Unveiling the Unknown: Unleashing the Power of Unknown to Known in Open-Set Source-Free Domain Adaptation

Supplementary Material

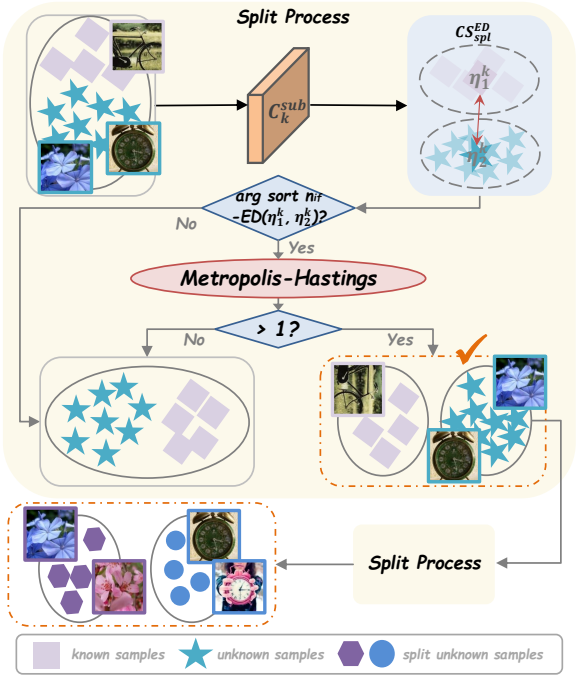

Figure 1. Split process during class space exploration. In the initial exploration of the target class space, some unknown and known samples are divided into one class, and these samples need to be separated through the splitting process. We select the first n_{if} distant subcluster-pairs based on the subcluster centers' distance $ED(\cdot,\cdot)$, which is integrated into the candidate set to be split. Then, the Hasting ratio is calculated on the candidate set to determine whether the split operation is accepted. Once the split operation is accepted in the class, which includes both the unknown and known samples, the unknown samples can be separated into a novel unknown class, while the known samples remain in one class. In this way, we can explore the accurate class space based on the initial target class space. More importantly, although the unknown samples are separated into a class, there still are samples belonging to different unknown classes. Thus, we also need to explore more accurate class space through the split process in the subsequent exploration process.

A. n_{if} Selection Criteria

The criteria for selecting n_{if} for merging or splitting are as follows. The target class number K is initialized by the known class number K_n, which is larger than 1 or 2, so the K we obtain is also large. During the merging process, the number of possibilities for all possible merging clusters is $[K(K-1)/2]$. Note that $[\cdot]$ is the floor function.

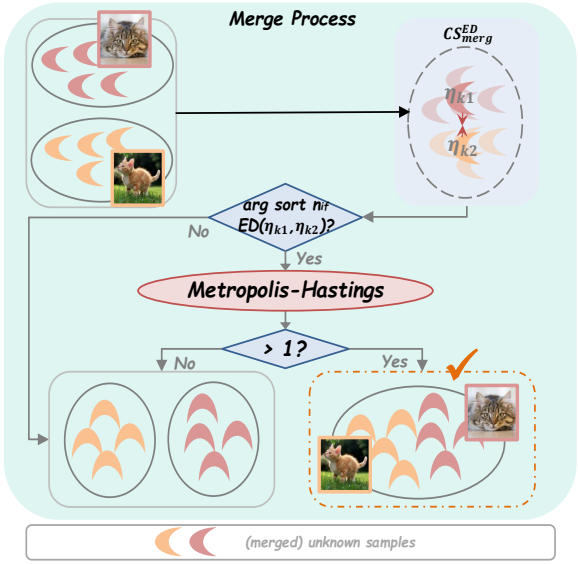

Figure 2. Merge process during class space exploration, whose candidate set obtaining and the determination of merging operation are similar to the one in the split process.

Based on the above, we set n_{if}(merge) to half of all possibilities as $[K(K-1)/4]$, so the conditions for obtaining candidate sets are not too loose. Then, by calculating the Hasting ratio from the M-H framework, we infer whether the (sub)clusters are truly merged, to make a more accurate judgment on the target class space. The criteria of selecting n_{if}(split) during the splitting process is the same as that during the merging process, and n_{if}(split) is set as $[K/2]$.

