CONTROLLABLE ADAPTIVE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

As deep learning enabled an unprecedented number of applications in versatile vision cognition tasks, researchers surged for the solutions of higher performance and more generalized algorithms, coming with expensive training and deployment to be applied in complex scenarios across domains. However, we argue that generalization and high performance are not always the ultimate goal in real-life with various applications and regulatory requirements. In this work, for the first time to our knowledge, we propose a Controllable Adaptive Learning (CAL) paradigm that allows the model to perform well on some data domains while performing poorly on others by control. We define the problem as a Controlled Multi-target Unsupervised Domain Adaptation (CMUDA) Task. Without the need to access labels in the target domain, we make the model have poor performance on certain target domains through a novel distribution different loss function design. We then introduced two easy-to-use control methods, namely implicit embedding-enabled controller and explicit prompt-based controller, to regain access of the high-performance result with little effort, without the need of retraining the entire network. Extensive experiments demonstrated the effectiveness of our approach. We believe that our CAL paradigm will lead to an emerging trend for future research. Our code is at *URL*.

1 INTRODUCTION

The era has witnessed the explosive growth of deep learning in a wide range of applications in computer vision tasks. Now, Artificial Intelligence as a Service (AIaaS) has become an emerging trend that underpins many applications for consumers used in daily life Lins et al. (2021). Under the trend, researchers have spent continuous efforts exploring higher-performance algorithms. Encouraging achievements have been made. For example, in nearly a decade, the classification accuracy of ImageNet has been elevated from around 60% to above 90% Krizhevsky et al. (2017); Liu et al. (2022), which is supported by factors like large dataset, growing size of the model, and lengthy and costly training.

However, we argue that high-performance is not always the right goal and cannot be the only goal in the context of AIaaS. The potential for misuse and abuse of AI services, e.g., the misuse of tools like DeepFake or facial recognition algorithms, has triggered the alarm for the massesCobbe & Singh (2021); Harwell (2022). Different standards in law enforcement in different regions regulating certain applications of AI, e.g., in the context of pornography, also call for AI service providers to carefully and selectively provide AI services Cobbe & Singh (2021). Moreover, there are scenarios with commercial purposes of limiting the performance of the AI algorithm for free users and unlocking the model's full potential for paid users for profit-making. All the above-mentioned examples derive the need for one demand – ***Controllable AI.***

Few efforts have been made to control the access of particular AI models in the realm of AI security. For example, methods like feature-based or trigger-based watermarking Uchida et al. (2017); Rouhani et al. (2018); Kuribayashi et al. (2020) and secure authorization approaches Alam et al. (2020) have been used to safeguard deep neural networks. Yet, these methods mainly focus on *Who is allowed to use the AI model*. However, we argue the AI model should also be controllable in terms of *What data is being used*, or namely *applicability authorization* Wang et al. (2021), to meet the demand in the above-mentioned scenarios, which is, however, a research gap in the field. The realm of domain adaptation and domain generalization can ensure the model's satisfactory performance by generalizing models with domain gaps filled Ahmed et al. (2021); Blanchard et al. (2011), but cannot

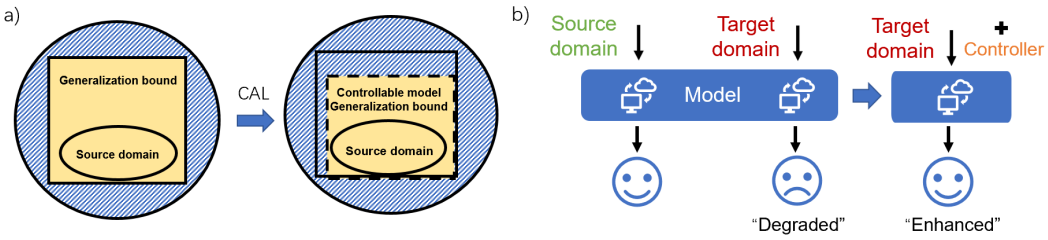


Figure 1: **The overall scheme of our Controllable Adaptive Learning (CAL).** a) Our CAL scheme aims to control the model performance by controlling the generalization bound. b) How users benefit from the CAL scheme. The model generalize to specific source domain data, if inputted by the target domain information, the model falls into "Degraded" mode with restricted performance (with a restricted generalization bound of the model). The "controller" can boost the performance of the model and bring back the satisfactory performance even with input from the target domain (with a enlarged generalization bound of the model).

be controlled to decrease the model performance to specific domains. Non-transfer Learning (NTL) Wang et al. (2021) proposes decreasing the model generalization in specifically selected domains, but cannot regain access to the high-performance model in the selected domains.

