Revisiting the Domain Shift and Sample Uncertainty in Multi-source Active Domain Transfer

Wenqiao Zhang [1†] Zheqi Lv [1,2†‡]
[1] Zhejiang University, China [2] National University of Singapore, Singapore
{wenqiaozhang, zheqilv}@zju.edu.cn

Abstract

Active Domain Adaptation (ADA) aims to maximally boost model adaptation in a new target domain by actively selecting a limited number of target data to annotate. This setting neglects the more practical scenario where training data are collected from multiple sources. This motivates us to extend ADA from a single source domain to multiple source domains, termed Multi-source Active Domain Adaptation (MADA). Not surprisingly, we find that most traditional ADA methods cannot work directly in such a setting, mainly due to the excessive domain gap introduced by all the source domains. Considering this, we propose a Detective framework that comprehensively considers the domain shift between multi-source domains and target domains to detect the informative target samples. Specifically, the Detective leverages a dynamic Domain Adaptation (DA) model that learns how to adapt the model's parameters to fit the union of multi-source domains. This enables an approximate single-source domain modeling by the dynamic model. We then comprehensively measure both domain uncertainty and predictive uncertainty in the target domain to detect informative target samples using evidential deep learning, thereby mitigating uncertainty miscalibration. Experiments demonstrate that our solution outperforms existing methods by a considerable margin on three domain adaptation benchmarks. Our project is available at https://github.com/wannature/MADA.

1. Introduction

Unsupervised Domain Adaptation (UDA) aims to transfer knowledge learned from labeled data in the original domain (source $\mathcal{D}_s$) to the new domain (target $\mathcal{D}_t$). Although UDA is capable of alleviating the poor generalization of learned deep neural networks when the data distribution significantly deviates from the original domain [44, 49, 54], the unavailability of target labels greatly hinders its performance. This presents a significant gap compared to its supervised counterpart [7, 43]. An appealing way to address this issue is by actively collecting informative target samples within an acceptable budget, thereby maximally benefiting the adaptation model. This promising adaptation paradigm integrates the idea of active learning [37] into traditional UDA, known as Active Domain Adaptation (ADA) [32].

While the existing ADA framework [9, 10, 31, 32, 38] presumes that all labeled training data share the same distribution from a single source domain $\mathcal{D}_s$, it conceals the practicality that data is typically collected from various domains ($\{\mathcal{D}_{s,i}\}_{i=1}^{M}$, with M representing the number of domains) in real-world scenarios [61]. This setting hinders the model's ability to learn from diverse domains, one of the most precious capacity of human that adapting knowledge acquired from varied environments to unseen fields. With this consideration, we introduce a challenging and realistic problem setting termed Multi-source Active Domain Adaptation (MADA). As captured in Figure 1(a), MADA posits the availability of several labeled source domains, capable of annotating a small portion of valuable target samples for maximally benefiting the adaptation process. The proposed MADA, while straightforward and promising, encounters multiple challenges when directly applying conventional ADA techniques to MADA task: **i) Multi-grained Domain Shift.** In ADA, acquiring target labels involves gauging the domain shift between a single $\mathcal{D}_s$ and the target domain $\mathcal{D}_t$ to identify informative target samples. However, the significantly variant data distributions of $\{\mathcal{D}_{s,i}\}_{i=1}^{M}$ also pose challenges for MADA, as it encompasses multiple source domains with disparate distributions, *i.e.*, the alignment among multiple source domains should be carefully addressed for a reliable MADA model. **ii) Uncertainty Miscalibration.** The uncertainty-aware selection criteria in current ADA methods typically rely on predictions from deterministic models, which are prone to miscalibration when faced with data exhibiting distribution shifts [12]. Relying solely on predictive uncertainty proves to be unreliable, as standard DNNs often give misplaced overconfidence in their predictions on target data (Figure 1(c)), potentially impair-

[*] †These authors contributed equally to this research.
[†] ‡Corresponding authors.

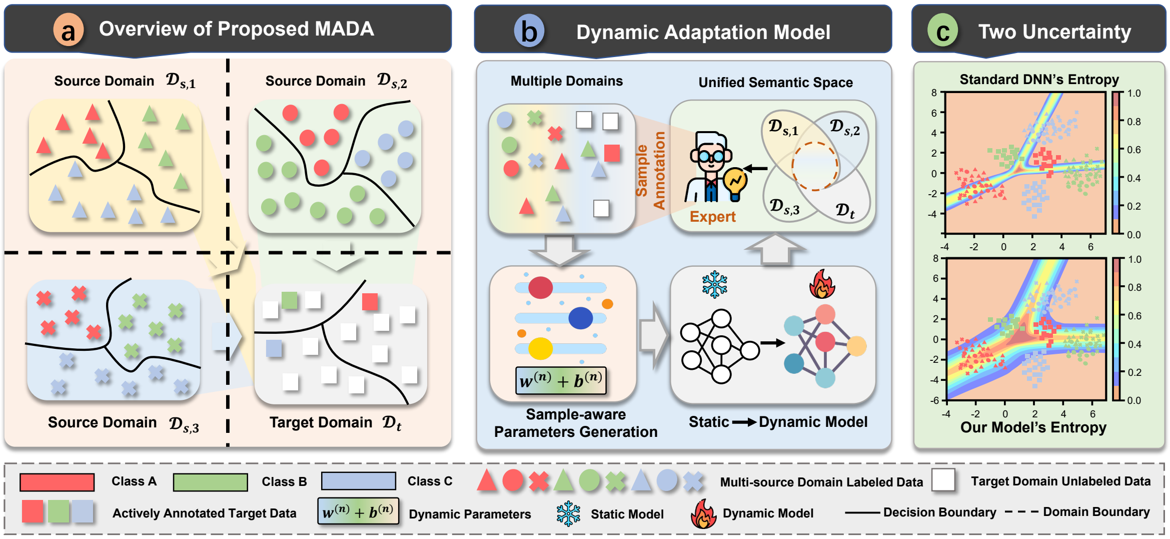


Figure 1. (a) illustrates an overview of the proposed MADA. (b) depicts a concise version of the dynamic adaptation and sample selection strategy. (c) compares the entropy of general DNNs with the entropy of our model, where both models are trained using source data.

ing MADA's performance. In summary, these shortcomings necessitate a thorough reevaluation of both MADA and its corresponding solutions.

To alleviate the aforementioned limitations, we propose a **D**ynamic int**e**gra**te**d un**c**er**t**a**i**nty **v**aluation fram**e**work, abbreviated as **Detective**. Detective comprises the following components: The first module, termed the Universal Dynamic Network (UDN), utilizes a Hypernetwork [13] to derive a domain-agnostic model spanning $\{\mathcal{D}_{s,i}\}_{i=1}^{M}$. As illustrated in Figure 1(b), the key insight is to fit the model towards the union space of $\{\mathcal{D}_{s,i}\}_{i=1}^{M}$ by adapting its parameters. Consequently, the model is divided into static layers (backbone) and dynamic layers (classifier). The static layers maintain fixed parameters, whereas the dynamic layers' parameters (classifier) are dynamically generated depending on multiple source samples. Employing this adaptable domain-agnostic model allows us to treat $\{\mathcal{D}_{s,i}\}_{i=1}^{M}$ as one-source domain. This significantly simplifies the alignment process between $\{\mathcal{D}_{s,i}\}_{i=1}^{M}$ and $\mathcal{D}_t$, as it is no longer necessary to pull all source samples together with the target samples. Next, we develop the integrated uncertainty selector (IUS) derived from Evidential Deep Learning (EDL) [35] to effectively measure the domain's characteristic to the target domain. In EDL, categorical predictions are construed as distributions, with a Dirichlet prior [36] applied to class probabilities, transforming the prediction from a point estimator to a probabilistic one. We regard the EDL's prediction as the *domain uncertainty* as it enables the detection of unfamiliar data instances [28] illustrated in Figure 1(c), *i.e.*, the miscalibration of deterministic's prediction can be mitigated by considering the spectrum of possible outcomes. By amalgamating the predictive uncertainty from a standard DNN, we integrate both domain and predictive uncertainties to select samples that are informative for the target domain. Moreover, we design a contextual diversity calculator (CDC) that evaluates the diversity of chosen target samples at the image-level, guaranteeing the information diversity of selected samples.

