2	87.0 ± 0.2	81.4 ± 0.3
	AADA [WACV 2020]	ADA	79.2 ± 0.6	72.3 ± 0.2	88.3 ± 0.4	87.5 ± 0.3	81.8 ± 0.4
	CLUE [ICCV 2021]	ADA	79.8 ± 0.4	74.0 ± 0.7	87.1 ± 0.1	88.2 ± 0.1	82.3 ± 0.3
	SDM [CVPR 2022]	ADA	79.5 ± 0.3	73.4 ± 0.4	87.8 ± 0.6	88.2 ± 0.2	82.2 ± 0.4
	D^{3}GU [WACV 2024]	ADA	78.2 ± 0.1	72.6 ± 0.8	87.4 ± 0.3	87.9 ± 0.5	81.5 ± 0.4
	GALA (Ours)	ADA	**84.2 ± 0.1**	**75.3 ± 0.4**	**88.6 ± 0.4**	**89.6 ± 0.1**	**84.4 ± 0.2**
Detective [CVPR 2024]		ADA	81.2 ± 1.4	73.5 ± 2.3	85.0 ± 0.3	86.6 ± 1.4	81.6 ± 0.8

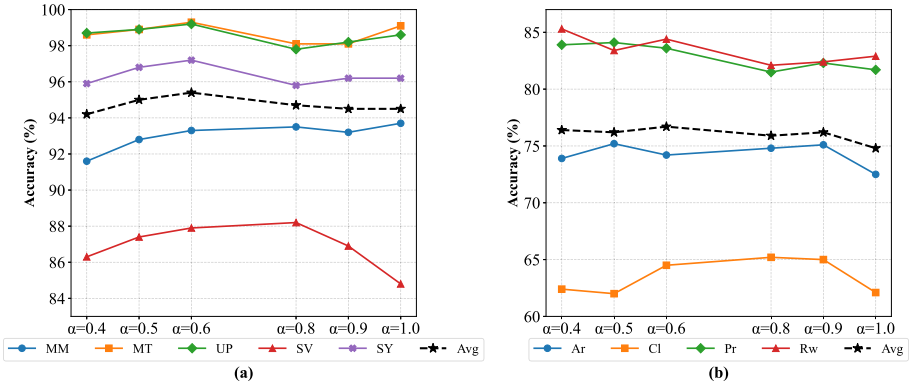

Fig. 5. Sensitive analysis of α introduced in Section 3.2: (a) Performance on different value of α for **Digit-Five**(b) Performance on different value of α for **Office-Home**.

evidence suggests that GALA achieves the best performance without any module of domain loss. By embedding GALA to ResNet, M^{3}SDA has a narrower margin of improvement compared to DANN. This can likely be attributed to that M^{3}SDA serves as a highly robust DA method for MSDA setting performs nearer to the theoretical ceiling and has limited space to enhance.

5.2. *Feature map ($\mathcal{M}\left(x_t^j\right)$) or gradient embedding ($g_c\left(x_t^j\right)$)?*

Gradient embedding ($g_c\left(x_t^j\right)$) is widely used in conventional active learning algorithm [43], which can be represented as the uncertainty of the sample. Also, feature map ($\mathcal{M}\left(x_t^j\right)$) enables the calculation of

Table 3
Overall accuracy (%) on **DomainNet** for SOTA methods. In all experiments, ResNet-101 serves as the underlying architecture and we set the active ratio as 1%.

