

Adversarial Experts Model for Black-box Domain Adaptation

Appendix

Anonymous Author(s)

A PRELIMINARY

The work most relevant to our work investigates the unsupervised domain adaptation problem, known as MCD (Maximum Classifier Discrepancy) [2]. Employing a semi-supervised learning framework, it utilizes source domain data and ground truth labels for supervised training. For the target domain, it maximizes the discrepancy between two predictions to update classifiers and minimizes the discrepancy to update feature extractors. To be specific, there are source domain data $X_s = \{(x_s^i)\}_{i=1}^{N_s}$ with corresponding ground truth labels $Y_s = \{(y_s^i)\}_{i=1}^{N_s}$, where N_s denotes the number of source domain samples. The unlabeled target domain data is represented as $X_t = \{(x_t^i)\}_{i=1}^{N_t}$, with number N_t samples. The shared label space is $C = \{1, 2, \dots, K\}$. MCD utilizes one feature extractor and two classifiers, denoted as $f(\cdot), h_1(\cdot), h_2(\cdot)$ respectively. The training process typically consists of three steps.

The first step involves training the model using source domain data and ground truth labels to ensure that the model can correctly classify the source domain. The cross entropy is used as follows:

$$\min_{f, h_1, h_2} \mathcal{L}_{ce}(X_s, Y_s). \quad (1)$$

In the second step, the feature extractor is fixed, and two classifiers are trained to maximize the difference between themselves. Concurrently, supervised training on the source domain data is also needed. This step helps identify challenging samples not supported by the source domain. The loss function is as follows:

$$\min_{h_1, h_2} \mathcal{L}_{ce}(X_s, Y_s) - \mathcal{L}_{adv}(X_t), \quad (2)$$

$$\mathcal{L}_{adv}(X_t) = \mathbb{E}_{x_t^i \sim X_t} [d(p_1(x_t^i), p_2(x_t^i))],$$

where $d(p_1, p_2) = \frac{1}{K} \sum_{k=1}^K |p_{1k} - p_{2k}|$ is a measurement of distribution divergence. In the third step, the two classifiers are fixed while the feature extractor is trained to minimize the discrepancy of the outputs. This encourages that the extracted features are indistinguishable for the two classifiers. The objective is as follows:

$$\min_f \mathcal{L}_{adv}(X_t) \quad (3)$$

Although MCD has made breakthrough progress in the problem of unsupervised domain adaptation (UDA), the highly reliance on accessing source domain data and labels is entirely unsuitable for black-box domain adaptation setting.

B FURTHER ANALYSIS

Due to page limitations, we present further experimental analyses in the appendix.

B.1 Visualization on GRAD-CAM.

The visualization on GRAD-CAM [3] are shown in Fig.1. Specifically, we randomly selected four images from different classes of

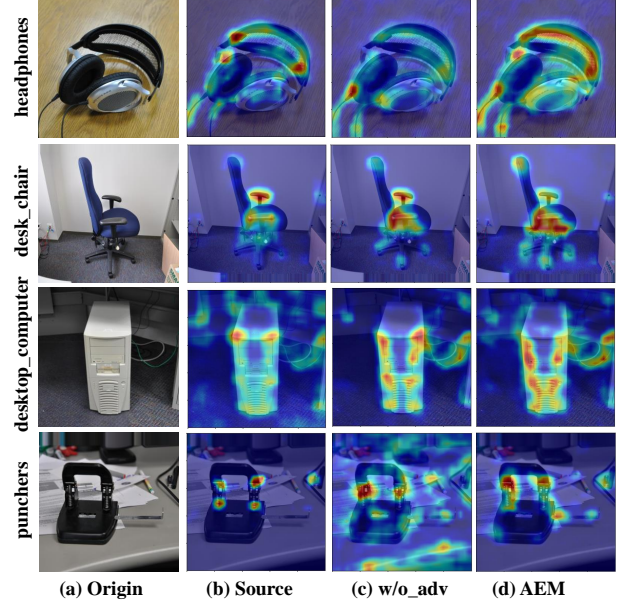


Figure 1: The GRAD-CAM visualization of A→D on Office-31 dataset.