To summarize, our contributions are fourfold: (1) We have designed a practical yet challenging task, Multi-source Active Domain Adaptation (MADA). (2) We propose a universal dynamic network capable of adaptively mapping samples from multi-source domains to the adaptation model's parameters. (3) We introduce a novel integrated uncertainty calculation strategy to tackle the MADA challenge. This strategy thoroughly assesses both domain and predictive uncertainties to select informative samples and is complemented by a contextual diversity calculator to enrich the information diversity of the selected data. (4) Experimental studies demonstrate that our method significantly improves upon the prevalent ADA methodologies for MADA.

2. Related Work

Active&Multi Domain Adaptation. To improve the generalization of the model, domain adaptation is receiving more and more attention from researchers [8, 23, 25, 31, 41]. Active Domain Adaptation (ADA) adapts a source model to an unlabeled target domain by using an oracle to obtain the labels of selected target instances. ALDA [32] samples instances based on model uncertainty and learned domain separators and applies them to text data sentiment classification. [38] and [10] introduce ADA into adversarial learning and identify domains through the resulting domain discriminators. However, they may give the same high score for most of the target data, so it is not sufficient to ensure that the selected samples represent the entire target distribution. [9, 31] improves this weakness by selecting active samples by clustering, but they still focus on the measurement of prediction uncertainty like existing ADA methods, which leads to sometimes being misunderstood on the target data. Multi-source domain adaptation(MSDA) aims to learn domain-invariant features across all domains, or leverage auxiliary classifiers trained with multi-source domain to ensemble a robust classifier for the target domain [39, 40]. [17] presents a framework to search for the best initial conditions for MSDA via meta-learning. [24] propose dynamic transfer which adapts the model's parameters for each sample to simplify the alignment between source domains and target domain and address domain conflicts. MDDA [62] considers the distances on not only the domain level but also the sample level, it selects source samples which is similar to the target to finetune the source-specific classifiers. In addition, multi-type data domain adaptation and fusion have also been greatly developed [18, 19, 56, 57]. Despite promising, the aforementioned methods can not directly transfer to the MADA, due to they fail to measure the multi-source domain to select valuable target samples with multi-grained domain shift.

HyperNetwork. HyperNetwork [14] originally aimed to achieve model compression by reducing the number of parameters a model needs to train. It then develops into a neural network that generates parameters for another neural network. Due to its fast and powerful advantages in generalization, research on HyperNetwork has gradually increased. In terms of the theory, [5] studies the parameter initialization of HyperNetwork. In terms of application, HyperNetwork research is widely used in various tasks, such as few-shot learning [4], continual learning [48], graph learning [55], recommendation system [53], device-cloud collaboration [26, 27], large models [42, 60], etc. To the best of our knowledge, we are the first to leverage HyperNetwork to the adaptation problem with unique challenges arising from the problem setup.

Evidential Deep Learning (EDL). Although deep learning models have made surprising progress [20–22, 58, 59], how to effectively evaluate the uncertainty of model predictions is still worth pondering. [35] first proposed to estimate reliable classification uncertainty via EDL. It places Dirichlet priors on class probabilities to interpret classification predictions as distributions and achieves significant advantages in out-of-distribution queries. As the potential of EDL is gradually being explored, [1] introduces the evidential theory to regression tasks by placing evidential priors during training such that the model is regularized when its predicted evidence is not aligned with the correct output. [2] propose a novel model calibration method to regularize the EDL training to estimate uncertainty for open-set action recognition. [6] introduces a class competition-free uncertainty score based on EDL to find potential unknown samples in universal domain adaptation. Differently, we introduce the EDL theory into MADA that effectively measures the domain uncertainty to boost the adaptation.

3. Methodology

3.1. Problem Formulation

Formally, we have access to M fully labeled source domains and a target domain with actively labeled target data within a pre-defined budget $\mathcal{B}$ in MSDA. The i-th source domain $\mathcal{D}_{s,i} = \{(x_{s,i}^{(j)}, y_{s,i}^{(j)})\}_{j=1}^{N_{s,i}}$ contains $N_{s,i}$ labeled data sampled from the source distribution $P_{s,i}(x,y)$, where $i \in M$. The target domain $\mathcal{D}_t$ contains unlabeled samples $\{x_{tu}^{(j)}\}_{j=1}^{N_{tu}}$ from the target distribution $P_t(x,y)$. Following the standard ADA setting [50], the size of budget $\mathcal{B}$ set to N_{tl}, the labeled target domain $\mathcal{D}_{tl}$=$\{x_{tl}^{(j)}\}_{j=1}^{N_{tl}}$, where $N_{tl} \ll N_{tu}$ and $N_{tl} \ll N_{s,i}$. $P_{s,i}(x,y) \neq P_t(x,y)$ and $P_{s,i}(x,y) \neq P_{s,j}(x,y)$ where $i \neq j$. The multiple source domains and target domain have the same label space $Y = \{1, 2, \cdots, K\}$ with K categories. We aim to learn an model $\mathcal{M}(\cdot)$ adapting from $\{\mathcal{D}_{s,i}\}_{i=1}^{M}$ to $\mathcal{D}_t$, *i.e.*, the model can generalize well on unseen samples from $\mathcal{D}_t$. In general, $\mathcal{M}(\cdot)$ consists of two functions as below:

$$\begin{aligned} &\underbrace{\mathcal{M}(\cdot;(\Theta_m,\Theta_a))}_{\text{MADA Model}} : \underbrace{\mathcal{M}_m((\{\mathcal{D}_{s,i}\}_{i=1}^{M}, \mathcal{D}_{tl});\Theta_m)}_{\text{Multi-domain Adaptation Model}} \\ &\leftrightarrow \underbrace{\mathcal{M}_a((\mathcal{D}_{tl}|\mathcal{D}_t);\Theta_a)}_{\text{Active Learning Model}} , \end{aligned} \tag{1}$$

where $\mathcal{M}_m(\cdot;\Theta_m)$ is the multi-domain learning model and $\mathcal{M}_a(\cdot;\Theta_a)$ is the annotation candidate selection through active learning model with parameters Θ_m and Θ_a.

