Overcoming Learning Imbalance with Fusing Vision-Language Model Knowledge for Black-Box Domain Adaptation

Zhixin Zeng (zengzx@nudt.edu.cn)
College of Computer Science and Technology, National University of Defense Technology
Changsha, China

Yusen Zhang (zhangys04@nudt.edu.cn)
College of Computer Science and Technology, National University of Defense Technology
Changsha, China

Abstract

Once the human brain learns a concept, it can easily transfer the learned knowledge across diverse environments without referring back to the original learning materials. Inspired by this cognitive process, Black-Box Domain Adaptation (BBDA) has been purposed to transfer the knowledge learned in a black-box source model to the target domain without any premise for source data or model parameters. Existing BBDA methods mainly rely on knowledge distillation or sample selection with pseudo labels, overlooking the different learning difficulties of classes. This results in easy-learning classes dominating the adaptation process and thus degrades adaptation performance. Motivated by the significant success of Vision-Language models (ViL model), we propose an novel method that integrates the knowledge of ViL model to achieve adaptation while mitigating learning imbalance. Experiments on various datasets demonstrate the effectiveness of the proposed method.

Keywords: Black-box; Domain Adaptation; Learning Imbalance

Introduction

Deep learning models achieves remarkable success in various applications. However, the model performance may be severely degraded when the environment is changed. The cause of this phenomenon lies in the fact that the new domain data (target domain) and the training data (source domain) do not adhere to independent identical distribution assumption, which is also referred to as domain shift (L. Zhang & Gao, 2022). Regarding humans, cognitive adaptation is a fundamental ability of human to perceive the world (Crisp & Turner, 2011; Dunwoody, Haarbauer, Mahan, & Marino, 2000). Once a concept has learned, human could leverage the concept to recognize object that similar but different one in new environment. Motivated by the cognitive adaptation of human, **U**nsupervised **D**omain **A**daptation (UDA) has been proposed to adapt deep learning models to the new domain.

UDA transfers knowledge from the sufficient labeled data of source domain to the unlabeled target domain (Zhou, Liu, Qiao, Xiang, & Loy, 2022; Oza, Sindagi, Sharmini, & Patel, 2023). Existing UDA methods are designed under the premise that the source data or source model parameters are always available for conducting target domain adaptation. However, in practical scenarios, the premise may be unable to fulfill due to privacy protection, storage issues or transmission overheads, etc. As to the premise for cognitive adaptation in humans, once a concept about an object has been learned, using the concept to recognize the object in a new environment requires only the learned knowledge. there is no necessity to reuse the previous learning materials. However, most of UDA methods become infeasible in the absence of source data or model parameters.

To circumvent the need for source data, **S**ource-**F**ree **D**omain **A**daptation (SFDA) was proposed and attempted to adapt the source model to the target domain without using source data (Li, Yu, Du, Zhu, & Shen, 2024; Fang, Yap, Lin, Zhu, & Liu, 2024). However, SFDA methods are devised under the premise that the internal output and parameters of source model are assumed to be accessible during the training. Therefore, there are still significant constraints on flexibility, and the issue of source privacy remains unresolved. Inspired by the ability of human cognition adaptation process, **B**lack-**B**ox **D**omain **A**daptation (BBDA) has been proposed. BBDA utilizes only the output prediction of source model to achieve adaptation to the new target domain, thereby avoiding the necessity for source data or model parameters, which is arguably the minimal requirement of domain adaptation (Liang, Hu, Feng, & He, 2022). Current BBDA methods mostly focus on distilling knowledge from source model, or selecting reliable unlabeled target domain samples by certain criterion for adaptation. However, none of these methods take into account the learning imbalance problem due to the different learning difficulty of each class during the adaptation process, which results in sub-optimal performance.

The learning imbalance problem primarily originates from the uneven numbers of learning samples among different classes, which results from the intrinsic cross-domain bias hidden in the source model. For target domain, there will be a large number of misclassified pseudo labels predicted by the source model, leading to an uneven quantity of samples for certain classes. We refer to the classes with a large number of samples as easy-learning classes, and the classes with few samples as hard-learning classes. This situation leads to a more rapid learning pace for easy-learning classes in contrast to the hard-learning classes. Consequently, the training process becomes biased towards the easy-learning classes, thereby failing to achieve entire target domain adaptation.

To tackle the learning imbalance problem, we propose **O**vercoming **L**earning **I**mbalance with Fusing **V**ision-Language Model Knowledge for Black-Box Domain **A**daptation (OLIVA) to achieve target domain adaptation. Specifically, we first utilize a novel entropy-weighted fusion

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

strategy to fuse knowledge from the black-box source model and vision-language model (ViL model) (Radford et al., 2021). The predictions from ViL model and black-box source model serve as generic knowledge, since the ViL model are pre-training on various data collecting from Internet. The predictions from black-box source model are regard as the task-specific knowledge. Introducing generic knowledge would mitigate the bias of the source model, while task-specific knowledge can complement the generic knowledge for the specific adaptation task. Then we leverage a Gaussian Mixture Model (GMM) (Reynolds, 2009) to divide clean and noisy samples. The clean/noisy division is unevenly distributed, the easy-learning classes have large numbers of samples, and the hard-learning classes have a relatively small number of samples. To overcome the imbalance learning, we propose learning degree-guided complementation to complement the hard-learning classes in the clean set from the noisy set. Finally, we achieve adaptation via contrastive negative learning by using the clean/noisy divided samples.