Method	Active learning strategy	AL or ADA	Cli	Inf	Pai	Qui	Rea	Ske	Avg
ResNet-101 [CVPR 2016]	None	–	47.6	13.0	38.1	13.3	51.9	33.7	32.9
	Random	AL	61.5	22.9	52.6	19.7	59.7	36.9	42.2
	BADGE [ICLR 2020]	AL	61.3	22.9	51.9	**20.0**	62.4	40.2	43.1
	APFWA [CVPR 2023]	AL	60.9	21.2	51.5	19.3	58.6	38.4	41.7
	AADA [WACV 2020]	ADA	60.8	20.7	49.3	15.4	61.7	42.6	41.8
	CLUE [ICCV 2021]	ADA	61.6	**23.1**	52.5	17.9	64.0	54.4	45.6
	SDM [CVPR 2022]	ADA	57.6	19.7	47.3	14.9	62.5	46.6	41.4
	DUC [ICLR 2023]	ADA	60.9	20.1	49.7	16.5	63.1	53.8	44.0
	D^3GU [WACV 2024]	ADA	59.2	19.9	48.1	16.2	62.4	47.8	42.3
	GALA (Ours)	ADA	**63.4**	22.7	**53.4**	18.7	**64.8**	**54.8**	**46.3**
DANN [JMLR 2016]	None	–	45.5	13.1	37.0	13.2	48.9	31.8	31.6
	Random	AL	61.9	24.5	53.1	21.1	66.1	53.0	46.6
	BADGE [ICLR 2020]	AL	64.3	23.9	55.8	23.4	65.3	54.6	47.9
	APFWA [CVPR 2023]	AL	63.1	22.7	53.2	22.9	63.2	54.9	46.7
	AADA [WACV 2020]	ADA	65.3	23.6	**56.1**	20.9	66.2	55.6	48.0
	CLUE [ICCV 2021]	ADA	66.1	28.4	54.7	21.8	61.5	57.2	48.3
	SDM [CVPR 2022]	ADA	63.1	28.4	53.1	28.6	63.3	58.4	49.2
	DUC [ICLR 2023]	ADA	63.7	28.6	53.3	29.1	63.7	58.6	49.5
	D^3GU [WACV 2024]	ADA	64.5	27.9	52.2	28.9	64.1	58.5	49.4
	GALA (Ours)	ADA	**67.9**	**29.5**	52.7	**29.4**	**67.0**	**59.7**	**51.0**
M^3SDA [ICCV 2019]	None	–	58.6	26.0	52.3	6.3	62.7	49.5	42.6
	Random	AL	65.9	29.6	60.1	26.7	72.4	60.1	52.5
	BADGE [ICLR 2020]	AL	70.6	26.3	63.9	30.8	76.3	61.7	54.9
	APFWA [CVPR 2023]	AL	70.1	27.6	63.6	31.6	76.9	59.1	54.8
	AADA [WACV 2020]	ADA	68.2	28.9	62.9	29.6	74.3	59.7	53.9
	CLUE [ICCV 2021]	ADA	69.0	29.1	65.4	32.6	72.9	60.2	54.9
	SDM [CVPR 2022]	ADA	72.1	32.4	64.3	37.1	72.8	**63.9**	57.1
	DUC [ICLR 2023]	ADA	71.6	34.5	65.7	40.1	73.0	62.4	57.9
	D^3GU [WACV 2024]	ADA	72.5	33.1	64.0	39.5	73.6	61.2	57.3
	GALA (Ours)	ADA	**73.5**	**36.2**	**69.8**	**48.2**	**79.6**	63.2	**61.5**
Detective [CVPR 2024]		ADA	69.1	25.7	62.4	36.1	76.2	54.0	53.9
Full-Supervised			71.6	36.7	68.6	69.6	81.8	65.8	65.7

Table 4
Computational complexity and sample selection time of different active selection methods.

Method	Complexity	Time
Random	$O(B)$	1.0×
BADGE	$O(NF + NCd + NBCd)$	4.6×
AADA	$O(NF + NC + ND)$	1.8×
CLUE	$O(NF + INBd)$	4.1×
SDM	$O(NF + NC + N\log N)$	1.5×
DUC	$O(NF + NC)$	2.3×
D^3GU	$O(NF + NG + INBd)$	6.8×
GALA	$O(NF + INBd + \alpha NBKd)$	3.8×

discrepancies linking specific instances or across distinct domains [22, 23]. Therefore, the gradient embedding and the feature map could both use in the global step and the local step. As shown in Tables 6 and 7, we can compare the performance between the feature map $(\mathcal{M}\left(x_t^j\right))$ and the gradient embedding $(g_c\left(x_t^j\right))$ for **Digit-Five** and **Office-Home**, respectively. Experimental results show that when we choose the gradient embedding for the global step and the feature map for the local step, the final average accuracy performs best.

5.3. Performance on different active ratios

Fig. 6(a) and (b) display average accuracy for different active ratios including 1%, 5%, and 10% for Digit-Five and Office-Home, respectively. It is evident that the model performance consistently increases as the active ratio rises, except for M^3SDA + GALA for Digit-Five where it shows no significant change or even slightly decreases. The performance of some transfer tasks exceeds fully-supervised learning, such as M^3SDA + GALA with just 5% labeled target data. Additionally, Fig. 7(a) and (b) illustrate the accuracy gain when increasing the active ratio from 1% to 5% and from 5% to 10% with different methods across various tasks for Digit-Five and Office-Home, respectively. We can observe that in most scenarios, increasing the active ratio leads to a corresponding steady improvement in performance. However, in some cases, increasing the annotation ratio actually results in decreased accuracy, for example, with M^3SDA in SV and SY tasks. Surprisingly, integrating GALA with other DA methods consistently outperforms other active learning or active DA methods with substantial gains and achieves comparable results to the fully supervised target-domain reference with only 1% annotation.