3.2. Universal Dynamic Network

The Universal Dynamic Network (UDN) (Figure 2(a)) can generate the dynamic parameters for the adaptive classifier conditioned on input samples from different domains, which elegantly regard the multi-domains as one single domain. We train a primary model with a backbone and a classifier for developing the global adaptation model. Given the M source domain samples $\{\{(x_{s,i}^{(j)}, y_{s,i}^{(j)})\}_{j=1}^{N_{s,i}}\}_{i}^{M}$, the proposed UDN can be described as below:

$$P(p|x(i)) = \Psi(\Omega(x_{s,i}^{(j)};\Theta_m^b);\Theta_m^c) , \tag{2}$$

where $\Omega(;\Theta_m^b)$ is the backbone extracting features from input samples and $\Psi(;\Theta_m^c)$ is the classifier. Θ_m^b and Θ_m^c are the learnable parameters for the backbone and classifier.

Here, we treat the backbone as "static layers" and the classifier as "dynamic layers" to achieve dynamic adaptation:

- **Static Layers.** The backbone with Θ_m^b learned from multi-domain data can accurately map the image into the feature space. We fixed the backbone as "static layers" to generate a generalized representation for any given sample concerning the multi-domain distribution.

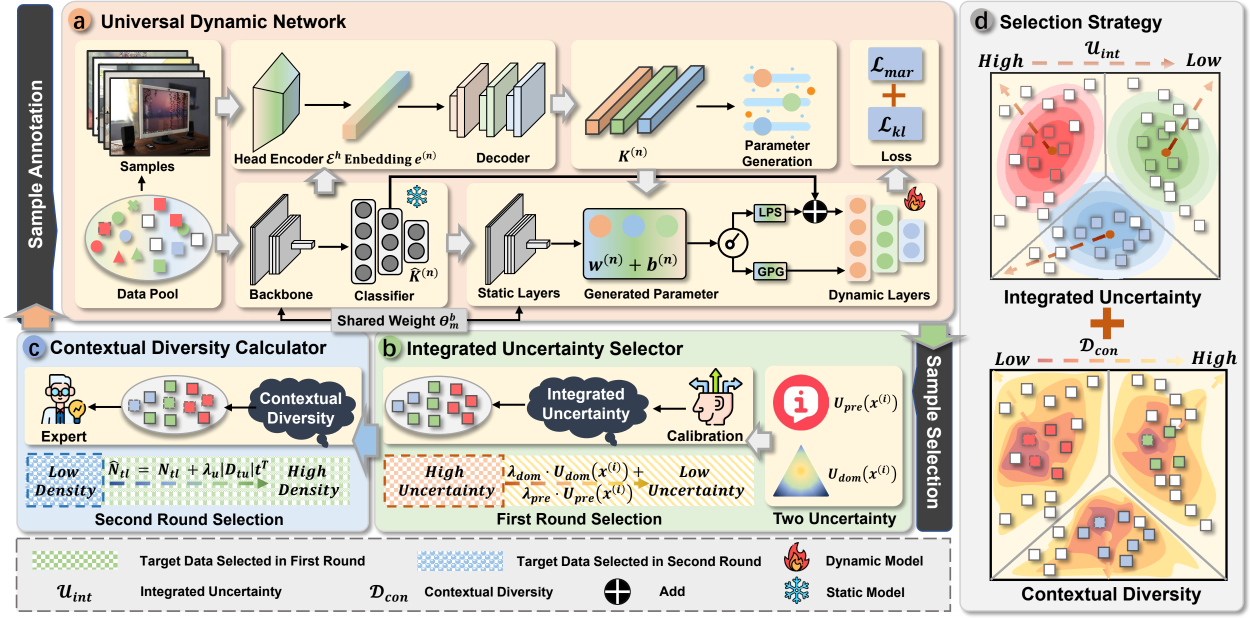


Figure 2. **Overview of Detective**. **(a)** Universal Dynamic Network (UDN) leveraging the Hypernetworks [13] that can generate the dynamic parameters for the adaptive classifiers conditioned on the different domain samples. **(b)** and **(c)** are the Integrated Uncertainty Selector (IUS) and Contextual Diversity Calculator (CDC) that respectively measure the uncertainty (first round selection) and diversity (second round selection) of target samples. **(d)** visualizes the two active learning strategies.

- **Dynamic Layers.** Depending on the image features, the dynamic layers learn adaptive classifier weights Θ_m^c. It learns to adapt the parameters to fit the model to the distribution formed by the union of source domains.

Retrospect of HyperNetwork. The core of dynamic layers are based on the HyperNetwork [13], we will outline the procedure for using a HyperNetwork to generate the parameters for another neural network. Specifically, HyperNetwork treats the parameters of the multi-layer perception (MLP) as a matrix $K^{(n)} \in \mathbb{R}^{N_{in} \times N_{out}}$, where N_{in} and N_{out} represent the number of input and output neurons of the n^{th} layer of MLP, respectively. N_{in} and N_{out} portray the structure of the MLP layers together. The generation process of $K^{(n)}$ can be regarded as a matrix factorization as below:

$$K^{(n)} = \xi(z^{(n)}; \Theta_p), \forall n = 1, \cdots, N_l\,, \tag{3}$$

where $z^{(n)}$ and $\xi(\cdot)$ are randomly initialized with parameters Θ_p in training procedure. The gradients are backpropagated to $z^{(n)}$ and $\xi(\cdot)$, which can help to update them. $z^{(n)}$ and $\xi(\cdot)$ will be saved instead of $K^{(n)}$.

HyperNetwork-based Dynamic Layers. However, as shown in Eq.(3), the original HyperNetwork only utilizes a randomly initialized $z^{(n)}$ to generate the parameters, which lacks the interaction between the parameter generation process and the current input. To measure the dynamic layers for different domains, we propose to model the parameters by replacing the $z^{(n)}$ with representations of the current input image. Specifically, given the feature $\Omega(x^{(i)}; \Theta_m^b)$ extracted from backbone of sample $x_{s,i}^{(j)}$, we first develop a layer-specific encoder $\mathcal{E}^h(\cdot)$ that encodes the $\Omega(x^{(i)}; \Theta_m^b)$ as $\boldsymbol{e}^h$. Then the HyperNetwork is used to convert the embedding $\mathbf{e}^{(n)}$ into parameters, *i.e.*, we input $\mathbf{e}^{(n)}$ into the following two MLP layers to generate parameters of dynamic layers:

$$\boldsymbol{w}^{(n)} = (W_1 \mathcal{E}^h(\Omega(x^{(i)}; \Theta_m^b)) + B_1) W_2 + B_2, \tag{4}$$

where the first and second MLP layers are respectively defined by their weights, W_1 and W_2, along with bias terms B_1 and B_2, encapsulating their unique characteristics.

By doing so, we present two dynamic layer learning strategies, local parameters shift (**LPS**) and global parameters generation (**GPG**). The LPS adopts a two-stage training strategy, *i.e.*, we first pretrain the classifier $\Psi(;\Theta_m^c)$ together with static layers, $\hat{K}^{(n)}$ denote the parameters for n-th layer. The matrices $\hat{K}^{(n)}$ can be seen as a basis for classifier weight space, although they are not necessarily linearly independent. Then we finetune the classifier with local parameters shift, which can be seen as the projections of the residual matrix in the corresponding weight subspaces. For GPG, the parameters are directly generated by the sample-aware learning manner, together with the static layers that are conditioned on the multi-source domain data.

$$K^{(n)} = \begin{cases} \hat{K}^{(n)} + \boldsymbol{w}^{(n)} + \boldsymbol{b}^{(n)}, & \text{if } \Psi(;\Theta_m^c) \text{ pretrained} \\ \boldsymbol{w}^{(n)} + \boldsymbol{b}^{(n)}, & \text{otherwise} \end{cases} \tag{5}$$

where $\forall n = 1, \cdots, N_l$, $K^{(n)}$ denotes the n^{th} layer parameters of dynamic layers.