In summary, this work tackles the critical yet understudied learning imbalance problem in BBDA, proposing OLIVA framework that integrates generic and task-specific knowledge from both ViL and source model for target adaptation. OLIVA consists of three core components: 1) An entropy-weighted knowledge fusio that reconciles task-specific predictions from black-box source models with generic knowledge of ViL model. 2) Target domain division with learning degree-guided complementation. 3) contrastive negative learning for adaptation. Multiple experiments are conduct on various benchmark datasets validate the superiority of our purposed method.

Related Works

Conventional and Source-free Domain Adaptation

Conventional UDA aims to utilize both labeled data of source domain and unlabeled target domain data to learn a target model for target domain adaptation. The main approaches of Conventional UDA mainly include explicitly reducing the domain discrepancy and learning domain-invariant features via adversarial training. Discrepancy minimization methods, such as those developed by (Y. Zhang, Liu, Long, & Jordan, 2019; Sun, 2015; Zellinger, Grubinger, Lughofer, Natschlger, & Saminger-Platz, 2017), focus on reducing the differences of feature space between the source and target domain. Adversarial based methods, such as (X. Chen, Wang, Long, & Wang, 2019; Long, Cao, Wang, & Jordan, 2018; Lee, Kim, Kim, & Jeong, 2019), employ adversarial training to learning domain-invariant feature representation for adaptation. However, these methods require simultaneous access to source and target data, raising privacy and storage concerns in real-world practice.

SFDA adapts to the target domain without source domain data. SHOT (Liang, Hu, Wang, He, & Feng, 2021) freezes the source feature extractor and optimizes mutual information between target features and predictions. NRC (S. Yang, Van de Weijer, Herranz, Jui, et al., 2021) enforces prediction consistency among neighbors in the feature space to leverage local semantic structures. GSFDA (S. Yang, Wang, Van De Weijer, Herranz, & Jui, 2021) activate domain-specific channels to enhance neighborhood consistency for adaptation. AaD (S. Yang, Jui, van de Weijer, et al., 2022) optimizes two regularization terms for an upper-bound of the neighborhood clustering objective to achieve adaptation. Although SFDA method achieve target domain adaptation without source data, it still requires the parameters or internal outputs of the source model, limiting its applicability in more practical scenarios.

Black-Box Domain Adaptation

BBDA methods conduct domain adaptation relying solely on black-box source predictor and unlabeled target data. LNL (H. Zhang, Zhang, Jia, & Zhang, 2021) adopt noisy label learning techniques, using iterative sample selection to refine pseudo-labels. DINE (Liang et al., 2022) pioneers a two-stage framework combining knowledge distillation, information maximization, and MixUp regularization to transfer source knowledge. BETA (J. Yang et al., 2023) divides the target domain into easy/hard subset and employs twin networks for mutual distillation, reducing confirmation bias. HD/SD-SHOT (Liang et al., 2021) adapt SHOT to BDA by leveraging source-predicted labels but struggle with minority class forgetting. RFC (S. Zhang, Shen, Lü, & Zhang, 2024) addresses this by introducing Selection Training (ST) to revisit forgotten classes and Neighborhood Clustering (NC) to balance feature learning. Despite the progress achieved by these methods, they fail to take into account the learning imbalance problem during the adaptation process, and thus resulting in sub-optimal performance.

Methodology

Problem Definition

BBDA aims to transfer knowledge from source domain to target domain without access to source data or model parameters. Concretely, we have a black-box source model $f_s(x) = h_s(g_s(x))$ that consists of a feature extractor $g_s : \mathcal{X}_s \to \mathbb{R}^d$ to map the input data to feature space and a classifier module $h_s : \mathbb{R}^d \to \mathbb{R}^K$. $f_s(x)$ is pre-trained on the labeled source domain data $\mathcal{D}_s$ containing n_s samples $\{(x_i^s, y_i^s)\}_{i=1}^{n_s}$. $\mathcal{D}_t = \{(x_i^t)\}_{i=1}^{n_t}$ denotes n_t unlabeled samples from the target domain. The distribution of source and target domain is not identical, i.e. $\mathcal{D}_s \neq \mathcal{D}_t$. Both the source and target domain have the same K classes, i.e., $Y_s = Y_t$. The objective of BBDA task is to learn a target model $f_t(x) = h_t(g_t(x))$ by utilizing unlabeled target domain $\mathcal{D}_t$ and noisy predictions $\tilde{Y}_t = f_s(\mathcal{X}_t)$ from the black-box source model, without access to source data $\mathcal{D}_s$ and internal output or parameters of $f_s(x)$.