5.4. Average or minimum for calculating d?

Fig. 8 shows the performance in **Office-Home** between the average and the minimum for d in Eq. (6) (see detail in Section 3.3). Empirical results indicates that the minimum method outperforms better than the average way in most cases. We select the minimum distance among $d_j = \min d_{k,j}, k = 1, \ldots, K$. If d_j is large, it means that the target sample (x_t^j) with high uncertainty is far away from any source domain, which maybe crucially important for the target domain. However, if we use the average distance of $d_j = 1/K \sum_{k=1}^{K} d_{k,j}$, although the d_j is large because the target sample is far away from some source domains, once a specific source domain (D_s^k) is small, the target sample (x_t^j) could still acquire great feature representations tailored to that specific source domain. To this end, we select the minimum distance as the domain gap among $d_{k,j}$ for each target sample x_t^j.

5.5. In-depth theoretical analysis

According to classic DA theory [55], the target error $\mathcal{E}_{\mathcal{T}}(h)$ of a hypothesis h satisfies the inequality: $(\mathcal{E}_{\mathcal{T}}(h) \leq \mathcal{E}_{\mathcal{S}}(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T}) + \lambda)$. Here, (i) $\mathcal{E}_{\mathcal{S}}(h)$ means the expected error associated with the source domain; (ii) $\mathcal{H}\Delta\mathcal{H}$-distance $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T})$, a distance metric reflecting the

Table 5
Overall accuracy (%) of distinct distance metrics across four standard DA benchmarks.

Method	Distance type	Digit-Five	Office-Home	DomainNet	Avg
ResNet	None	77.7	68.8	32.9	59.8
GALA	μ	90.9	73.5	42.8	69.1
GALA	Wasserstein	92.6	74.2	46.1	71.0
GALA	$\mu/\sqrt{\sigma^2+\epsilon}$	**93.3**	**76.8**	**47.0**	**72.4**
DANN	None	83.5	69.4	31.6	61.5
DANN + GALA	μ	93.1	74.4	46.8	71.4
DANN + GALA	Wasserstein	94.2	75.9	50.6	73.6
DANN + GALA	$\mu/\sqrt{\sigma^2+\epsilon}$	**94.5**	**76.7**	**51.0**	**74.1**
M^{3}SDA	None	87.6	71.4	42.6	67.2
M^{3}SDA + GALA	μ	94.3	77.0	59.3	76.9
M^{3}SDA + GALA	Wasserstein	94.8	80.0	61.9	78.9
M^{3}SDA + GALA	$\mu/\sqrt{\sigma^2+\epsilon}$	**95.4**	**80.1**	**61.5**	**79.0**
Full-Supervised		95.5	None	65.7	None

Table 6
Performance comparisons between the feature map ($\mathcal{M}(x_t^j)$) and the gradient embedding ($g_c(x_t^j)$) for **Digit-Five**. All methods use the LeNet as the backbone.

Method	Global		Local		MM	MT	UP	SV	SY	Avg
	$\mathcal{M}(x_t^j)$	$g_c(x_t^j)$	$\mathcal{M}(x_t^j)$	$g_c(x_t^j)$						
LeNet + GALA	✓		✓		88.6	98.9	98.5	83.9	94.9	93.0
	✓			✓	89.6	98.3	97.9	84.0	92.5	92.5
		✓	✓		87.6	99.3	99.0	85.6	95.1	93.3
		✓		✓	86.9	98.5	98.2	85.3	95.2	92.8
DANN + GALA	✓		✓		90.0	99.0	99.0	87.1	93.3	93.7
	✓			✓	90.5	98.4	98.7	86.9	94.9	93.9
		✓	✓		91.9	99.2	99.1	87.0	95.3	94.5
		✓		✓	90.1	98.9	98.4	87.2	94.8	93.9
M^{3}SDA + GALA	✓		✓		92.9	99.1	98.8	84.2	96.0	94.2
	✓			✓	93.0	89.7	99.0	86.3	96.8	93.0
		✓	✓		93.3	99.3	99.2	87.9	97.2	95.4
		✓		✓	93.5	99.0	98.4	87.1	96.3	94.9

Table 7
Performance comparisons contrasting the efficacy of feature map ($\mathcal{M}(x_t^j)$) versus the gradient embedding ($g_c(x_t^j)$) for **Office-Home**. All methods use the ResNet-50 as the underlying framework.