To fill the research gap, we propose a new **Controllable Adaptive Learning (CAL)** paradigm, which enables the control of model performance for different data domains. Specifically, our approaches and contributions are as follows:

- We introduce a Controlled Multi-target Unsupervised Domain Adaptation (CMUDA) Task. Without the need to access the label in the target domain, we deliberately prevent the knowledge from being transferred from the source domain to one or more target domains, thus decreasing the performance in specific domains.
- We propose two methods to regain access to high-performance in the decreased-performance domains, namely implicit embedding-enabled controller and explicit prompt-based controller. Retraining the neural network is no longer needed.
- We perform extensive experiments in two widely-used vision datasets including Digits and Multi-PIE, and demonstrate the effectiveness of our approach. The model performs well in the source domain while badly in the selected target domain. The access to high performance can be effectively regained using the two controllers.

To the best of our knowledge, our work pioneers to control the model performance adaptively based on the image data feed in. We believe that our work tailors a viable path in solving some AI-ethics problems and enabling new commercial possibilities.

2 Task Description

We propose to realize our Controllable Adaptive Learning (CAL) paradigm under a Controlled Multi-target Unsupervised Domain Adaptation (CMUDA) Task. We consider a labeled source domain $\mathcal{D}_s = \{x_i, y_i\}_{i=1}^{N^s}$ and multiple target domains $\{\mathcal{D}_t\}_{t=1}^{T}$, where $\mathcal{D}_t = \{x_j\}_{j=1}^{N_t}$. Conventional unsupervised domain adaptation approaches transfer knowledge from the source domain to the target domains to boost the model's generalization capability. We, however, make the process *controllable* – the performance on the target domains can be" Enhanced" or "Degraded" by control, with an explicit generalization bound. Specifically, in the "Degraded" mode, we deliberately stopped the knowledge learned from source domains leaked to the targets, leading to degraded performance, the "Enhanced" mode vice versa. The two modes can be freely switched without the need to costly retrain the entire model or deploy different models, serving the real-world demands with many conveniences.

3 METHOD

3.1 DISTRIBUTION DIFFERENCE EXPANDING LOSS WITH MAXIMUM MEAN DISCREPANCY MEASUREMENT

We make possible the CAL paradigm by firstly turning the target domains into "Degraded" mode with an explicit and narrow generalization bound that fits the source domain data but not the target domain data. That goal can convert to the effort of enlarging the distance between the feature distributions of source and target domains. The distance should be calculated with the appropriate measurement. Here, we employ the Maximum Mean Discrepancy (MMD) to measure the distance between every two distributions a and b, which is a kernel two-sample test and can be formulated as:

$$d_{a,b} = ||\mathbb{E}_{f\sim a}[\mathcal{H}(f, f^{'})] - \mathbb{E}_{f^{'}\sim b}[\mathcal{H}(f, f^{'})]||^2, \tag{1}$$

where f and $f^{'}$ denotes two feature distributions. $H(f, f^{'})$ refers to the reproducing kernel computed as $H(f, f^{'}) = e^{-||f-f^{'}||^2}$. $d_{a,b}$ measures the distance of a and b, with a larger $d_{a,b}$ reflecting greater similarity and vise versa. We argue that compared to the other similarity measurements like KL divergence, the MMD is more effective in measuring the inter-domain shared information Wang et al. (2021).

Based on the MMD, a novel distribution difference expanding (DDE) loss is designed to enlarge the distance between the feature distributions of source and target domains, enabling the model to have satisfactory performance in the source domain but degraded performance in the target domain.

We notice that the naive cross-domain training would make similar-to-source-domain data in target domains have higher performance, which contradicts our goal to establish a more explicit generalization bound. Therefore, inspired by the success of loss functions that pay attention to the hard examples mining, we propose to emphasize domain differences with different degrees when computing the overall loss, forcing the optimizer to focus more on the domains with high performance to take it down. We follow works in domain adaptation [cite] and semi-supervised learning [cite] and employ a confidence measurement to evaluate individual samples. Specifically, for each domain t, we get the pseudo label for all samples within it, and then compute their average confidence c_t. Then for domains that have better performance with higher c, a higher weight is assigned.