Entropy-Weighted Knowledge Fusion

The black-box source model is trained on the source domain. Inevitably, there inherently exists domain bias when making

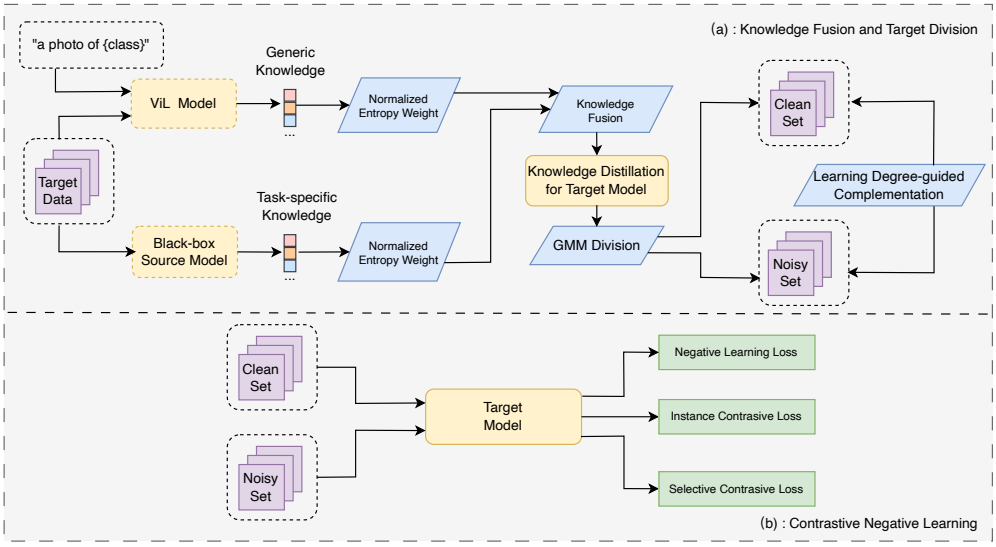


Figure 1: The OLIVA framework. We first use entropy-weighted knowledge fusion to fuse both the generic knowledge and task-specific knowledge. Then the fused knowledge is leveraged to warm up the target model. Subsequently, we employ a GMM to divide the samples of target domain into clean/noisy set. Learning degree-guided complementation strategy is devised for mitigating the imbalance in the division. Lastly, the complemented clean/noisy sets is utilized for target adaptation.

predictions on target domain data. The domain bias would cause a large number of noisy pseudo-label on target samples, and increasing the learning imbalance.

Inspired by the recent success of the ViL foundation model in various tasks, we integrate the knowledge of the ViL model to mitigate the domain bias. The ViL model is trained with a large amount of image-text pairs data collected from the Internet. Thus, it has less bias and can provide more general knowledge, yet it lacks task-specific knowledge regarding the target domain. We adopt the most representative ViL model CLIP in our proposed method, given the fact that it has become a widely used foundation ViL model and has empowered numerous works. On the other hand, despite the significant domain bias hidden in the source model, it contains task-specific knowledge for adaptation.

To combine the advantages of the source model and the ViL model, we present the entropy-based knowledge Fusion. Specifically, we obtain the predictions of the source model and the ViL model for each target sample. For the ViL model, it consists of an image encoder and a text encoder, parameterized by θ_V and θ_L. The image and text encoder take an image x_i and text prompt t_k as input, and output corresponding encoded image feature $\theta_V(x_i)$ and prompt feature $\theta_L(t_k)$ respectively. The standard text prompt "a photo of $\{$class$\}$" is utilized as text prompt for text encoder, the $\{$class$\}$ is depends on the name of classes in the dataset. The prediction of a target sample is calculated by the cosine similarity between encoded image and text prompt feature. For the source model, we take the output probability as prediction for each samples. Then we calculate the normalized entropy values of predictions from source and ViL model by:

$$H_{norm} = \frac{-\sum_{i=1}^{K} y_i \log(y_i)}{\log(K)} \tag{1}$$

Next, we leverage both the normalized source H^s_{norm} and ViL entropy $H^v_{norm}$to obtain the fuse weight:

$$w_s = 1 - \frac{H^s_{norm}}{H^s_{norm} + H^v_{norm}}, w_v = 1 - \frac{H^s_{norm}}{H^s_{norm} + H^v_{norm}} \tag{2}$$

$$\tilde{y}_{f,i} = w_s \tilde{y}_{s,i} + w_v \tilde{y}_{v,i} \tag{3}$$

The fuse weight w_s and w_v is employed to adjust the proportion of knowledge from the corresponding model for knowledge fusion. The fused knowledge $\tilde{y}_{f,i}$ is regarded as the final predictions for each target sample.