Method	Global		Local		Ar	Cl	Pr	Rw	Avg
	$\mathcal{M}(x_t^j)$	$g_c(x_t^j)$	$\mathcal{M}(x_t^j)$	$g_c(x_t^j)$					
ResNet-50 + GALA	✓		✓		73.2	61.8	84.2	84.5	75.9
	✓			✓	73.6	59.1	84.7	83.9	75.3
		✓	✓		74.9	62.0	84.9	85.4	76.8
		✓		✓	72.7	62.7	83.9	85.6	76.2
DANN + GALA	✓		✓		73.9	61.8	83.5	85.3	76.1
	✓			✓	74.8	65.6	83.4	84.7	77.1
		✓	✓		74.2	64.5	83.6	84.4	76.7
		✓		✓	74.0	63.9	83.2	84.1	76.7
M^{3}SDA + GALA	✓		✓		75.6	65.0	87.8	85.1	78.4
	✓			✓	74.3	66.5	86.9	87.4	78.8
		✓	✓		76.8	67.9	87.5	88.3	80.1
		✓		✓	76.2	66.3	86.1	88.6	79.3

divergence between two domains via the disagreement of hypotheses $h, h' \in \mathcal{H}\Delta\mathcal{H}$; and (*iii*) λ denotes the error of the ideal joint hypothesis shared by both source and target domains, which reflects the intrinsic adaptability between domains. In practice, it is commonly approximated using the joint error of a classifier trained on both domains. Inspired by classic domain adaptation theory, we provide an intuitive empirical interpretation of why GALA improves transferability in the MS-ADA setting. This discussion is not intended as a formal theoretical proof, but rather as a qualitative analysis connecting the behavior of GALA with commonly used DA generalization concepts such as the $\mathcal{H}\Delta\mathcal{H}$-distance and the shared-error term. As depicted in Fig. 9(a) and (b), the values of $\mathcal{H}\Delta\mathcal{H}$-distance and λ are visualized, which demonstrate GALA effectively minimizes both the $\mathcal{H}\Delta\mathcal{H}$-distance and the λ, leading a reduction in generalization error and suggesting that GALA reduces cross-domain discrepancy and improves the consistency between source and target predictions. Furthermore, it suggests that GALA successfully harmonizes the model's classification capability with its transferability.

In practice, the $\mathcal{H}\Delta\mathcal{H}$-distance is approximated through domain discrimination performance, while the shared-error term is only analyzed qualitatively since it is not directly computable in the unsupervised target setting.

5.6. Robustness to noisy pseudo-labels

The global clustering step in GALA relies on pseudo-labels to construct gradient embeddings. Therefore, extremely noisy pseudo-labels may potentially affect the clustering quality under severe domain shift. To quantitatively evaluate this issue, we conduct additional stress-test experiments on DomainNet, the most challenging benchmark in

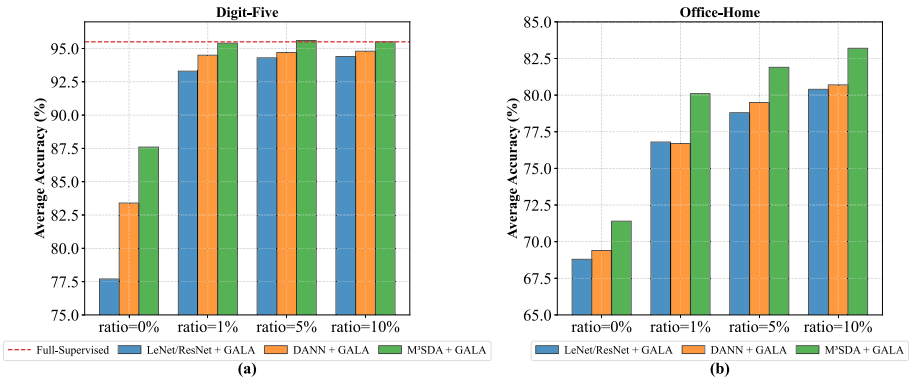

Fig. 6. Average accuracy on different ratios including 1%, 5%, 10% for (a) **Digit-Five** and (b) **Office-Home**.