In summary, the UDN $\mathcal{M}_m(\cdot;\Theta_m)$ learns to adapt the parameters to fit the model to the distribution formed by the union of source domains. The target domain is not required to be aligned with any specific domains and there are no rigid domain boundaries. The model parameters Θ_m can be similar for examples from different domains and different for examples from the same domain.

3.3. Integrated Uncertainty Selector

Evidential Deep Learning with DNN. The above learned multi-domain adaptation model $\mathcal{M}_m(\cdot;\Theta_m)$ explicitly modifies the model's weights according to the samples' distribution from different domains. However, this model has a vague understanding of the target data distribution for out-of-source domain. Therefore, we develop the Integrated Uncertainty Selector to choose the informative samples to facilitate target domain adaptation. Traditional active learning often adopts the prediction of class probability p vector to measure the uncertainty as the criteria of sample selection. For active domain adaptation (ADA), such a manner essentially gives a point estimate of p and can be easily miscalibrated on data with distribution shift [12]. Motivated by Evidential Deep Learning (EDL) [35], unlike the traditional active learning, we predict a high-order Dirichlet distribution $Dir(p|\alpha)$ which is a conjugate prior of the lower-order categorical likelihood. Specifically, a Dirichlet distribution is placed over p to represent the probability density of each variable p. Given sample $x^{(i)}$, the probability density function of p is denoted as $P(p|x^{(i)}) = Dir(p|\alpha^{(i)})$.

$$Dir(p|\alpha^{(i)}) = \begin{cases} \frac{\Gamma(\sum_{d=1}^{D}\alpha_d^{(i)})}{\prod_{d=1}^{D}\Gamma(\alpha_d^{(i)})}\prod_{d=1}^{D} p_d^{\alpha_d^{(i)}-1}, & \text{if } p \in \Delta^D \\ 0 \qquad\qquad\qquad\qquad\quad, & \text{otherwise} \end{cases} \tag{6}$$

where $\alpha_d^{(i)}$ is the parameters of the Dirichlet distribution for sample $x^{(i)}$, $\Gamma(\cdot)$ is the Gamma function and Δ^D the D-dimensional unit simplex, where Δ^D=$\{\sum_{d=1}^{D} p_d = 1\&0 \leq p_d \leq 1\}$, $\alpha^{(i)} = g(f(x^{(i)}, \Theta))$, and $g(\cdot)$ is a function (*e.g.*, exponential function) to keep $\alpha^{(i)}$ positive. In this way, the prediction of each sample is interpreted as a distribution over the probability simplex, rather than the simple predictive uncertainty. And we can mitigate the uncertainty miscalibration by considering all possible predictions rather than unilateral predictions.

After applying the EDL to a standard DNN with softmax for the classification task, the predicted probability for class k can be denoted as Eq. (6), by marginalizing over p. The details of derivation in the supplementary material.

$$\begin{aligned} \hat{P}(y=k|x^{(i)}) &= \int P(y=k|p)P(p|x^{(i)})dp \\ &= \frac{g(f_k(x^{(i)}))}{\sum_{c=1}^{K} g(f_c(x^{(i)}))} = \mathbb{E}[Dir(p_k|\alpha^{(i)})]\,, \end{aligned} \tag{7}$$

where $\hat{P}$ is the prediction. If $g(\cdot)$ adopts the exponential function, then softmax-based DNNs can be regarded as predicting the expectation of Dirichlet distribution.

Domain & Predictive Uncertainty Integration. For the evidential model supervised with multi-source data, if target samples are obviously distinct from the source domain, *e.g.*, the realistic *v.s.* clipart style, the evidence collected for these target samples may be insufficient, because the model lacks the knowledge about this kind of data. To solve this problem, we use the uncertainty obtained from the lack of evidence, *i.e.*, domain uncertainty, to measure the target domain's characteristics. Specifically, the domain uncertainty (Figure 2(d)) $\mathcal{U}_{dom}$ of sample $x^{(i)}$ is defined as:

$$\begin{aligned} &\mathcal{U}_{dom}(x^{(i)}) = \sum_{k=1}^{K}\hat{P}(y=k|x^{(i)})(\Phi(\alpha_k^{(i)}+1) - \\ &\Phi(\sum_{c=1}^{K}\alpha_c^{(i)}+1)) - \sum_{k=1}^{K}\hat{P}(y=k|x^{(i)})\log\hat{P}(y=k|x^{(i)}), \end{aligned} \tag{8}$$

where Φ is the digamma function. Here, we use mutual information to measure the spread of Dirichlet distribution on the simplex like [35]. Higher $\mathcal{U}_{dom}(x^{(i)})$ indicates larger domain uncertainty, *i.e.*, the Dirichlet distribution is broadly spread on the probability simplex.

We also utilize the predictive entropy to quantify predictive uncertainty, which is denoted as the expected entropy of all possible predictions. Specifically, given sample $x^{(i)}$, the predictive uncertainty (Figure 2(d)) $\mathcal{U}_{pre}(x^{(i)})$ is:

$$\begin{aligned} &\mathcal{U}_{pre}(x^{(i)}) = \mathbb{E}[H[P(y|p)]] = \\ &\sum_{k=1}^{K} P(y=k|x^{(i)})(\Phi(\sum_{c=1}^{K}\alpha_c^{(i)}+1) - \Phi(\alpha_{k^{(i)}}+1))\,. \end{aligned} \tag{9}$$

With the domain uncertainty $\mathcal{U}_{dom}$ and predictive uncertainty $\mathcal{U}_{pre}$, we comprehensively consider the two uncertainties to select the target domain samples. We use the mix-up strategy (Figure 2(b)) to fuse the $\mathcal{U}_{dom}$ and $\mathcal{U}_{pre}$: $\mathcal{U}_{int}(x^{(i)}) = \lambda_{dom}\cdot\mathcal{U}_{dom}(x^{(i)}) + \lambda_{pre}\cdot\mathcal{U}_{pre}(x^{(i)})$, where coefficients λ_{dom} and λ_{pre} are pre-defined hyperparameters. The bigger value of λ_{dom} means the higher influence of domain uncertainty and vice versa.

3.4. Contextual Diversity Calculator

The selected target samples based on the aforementioned integrated uncertainty $\mathcal{U}_{int}$ may have similar semantics in the implicit feature space. We argue that when considering the potential similarity of selected samples in the learned feature space, not all selected samples contribute highly to the performance of MADA but instead increase the labeling cost and computational expense. To enhance the diversity of sample selection, we measure the contextual diversity (Figure 2(c)) based on feature space coverage. Specifically,

Table 1. **Accuracy(%) comparison on `Office-Home` using Resnet50 as backbone.** Acronym of each model can be found in Section 4.1. We color each row as the **best** , **second best** , and **third best** . ∪ represents the rest of the domains.