Target Division with Learning Degree-guided Complementation

The fused pseudo-labels are naturally noisy due to the domain distribution gap between source and target domain. Previous studies (J. Yang et al., 2023) indicate that the loss value of clean samples tends to be smaller than that of noisy samples during the initial training stages, and the loss value distribution of all target samples exhibits a bi-modal pattern. In light of this property, we are motivated to divide the pseudo-labeled target domain data into clean and noisy subsets by utilizing the distribution of loss values. Firstly, we leverage knowledge distillation from (Liang et al., 2022) to warm up the target model based on the fused predictions:

$$\mathcal{L}_{kd}\left(f_t; \mathcal{X}_t, f_s\right) = \mathbb{E}_{x_t \in \mathcal{X}_t} \mathcal{D}_{kl}\left(\tilde{y}_{f,i} \,||\, f_t\left(x_t\right)\right) \tag{4}$$

$$\mathcal{L}_{im} = h\left(\mathbb{E}_{x_t \in \mathcal{X}_t} f_t\left(x_t\right)\right) - \mathbb{E}_{x_t \in \mathcal{X}_t} h\left(f_t\left(x_t\right)\right) \tag{5}$$

where the $\mathcal{L}_{kd}$ and $\mathcal{L}_{im}$ is the knowledge distillation and information maximization loss. The $\mathcal{D}_{kl}$ denotes the Kullback-Leibler divergence, and h denotes $h(p_i) = -\sum y_i \log y_i$. After distillation, we adopt cross entropy loss to calculate the loss value of target samples to pseudo label:

$$\mathcal{L}_{ce}(\mathbf{x}_i^t) = -\sum_{k=1}^{K} \tilde{y}_i^k \log\left(h_t^k\left(g_t(x_i^t)\right)\right) \tag{6}$$

where h_t^k is the softmax probability for class k output by target model. After calculating the loss value of target samples, we then fit the loss distribution via GMM.

By using the fitted GMM, we can assign a probability of belonging to the clean cluster based on the loss value of each target sample. Concretely, the probability of a sample belonging to the clean cluster is equivalent to the posterior probability $\tilde{p}(c|l_i(\mathbf{x}_i^t))$, where c indicates the Gaussian component associated with a lower loss value. Then the clean and noisy set can be divided by the a given threshold τ of clean probability:

$$\begin{aligned} \mathcal{X}_{clean} &= \{\mathbf{x}_i|\mathbf{x}_i \in \mathcal{X}_t, \rho_i^c \geq \tau\}, \\ \mathcal{X}_{noisy} &= \{(\mathbf{x}_i|\mathbf{x}_i \in \mathcal{X}_t, \rho_i^c < \tau\} \end{aligned} \tag{7}$$

The clean set is divided based on the loss value, the majority of classes within it belong to the easy-learning classes, whereas the noisy set mainly consists of hard-learning classes. This deteriorates the adaptation of target domain.

To overcoming the learning imbalance problem, we purpose learning degree-guided complementation strategy to complement the hard-learning classes in clean set. In detail, the global learning degree γ_g is defined as the proportion of clean samples in the target domain n_t, which reflects the current target model learning degree with respect to the entire target domain:

$$\gamma_g = \frac{n_{clean}}{n_t}, \gamma_l = \frac{Max\left(n_i^k, 1\right)}{Mean\left(n_{clean}\right)} \tag{8}$$

The local learning degree γ_l is defined as the ratio of samples in each class relative to the average value of samples across all classes in the clean set. Note that the max operation $Max(\cdot)$ is used to avoid the numerator being zero value. The local learning degree indicates the concrete learning degree of each class in the clean set, which reflects how well the target model learns the specific class. Classes with sample numbers below the average value across all classes in the clean set are identified as hard-learning classes and will be complemented from noisy set according to their global and local learning degrees.

With the global and local learning degrees, we utilize the Beta distribution to determine the number of samples to be complement for the hard-learning classes. The Beta distribution involves two parameters, α and μ. The larger alpha is relative to gamma, the closer the sampling expectation will be to 1. Conversely, the closer it will be to 0. The complement co-efficient η is calculated as:

$$\eta = Round(Beta(1-\gamma_l, \gamma_g)),\ Beta(\alpha,\mu) = \frac{\Gamma(\alpha)\Gamma(\mu)}{\Gamma(\alpha+\mu)} \tag{9}$$

Where $\Gamma(\cdot)$ is Gamma function. In (9), if the α is greater than the μ, it indicates that the number of samples in this hard-learning class is much less than the average number of all classes in the clean set. Therefore, more samples of this class need to be complemented. When the μ is greater than the α, it means that most of the samples in the target domain have been learned, and the remaining samples in the noisy set are highly likely to be mislabeled. In this case, only a limited number of samples should be complemented.