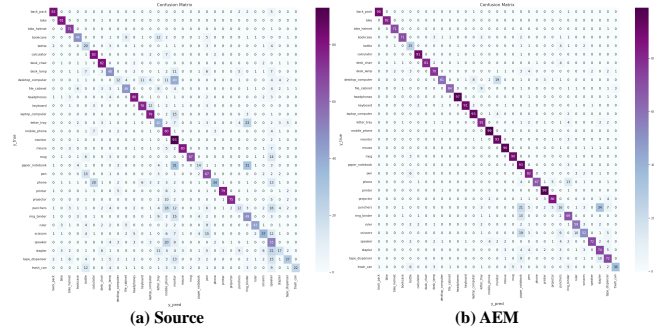


Figure 2: Confusion matrix on task D→A of Office-31 dataset.

Table 1: Comparison results of FLOPs and model parameters. K: class number.

	K=31		K=65	
	FLOPs(M)	Params(M)	FLOPs(M)	Params(M)
CLIP	63017.58	84.23	128901.71	84.23
AEM	4132.23	24.04	4123.24	24.05

Office-31 dataset for visualization. From left to right, the visualization results represent the original image, source domain (black-box)

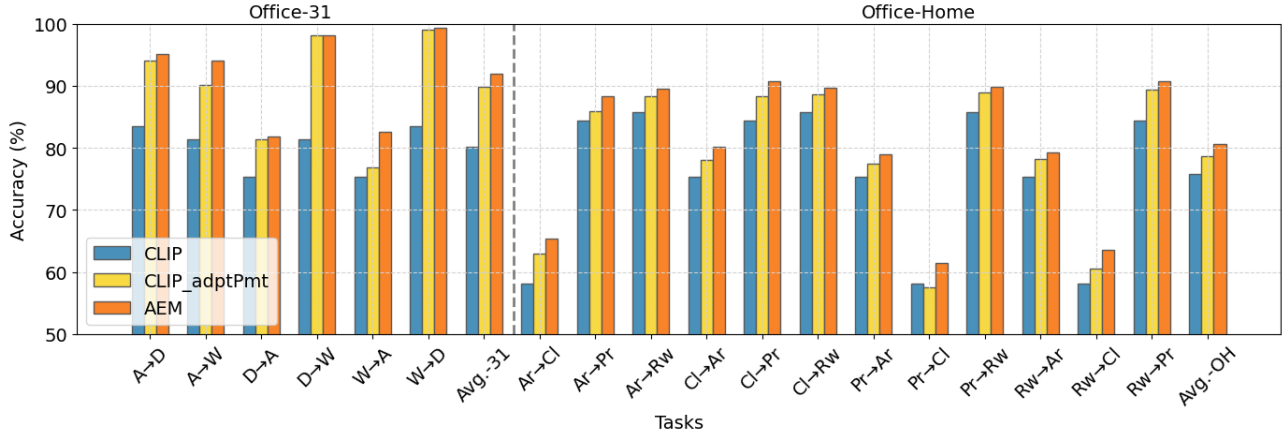


Figure 3: Accuracy comparison of CLIP, CLIP_adptPmt and AEM on Office-31 and Office-Home datasets.

model, AEM variant without adversarial (w/o_adv), and AEM. Compared to the source model, the introduction of external high-level semantic information in the w/o_adv variant demonstrates slightly better performance. However, solely utilizing consistency learning enables the w/o_adv variant to additionally focus attention on objects, building upon the foundation of the source model. AEM conducts adversarial learning between the feature extractor and the classifier. Consequently, compared to the non-adversarial variant, our model better integrates the knowledge from black-box experts and ViL experts. This is specifically reflected in the visualization by introducing attention that does not exist in the black-box model and demonstrating a more continuous phenomenon of attention.

B.2 Visualization of confusion matrix.

The confusion matrix for the task D→A on the Office-31 dataset is presented in Fig.3. The sum of each row represents the total number of instances in the true class, and the sum of each column represents the total number of instances in the predicted class. It can be observed that comparing with source (black-box) model, the number of correct predictions by AEM has been significantly improved.

B.3 Comparison with CLIP.