DA Setting	Methods	`Office-Home` Dataset				
		$\cup \rightarrow$ **A**	$\cup \rightarrow$ **P**	$\cup \rightarrow$ **C**	$\cup \rightarrow$ **R**	**Mean**
MSDA	MFSAN [65]	72.1	80.3	62.0	81.8	74.1
	MDDA [62]	66.7	79.5	62.3	79.6	71.0
	SImpAI [46]	70.8	80.2	56.3	81.5	72.2
	MADAN [63]	66.8	78.2	54.9	81.5	70.4
ADA	CLUE [31]	75.61	86.48	67.84	84.07	78.50
	TQS [10]	78.29	86.62	68.74	86.07	79.93
	DUC [51]	**81.21**	**89.46**	**71.63**	**88.18**	**82.62**
	EADA [50]	**81.74**	**89.15**	**71.75**	**89.18**	**83.03**
MADA	**Detective (Ours)**	**86.03**	**91.78**	**86.91**	**95.22**	**89.99**

we calculate the image-level density of selected samples $\mathcal{S}_t = \{x_{tu}^{(i)}\}_{j=1}^{J}$, the j-th feature extracted by the backbone. Finally, we remove $\hat{N}_{tl}$ - N_{tl} images with the largest image-level density value to reach the desired labeling budget N_{tl}. Note that N_{tl} is the actual labeling budget and $\hat{N}_{tl}$ is a hyperparameter. Obviously, the $\hat{N}_{tl}$ should be gradually increased during training, $\hat{N}_{tl}$ can be computed as:

$$\hat{N}_{tl} = N_{tl} + \lambda_u|\mathcal{D}_{tu}|t^{\tau} \tag{10}$$

where λ_u is a hyperparameter. t^{τ} denotes its temperature. By modulating t^{τ}, we can control the uncertainty-diversity tradeoff to further boost MADA performance.

3.5. Detective Training

To get reliable and consistent opinions for labeled data, the evidential model is trained to generate sharp Dirichlet distribution located at the corner of these labeled data. Concretely, we train the model by minimizing the negative logarithm of the marginal likelihood ($\mathcal{L}_{mar}$) which is minimized to ensure the correctness of prediction.

$$\mathcal{L}_{mar} = \sum_{x^{(i)}\in[\{\mathcal{D}_{s,i}\}_{i=1}^{M},\mathcal{D}_t]} \beta_k^{(i)}(\log(\sum_{k=1}^{k}\alpha_k^{(i)}) - log\alpha_k^{(i)})\,, \tag{11}$$

where [,] is the concatenation of two sets. $\beta_k^{(i)}$ is the k-th element of the one-hot label vector of sample $x^{(i)}$.

We also minimize the KL-divergence between two Dirichlet distributions ($\mathcal{L}_{kl}$) of source domains and target domain.

$$\mathcal{L}_{kl} = \sum_{x^{(i)}\in[\{\mathcal{D}_{s,i}\}_{i=1}^{M},\mathcal{D}_t]} KL[Dir(p|\hat{\alpha^{(i)}})|Dir(p|1)] \tag{12}$$

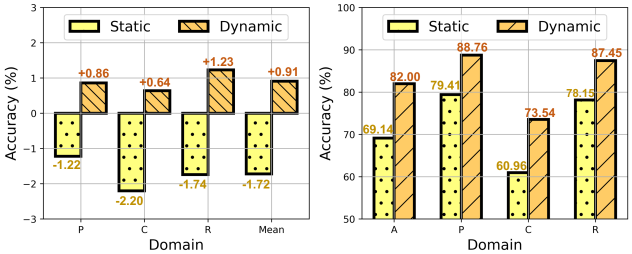
where $\hat{\alpha^{(i)}} = \beta_k^{(i)} + (1-\beta_k^{(i)})\cdot\alpha^{(i)}$. $\hat{\alpha^{(i)}}$ can be seen as removing the evidence of ground-truth class. Minimizing $\mathcal{L}_{kl}$ will force the evidences of other classes to reduce, avoiding the collection of mis-leading evidences.

Table 2. **Ablation study of the effect of individual module.**

Methods	`Office-Home Dataset`				
	$\cup \rightarrow$ **A**	$\cup \rightarrow$ **p**	$\cup \rightarrow$ **c**	$\cup \rightarrow$ **R**	**Mean**
-UDN	82.82	89.77	72.97	89.88	83.86
-IUS	81.71	87.41	84.19	91.81	86.28
-CDC	85.12	90.63	85.34	94.10	88.80
Detective (Ours)	**86.03**	**91.78**	**86.91**	**95.22**	**89.99**

Table 3. **LPS vs. GPG on `Office-Home` dataset.**

Methods	`Office-Home Dataset`				
	$\cup \rightarrow$ **A**	$\cup \rightarrow$ **p**	$\cup \rightarrow$ **c**	$\cup \rightarrow$ **R**	**Mean**
LPS	85.60	90.13	85.91	93.78	88.86
GPG	**86.03**	**91.78**	**86.91**	**95.22**	**89.99**



(a) Performance on Source Domain after Adaptation.

(b) MSDA performance of Static and Dynamic Models.

Figure 3. **Performance of static and dynamic models.**

To sum up, the Detective is learned with a combination of two losses: $\mathcal{L}_{sum} = \lambda_{mar}\cdot\mathcal{L}_{mar} + \lambda_{kl}\cdot\mathcal{L}_{kl}$, where λ_{mar} and λ_{kl} are hyperparameters.

4. Experiments

4.1. Dataset and Setting

Benchmark Datasets. `Office-Home` [47] benchmark contains 65 classes, with 12 adaptation scenarios constructed from 4 domains (*i.e.*, **R**: Real world, **C**: Clipart, **A**: Art, **P**: Product). `miniDomainNet` is a subset of DomainNet [30] and contains 140,006 96×96 images of 126 classes from four domains: Clipart, Painting, Real, and Sketch (abbr. **R**, **C**, **P** and **S**). `Digits-five` contains five-digit sub-datasets: MNIST (mt) [16] , Synthetic (sy) [11], MNIST-M(mm) [11], SVHN (sv) [29], and USPS (up) [11]. Each sub-dataset contains images of numbers ranging from 0 to 9.

Implementation Details. We employ the ResNet [15] as the backbone model on three datasets. We train Detective with the SGD [3] optimizer in all experiments. Besides, we use an identical set of hyperparameters (B=64, M_o=0.9, W_d=0.00005, L_r=0.0004)[1] across all datasets. Following [50], the total labeling budget $\mathcal{B}$ is set as 5% of target samples, which is divided into 5 selection steps, *i.e.*, the labeling budget in each round is b = $\mathcal{B}$/5.

[1] B, M_o, W_d and L_r refer to batch size, momentum, weight decay and learning rate in SGD optimizer.

Comparison of Methods. For quantifying the efficacy of the proposed framework, we compare Detective with previous SoTA MSDA and ADA approaches. For MSDA methods, wo choose MCD [33], DCTN [52], M^{3}SDA [30], MME [34], DEAL [64], MFSAN [65], MDDA [62], SImpAI [46], MADAN [63] as baselines. The ADA methods include CLUE [31], TQS [10], DUC [51] and EADA [50][2].