After obtaining η, we calculate the entropy of each hard-learning classes in the noisy set. Then, we sort entropy value in ascending order and select the first η% samples as complemented samples:

$$U_{comp}^h = \text{argsort}_{\text{ascending}}\left(H_j\right) * [0 : \alpha * n_{hard}^j], j \in K_{hard} \tag{10}$$

The complemented clean/noise set division after complementation will be leveraged for the adaptation training.

Contrasitve Negative Learning for Adaptation

After obtaining clean/noisy division of target domain, contrastive negative learning is devised to conduct adaptation. Conventional contrastive learning (T. Chen, Kornblith, Norouzi, & Hinton, 2020) that learning representation with all samples may introduce noise into feature space. Considering that the divided sets have distinct label qualities, it is reasonable to treat them in different ways. Following this spirit, we introduce contrastive negative adaptation method designed to learn representations from each set with differentiated treatments, thus preventing noise accumulation.

To this end, we devise contrastive negative learning to learn representation from clean and noisy sets respectively. Our contrasitive negative learning consists of two components tailored for clean and noisy sets. Given the fact that the clean set have high quality labels, we leverage instance contrastive learning with negative learning to learn target domain with clean set. Specifically, we employ instance contrastive learning with the weak and strong augmented versions of a sample:

$$\mathcal{L}_{icon} = -\log \frac{\exp\left(x_{w,i}^t, x_{s,i}^t\right)}{\exp\left(x_{w,i}^t, x_{s,i}^t\right) + \sum_{j=1}^{K-1} \exp\left(x_{w,j}^t, x_{s,j}^t\right)} \tag{11}$$

Where the weak augmented sample $x_{w,i}$ utilizes its corresponding strong augmented view $x_{s,i}$ as a positive pair, while the remaining samples within the batch are regarded as the negative pairs. Then, negative learning with class balance regularization is introduced for target domain adaptation. Mixup (Li, Socher, & Hoi, 2019) and Co-teaching (Qiao, Shen, Zhang, Wang, & Yuille, 2018) is also applied along with negative learning. The negative learning formula is written as :

$$\mathcal{L}_{nl} = -\mathrm{E}_{x_t \in \mathcal{X}_t}\left[\sum_{c=1}^{C} \tilde{y}^c log\left(1 - p_t^c\right)\right] \tag{12}$$

Due to the noisy nature of the noisy set, directly learning from its noisy pseudo labels would result in error accumu-

lation, thereby degrading the adaptation performance. However, these still exist valuable information of target domain in the noisy set. To harness this information to benefit adaptation, we propose selective contrastive learning to learn representation while preventing error accumulation introduced by noisy labels.

In detail, the prediction of a sample signifies the degree of certainty with which the target model assigns labels to the target sample. Although the label in the noisy set is often inaccurate, the confidence score still reflects the general learning trend. The lower the confidence score of a class is, the less likely a sample belongs to that class. To this end, σ percentage of classes with the lowest confidence are selected as the negative classes for class-wise selective contrastive learning. Formally, we select σ percent of samples from the class with the minimal confidence as negative classes by:

$$U_{neg} = \text{argsort}_{\text{ascending}}\left(\tilde{y}_i\right)[0 : \sigma * K_{num}] \tag{13}$$

The $\sigma * K_{num}$ denotes the percentage of class numbers. $\tilde{y}_i$ denotes the probability output by the target model. We sort the probabilities of a target sample in ascending order and select the top portion by $\sigma * K_{num}$. With the selected negative classes, we regard those classes in the batch as negative classes for conducting contrastive learning. We leverage the strong augmented view corresponding to the sample as the positive pair, and all the weak and strong views of the negative classes are regarded as negative pairs. The selective contrastive learning is formulated as:

$$\mathcal{L}_{ncon} = -\log \frac{\exp\left(x^t_{w,i}, x^t_{s,i}\right)}{\exp\left(x^t_{w,i}, x^t_{s,i}\right) + \sum_{j=1}^{U_{neg}} \exp\left(x^t_{w,j}, x^t_{s,j}\right)} \tag{14}$$

By Eq.(14), the representations learned from selected negative classes would facilitate the noise sample closer to the potentially correct classes, and vice versa for negative classes. The overall objective of contrastive negative learning is:

$$\mathcal{L}_{psca} = \mathcal{L}_{nl} + \lambda * (\mathcal{L}_{icon} + \mathcal{L}_{ncon}) \tag{15}$$

where λ is a hyper-parameter to control the magnitude of contrastive representation learning.

Experiments

Datasets and Evaluation Protocol

To verify our proposed method, we conduct experiments on three representative domain adaptation datasets: Office31, OfficeHome, and VisDA. **Office-31**(Saenko, Kulis, Fritz, & Darrell, 2010) consists of three different domains: Amazon (A), DSLR (D), and Webcam (W). Each domain has 31 classes and 4652 images in total. **Office-Home**(Venkateswara, Eusebio, Chakraborty, & Panchanathan, 2017) contains 15,500 images and shares 65 classes for four domains: Art (Ar), Clipart (Cl), Product (Pr), and Real-world (Rw). **VisDA**(Peng et al., 2018) is the most challenging dataset among all, comprising 152k synthetic images as source domain and 55k real images as the target domain. Adhering to the standard BBDA evaluation protocol, we compare the top-1 accuracy on the target domain. For fair comparison, we directly report the results of all baselines from their original papers.

