Supplementary Material:
Preserving Clusters in Prompt Learning for Unsupervised Domain Adaptation

Tung-Long Vuong[1], Hoang Phan[2], Vy Vo[1], Anh Bui[1], Thanh-Toan Do[1], Trung Le[1], Dinh Phung[1]
[1]Monash University, [2]New York University
{Tung-Long.Vuong,v.vo,tuananh.bui,toan.do,trunglm,dinh.phung}@monash.edu
hvp2011@nyu.edu

In this supplementary material, we present the experimental settings and implementation details in Section A, followed by additional discussions and experiments in Section B. The limitations of our current work are outlined in Section C, and the proof of Lemma 1 is provided in Section D.

A. Experimental Settings

Datasets ImageCLEF is a small-scaled dataset with 1,800 images across 12 object categories from three domains: ImageNet ILSVRC 2012 (I), Pascal VOC 2012 (P), and Caltech-256 (C). Office-Home is a medium-scaled dataset containing approximately 15,500 images from 65 categories in four domains: Art, Clipart, Product, and Real World. DomainNet is the largest dataset, comprising around 600,000 images from 345 categories across six domains: Clipart, Infograph, Painting, Quickdraw, Real, and Sketch.

Baselines. Regarding prompt-based baselines, we compare our method with MPA [1], DAPL [4], Simple Prompt [1], PGA [8], and Zero-shot CLIP [9]. To ensure a comprehensive evaluation, we also include comparisons with various non-prompt methods such as DCTN [10], MDDA [11], MFSAN [14], T-SVDNet [7], and PFSA [3]. As we follow the same settings as in [1, 4, 8], the results for these baselines are reproduced based on their public implementations and hyperparameters to ensure consistency. Note that while DAPL, MPA, PGA, and our method employ CoOp [12] with text-end soft prompts, other methods fine-tune the transformer block [2], both image and text-end soft prompts [6], or the entire encoders [5, 13]. Since these alternative methods typically fine-tune many more parameters, we exclude them from the experiments to ensure a fair comparison.

Metrics. We use the top-1 accuracy for each target domain and the average accuracy across all domains as the evaluation metric. Following [8], we conduct experiments in two standard settings: a *source-combined* setting, where data from all source domains are merged, and a *multi-source* setting, which utilizes individual domain identifications. We also provide pair-wise single-source domain adaptation results for the Office-Home dataset.

Implementation details For fair comparisons, we use ResNet50 as our backbone on the Image-CLEF and Office-Home datasets and ResNet101 on DomainNet. The weights are initialized from a pre-trained CLIP model and kept frozen during training. Following previous baselines [1, 4, 8], prompts are trained using the mini-batch SGD optimizer with a learning rate of 0.005. We use a batch size of 32 and apply a cosine learning rate scheduler. For hyperparameters, token lengths M_1 and M_2 are both set to 16. Additionally, we do not require a pseudo-label threshold τ for label generation as we combine base-prompt and source-prompt to generate enhanced soft pseudo-labels. For our specific parameter λ_T and $\lambda_{\mathcal{W}}$, we simply set $\lambda_T = \lambda_{\mathcal{W}} = 0.5$ for all experiments. **Code available at**: https://github.com/VuongLong/Clustering-Reinforcement-Prompt-Learning/

B. Additional Experiments

B.1. Distance-Aware Pseudo-Label

As discuss in previous section, different transferability between domains motivate us a distance aware pseudo-labels scheme. Specifically, we calculate the weighted average cosine distance from the visual embedding $\boldsymbol{z} = f_v(\boldsymbol{x})$ to the text embeddings of each class across all source domains. We recap Eq. (**??**) in the main for easier readability:

$$\boldsymbol{\tau}^k_{ave}(x) = \frac{1}{2}\boldsymbol{\tau}^k_{base} + \frac{w_{k,i}(x)}{2}\sum_{i=1}^{N}\boldsymbol{\tau}^k_{S_i} \tag{1}$$

and $w_{i,k}$ is the important weight for each domain-class.