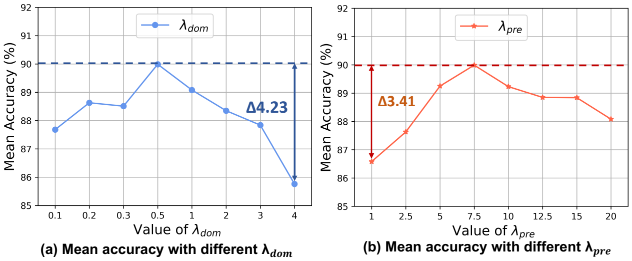
4.2. Overall Performance

Table 1 summarizes the quantitative MADA results of our framework and baselines on `Office-Home` (The experimental results of `miniDomainNet` and `Digits-five` in the Appendix). We make the following observations: 1) In general, irrespective of the different shot scenarios, compared to SoTAs, Detective achieves the best performance on almost all the adaptation scenarios. In particular, Detective outperforms other baselines in terms of mean accuracy by a large margin (`Office-Home`: **6.96% ∼ 19.59%**) for the MADA task. 2) It is worth noting that almost all ADA models generally outperform MSDA baselines by employing an appropriate annotation strategy, showing the necessity of considering informative target supervision. The two tables also suggest that the clustering-based ADA methods (*e.g.*, CLUE) seem to be less effective than uncertainty-based methods (*e.g.*, TQS, DUC, EADA) on the large-scale dataset. This may be because clustering becomes more difficult with multi-source domain modeling. 3) Benefiting from the carefully designed Integrated Uncertainty Selector (INS), our Detective achieves better estimation of uncertainty by incorporating the distribution interpretation into prediction, while other uncertainty-aware methods can easily be miscalibrated. Further, combined with dynamic multi-domain modeling and contextual diversity enhancement, we obtain the best MADA performance. Overall, these results strongly demonstrate the effectiveness of our proposed Detective under the practical multi-source active domain transfer setting.

4.3. Ablation Study

Effectiveness of Each Component. We conduct an ablation study, as illustrated in Table 2, to demonstrate the effectiveness of each component. Comparing Detective and Detective(-UDN) (Row 1 *vs.* Row 4), the UDN contributes a 6.13% improvement in mean accuracy. The results in Row 2 show that the IUS leads to a 3.71% increase in mean accuracy. Meanwhile, Row 3 indicates a noticeable performance degradation of 1.19% without the CDC. In summary, each module's contribution to improvement is distinct. When combining all components, our Detective framework exhibits steady improvement over the baselines.

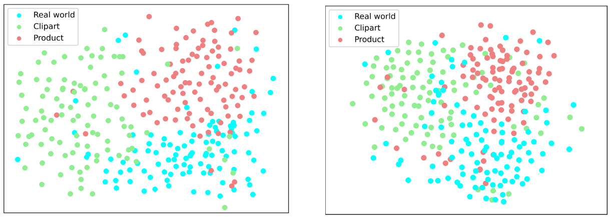
Integration Coefficient λ_{dom} and λ_{pre}. We investigate the impact of integration coefficient $\lambda_{dom}$and λ_{pre}, which controls the uncertainty integration ratio to select valuable samples. The mean accuracy of different coefficients of λ_{dom} and λ_{pre} on `Office-Home` dataset is shown in Figure 4. This figure suggests that the optimal choices of λ_{dom} and λ_{pre} are approximately 7.5 and 0.5, respectively. This phenomenon implies that domain uncertainty predominates in uncertainty estimation, which demonstrates the importance of Evidential Deep Learning (EDL) for sample selection in the presence of domain discrepancies.



(a) Mean accuracy with different λ_{dom}

(b) Mean accuracy with different λ_{pre}

Figure 4. **Ablation with respect to λ_{dom} and λ_{pre}.**



(a) t-Sne Visualization of Static Model

(b) t-Sne Visualization of Dynamic Model

Figure 5. **t-SNE Visualization of multi-source embeddings.**

LPS *vs.* GPG. We evaluate two dynamic layer learning approaches: local parameters shift (LPS) and global parameters generation (GPG), as shown in Table 3. The results in the table indicate that utilizing GPG yields superior MADA outcomes. Our intuition behind this observation is that GPG operates the dynamic parameters within a larger semantic space compared to LPS, thereby benefiting the modeling of complex multiple domains.

4.4. In-Depth Analysis

We further validate several vital issues of the three modules, in proposed Detective by answering the following questions. **Q1: Can the UDN appropriately model the multi-domain by a single model?** To build the insight on the effectiveness of the dynamic multi-domain learning in Detective, for the ∪ → **A** on `Office-Home` dataset, we first evaluate the performance of multi-source domain after adaptation in Figure 3(a). A clear performance degradation (1.72% in average) exists when using the static model, indicating the static transfer essentially averages the domain conflicts and thus the performance drops on each source domain. In contrast, our proposed dynamic model handles the domain shifts well, *i.e.*, adapting to the target domain further improves the accuracy on the source domains. Figure 3(a) reports MSDA

[2] For ADA methods, we only compare with the approaches that have open source code and run them on our MADA setting.

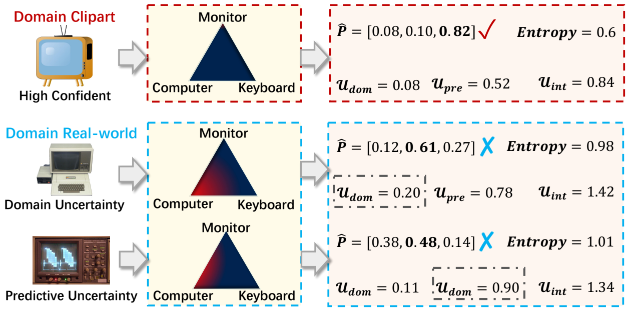

Domain Clipart
High Confident
Monitor
Computer
Keyboard
$\hat{P} = [0.08, 0.10, \mathbf{0.82}]$ ✓
$Entropy = 0.6$
$\mathcal{U}_{dom} = 0.08$
$\mathcal{U}_{pre} = 0.52$
$\mathcal{U}_{int} = 0.84$
Domain Real-world
Domain Uncertainty
Monitor
Computer
Keyboard
$\hat{P} = [0.12, \mathbf{0.61}, 0.27]$ ✗
$Entropy = 0.98$
$\mathcal{U}_{dom} = 0.20$
$\mathcal{U}_{pre} = 0.78$
$\mathcal{U}_{int} = 1.42$
Predictive Uncertainty
Monitor
Computer
Keyboard
$\hat{P} = [0.38, \mathbf{0.48}, 0.14]$ ✗
$Entropy = 1.01$
$\mathcal{U}_{dom} = 0.11$
$\mathcal{U}_{dom} = 0.90$
$\mathcal{U}_{int} = 1.34$

Figure 6. **Case study with respect to $\mathcal{U}_{dom}$ and $\mathcal{U}_{pre}$.**

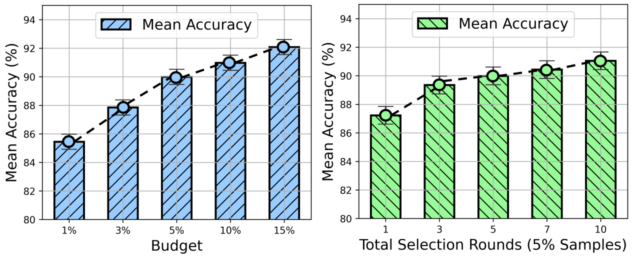
results without target supervision, which shows that our dynamic transfer obtains a superior adaptation result against the static model, even comparable compared with ADA models. We also visualize the samples with one specific class ("chair") on the other source domains. This figure suggests the UDN tends to group the samples with the same class in one cluster, which validates our claim that adapting the model across domains can be achieved by adapting the dynamic model to samples. However, the sample density using the static model is relatively low, since forcing a static model to leverage multi-domains may degrade the performance.