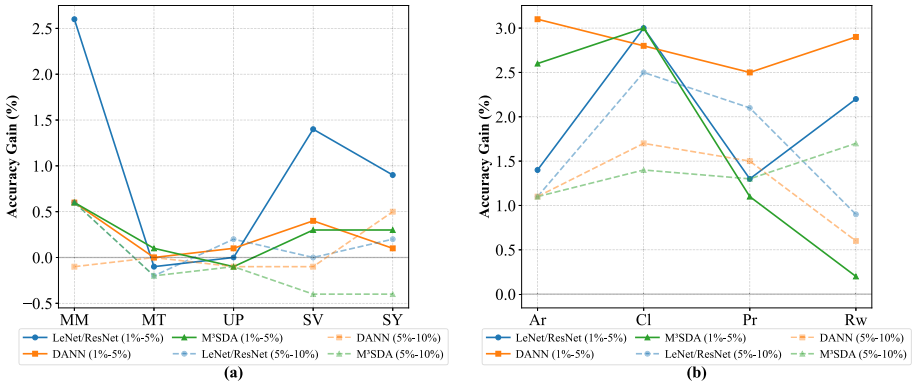

Fig. 7. Accuracy gain when increasing active ratio including 1%–5% and 5%–10% for (a) **Digit-Five** and (b) **Office-Home**.

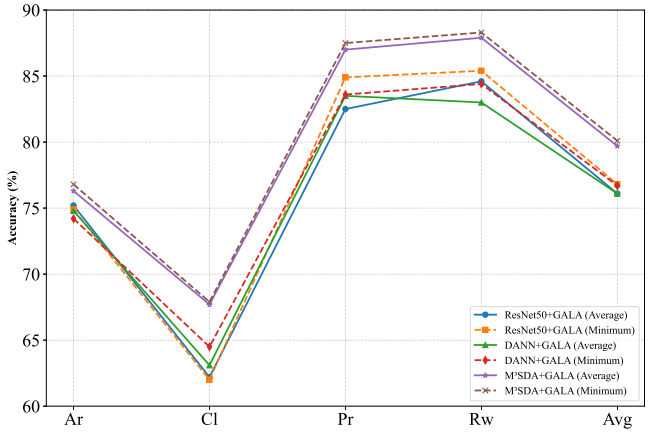

Fig. 8. Comparisons in **Office-Home** between the average and the minimum for d in Eq. (6). All methods use the ResNet-50 as the backbone.

Table 8
Stress-test results on DomainNet under highly noisy pseudo-label settings.

Setting	Method	Inf	Qui
M³SDA + ResNet-50	None	14.6	4.5
	Random	22.4	19.6
	BADGE	22.7	20.4
	AADA	23.6	20.7
	CLUE	23.8	22.0
	SDM	25.7	23.3
	GALA	**30.4**	**32.1**
M³SDA + Gaussian noise	None	19.6	5.9
	Random	24.3	21.4
	BADGE	23.6	23.6
	AADA	24.7	26.0
	CLUE	27.1	26.8
	SDM	27.6	28.1
	GALA	**32.5**	**39.5**

our experiments. We report the accuracy (%) on two difficult target domains, Inf and Qui, with a 1% active annotation budget. "ResNet-50" denotes the setting with a weaker backbone, and "Gaussian noise" denotes the setting where target images are corrupted by $I_{\text{noisy}} = I + \mathcal{N}(0, \sigma^2)$ with $\sigma = 10$ before active selection.

As shown in Table 8, GALA consistently outperforms representative AL and ADA methods under both degraded settings. With ResNet-50, GALA achieves 30.4% and 32.1% accuracy on Inf and Qui, respectively, improving over the strongest competing method SDM by 4.7% and 8.8%. Under Gaussian-noise corruption, GALA achieves 32.5% and 39.5% accuracy, outperforming SDM by 4.9% and 11.4%. These results indicate that although noisy pseudo-labels may affect the gradient embedding, the proposed global-local strategy remains reasonably robust in heavily misaligned settings. This robustness mainly comes from two aspects: the global clustering step encourages diverse sample selection rather than relying only on pseudo-label confidence, while the local selection step further considers uncertainty and multi-source domain relationships, which helps reduce the influence of individual noisy pseudo-labels.