Table 1: Classification accuracies (%) on Office-31 dataset for BBDA. (The right arrow $\rightarrow$ denotes 'adapt to')

Method	A→D	A→W	D→A	D→W	W→A	W→D	Avg.
LNL-OT(Asano, Rupprecht, & Vedaldi, 2020)	88.8	85.5	64.6	95.1	66.7	98.7	83.2
LNL-KL(H. Zhang et al., 2021)	89.4	86.8	65.1	94.8	67.1	98.7	83.6
HD-SHOT(Liang et al., 2021)	86.5	83.1	66.1	95.1	68.9	98.1	83.0
SD-SHOT(Liang et al., 2021)	89.2	83.7	67.9	95.3	71.1	97.1	84.1
DINE(Liang et al., 2022)	91.6	86.8	72.2	96.2	73.3	98.6	86.4
BiMem(J. Zhang, Huang, Jiang, & Lu, 2023)	92.8	88.2	73.9	96.8	75.3	**99.4**	87.7
BETA(J. Yang et al., 2023)	93.6	88.3	76.1	95.5	76.5	99.0	88.2
RFC(S. Zhang et al., 2024)	**94.4**	**93.0**	76.7	95.6	77.5	98.1	89.2
OLIVA (Ours)	94.2	91.6	**81.5**	**98.2**	**83.0**	97.8	**91.1**

Implementation Details

We follow the pioneer BBDA works (Liang et al., 2022) to choose the same backbone network and learning rate for training source model. The ViL model CLIP is directly adopted from their public repository(Radford et al., 2021). For training OLIVA, we use the same learning rate setting in source model training. The values of the hyper parameters τ, σ and λ are set to 0.9, 0.2 and 0.3 respectively on all tasks. The training epoch is set to 30 for Office31 and OfficeHome, and 10 for VisDA. The code will be available soon.

Main Results

Results on Office31. From Table 1, OLIVA demonstrates superior performance on the Office31 dataset, achieving the highest average accuracy of 91.1% among all compared methods. This represents a notable improvement over previously state-of-the-art methods. In challenging adaptation tasks like D→A and W→A, OLIVA outperforms the second best method RFC by a large margin (+4.8% and +5.5% respectively), which indicates OLIVA performs well in tasks with a large domain gap. **Results on OfficeHome** As shown in Table2, OLIVA achieves the highest average performance among 12 tasks compared with other BBDA baselines. Since the OfficeHome dataset has 65 classes in each domain, this superior result suggests that OLIVA is effective in handling BBDA scenarios with a large number of classes, implies its potential in real-world scenarios where multi-class and cross-domain data are prevalent. **Results on VisDA.** For the VisDA dataset, we report the accuracy of each class and per-class accuracy, following the standard BBDA evaluation protocol. Table 3 presents the adaptation result on VisDA. OLIVA achieves high per-class accuracy of 85.3% on VisDA, outperforming compared baselines across multiple classes, highlighting its effectiveness in handling diverse object categories in large-scale domain shift.

Ablation Study

Components Contribution Table 4 present ablation study on the A→D task in Office31, the performance of the OLIVA

Table 2: Classification accuracies (%) on OfficeHome dataset for BBDA. (The right arrow → denotes 'adapt to')

Method	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg.
LNL-OT(Asano et al., 2020)	49.1	71.7	77.3	60.2	68.7	73.1	57.0	46.5	76.8	67.1	52.3	79.5	64.9
LNL-KL(H. Zhang et al., 2021)	49.0	71.5	77.1	59.0	68.7	72.9	56.4	46.9	76.6	66.2	52.3	79.1	64.6
HD-SHOT(Liang et al., 2021)	48.6	72.8	77.0	60.7	70.0	73.2	56.6	47.0	76.7	67.5	52.6	80.2	65.3
SD-SHOT(Liang et al., 2021)	50.1	75.0	78.8	63.2	72.9	76.4	60.0	48.0	79.4	69.2	54.2	81.6	67.4
DINE(Liang et al., 2022)	52.2	78.4	81.3	65.3	76.6	78.7	62.7	49.6	82.2	69.8	55.8	84.2	69.7
BiMem(J. Zhang et al., 2023)	54.5	78.8	81.4	66.7	78.7	79.6	65.9	53.6	82.3	73.6	57.8	84.9	71.5
BETA(J. Yang et al., 2023)	57.2	78.5	82.1	68.0	78.6	79.7	67.5	56.0	83.0	71.9	58.9	84.2	72.1
RFC(S. Zhang et al., 2024)	57.4	80.0	82.8	67.0	80.6	80.2	68.3	57.8	82.8	72.8	59.3	85.9	72.9
OLIVA (Ours)	**61.4**	**84.4**	**83.8**	**72.0**	**82.0**	**81.1**	**70.5**	**60.8**	**84.1**	**75.2**	**63.9**	**86.2**	**75.5**