Here we use the l_2 distance between the visual embedding z^{pre} (the unnormalized visual embedding of x, where $z = \frac{z^{pre}}{\|z^{pre}\|_2}$) and the centroid of class k in domain i. Let

$C^i = \{c_k^i\}_{k=1}^K$ represent the set of class centroids for the i-th domain i.e., $c_k^i = \frac{1}{\sum_{j=1}^{N_{S_i}} \mathbb{I}_{(y_j^i=k)}} \sum_{j=1}^{N_{S_i}} \mathbb{I}_{(y_j^i=k)} \boldsymbol{z}_j^{pre,i}$. The weight for class k is then calculated by applying the softmax function:

$$w_{i,k}(z) = \frac{\exp\left(\left\|z_{pre} - c_k^i\right\|_2\right)}{\sum_{i'=1}^{N} \exp\left(\left\|z - c_k^{i'}\right\|_2\right)} \tag{2}$$

It is important to highlight that the weights $w_{i,k}(z)$ are applied to compute the combined prompts $\boldsymbol{\tau}_{\text{ave}}^k$, rather than being applied directly to the outputs of the softmax (i.e., the predictions from individual prompts). Specifically, the enhanced pseudo-labels are computed as described in Eq. (**??**) in the main paper:

$$\hat{y}[k] = \frac{\exp\left(\langle \boldsymbol{z}, \boldsymbol{\tau}_{ave}^k((x))\rangle/\gamma\right)}{\sum_{k'=1}^{K} \exp\left(\langle \boldsymbol{z}, \boldsymbol{\tau}_{ave}^{k'}(x)\rangle/\gamma\right)} \tag{3}$$

where

$$\langle \boldsymbol{z}, \boldsymbol{\tau}_{ave}^k((x))\rangle = \frac{1}{2}\langle \boldsymbol{z}, \boldsymbol{\tau}_{base}^k\rangle + \frac{w_{k,i}(x)}{2}\sum_{i=1}^{N}\langle \boldsymbol{z}, \boldsymbol{\tau}_{S_i}^k\rangle \tag{4}$$

It can be observed that this design preserves semantic similarity, as the magnitude of cosine similarity between the visual embedding and the text embedding of each individual prompt is taken into account in the final pseudo-label. Meanwhile, the weights $w_{k,i}(x)$ represent the spatial relationship between the visual embedding and the text embedding.

Distance	Ar	Cl	Pr	Rw	Average
Average	76.0	62.6	87.0	87.5	78.3
cosine	76.6	56.2	88.2	86.7	76.9
L2	**76.8**	**63.5**	**87.5**	**87.6**	**78.9**

Table 1. Different metric for Distance-aware Pseudo-Labels on Office-Home dataset.

To evaluate the effectiveness of this design, we compare the performance of the proposed distance-aware pseudo-labeling scheme with two alternative strategies: a simple averaging of source prompts and a distance-aware approach using cosine distance instead of L_2. For the cosine distance, we use z rather than z_{pre} to compute weights and class centroids. As shown in Table 1, the L_2 distance achieves the best performance, while the cosine distance performs the worst. This discrepancy may be due to overlapping information between the cosine similarity of visual embeddings with text embeddings of prompts and the cosine similarity of visual embeddings with class centroids.

B.2. Performance with ViT models

In the main paper, we conduct experiments using the ResNet backbone. Additionally, we perform experiments with the ViT-14L model. The results presented in Table 2, comparing our approach with the most recent SOTA method, PGA, demonstrate that our approach outperforms previous methods.

Setting	Ar	Cl	Pr	Rw	Ave
Zero-Shot	86.6	72.4	92.8	92.7	86.1
PGA	87.5	77.2	94.5	94.5	88.4
ours	**89.7**	**83.2**	**95.9**	**95.5**	**91.1**

Table 2. Performance of ViT-14L on OfficeHome dataset.

B.3. Performance of corrupted dataset

To further assess the robustness of our method, we performed experiments using corrupted data generated from the Office-Home dataset (for details on the corruptions, refer to https://github.com/hendrycks/robustness).

Corruption	Method	Ar	Cl	Pr	Rw	Ave
No corruption	Zero-Shot	71.2	50.4	81.4	82.6	71.4
	PGA	74.8	56.0	85.2	86.0	75.5
	ours	**76.8**	**63.5**	**87.5**	**87.6**	**78.9**
Defocus_blur	Zero-Shot	60.2	40.7	73.8	76.8	62.9
	PGA	66.2	50.0	77.6	80.3	68.5
	ours	**69.1**	**61.3**	**78.1**	**81.1**	**72.4**
elastic_transform	Zero-Shot	635	411	74.2	77.3	64.0
	PGA	69.4	48.1	80.7	82.3	70.1
	ours	**69.3**	**60.5**	**82.1**	**83.2**	**73.8**
Gaussian_noise	Zero-Shot	62.0	52.8	73.7	75.2	65.9
	PGA	67.3	48.9	**75.8**	79.2	67.8
	ours	**68.1**	**60.2**	75.2	**80.6**	**71.0**
Speckle_npose	Zero-Shot	62.1	38.5	65.3	71.2	59.2
	PGA	67.3	47.0	**74.2**	78.3	66.7
	ours	**67.7**	**58.0**	**73.7**	**78.8**	**69.5**

Table 3. Performance on Corrupted OfficeHome dataset.