Q2: How does the IUS calibrate the uncertainty to select informative samples? The key insight of uncertainty calibration is to use the EDL to measure the "targetness", *i.e*, domain uncertainty. We plot the distribution of $\mathcal{U}_{dom}$ in Figure.1 (in the Appendix), where the model is trained on the multi-source domain. We see that the $\mathcal{U}_{dom}$ of target data is noticeably biased from multi-source domain. Such results show that our $\mathcal{U}_{dom}$ can play the role of domain discriminator without introducing it. We also give an intuitive case study in Figure 6: given a "monitor", for the two images from the Real-World domain, the general prediction entropy cannot distinguish them, whereas $\mathcal{U}_{dom}$ and $\mathcal{U}_{pre}$ calculated based on the prediction distribution can reflect what contributes more to their uncertainty and be utilized to guarantee the informativeness of selected data. Additionally, we study the effect of $\mathcal{U}_{dom}$ and $\mathcal{U}_{pre}$ in IUS, we train the model with the individual uncertainty under MADA. As shown in Table. 4, both the $\mathcal{U}_{dom}$ and $\mathcal{U}_{pre}$ play significant roles in IUS. Notably, simply using $\mathcal{U}_{pre}$ results in a higher performance decay compared to only using $\mathcal{U}_{dom}$, which implies a standard DNN can easily produce overconfident but wrong predictions for target data, making the estimated predictive uncertainty unreliable. On the contrary, reasonably integrating the $\mathcal{U}_{dom}$ and $\mathcal{U}_{pre}$ can calibrate uncertainty to choose the target samples, thereby facilitating MADA.

Q3: How does the CDC further benefit the MADA? Section 4.3 of the ablation study has initially shown that the CDC can boost the MADA performance on `Office-Home` dataset. To further analyze the source of the CDC's improvements, we visualize our method's sample selection behavior using t-SNE [45] and report the performance after the first round in Figure.2 (in the Appendix). This visualization result suggests that our Detective tends to select representative uncertain samples with lower image-level density compared with Detective (w/o CDC). These samples can propagate representative label information to nearby samples, avoiding the curse of scale by leveraging uncertainty to obtain preliminary samples, thereby boosting MADA performance. We also perform the hyperparameter analysis of λ_u and $t^{\mathcal{T}}$ in the Appendix. The observations and analysis verify the effectiveness of the CDC in improving the diversity of sample selection and thus boosting MADA performance.

Table 4. **Two uncertainties in IUS.**

Uncertainty		`Office-Home Dataset`				
$\mathcal{U}_{dom}$	$\mathcal{U}_{pre}$	$\cup \to$ **A**	$\cup \to$ **P**	$\cup \to$ **C**	$\cup \to$ **R**	**Mean**
	✓	83.81	88.36	84.43	91.13	86.93
✓		85.39	90.63	86.05	93.77	88.96
✓	✓	**86.03**	**91.78**	**86.91**	**95.22**	**89.99**



(a) Performance using different budget. **(b) Performance with different selection rounds.**

Figure 7. **Ablation study of different AL setting.**

5. Conclusion

We propose a novel task, termed Multi-source Active Domain Adaptation (MADA), and have correspondingly developed **Detective**. This framework comprehensively considers domain and predictive uncertainty, and context diversity to select informative target samples, thereby boosting multi-source domain adaptation. We hope that Detective can provide new insights into the domain adaptation community.

6. Acknowledgments

This work was supported by NSFC(No. 62272411), Key Research and Development Projects in Zhejiang Province(No. 2024C01106), National Key Research and Development Project of China(2018AAA0101900), Scientific Research Fund of Zhejiang Provincial Education Department (Y202353679), and Ant Group.

References

[1] Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. *Advances in Neural Information Processing Systems*, 33:14927–14937, 2020.

[2] Wentao Bao, Qi Yu, and Yu Kong. Evidential deep learning for open set action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13349–13358, 2021.

[3] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

[4] Andrew Brock, Theodore Lim, James M. Ritchie, and Nick Weston. SMASH: one-shot model architecture search through hypernetworks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[5] Oscar Chang, Lampros Flokas, and Hod Lipson. Principled weight initialization for hypernetworks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[6] Liang Chen, Yihang Lou, Jianzhong He, Tao Bai, and Minghua Deng. Evidential neighborhood contrastive learning for universal domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6258–6267, 2022.

[7] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3339–3348, 2018.

[8] Zhengyu Chen, Teng Xiao, Kun Kuang, Zheqi Lv, Min Zhang, Jinluan Yang, Chengqiang Lu, Hongxia Yang, and Fei Wu. Learning to reweight for generalizable graph neural network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8320–8328, 2024.

[9] Antoine de Mathelin, Francois Deheeger, Mathilde Mougeot, and Nicolas Vayatis. Discrepancy-based active learning for domain adaptation. *arXiv preprint arXiv:2103.03757*, 2021.

[10] Bo Fu, Zhangjie Cao, Jianmin Wang, and Mingsheng Long. Transferable query selection for active domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7272–7281, 2021.

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

[12] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

[13] David Ha, Andrew Dai, and Quoc V Le. Hypernetworks. *arXiv preprint arXiv:1609.09106*, 2016.

[14] David Ha, Andrew M. Dai, and Quoc V. Le. Hypernetworks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[15] 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.

[16] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[17] Da Li and Timothy Hospedales. Online meta-learning for multi-source and semi-supervised domain adaptation. In *European Conference on Computer Vision*, pages 382–403. Springer, 2020.

[18] Juncheng Li, Minghe Gao, Longhui Wei, Siliang Tang, Wenqiao Zhang, Mengze Li, Wei Ji, Qi Tian, Tat-Seng Chua, and Yueting Zhuang. Gradient-regulated meta-prompt learning for generalizable vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2551–2562, 2023.

[19] Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Hanwang Zhang, Wei Ji, Wenqiao Zhang, Tat-Seng Chua, Siliang Tang, and Yueting Zhuang. Fine-tuning multimodal llms to follow zero-shot demonstrative instructions. In *The Twelfth International Conference on Learning Representations*, 2023.

[20] Mengze Li, Kun Kuang, Qiang Zhu, Xiaohong Chen, Qing Guo, and Fei Wu. Ib-m: A flexible framework to align an interpretable model and a black-box model. In *2020 IEEE international conference on bioinformatics and biomedicine (BIBM)*, pages 643–649. IEEE, 2020.

[21] Mengze Li, Han Wang, Wenqiao Zhang, Jiaxu Miao, Zhou Zhao, Shengyu Zhang, Wei Ji, and Fei Wu. Winner: Weakly-supervised hierarchical decomposition and alignment for spatio-temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23090–23099, 2023.

[22] Mengze Li, Tianbao Wang, Jiahe Xu, Kairong Han, Shengyu Zhang, Zhou Zhao, Jiaxu Miao, Wenqiao Zhang, Shiliang Pu, and Fei Wu. Multi-modal action chain abductive reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4617–4628, 2023.