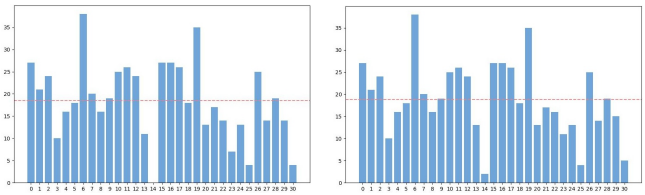
Table 3: Classification accuracies (%) on VisDA dataset for BBDA.

Method	plane	bicycle	bus	car	horse	knife	motocycle	person	plant	skateboard	train	truck	Per-class
LNL-OT(Asano et al., 2020)	82.6	84.1	76.2	44.8	90.8	39.1	76.7	72.0	82.6	81.2	82.7	50.6	72.0
LNL-KL(H. Zhang et al., 2021)	82.7	83.4	76.7	44.9	90.9	38.5	78.4	71.6	82.4	80.3	82.9	50.4	71.9
HD-SHOT(Liang et al., 2021)	75.8	85.8	78.0	43.1	92.0	41.0	79.9	78.1	84.2	86.4	81.0	**65.5**	74.2
SD-SHOT(Liang et al., 2021)	79.1	85.8	77.2	43.4	91.6	41.0	80.0	78.3	84.7	86.8	81.1	65.1	74.5
DINE(Liang et al., 2022)	81.4	86.7	77.9	55.1	92.2	34.6	80.8	79.9	87.3	87.9	84.3	58.7	75.6
BETA(J. Yang et al., 2023)	**96.2**	83.9	82.3	71.0	95.3	73.1	88.4	80.6	95.5	90.9	88.3	45.1	82.6
RFC(S. Zhang et al., 2024)	95.6	**89.7**	**87.8**	75.8	**96.5**	**96.5**	**90.4**	82.8	**96.0**	70.0	85.7	55.1	85.2
OLIVA (Ours)	93.4	82.4	84.6	**76.0**	94.5	93.9	89.2	**85.0**	94.2	**93.0**	**90.3**	47.6	**85.3**

Table 4: Result of ablation study on A→D task in Office31.

Method	Result (%)
OLIVA (Full)	94.2
w/o Entropy-weighted Knowledge Fusion	92.6
w/o Learning Degree-guided Complementation	93.3
w/o Instance Contrastive Learning	93.6
w/o Selective Contrastive Learning	93.8

is compared with variants lacking specific components. Removing Entropy-weighted Knowledge Fusion leads to a decrease to 92.6%, indicating the importance of fusing knowledge of ViL model. Without Learning Degree-guided Complementation, the result drops to 93.3%. Similarly, eliminating Instance Contrastive Learning and Selective Contrastive Learning results in scores of 93.6% and 93.8% respectively. These results demonstrate that each component contributes to the overall adaptation performance of the purposed method.



(a) Before complementation (b) After complementation

Figure 2: Visualization of learning degree-guided complementation at epoch 5 in A→W.

Visualization Fig. 2 shows the before and after of complementation for hard-learning classes in the A→W tasks during adaptation. It can be seen that the complementation mechanism of OLIVA has complemented samples for hard-learning

Figure 3: The t-SNE visualization of VisDA feature space. Each color represents a class.

classes. Especially for class 14, the number of samples was 0 before the complementation. If conducting adaptation without complementation, it will lead to a bias towards othe easy-learning classes. After complementation, it is possible to correct the deviation in the adaptation process. Fig. 3 presents the feature space of the target model on VisDA. As can be observed, each class has learned a distinct decision boundary. This represents that OLIVA can effectively learn to adapt to target domain via contrastive negative learning.

Conclusion

In this paper, we investigate the learning imbalance problem in BBDA. This learning imbalance issue mainly stems from the domain bias inherent in the source model. To surmount this problem, we propose OLIVA, which fuses knowledge from the ViL model, devises a complementation strategy to mitigate the imbalance during adaptation, and then employs contrastive negative learning for target adaptation. Extensive experiments are conducted on multiple benchmark datasets validate the effectiveness of OLIVA.

References

Asano, Y. M., Rupprecht, C., & Vedaldi, A. (2020). Self-labelling via simultaneous clustering and representation learning. In *International conference on learning representations (iclr)*.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (pp. 1597–1607).

Chen, X., Wang, S., Long, M., & Wang, J. (2019). Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In *International conference on machine learning* (pp. 1081–1090).

Crisp, R. J., & Turner, R. N. (2011). Cognitive adaptation to the experience of social and cultural diversity. *Psychological bulletin*, *137*(2), 242.