6. Conclusions

In this work, we propose GALA, a simple yet effective GlobAl-LocAl strategy for Multi-Source Active Domain Adaptation (MS-ADA). By combining global k-means clustering with cluster-wise local selection, GALA selects target samples that are both diverse and informative

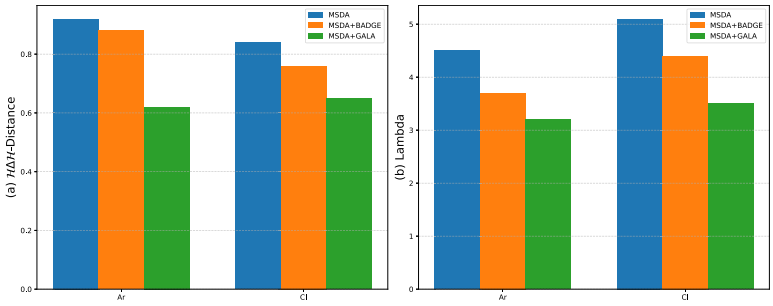

Fig. 9. In-depth theoretical analysis of $\mathcal{H}\Delta\mathcal{H}$-distance and λ.

under multi-source domain variations. GALA is plug-and-play, introduces no additional trainable parameters, and can be seamlessly integrated into different DA architectures. Experiments on three benchmarks demonstrate that GALA consistently improves adaptation performance using only 1% target annotations.

Although GALA shows strong performance across multiple benchmarks and stress-test settings, it still depends on the quality of the feature space and pseudo-labels produced by the current DA model. As shown in Section 5.6, GALA remains reasonably robust when using a weaker ResNet-50 backbone or when target images are corrupted by Gaussian noise. However, if the initial model fails to learn meaningful target-domain representations, highly unreliable pseudo-labels may affect the stability of the global clustering step. Future work will explore pseudo-label denoising, confidence-aware gradient embeddings, and source-free uncertainty calibration to further improve MS-ADA under extreme domain shifts.

CRediT authorship contribution statement

Juepeng Zheng: Writing – original draft, Methodology, Data curation. **Yibin Wen:** Visualization, Data curation. **Peifeng Zhang:** Validation, Investigation. **Zurong Mai:** Writing – review & editing, Validation. **Qingmei Li:** Writing – review & editing, Supervision, Conceptualization. **Yang Zhang:** Validation, Investigation. **Haohuan Fu:** Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant T2125006 and Grant 42401415; in part by Shenzhen Science and Technology Program under Grant KCXFZ20240903093759004 and Grant KJZD20230923115106012; in part by Guangdong Science & Technology Program under Grant 2025B-0101080001.

Data availability

The dataset used in this paper is publicly available.

References

[1] Q. Li, Y. Zhang, J. Zheng, Y. Zhang, J. Huang, H. Fu, Boosting universal domain adaptation in remote sensing with dual-classifiers consistency discrimination and cross-domain feature mixup, IEEE Trans. Geosci. Remote Sens. (2025).

[2] Y. Liang, S. Cao, J. Zheng, X. Zhang, J. Huang, H. Fu, Low saturation confidence distribution-based test-time adaptation for cross-domain remote sensing image classification, Int. J. Appl. Earth Obs. Geoinf. 139 (2025) 104463.

[3] J. Guo, J. Zheng, Y. Liang, Y. Wen, B. Yu, J. Hu, J. Huang, H. Fu, Towards cross-regional land cover Mapping Using Regional consistency and certainty-based active domain adaptation with limited annotations, IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. (2026).

[4] Y. Fu, H. Fu, Y. Lu, M. Zhao, Deep non-convex tensor and higher-order graph embedding for multi-source domain adaptation, Pattern Recognit. 176 (2026) 113170, http://dx.doi.org/10.1016/j.patcog.2026.113170, URL https://www.sciencedirect.com/science/article/pii/S0031320326001354.

[5] H. Guo, R. Pasunuru, M. Bansal, Multi-source domain adaptation for text classification via distancenet-bandits, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 7830–7838.

[6] S. Zhao, B. Li, X. Yue, Y. Gu, P. Xu, R. Hu, H. Chai, K. Keutzer, Multi-source domain adaptation for semantic segmentation, Adv. Neural Inf. Process. Syst. 32 (2019).

[7] S. Wang, Y. Chang, Q. Li, C. Wang, G. Li, M. Mao, Pose-robust personalized facial expression recognition through unsupervised multi-source domain adaptation, Pattern Recognit. 150 (2024) 110311, http://dx.doi.org/10.1016/j.patcog.2024.110311, URL https://www.sciencedirect.com/science/article/pii/S0031320324000621.

[8] H. Zhao, S. Zhang, G. Wu, J.M. Moura, J.P. Costeira, G.J. Gordon, Adversarial multiple source domain adaptation, Adv. Neural Inf. Process. Syst. 31 (2018).