The results in Table 3 demonstrate that our method remains robust even as CLIP's zero-shot performance declines.

C. Limitations

One of our main contributions is the enhancement of pseudo-labels for target domains by effectively leveraging information from source domains. Specifically, we analyze the relationship between the target and source domains, considering both semantic similarity (via cosine distance between visual embeddings and text embeddings of prompts) and spatial relationships (via L_2 distance in the pre-normalized embedding space), to appropriately weight references from source domains. Additionally, we demonstrate the equivalence between the references provided by the base prompt and those from source prompts to the target prompt. However, due to the lack of access to the data used for training the base prompt, we are unable to directly compare the quality of references between the base prompt and source prompts. This limitation results in suboptimal utilization of these references, as we currently assign equal weight to the base prompt and the weighted source prompts to balance their contributions, as described in Eq. (1). Addressing this limitation is a key focus for future work.

D. Proof of Lemma 1

We propose minimizing the Wasserstein distance between the target prompts' text embeddings and the visual embeddings from the target domain. Specifically, denote $\mathcal{T} = \{\boldsymbol{\tau}_T^k\}_{k=1}^K$ where $\boldsymbol{\tau}_T^k$ represents the text embeddings of the context prompt $[\boldsymbol{P}_{sh}^k][\boldsymbol{P}_T][\text{CLASS}_k]$ for class k. Let

$$\mathbb{P}_{\tau,\pi} = \sum_{k=1}^{K} \pi_k \delta_{\boldsymbol{\tau}_T^k} \tag{5}$$

be the discrete distribution over the set of text embeddings $\mathcal{T}$ for the target domain, where the category probabilities $\pi \in \Delta_K = \{\alpha \geq 0 : \|\alpha\|_1 = 1\}$ lie in the K-simplex. Additionally, let

$$\mathbb{P}^T = \frac{1}{N_T} \sum_{j=1}^{N_T} \delta_{\boldsymbol{z}_j}$$

represent the visual embedding distribution of the target domain, where δ is the Dirac delta function and $\boldsymbol{z} = f_v(\boldsymbol{x})$ for the target image $\boldsymbol{x}$. The clustering assumption is then enforced by the following objective:

$$\mathcal{L}_{\mathcal{W}} = \mathcal{W}_{d_z}\left(\mathbb{P}_{\tau,\pi}, \mathbb{P}^T\right) \tag{6}$$

where d_z represents a metric function. We use the cosine distance $d_z(a,b) = 1 - \langle a, b\rangle$ since the visual embeddings and text embeddings already lie in the unit hypersphere. The following lemma demonstrates the behavior of π and explains how the Wasserstein term helps target prompts enforce clustering properties.

Lemma 1 ***(Lemma 1 in the main paper)*** *Let $\mathcal{T}^* = \left\{\boldsymbol{\tau}_T^{k,*}\right\}_{k=1}^K$ be the optimal solution of the OP in Eq. (6), then $\mathcal{T}^*$ is also the optimal solution of the following OP:*

$$\min_{\mathcal{T},\pi} \min_{\sigma \in \Sigma_\pi} \mathbb{E}_{z \sim \mathbb{P}^T}\left[d_z\left(\boldsymbol{z}, \tau_T^{\sigma(z)}\right)\right], \tag{7}$$

where Σ_π is the set of assignment functions $\sigma : \mathcal{Z} \to \{1,...,K\}$ such that the cardinalities $|\sigma^{-1}(k)|, k = 1,...,K$ are proportional to $\pi_k, k = 1,...,K$. Moreover, given the set of text embeddings $\mathcal{T}$, the optimal σ of the inner minimization is the nearest assignment: $\sigma^{-1}(k) = \{\boldsymbol{z} \mid k = argmin_m d_z(\boldsymbol{z}, \boldsymbol{\tau}_T^m)\}$ is set of visual embeddings which are quantized to k^{th} text embedding $\boldsymbol{\tau}_T^k$.