Eventually, the overall DDE loss can be formulated as:

$$\mathcal{L}_{DDE} = \sum_{t=1}^{T} w_t d_{s,t} = \sum_{t=1}^{T} \frac{c_t}{\sum_{t=1}^{T} c_t} ||\mathbb{E}_{f\sim a}[\mathcal{H}(f, f^{'})] - \mathbb{E}_{f^{'}\sim b}[\mathcal{H}(f, f^{'})]||^2. \tag{2}$$

3.2 THEORETICAL ANALYSIS

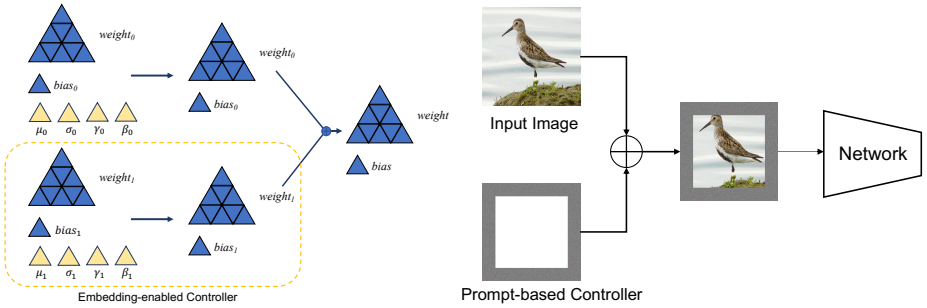
The above-mentioned design can be supported by theoretical analysis. We consider a general domain adaptation theory to show how our method work.Ben-David et al. (2010) Given a source domain D_{Source} and a target domain D_{Target}, let K be a hypothesis space (of a particular VC dimension) for any $k' \in K$:

$$\begin{aligned} R_{Target}(k) \leq R_{Source}(k) + \frac{1}{2} d_{k\Delta k}(D_{Source}, D_{Target}) \\ + \min_{k'\in K}(R_{Target}(k') + R_{Source}(k')), \end{aligned} \tag{3}$$

where R_{Target} and R_{Source} denote the expected source and target errors respectively. $d_{k\Delta k}$ denotes the divergence measuring maximal discrepancy between two distributions. The implementation of DDE loss results in a comparable source error and leads to a significantly greater divergence term, which would eventually bring a much looser upper bound for the target error, as indicated in Equation 3. Such a loosened upper bound results in a significant increase in target error, effectively preventing knowledge transfer into the target domain, thus realizing our goal.

3.3 CONTROLLER FOR STATE-SWITCHING

The above-mentioned method has made a narrow and explicit generalization bound possible, enabling precise control of model performance and making the model in the target domain fall into the "De-



(a) Embedding-enabled Controller (b) Prompt-based Controller

Figure 2: **The illustration of two controllers.** (a) The embedding-enabled controller is performed by modifying the model parameters without adding extra computation burden. (b) the prompt-based controller is performed by adding a learned prompt to the image and feed into the network.

graded" mode. This section illustrates how we control the model to achieve satisfactory performance in its "Enhanced" mode.

We argue that the design of the controller should be **domain-specific** – one controller has to be used in one particular domain to meet the data-specific design of CAL, and **inter-domain general** – the controller should work well on all data in one specific domain. Based on such principles, we design two kinds of controllers: 1) *Implicit Embedding-enabled Controller* that enables the transfer of source-related representations and 2) *Explicit Prompt-based Controller* that modifies the input image to explicitly transfer the domain properties. Their structures are presented in Fig. 2.

Implicit Embedding-enabled Controller. To make possible the performance enhancement in a particular target domain, the model should be controlled to utilize the target-specific information. We realized the information utilization by adding a sub-branch of the network without changing the main trained network that is only trained on the source domain data. The goal is to transfer the target-like embedding to source-like ones and weaken the effect of domain-specific features.

Concretely, as most CNNs composed of the convolution layers and followed by a batch normalization (BN) and a non-linear activation, we consider each layer as a set of parameters, in which $\{W, b\}$ denotes the convolution layer and $\{\mu, \sigma, \gamma, \beta\}$ denotes the BN layer. We add a parallel target-specific layers with the parameters $\{W_t, b_t\}$ and $\{\mu_t, \sigma_t, \gamma_t, \beta_t\}$. As a result, as shown in Fig. 2(a), the new operation for the l-th layer can be formulated as:

$$\begin{aligned} f_{l+1} = &\,\mathrm{BN}\left(\mathrm{Conv}\left(f_l; W, b\right); \mu, \sigma, \gamma, \beta\right) \\ &+\mathrm{BN}\left(\mathrm{Conv}\left(f_l; W_t, b_t\right); \mu_t, \sigma_t, \gamma_t, \beta_t\right). \end{aligned} \tag{4}$$