[9] S. Zhao, G. Wang, S. Zhang, Y. Gu, Y. Li, Z. Song, P. Xu, R. Hu, H. Chai, K. Keutzer, Multi-source distilling domain adaptation, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 12975–12983.

[10] J.-C. Su, Y.-H. Tsai, K. Sohn, B. Liu, S. Maji, M. Chandraker, Active adversarial domain adaptation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 739–748.

[11] M. Xie, Y. Li, Y. Wang, Z. Luo, Z. Gan, Z. Sun, M. Chi, C. Wang, P. Wang, Learning distinctive margin toward active domain adaptation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 7993–8002.

[12] B. Xie, L. Yuan, S. Li, C.H. Liu, X. Cheng, G. Wang, Active learning for domain adaptation: An energy-based approach, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, 2022, pp. 8708–8716.

[13] W. Zhang, Z. Lv, H. Zhou, J.-W. Liu, J. Li, M. Li, Y. Li, D. Zhang, Y. Zhuang, S. Tang, Revisiting the domain shift and sample uncertainty in multi-source active domain transfer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 16751–16761.

[14] W. Chen, Y. Wen, J. Zheng, J. Huang, H. Fu, Ban: A universal paradigm for cross-scene classification under noisy annotations from rgb and hyperspectral remote sensing images, IEEE Trans. Geosci. Remote Sens. 63 (2025) 1–13.

[15] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, V. Lempitsky, Domain-adversarial training of neural networks, J. Mach. Learn. Res. 17 (59) (2016) 1–35.

[16] J. Zheng, G. Li, Y. Wen, J. Zhang, R. Dong, H. Fu, Evidential graph contrastive alignment for source-free blending-target domain adaptation, IEEE Trans. Neural Netw. Learn. Syst. (2025).

[17] Q. Li, Y. Wen, J. Zheng, Y. Zhang, H. Fu, Hyunida: Breaking label set constraints for universal domain adaptation in cross-scene hyperspectral image classification, IEEE Trans. Geosci. Remote Sens. 62 (2024) 1–15.

[18] Y. Zou, Z. Yu, B. Kumar, J. Wang, Unsupervised domain adaptation for semantic segmentation via class-balanced self-training, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 289–305.

[19] Y. Wen, D. Gu, J. Zheng, G-SFADA: Gradient-inspired source-free active domain adaptation for semantic segmentation, in: ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2026, pp. 5756–5760, http://dx.doi.org/10.1109/ICASSP55912.2026.11462892.

[20] Y. Guo, H. Yu, S. Xie, L. Ma, X. Cao, X. Luo, DSCA: A dual semantic correlation alignment method for domain adaptation object detection, Pattern Recognit. 150 (2024) 110329, http://dx.doi.org/10.1016/j.patcog.2024.110329, URL https://www.sciencedirect.com/science/article/pii/S0031320324000803.

[21] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Adv. Neural Inf. Process. Syst. 27 (2014).

[22] M. Long, Z. Cao, J. Wang, M.I. Jordan, Conditional adversarial domain adaptation, in: Advances in Neural Information Processing Systems, 2018, pp. 1640–1650.

[23] M. Long, Y. Cao, J. Wang, M. Jordan, Learning transferable features with deep adaptation networks, in: International Conference on Machine Learning, PMLR, 2015, pp. 97–105.

[24] M. Long, H. Zhu, J. Wang, M.I. Jordan, Deep transfer learning with joint adaptation networks, in: Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org, 2017, pp. 2208–2217.

[25] Y. Lu, H. Huang, W.K. Wong, X. Hu, Z. Lai, X. Li, Adaptive dispersal and collaborative clustering for few-shot unsupervised domain adaptation, IEEE Trans. Image Process. (2025).

[26] Y. Lu, Y. Yang, W.K. Wong, A. Toomey, Z. Lai, X. Li, Energy-driven explicit alignment network: A blended-target domain adaptation approach, IEEE Trans. Multimed. (2026).

[27] Y. Lu, Y. Lan, H. Yang, Z. Lai, X. Li, Exploring generic knowledge and reactivating source model for source-free universal domain adaptation, IEEE Trans. Multimed. (2026).

[28] R. Xu, Z. Chen, W. Zuo, J. Yan, L. Lin, Deep cocktail network: Multi-source unsupervised domain adaptation with category shift, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3964–3973.

[29] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, B. Wang, Moment matching for multi-source domain adaptation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1406–1415.