Proof:

It is clear that

$$\mathbb{P}_{\tau,\pi} = \sum_{k=1}^{K} \pi_k \delta_{\tau_T^k}.$$

Therefore, we reach the following OP:

$$\min_{\mathcal{T},\pi} \mathcal{W}_{d_z}\left(\frac{1}{N}\sum_{n=1}^{N_T}\delta_{z_n}, \sum_{k=1}^{K}\pi_k\delta_{\tau_T^k}\right).$$

By using the Monge definition, we have

$$\begin{aligned}
&\mathcal{W}_{d_z}\left(\frac{1}{N_T}\sum_{n=1}^{N_T}\delta_{z_n}, \sum_{k=1}^{K}\pi_k\delta_{\tau_T^k}\right)\\
&= \min_{T:T\#\mathbb{P}^T=\mathbb{P}_{\tau,\pi}} \mathbb{E}_{z\sim\mathbb{P}^T}\left[d_x\left(z, T\left(z\right)\right)\right]\\
&= \frac{1}{N_T}\min_{T:T\#\mathbb{P}^T=\mathbb{P}_{\tau,\pi}} \sum_{n=1}^{N_T} d_z\left(z_n, T\left(z_n\right)\right).
\end{aligned}$$

Since $T\#\mathbb{P}^T = \mathbb{P}_{\tau,\pi}$, $T(z_n) = \tau_T^k$ for some k. Additionally, $\left|T^{-1}\left(\tau_T^k\right)\right|, k = 1,...,K$ are proportional to $\pi_k, k = 1,...,K$. Denote $\sigma : \{1,...,N_T\} \to \{1,...,K\}$ such that $T(z_n) = \tau_T^{\sigma(n)}, \forall i = 1,...,N$, we have $\sigma \in \Sigma_\pi$. It follows that

$$\begin{aligned}
&\mathcal{W}_{d_z}\left(\frac{1}{N}\sum_{n=1}^{N_T}\delta_{z_n}, \sum_{k=1}^{K}\pi_k\delta_{\tau_T^k}\right)\\
&= \frac{1}{N_T}\min_{\sigma\in\Sigma_\pi}\sum_{n=1}^{N_T} d_z\left(z_n, \tau_T^{\sigma(n)}\right).
\end{aligned}$$

Finally, the the optimal solution of the OP in Eq. (6) is equivalent to

$$\min_{\mathcal{T},\pi}\min_{\sigma\in\Sigma_\pi}\sum_{n=1}^{N_T} d_z\left(z_n, \tau_T^{\sigma(n)}\right),$$

which directly implies the conclusion.

References

[1] Haoran Chen, Zuxuan Wu, and Yu-Gang Jiang. Multi-prompt alignment for multi-source unsupervised domain adaptation. *Neural Information Processing Systems*, 2022. 1

[2] Zhekai Du, Xinyao Li, Fengling Li, Ke Lu, Lei Zhu, and Jingjing Li. Domain-agnostic mutual prompting for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23375–23384, 2024. 1

[3] Yangye Fu, Ming Zhang, Xing Xu, Zuo Cao, Chao Ma, Yanli Ji, Kai Zuo, and Huimin Lu. Partial feature selection and alignment for multi-source domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16654–16663, 2021. 1

[4] Chunjiang Ge, Rui Huang, Mixue Xie, Zihang Lai, Shiji Song, Shuang Li, and Gao Huang. Domain adaptation via prompt learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 1

[5] Zhengfeng Lai, Noranart Vesdapunt, Ning Zhou, Jun Wu, Cong Phuoc Huynh, Xuelu Li, Kah Kuen Fu, and Chen-Nee Chuah. Padclip: Pseudo-labeling with adaptive debiasing in clip for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16155–16165, 2023. 1

[6] Zhengfeng Lai, Haoping Bai, Haotian Zhang, Xianzhi Du, Jiulong Shan, Yinfei Yang, Chen-Nee Chuah, and Meng Cao. Empowering unsupervised domain adaptation with large-scale pre-trained vision-language models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2691–2701, 2024. 1

[7] Ruihuang Li, Xu Jia, Jianzhong He, Shuaijun Chen, and Qinghua Hu. T-svdnet: Exploring high-order prototypical correlations for multi-source domain adaptation. *ICCV*, 2021. 1

[8] Hoang Phan, Lam Tran, Quyen Tran, and Trung Le. Enhancing domain adaptation through prompt gradient alignment. *Neural Information Processing Systems*, 2024. 1

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

[10] Ruijia Xu, Ziliang Chen, W. Zuo, Junjie Yan, and Liang Lin. Deep cocktail network: Multi-source unsupervised domain adaptation with category shift. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 1

[11] 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 *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 12975–12983. AAAI Press, 2020. 1

[12] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 1

[13] Wenlve Zhou and Zhiheng Zhou. Unsupervised domain adaption harnessing vision-language pre-training. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 1

[14] Yongchun Zhu, Fuzhen Zhuang, and Deqing Wang. Aligning domain-specific distribution and classifier for cross-domain classification from multiple sources. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 5989–5996. AAAI Press, 2019. 1