We found that satisfactory performance can be achieved even when editing layers only at the first stage of the ResNet backbone. The controller with parameter set rc_t can be added or eliminated by choice to select a domain t that falls into the "Enhanced" or "Degraded" status. We make the addition of rc_t at test time and add no extra computation effort. We denote two parallel convolution layers with weight W_0, W_1 and bias b_0, b_1, $Norm_0 = \{\mu_0, \sigma_0, \gamma_0, \beta_0\}$ and $Norm_1 = \{\mu_1, \sigma_1, \gamma_1, \beta_1\}$ denote the mean, variance, weight and bias of two normalization layers $Norm_0$ and $Norm_1$, the calculation is performed with merged convolution represented as one with weight W and bias b:

$$\begin{aligned} x &= Norm_0(W_0 x + b_0) + Norm_1(W_1 x + b_1) \\ &= \gamma_0 \frac{(W_0 x + b_0 - \mu_0)}{\sigma_0} + \gamma_1 \frac{(W_1 x + b_1 - \mu_1)}{\sigma_1} + \beta_0 + \beta_1 \\ &= (\sum_{i=0}^{1} \frac{\gamma_i W_i}{\sigma_i}) x + (\sum_{i=0}^{1} \frac{\gamma_i b_i - \gamma_i \mu_i}{\sigma_i} + \beta_i) \\ &= Wx + b. \end{aligned} \tag{5}$$

Concretely, if rc_t is not added, images for the t-th domain will only be processed by the source-specific parameters, resulting in unsatisfactory performance. While after rc_t is added, by receiving the domain-transfer representation from the extra parameters, performance can be improved, achieving the result of domain adaptation.

Explicit Prompt-based Controller. Besides adding extra parameters to the network, we also offer a control approach that applies to only the data itself, namely data-editing, which is applicable in many real-world applications where the deployed model is fixed and network-editing cannot be performed. Specifically, we introduced a prompt-based approach to make data editing possible. For each domain t, a visual prompt pc_t is learned and added on the images within t. Thus, the prompted image can be formulated as $\hat{x}_t = x + pc_t$. The prompt pc_t is domain-specific and is obtained in training by minimizing the overall loss function, in which the gradient updates are also applied to the prompt parameters. After the prompt have been well-trained, we add the specifically trained visual prompt to the image of the corresponding domain as the input of the network at the test time.

Specifically, the pc_t can be in various forms, for example, pixel patches in random locations, pixel patches in fixed locations, or paddings around the image. We have tested these settings and found that padding achieves the best performance. The spatial information distribution can explain this – we found that in many cases, the information that determines labels is usually in the central part of the image. At the same time, the background contains more domain properties. The padding prompt that changes the boundary region can better control the domain adaptation.

Compared to the embedding-enabled controller, such a prompt-based controller, as illustrated in Fig. 2(b), can be more convenient in real-world applications. Users can use pc_t to switch the domain to the "Enhanced" mode without the need to access and modify the network model parameters.

3.4 Loss Function

With the method to control the network into "Degraded" or "Enhanced" mode, the training of the CAL model is targeted for two goals: 1) enlarging the distance between the source and target domains without the controller, and 2) bringing the distance between source and target closer using controllers. For the first goal, the proposed DDE loss for adjusting the inter-domain distribution distance is applied. For the second goal, we use the opposite of DDL loss. The final loss for training the CAL model can be formulated as:

$$\mathcal{L}_{CAL} = \mathcal{L}_{DDE}\left(\mathcal{D}_s, \mathcal{D}_t; \theta\right) - \mathcal{L}_{DDE}\left(\mathcal{D}_s, \mathcal{D}_t; \theta, \{c_t\}_{t=1}^{T}\right) + \mathcal{L}_{CE}, \tag{6}$$

where $\mathcal{L}_{CE}$ refers to the cross entropy loss for source data training. c_t denotes the controller for the target domain t. It can be either the embedding-enabled controller rc_t or the prompt-based controller pc_t. It is also an option to use both the representation and prompt-based controllers for training, as we have evaluated and demonstrated the result in the following part.

4 Experiments

4.1 Dataset

We perform experiments on two commonly-used domain adaptation datasets: **Digits** and **Multi-PIE**. Digits is a combination of 4 different digits datasets including MNIST (**mt**), MNIST-M (**mm**), SVNH (**sv**) and USPS (**up**). Each dataset contains 10 classes that represent 10 different digits. The Multi-PIE dataset contains face images for 337 individuals with different views, expressions and illumination conditions. Specifically in our experiments, following Gholami et al. (2020), images from different camera views including **C05**, **C08**, **C09**, **C13** and **C14** are regarded as 5 different domains, and each one contain 5 different face expressions (normal, smile, surprise, squint, disgust, scream) as the classes.

4.2 Implementation Details

We use Adam as the optimizer for training, with the learning rate of 0.0002 and the momentum parameters being 0.5 and 0.999. The batch size is 16 for each domain and the input images were mean-centered/rescaled to range from -1 to 1. We use ResNet50 as the backbone network for all experiments.

Table 1: Classification accuracy on the Digits dataset, where mt, mm, sv and up denote four different sub-datasets including MNIST, MNIST-M, SVNH and USPS, respectively. The left side of each arrow denotes the source domain and the right side denotes the tested target domain. The first column shows the performance of the baseline setting, which indicates training a model only on the source domain and directly evaluating on the target domains. EC and PC refer to the embedding-enabled controller and prompt-based controller, respectively.