[30] Y. Li, L. Yuan, Y. Chen, P. Wang, N. Vasconcelos, Dynamic transfer for multi-source domain adaptation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 10998–11007.

[31] Y. Fu, M. Zhang, X. Xu, Z. Cao, C. Ma, Y. Ji, K. Zuo, H. Lu, Partial feature selection and alignment for multi-source domain adaptation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 16654–16663.

[32] Z. Li, R. Cai, G. Chen, B. Sun, Z. Hao, K. Zhang, Subspace identification for multi-source domain adaptation, Adv. Neural Inf. Process. Syst. 36 (2023) 34504–34518.

[33] Z. Cai, Y. Huang, T. Zhang, Y. Zheng, D. Yue, Multi-source domain adaptation by causal-guided adaptive multimodal diffusion networks, Int. J. Comput. Vis. 133 (7) (2025) 4623–4645.

[34] Z. Cai, J. Liu, Y. Huang, C. Hu, X. Wu, X.-Y. Jing, L. Shao, Make source-free transfer great: Dynamic uncertainty-driven confident fusion for multi-source-free domain adaptation, Inf. Fusion (2025) 103859.

[35] Y. Lu, H. Huang, X. Hu, Z. Lai, Multiple adaptation network for multi-source and multi-target domain adaptation, IEEE Trans. Multimed. 27 (2025) 5731–5745.

[36] Y. Lu, H. Huang, X. Hu, Style adaptation and uncertainty estimation for multi-source blended-target domain adaptation, Adv. Neural Inf. Process. Syst. 37 (2024) 87042–87060.

[37] J. Peng, M. Sun, E.G. Lim, Q. Wang, J. Xiao, Prototype guided pseudo labeling and perturbation-based active learning for domain adaptive semantic segmentation, Pattern Recognit. 148 (2024) 110203.

[38] J. Chang, Y. Kang, W.X. Zheng, Y. Cao, Z. Li, W. Lv, X.-M. Wang, Active domain adaptation with application to intelligent logging lithology identification, IEEE Trans. Cybern. 52 (8) (2022) 8073–8087.

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

[40] H. Wang, J. Chen, S. Zhang, Y. He, J. Xu, M. Wu, J. He, W. Liao, X. Luo, Dual-reference source-free active domain adaptation for nasopharyngeal carcinoma tumor segmentation across multiple hospitals, IEEE Trans. Med. Imaging 43 (12) (2024) 4078–4090.

[41] H. Wang, W. Chen, X. Luo, Z. Xing, L. Liu, J. Qin, S. Wu, L. Zhu, Toward fair and accurate cross-domain medical image segmentation: A vlm-driven active domain adaptation paradigm, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 24102–24112.

[42] J. Li, H. Wang, W. Wang, J. Qin, Q. Wang, L. Zhu, Source-free active domain adaptation for efficient medical video polyp segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2025, pp. 499–509.

[43] J.T. Ash, C. Zhang, A. Krishnamurthy, J. Langford, A. Agarwal, Deep batch active learning by diverse, uncertain gradient lower bounds, in: International Conference on Learning Representations, 2020.

[44] L. Zhang, L. Xu, S. Motamed, S. Chakraborty, F. De la Torre, D3GU: Multi-target active domain adaptation via enhancing domain alignment, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 2577–2586.

[45] C. Zhou, Z. Wang, B. Du, Y. Luo, Cycle self-refinement for multi-source domain adaptation, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol 38, 2024, pp. 17096–17104.

[46] H. Venkateswara, J. Eusebio, S. Chakraborty, S. Panchanathan, Deep hashing network for unsupervised domain adaptation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5018–5027.

[47] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, Adv. Neural Inf. Process. Syst. 32 (2019).

[48] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324.

[49] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[50] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252.

[51] J.G. Tejero, M.S. Zinkernagel, S. Wolf, R. Sznitman, P. Márquez-Neila, Full or weak annotations? An adaptive strategy for budget-constrained annotation campaigns, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 11381–11391.

[52] M. Xie, S. Li, R. Zhang, C.H. Liu, Dirichlet-based uncertainty calibration for active domain adaptation, in: The Eleventh International Conference on Learning Representations, 2023.

[53] C. Villani, The wasserstein distances, in: Optimal Transport, Springer, 2009, pp. 93–111.

[54] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: International Conference on Machine Learning, 2015, pp. 448–456.

[55] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, J.W. Vaughan, A theory of learning from different domains, Mach. Learn. 79 (1–2) (2010) 151–175.