Comparison on accuracy. We further conduct experimental analysis on CLIP. As shown in Fig.??, we compared three models: using only the pre-trained CLIP model (CLIP) [1], the pre-trained CLIP model with target domain adapted prompts (CLIP_adptPmt), and our adapted target domain model AEM. Overall, AEM has demonstrated the best performance across all tasks. To be specific, comparing CLIP and CLIP_adptPmt, the average accuracy of CLIP_adptPmt on Office-31 and Office-Home are higher by 9.8% and 2.9% than CLIP, respectively. CLIP_adptPmt outperforms CLIP on almost all sub-tasks. In particular, on the D→W and W→D of Office-31 dataset, the accuracy of CLIP_adptPmt increased by 16.7% and 15.5%, respectively. This is because in the knowledge feedback stage, we use the target domain information to update the prompt, making the ViL model more tailored to the current task. Comparison between CLIP_adptPmt and AEM shows that AEM achieves an

average accuracy higher by 2.0% and 1.9% on Office-31 and Office-Home, respectively. Because we use the outputs of the ViL-guided classifier as the final results, the prediction accuracy upper is limited by the constraint of the corresponding expert. However, in AEM, the adversarial learning between the feature extractor and the two classifiers enables our feature extractor to better extract discriminative features suitable for the target domain. Also, the target model implicitly integrates knowledge from the black-box model. Therefore, AEM achieves a slightly higher accuracy than CLIP_adptPmt.

Comparison of model parameters. Table1 shows the comparison results of FLOPs (floating point operations) and model parameters between the trained AEM model and the CLIP model. The table provides the following insights. First, as the number of categories increases, the FLOPs of CLIP also increase. This is because the ViL model connects images with text, necessitating more computational resources to process category label text. In contrast, for AEM, since the classifier is a fully connected layer, variations in the number of categories have minimal impact on the model's computational complexity. Vertically, the FLOPs of CLIP are at least 10 times larger than that of AEM, with its parameter count around 3.5 times that of AEM. This signifies a greater demand for computational resources in the ViL model. Moreover, as depicted in Fig.??, AEM achieves approximately 2% higher average accuracy than CLIP_adptPmt. This observation shows the necessity of research on the black-box domain adaptation.

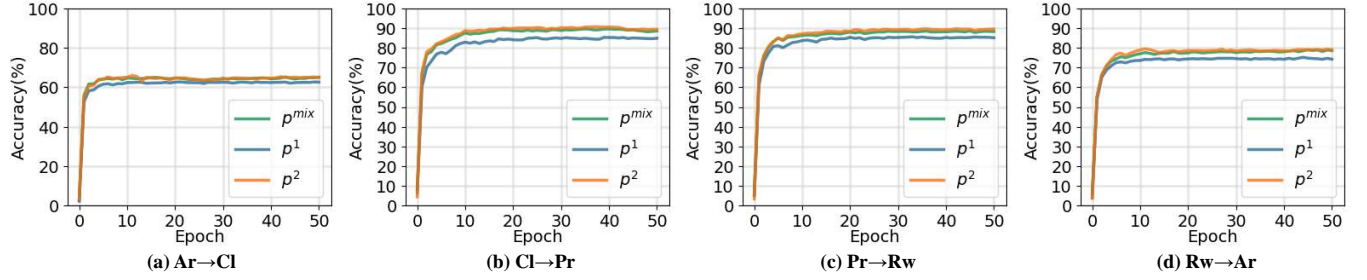
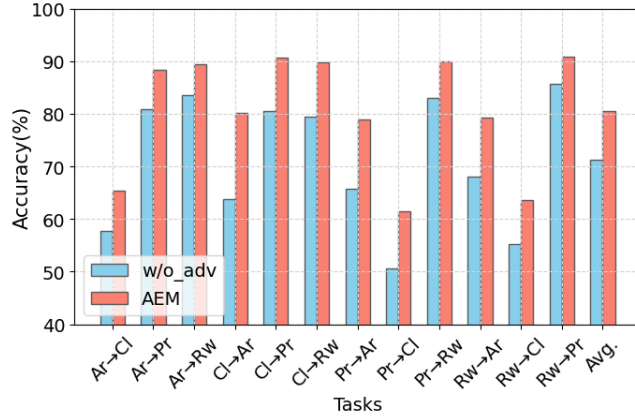
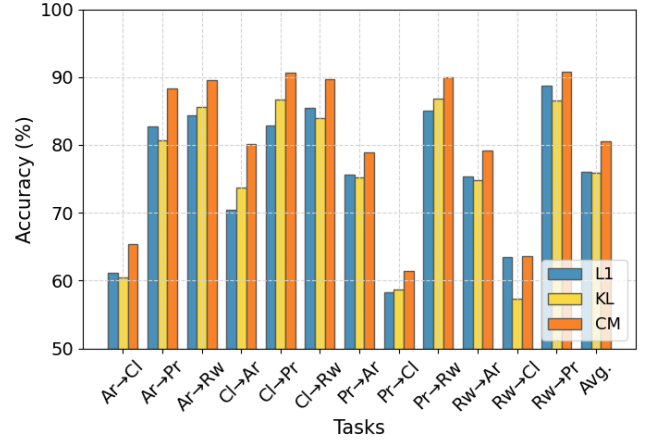
B.4 More ablation studies.