Setting	mt → mm	mt → sv	mt → up	sv → mt	sv → mm	sv → up
Baseline	59.12 ± 0.55	35.38 ± 0.70	80.05 ± 0.44	66.18 ± 0.58	44.52 ± 0.77	43.80 ± 0.71
"Degraded" Status	9.13 ± 0.42	5.80 ± 0.60	13.75 ± 0.41	9.85 ± 0.50	7.10 ± 0.55	6.90 ± 0.42
"Enhanced" Status w/ EC	87.94 ± 0.30	52.19 ± 0.25	92.95 ± 0.41	91.37 ± 0.33	63.90 ± 0.32	65.75 ± 0.29
"Enhanced" Status w/ PC	82.30 ± 0.32	49.77 ± 0.29	88.58 ± 0.38	86.60 ± 0.50	60.12 ± 0.40	61.11 ± 0.29
"Enhanced" Status w/ EC & PC	88.50 ± 0.39	53.90 ± 0.41	93.78 ± 0.46	92.91 ± 0.35	65.45 ± 0.43	67.06 ± 0.33

Table 2: Classification accuracy on the Multi-PIE dataset, where C05, C08, C09, C13 and C14 refer to five different views representing different domains. The left side of each arrow denotes the source domain and the right side refers to the tested target domain. The first column shows the performance of the baseline setting, which indicates training a model only on the source domain and directly evaluating on the target domains. EC and PC refer to the embedding-enabled controller and prompt-based controller respectively.

Setting	C13 → C05	C13 → C08	C13 → C09	C13 → C14
Baseline	53.29 ± 0.34	49.18 ± 0.50	44.70 ± 0.42	63.91 ± 0.25
"Degraded" Status	6.12 ± 0.39	5.77 ± 0.44	5.20 ± 0.60	7.39 ± 0.29
"Enhanced" Status w/ EC	76.62 ± 0.44	64.82 ± 0.20	68.65 ± 0.35	87.83 ± 0.30
"Enhanced" Status w/ EC	72.19 ± 0.40	60.50 ± 0.25	65.02 ± 0.29	83.95 ± 0.33
"Enhanced" Status w/ EC & PC	77.69 ± 0.38	65.92 ± 0.20	69.11 ± 0.31	88.59 ± 0.23
Setting	C14 → C05	C14 → C08	C14 → C09	C14 → C13
Baseline	64.73 ± 0.45	39.88 ± 0.53	45.29 ± 0.38	65.13 ± 0.44
"Degraded" status	7.87 ± 0.19	4.39 ± 0.37	5.50 ± 0.30	7.99 ± 0.22
"Enhanced" status w/ EC	86.57 ± 0.26	60.92 ± 0.17	58.37 ± 0.44	84.71 ± 0.19
"Enhanced" status w/ PC	82.07 ± 0.30	56.20 ± 0.21	56.03 ± 0.32	80.91 ± 0.22
"Enhanced" status w/ EC & PC	88.00 ± 0.21	62.05 ± 0.35	59.60 ± 0.40	85.94 ± 0.25

4.3 Main Results

In our task, each target domain can switch between the "Enhanced" and the "Degraded" statuses by using the controllers or not. We report the experimental result for performance degradation and enhancement using the embedding-enabled controller and prompt-based controller. We perform experiments on both the Digits and Multi-PIE datasets. For Digits, we test 2 different settings, in which the MNIST (**mt**) and SVNH (**sv**) datasets are used as the source, and other datasets are the target domains. For Multi-PIE, we also evaluate two conditions where $C13$ and $C14$ are the source domain, respectively and others are the target domains. Table. 1 shows the result for the Digits dataset and Table. 2 shows the result for the Multi-PIE dataset.

Baseline Results. As a baseline, we first train a supervised task only on the source domain, and then evaluate its results on other domains that do not participate in the training. The results are shown in the first column of Table. 1 and Table. 2 for Digits and Multi-PIE respectively. As can be observed, while only training using source data, since the feature distributions are not explicitly enlarged, the model can still achieve considerable performance on the unseen target domains due to the semantic similarity. Thus in this case, the unauthorized users can still apply the trained model to their own domains, jeopardizing the property rights.

Results for the "Degraded" Status. We further test the performance after using the proposed methods. With the controller not applied, training the network solely on the target domain with DDE loss makes it hard to make the correct classification decisions for the network to perform in the target domain. As can be seen from the second column in Table. 1 and Table. 2 for Digits and Multi-PIE

Table 3: Ablation results for the DDE loss on the Digits dataset.