Dunwoody, P. T., Haarbauer, E., Mahan, R. P., & Marino. (2000). Cognitive adaptation and its consequences: A test of cognitive continuum theory. *Journal of Behavioral decision making*, *13*(1), 35–54.

Fang, Y., Yap, P.-T., Lin, W., Zhu, H., & Liu, M. (2024). Source-free unsupervised domain adaptation: A survey. *Neural Networks*, 106230.

Lee, S., Kim, D., Kim, N., & Jeong, S.-G. (2019). Drop to adapt: Learning discriminative features for unsupervised domain adaptation. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 91–100).

Li, J., Socher, R., & Hoi, S. C. (2019). Dividemix: Learning with noisy labels as semi-supervised learning. In *International conference on learning representations.*

Li, J., Yu, Z., Du, Z., Zhu, L., & Shen, H. T. (2024). A comprehensive survey on source-free domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Liang, J., Hu, D., Feng, J., & He, R. (2022). Dine: Domain adaptation from single and multiple black-box predictors. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 8003–8013).

Liang, J., Hu, D., Wang, Y., He, R., & Feng, J. (2021). Source data-absent unsupervised domain adaptation through hypothesis transfer and labeling transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(11), 8602–8617.

Long, M., Cao, Z., Wang, J., & Jordan, M. I. (2018). Conditional adversarial domain adaptation. *Advances in neural information processing systems*, *31*.

Oza, P., Sindagi, V. A., Sharmini, V. V., & Patel, V. M. (2023). Unsupervised domain adaptation of object detectors: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Peng, X., Usman, B., Kaushik, N., Wang, D., Hoffman, J., & Saenko, K. (2018). Visda: A synthetic-to-real benchmark for visual domain adaptation. In *Proceedings of the ieee conference on computer vision and pattern recognition workshops* (pp. 2021–2026).

Qiao, S., Shen, W., Zhang, Z., Wang, B., & Yuille, A. (2018). Deep co-training for semi-supervised image recognition. In *Proceedings of the european conference on computer vision (eccv)* (pp. 135–152).

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... others (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763).

Reynolds, D. A. (2009). Gaussian mixture models. *Encyclopedia of biometrics*, *741*(659-663).

Saenko, K., Kulis, B., Fritz, M., & Darrell, T. (2010). Adapting visual category models to new domains. In *Computer vision–eccv 2010: 11th european conference on computer vision, heraklion, crete, greece, september 5-11, 2010, proceedings, part iv 11* (pp. 213–226).

Sun, S. K., B. (2015). Deep coral: Correlation alignment for deep domain adaptation. In *International conference on machine learning* (pp. 1180–1189).

Venkateswara, H., Eusebio, J., Chakraborty, S., & Panchanathan, S. (2017). Deep hashing network for unsupervised domain adaptation. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 5018–5027).

Yang, J., Peng, X., Wang, K., Zhu, Z., Feng, J., Xie, L., & You, Y. (2023). Divide to adapt: Mitigating confirmation bias for domain adaptation of black-box predictors..

Yang, S., Jui, S., van de Weijer, J., et al. (2022). Attracting and dispersing: A simple approach for source-free domain adaptation. *Advances in Neural Information Processing Systems*, *35*, 5802–5815.

Yang, S., Van de Weijer, J., Herranz, L., Jui, S., et al. (2021). Exploiting the intrinsic neighborhood structure for source-free domain adaptation. *Advances in neural information processing systems*, *34*, 29393–29405.

Yang, S., Wang, Y., Van De Weijer, J., Herranz, L., & Jui, S. (2021). Generalized source-free domain adaptation. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 8978–8987).

Zellinger, W., Grubinger, T., Lughofer, E., Natschlger, T., & Saminger-Platz, S. (2017). Central moment discrepancy (cmd) for domain-invariant representation learning.

Zhang, H., Zhang, Y., Jia, K., & Zhang, L. (2021). Unsupervised domain adaptation of black-box source models. *arXiv preprint arXiv:2101.02839*.

Zhang, J., Huang, J., Jiang, X., & Lu, S. (2023). Black-box unsupervised domain adaptation with bi-directional atkinson-shiffrin memory. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 11771–11782).

Zhang, L., & Gao, X. (2022). Transfer adaptation learning: A decade survey. *IEEE Transactions on Neural Networks and Learning Systems*.

Zhang, S., Shen, C., Lü, S., & Zhang, Z. (2024, Mar.). Reviewing the forgotten classes for domain adaptation of black-box predictors. *Proceedings of the AAAI Confer-*

ence on Artificial Intelligence, *38*, 16830-16837. doi: 10.1609/aaai.v38i15.29624

Zhang, Y., Liu, T., Long, M., & Jordan, M. (2019). Bridging theory and algorithm for domain adaptation. In *International conference on machine learning* (pp. 7404–7413).

Zhou, K., Liu, Z., Qiao, Y., Xiang, T., & Loy, C. C. (2022). Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(4), 4396–4415.