In this part, we conduct ablation studies on *Office-Home* dataset. Setup details have been provided in the Section 4.3 of main text.

Ablation study on knowledge feedback strategy. As mentioned in Section 4.3, we set three additional variants: training without knowledge feedback strategy (w/o_bv), training without updating the black-box labels (w/o_b), and training without updating the prompt of the ViL model (w/o_v). Let p^1 and p^2 denote predictions of classifiers h^1 and h^2 guided by black-box and ViL model, respectively. p^{mix} denotes the prediction obtained by averaging the outputs of both classifiers. The values in the table

Table 2: Ablation study on knowledge feedback strategy on tasks Ar→Cl, Cl→Pr, Pr→Rw and Rw→Ar in *Office-Home* dataset.

	Ar→Cl				Cl→Pr				Pr→Rw				Rw→Ar			
	p^1	p^2	p^{mix}	Avg.	p^1	p^2	p^{mix}	Avg.	p^1	p^2	p^{mix}	Avg.	p^1	p^2	p^{mix}	Avg.
w/o_bv	51.3	60.2	59.2	56.9	74.6	86.1	84.2	81.6	79.8	86.6	85.3	83.9	69.8	77.1	73.8	73.6
w/o_b	52.4	60.6	57.0	56.7	75.8	86.3	84.4	82.2	80.3	87.0	84.7	84.0	70.8	78.2	76.6	75.2
w/o_v	52.9	63.9	60.3	59.0	79.3	86.3	83.9	83.2	83.2	87.9	86.1	85.7	70.7	77.3	75.8	74.6
AEM	61.9	65.8	64.4	64.0	85.0	90.7	89.7	88.5	85.1	89.9	88.7	87.9	73.8	78.9	77.5	76.7

**Figure 4: Accuracy curves of different predictions on 4 tasks for *Office-Home* dataset.****Figure 5: Ablation study on adversarial learning for *Office-Home* dataset. w/o_adv and AEM represent training without and with adversarial learning, respectively.****Figure 6: Accuracy using different classifier consistency loss on *Office-Home* dataset.**

represent the prediction accuracy under different updating strategy. The results are shown in Table 2. Compared to the three variants, AEM demonstrated the best performance across all four tasks, with accuracy surpassing the second-best by 5.0%, 5.3%, 2.2%, and 1.5%, respectively. Due to the lack of updating black-box noisy labels and prompts for ViL, the performance of w/o_bv is the worst. Updating either expert will lead to an improvement. Furthermore, AEM achieved the best results, demonstrating the effectiveness of the proposed knowledge feedback strategy.

Ablation study on different classifier combination. Fig.4 shows the accuracy curve of three different predictions, i.e., p^1 , p^2 and p^{mix} . The accuracy curves for all 4 tasks reach the peak around the 10th epoch and then converge. Overall, classifier guided by the

ViL model shows the best performance. This is attributed to the ViL-guided classifier harboring richer high-level semantic information.

Ablation study on adversarial learning. Fig.5 shows the comparison of accuracy between AEM without adversarial learning and AEM. The average accuracy of AEM is higher by 9.4%. This is because solely employing consistency loss fails to address the issues of ambiguous samples and decision boundaries. AEM alleviates this problem through adversarial training between the feature extractor and the classifier.

Ablation study on classifier consistency loss. From Fig.6, it can be observed that the proposed CM loss achieves the best performance on 12 specific tasks. In terms of average accuracy, CM loss outperforms L1 loss and KL loss by 4.5% and 4.7%, respectively.

REFERENCES

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Aspell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.

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

[3] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.