Setting	mt → mm	mt → sv	mt → up
Baseline	59.12 ± 0.55	35.38 ± 0.70	80.05 ± 0.44
"Degraded" Status w/ Weight Strategy	9.13 ± 0.42	5.80 ± 0.60	13.75 ± 0.41
"Degraded" Status w/o Weight Strategy	13.59 ± 0.45	9.39 ± 0.30	17.08 ± 0.44
"Enhanced" Status w/ EC w/ Weight Strategy	87.94 ± 0.30	52.19 ± 0.25	92.95 ± 0.41
"Enhanced" Status w/ EC w/o Weight Strategy	82.18 ± 0.33	49.04 ± 0.20	87.56 ± 0.46
"Enhanced" Status w/ PC w/ Weight Strategy	82.30 ± 0.32	49.77 ± 0.29	88.58 ± 0.38
"Enhanced" Status w/ PC w/o Weight Strategy	77.91 ± 0.32	44.99 ± 0.22	85.19 ± 0.40

respectively, compared to the baseline performance where the model was trained on source data and directly used for testing on target data, classification accuracy for the "Degraded" status in our method is significantly decreased, showing the data is apparently out of the generalization bound of the model. The maximal accuracy drop occurs in the mt-up setting, where the accuracy decreases from 80.05 to 13.75 after using our method. The results verify the effectiveness of our method in preventing unauthorized data been utilized by the network.

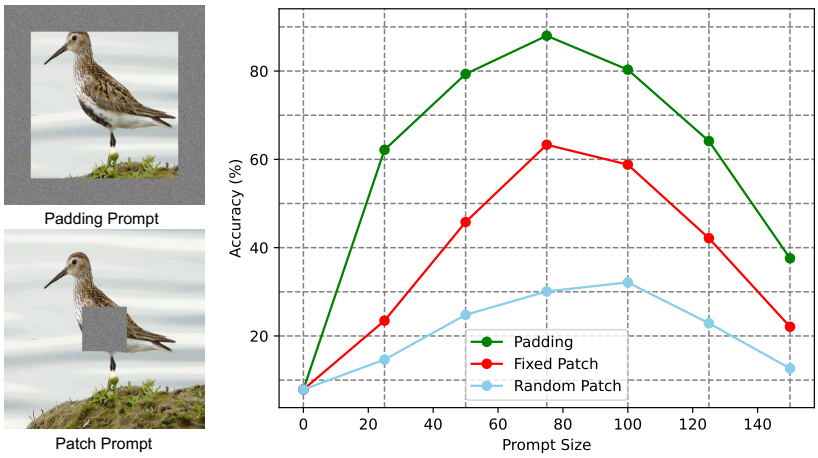
Results with the Embedding-enabled Controller. The controllers can transfer embedding information from the target domain, switching the model from "Degraded" to the "Enhanced" status. The results are shown in the third column of Table. 1 for Digits dataset and Table. 2 for Multi-PIE dataset. It can be found that with the embedding-enabled controller, the average accuracy over all target domains can be significantly improved, with the maximum increase reaching 80%, showing the effectiveness of the embedding-enabled controller.

Results with the Prompt-based Controller. We further evaluate the effectiveness of the prompt-based controller. The results are presented in the fourth column of Table. 1 for Digits dataset and Table. 2 for Multi-PIE dataset. It can also observe a significant accuracy improvement after using the prompt-based controller. We find that the embedding-enabled controller demonstrated better capability than the prompt-based controller in our experiment. We hypothesize the reason could be that the embedding-enabled controller can deliver the effective domain-transferred information in each layer, while by only modifying the input image, the model will still struggle to extract the useful information.

Results for using Both Embedding-enabled and Prompt-based Controllers. In addition to using the embedding-enabled and prompt-based controllers individually, it is also an option to apply them together, and the results are shown in the fifth column of Table. 1 and Table. 2 for Digits and Multi-PIE respectively. The results under this setting are slightly better than only using any individual controllers, with the average accuracy increasing 2.16% for Digits and 1.93% for Multi-PIE.