[23] Mengze Li, Haoyu Zhang, Juncheng Li, Zhou Zhao, Wenqiao Zhang, Shengyu Zhang, Shiliang Pu, Yueting Zhuang, and Fei Wu. Unsupervised domain adaptation for video object grounding with cascaded debiasing learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3807–3816, 2023.

[24] Yunsheng Li, Lu Yuan, Yinpeng Chen, Pei Wang, and Nuno Vasconcelos. Dynamic transfer for multi-source domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10998–11007, 2021.

[25] Zheqi Lv, Feng Wang, Shengyu Zhang, Wenqiao Zhang, Kun Kuang, and Fei Wu. Parameters efficient fine-tuning for long-tailed sequential recommendation. In *CAAI International Conference on Artificial Intelligence*, pages 442–459. Springer, 2023.

[26] Zheqi Lv, Wenqiao Zhang, Shengyu Zhang, Kun Kuang, Feng Wang, Yongwei Wang, Zhengyu Chen, Tao Shen, Hongxia Yang, Beng Chin Ooi, and Fei Wu. Duet: A tuning-free device-cloud collaborative parameters generation framework

for efficient device model generalization. In *Proceedings of the ACM Web Conference 2023*, 2023.

[27] Zheqi Lv, Wenqiao Zhang, Zhengyu Chen, Shengyu Zhang, and Kun Kuang. Intelligent model update strategy for sequential recommendation. In *Proceedings of the ACM Web Conference*, pages 1–12, 2024.

[28] Jay Nandy, Wynne Hsu, and Mong Li Lee. Towards maximizing the representation gap between in-domain & out-of-distribution examples. *Advances in Neural Information Processing Systems*, 33:9239–9250, 2020.

[29] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

[30] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019.

[31] Viraj Prabhu, Arjun Chandrasekaran, Kate Saenko, and Judy Hoffman. Active domain adaptation via clustering uncertainty-weighted embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8505–8514, 2021.

[32] 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.

[33] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3723–3732, 2018.

[34] 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.

[35] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

[36] Jayaram Sethuraman. A constructive definition of dirichlet priors. *Statistica sinica*, pages 639–650, 1994.

[37] Burr Settles. Active learning literature survey. 2009.

[38] 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.

[39] Qian Sun, Rita Chattopadhyay, Sethuraman Panchanathan, and Jieping Ye. A two-stage weighting framework for multi-source domain adaptation. *Advances in neural information processing systems*, 24, 2011.

[40] Shiliang Sun, Honglei Shi, and Yuanbin Wu. A survey of multi-source domain adaptation. *Information Fusion*, 24: 84–92, 2015.

[41] Zihao Tang, Zheqi Lv, Shengyu Zhang, Yifan Zhou, Xinyu Duan, Fei Wu, and Kun Kuang. Aug-kd: Anchor-based mixup generation for out-of-domain knowledge distillation. In *The Twelfth International Conference on Learning Representations*.

[42] Zihao Tang, Zheqi Lv, Shengyu Zhang, Fei Wu, and Kun Kuang. Modelgpt: Unleashing llm's capabilities for tailored model generation. *arXiv preprint arXiv:2402.12408*, 2024.

[43] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7472–7481, 2018.

[44] 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.

[45] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[46] Naveen Venkat, Jogendra Nath Kundu, Durgesh Singh, Ambareesh Revanur, et al. Your classifier can secretly suffice multi-source domain adaptation. *Advances in Neural Information Processing Systems*, 33:4647–4659, 2020.

[47] 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.

[48] Johannes von Oswald, Christian Henning, João Sacramento, and Benjamin F. Grewe. Continual learning with hypernetworks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

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

[50] Binhui Xie, Longhui Yuan, Shuang Li, Chi Harold Liu, Xinjing Cheng, and Guoren Wang. Active learning for domain adaptation: An energy-based approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8708–8716, 2022.

[51] Mixue Xie, Shuang Li, Rui Zhang, and Chi Harold Liu. Dirichlet-based uncertainty calibration for active domain adaptation. *arXiv preprint arXiv:2302.13824*, 2023.

[52] Ruijia Xu, Ziliang Chen, Wangmeng Zuo, Junjie Yan, and Liang Lin. Deep cocktail network: Multi-source unsupervised domain adaptation with category shift. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3964–3973, 2018.

[53] Bencheng Yan, Pengjie Wang, Kai Zhang, Feng Li, Jian Xu, and Bo Zheng. Apg: Adaptive parameter generation network for click-through rate prediction. In *Advances in Neural Information Processing Systems*, 2022.

[54] Kaichao You, Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Universal domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2720–2729, 2019.

[55] Chris Zhang, Mengye Ren, and Raquel Urtasun. Graph hypernetworks for neural architecture search. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[56] Wenqiao Zhang, Siliang Tang, Yanpeng Cao, Shiliang Pu, Fei Wu, and Yueting Zhuang. Frame augmented alternating attention network for video question answering. *IEEE Transactions on Multimedia*, 22(4):1032–1041, 2019.

[57] Wenqiao Zhang, Haochen Shi, Siliang Tang, Jun Xiao, Qiang Yu, and Yueting Zhuang. Consensus graph representation learning for better grounded image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3394–3402, 2021.

[58] Wenqiao Zhang, Lei Zhu, James Hallinan, Shengyu Zhang, Andrew Makmur, Qingpeng Cai, and Beng Chin Ooi. Boostmis: Boosting medical image semi-supervised learning with adaptive pseudo labeling and informative active annotation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20666–20676, 2022.

[59] Wenqiao Zhang, Changshuo Liu, Lingze Zeng, Bengchin Ooi, Siliang Tang, and Yueting Zhuang. Learning in imperfect environment: Multi-label classification with long-tailed distribution and partial labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1423–1432, 2023.

[60] Wenqiao Zhang, Tianwei Lin, Jiang Liu, Fangxun Shu, Haoyuan Li, Lei Zhang, He Wanggui, Hao Zhou, Zheqi Lv, Hao Jiang, et al. Hyperllava: Dynamic visual and language expert tuning for multimodal large language models. *arXiv preprint arXiv:2403.13447*, 2024.

[61] Han Zhao, Shanghang Zhang, Guanhang Wu, José MF Moura, Joao P Costeira, and Geoffrey J Gordon. Adversarial multiple source domain adaptation. *Advances in neural information processing systems*, 31, 2018.

[62] Sicheng Zhao, Guangzhi Wang, Shanghang Zhang, Yang Gu, Yaxian Li, Zhichao Song, Pengfei Xu, Runbo Hu, Hua Chai, and Kurt Keutzer. Multi-source distilling domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12975–12983, 2020.

[63] Sicheng Zhao, Bo Li, Pengfei Xu, Xiangyu Yue, Guiguang Ding, and Kurt Keutzer. Madan: multi-source adversarial domain aggregation network for domain adaptation. *International Journal of Computer Vision*, 129(8):2399–2424, 2021.

[64] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain adaptive ensemble learning. *IEEE Transactions on Image Processing*, 30:8008–8018, 2021.

[65] Yongchun Zhu, Fuzhen Zhuang, and Deqing Wang. Aligning domain-specific distribution and classifier for cross-domain classification from multiple sources. In *Proceedings of the AAAI conference on artificial intelligence*, pages 5989–5996, 2019.