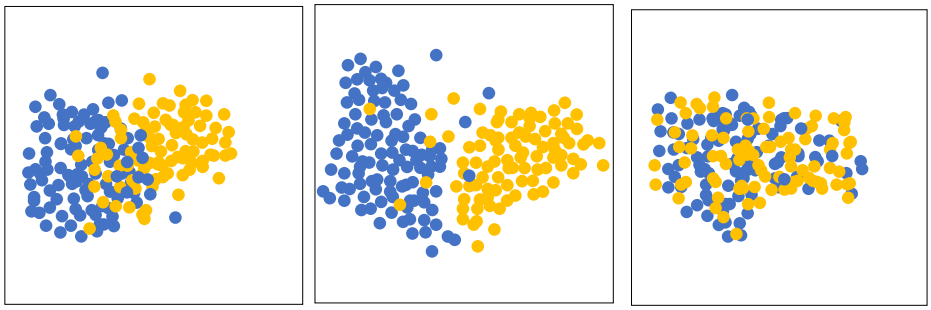
4.4 Ablation Studies

Ablation of DDE Loss. Here we conduct experiments to verify the effectiveness of the proposed DDE loss. In our method, the DDE loss is employed to enlarge the inter-domain feature distributions. Thus, it can encourage target domains to have a bad performance by increasing their distance to the source one. And by using its opposite as the optimization goal, the performance for target domains after using controllers can be pushed to be better. One crucial design in the DDE loss is the confidence-based weight, which encourages the optimizer to pay more attention to domains with better performance. From the experimental results performed on the Digits dataset and presented in Table. 3, we observe that the domain imbalance can negatively affect the performance if the weight strategy is not applied, with the average accuracy for targets under degraded status increasing from 9.56% to 13.35%, the average accuracy for targets under enhanced status decreasing from 77.69% to 72.93% when using the embedding-enabled controller, and decreasing from 73.55% to 69.36%. The results demonstrate that such a weight strategy is effective for both the non-adaptive target performance degrading and the adaptive target performance improvement.



(a) (b)

Figure 3: (a) An example of the padding prompt and patch prompt. (b) The ablation results of different prompt design with different prompt sizes.



(a) Baseline (b) Degraded Status (c) Enhanced Status

Figure 4: **Visualization of sample feature distributions under different situations.** (Best view in color) The blue and orange dots denote the samples for source and target domains respectively within the same class. Baseline refers to training a model on source and then testing on the target domain. The experiments are conducted on the C14-C05 setting for the Multi-PIE dataset.

Ablation of Prompt-based Controllers. The prompt used for constructing the controllers can be designed in multiple forms. As shown in Fig. 3, besides the padding way we have adopted in our method, we can also use pixel patches at random locations and pixel patches at fixed locations for the visual prompt. Here we perform experiments on the Multi-PE dataset C14-C05 setting to validate the effectiveness of different forms, with the prompt size p ranging from 0 to 150. We observe that the padding prompt with the size of 75 can achieve the best performance. For classification task, information that determines labels is usually in the central part of the image, while the background contains more domain properties. Thus padding prompt that changes the boundary region can better control the domain adaptation.

4.5 VISUALIZATION

We provide the visualization results of feature distribution under different statuses in Fig. 4 (best view in color), where the blue and orange dots denote the samples for source and target domains within the same class. Fig.4 (a) shows the baseline where the model is trained on the source domain and then tested on the target domain. Despite the significant difference, there are still overlapping areas for the source and target domain distributions in this case. Thus, the learned decision boundary on the source can still cover some areas of the target region, thus achieving the target classification

accuracy that is not very low. Fig.4 (b) shows the degraded status after using our method. As can be observed, two distributions for the source and target are separated, i.e., the generalization bound "shrank." Thus, the learned decision boundary for the source domain cannot cover the target domain, resulting in degraded performance. Fig.4 (c) presents the "enhanced" status after using our proposed control method, where two distributions for source and target almost completely overlap, i.e., the generalization bound "enlarged." In this case, the source decision boundary can be shared with the target domain, resulting in a satisfactory performance. These visualization results further demonstrate the effectiveness of our method.

5 Related Work

The most related to our CAL approach is the realm of domain adaptation (DA) and domain generalization (DG). When the model is trained solely on the source domain, the performance will degrade in the target domain outside the training distribution. The objective is to boost the performance in the target domain through enhanced model generalization Ben-David et al. (2010). In domain adaptation, target data can be accessed, while domain generalization has no access to the target data Xu et al. (2021); Dong et al. (2021). For domain adaptation, the popular approach is to transfer the knowledge from a source domain to the target domain. Attempts have been proposed to minimize the Maximum Mean Discrepancy (MMD) Muandet et al. (2013); Ghifary et al. (2016a), to learn an auxiliary reconstruction task Ghifary et al. (2016b); Hoffman et al. (2018), to design and implement a gradient reversal layer Tzeng et al. (2017), to focus on classifier discrepancy and align source and target features Lee et al. (2019); Saito et al. (2018), or to implement self-training Zhu et al. (2017); Zou et al. (2019). Among the abovementioned approaches, many works are unsupervised, in which the label of the target domain is unknown. In this work, we follow the unsupervised settings, which are closer to real-world scenarios.

Contrary to DA & DG, Wang et al. (2021) propose a Non-Transferable Learning (NTL) paradigm that can reduce the generalization bound for AI models aiming for model security and intellectual property (IP) protection. NTL resolves the issue that previous IP protection method Alam et al. (2020) can only prevent unauthorized users to access the network, but cannot forbid the access of unauthorized data. However, it is a supervised approach that relies on large amount of labeled data from the target domain, which are hard to obtain. Moreover, after the model well-trained, users cannot reclaim the access of the unauthorized domain.

As mentioned in the introduction, the main difference between our CAL approach and DA & DG & NTL is that we aim to control the generalization bound toward different domains in an unsupervised setting, which, to the best of our knowledge, is still a research gap to be filled.

6 Conclusion

In this work, we propose a novel Controllable Adaptive Learning (CAL) scheme, which can control the model performance based on the data feed-in, enabled by manipulating the generalization bound of the model. We provide theoretical analysis and extensive experiments to demonstrate how our method works. Specifically, we propose a novel Distribution Difference Expanding (DDE) Loss to add convergence terms in the target domain and restrict the information transfer from the target to the source domain (making the model falls into a "Degraded" mode). We then introduced two controllers, an implicit embedding-enabled controller, and an explicit prompt-based controller, to transfer relevant information and realize the effect of domain adaptation (making the model fall into an "Enhanced" mode). We believe our newly introduced paradigm will broadly impact the IP protection AI model, solving AI-related ethics problems and a wide range of commercial purposes under AIaaS in the near future.

REFERENCES

Sk Miraj Ahmed, Dripta S Raychaudhuri, Sujoy Paul, Samet Oymak, and Amit K Roy-Chowdhury. Unsupervised multi-source domain adaptation without access to source data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10103–10112, 2021.

Manaar Alam, Sayandeep Saha, Debdeep Mukhopadhyay, and Sandip Kundu. Deep-lock: Secure authorization for deep neural networks. *arXiv preprint arXiv:2008.05966*, 2020.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. *Advances in neural information processing systems*, 24, 2011.

Jennifer Cobbe and Jatinder Singh. Artificial intelligence as a service: Legal responsibilities, liabilities, and policy challenges. *Computer Law & Security Review*, 42:105573, 2021.

Jiahua Dong, Yang Cong, Gan Sun, Zhen Fang, and Zhengming Ding. Where and how to transfer: knowledge aggregation-induced transferability perception for unsupervised domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Muhammad Ghifary, David Balduzzi, W Bastiaan Kleijn, and Mengjie Zhang. Scatter component analysis: A unified framework for domain adaptation and domain generalization. *IEEE transactions on pattern analysis and machine intelligence*, 39(7):1414–1430, 2016a.

Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li. Deep reconstruction-classification networks for unsupervised domain adaptation. In *European conference on computer vision*, pp. 597–613. Springer, 2016b.

Behnam Gholami, Pritish Sahu, Ognjen Rudovic, Konstantinos Bousmalis, and Vladimir Pavlovic. Unsupervised multi-target domain adaptation: An information theoretic approach. *IEEE Transactions on Image Processing*, 29:3993–4002, 2020.

Drew Harwell. This facial recognition website can turn anyone into a cop—or a stalker. In *Ethics of Data and Analytics*, pp. 63–67. Auerbach Publications, 2022.

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

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

Minoru Kuribayashi, Takuro Tanaka, and Nobuo Funabiki. Deepwatermark: Embedding watermark into dnn model. In *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1340–1346. IEEE, 2020.

Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. Sliced wasserstein discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10285–10295, 2019.

Sebastian Lins, Konstantin D Pandl, Heiner Teigeler, Scott Thiebes, Calvin Bayer, and Ali Sunyaev. Artificial intelligence as a service. *Business & Information Systems Engineering*, 63(4):441–456, 2021.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12009–12019, 2022.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pp. 10–18. PMLR, 2013.

Bita Darvish Rouhani, Huili Chen, and Farinaz Koushanfar. Deepsigns: A generic watermarking framework for ip protection of deep learning models. *arXiv preprint arXiv:1804.00750*, 2018.

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*, pp. 3723–3732, 2018.

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*, pp. 7167–7176, 2017.

Yusuke Uchida, Yuki Nagai, Shigeyuki Sakazawa, and Shin'ichi Satoh. Embedding watermarks into deep neural networks. In *Proceedings of the 2017 ACM on international conference on multimedia retrieval*, pp. 269–277, 2017.

Lixu Wang, Shichao Xu, Ruiqi Xu, Xiao Wang, and Qi Zhu. Non-transferable learning: A new approach for model ownership verification and applicability authorization. In *International Conference on Learning Representations*, 2021.

Shichao Xu, Lixu Wang, Yixuan Wang, and Qi Zhu. Weak adaptation learning: Addressing cross-domain data insufficiency with weak annotator. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8917–8926, 2021.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5982–5991, 2019.