B. Splitting and Merging Decision

Following [1], the Hasting ratio for splitting is computed on the candidate master cluster to be split:

$$H^s_k = \frac{\alpha \prod_{c_s \in \{k_{sub1}, k_{sub2}\}} \Gamma(n_{c_s})\, f_x\left(x_{\{c_s\}}; \lambda\right)}{\Gamma(n_k)\, f_x\left(x_{\{k\}}; \lambda\right)}, \quad (1)$$

where k_{sub1} and k_{sub2} are the two sub-clusters of k-th master cluster, and c_s represents the candidate cluster pairs to be split. n_k is the number of features belonging to the k-th cluster and n_{c_s} is the number of features belonging to the candidate cluster pairs to be split. $\Gamma(\cdot)$ is the Gamma function. $f_x(\cdot;\lambda)$ is the marginal likelihood with λ representing the posterior hyperparameters, whose calculating details are following [2]. Thus, the split on the k-th master cluster can

be accepted if $H_{\mathrm{k}}^{s} > 1$. Similarly, the Hasting ratio for merging is computed on the candidate cluster to be merged:

$$H_{\mathrm{k_1 k_2}}^{m} = \frac{\Gamma\left(n_{\hat{k}}\right) f_x\left(x_{\{\hat{k}\}}; \lambda\right)}{\alpha \prod_{c_m \in \{k_1, k_2\}} \Gamma\left(n_{c_m}\right) f_x\left(x_{\{c_m\}}; \lambda\right)}, \tag{2}$$

where k_1 and k_2 are the two clusters with the potential to be merged, and c_m represents the candidate cluster pairs to be merged. $\hat{k}$ represents the cluster after merging the clusters k_1 and k_2. $n_{\hat{k}}$ is the number of features belonging to the $\hat{k}$-th cluster and n_{c_m} is the number of features belonging to the candidate cluster pairs to be merged. Thus, the merge on the k_1-th and k_2-th clusters can be accepted if $H_{\mathrm{k_1 k_2}}^{m} > 1$. Note that, when the merge clusters are finally determined according to Eq. (2).

By calculating the Hasting ratio for splitting and merging, we can infer a more accurate target class number K, which is beneficial to exploring the target class space. After that, we optimize the unknown diffuser over the wider class space, which is mentioned in our paper, by taking advantage of the reliable known knowledge and cluster alignment on the explored space.

In such an optimization based on the explored wider class space, we first select the high-confidence known samples as the reliable known knowledge. Then, we hope to use the hard pseudo-labels of the high-confidence known samples as supervision information, thereby leveraging them for better knowledge transfer on the explored class space. Specifically, for a known sample, we take the maximum logit output by the classifier in the pre-trained model as its confidence. And the known sample is considered as the high-confidence known sample when it is greater than a manually-set confidence threshold, whose hard pseudo-label is represented by the one-hot paradigm $\boldsymbol{y}_i^h$ ($x_i^h \in \mathcal{H}$). Note that $\mathcal{H}$ is the set of the high-confidence known samples. Therefore, we can perform better knowledge transfer on the explored class space by leveraging the reliable known knowledge $\boldsymbol{y}_i^h$ ($x_i^h \in \mathcal{H}$) obtained on the pre-trained model. In addition, we also perform cluster alignment on the explored class space, similar to Equation (2) in our paper ($\sum_{i=1}^{n} L_{alg}(x_i^t)$), to complete the transfer on known samples further and achieve generalization on unknown samples. Finally, we leverage the supervision of the high-confidence known sample labels and the constraint of the cluster alignment, to jointly achieve superior knowledge transfer on known classes and generalization on unknown classes over the wider class space:

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{kt} + \mathcal{L}_{alg} &= \sum_{i}^{\mathcal{H}} L_{kt}(\widetilde{\boldsymbol{z}}_i, \boldsymbol{y}_i^h) + \sum_{i=1}^{n} L_{alg}(x_i^t) \\ &= \underbrace{\sum_{i}^{\mathcal{H}} L_{kt}(\widetilde{\boldsymbol{z}}_i, \boldsymbol{y}_i^h) + \sum_{j_1=1}^{n_1} L_{alg}(x_{j_1}^t)}_{known} + \underbrace{\sum_{j_2=1}^{n_2} L_{alg}(x_{j_2}^t)}_{unknown},\end{aligned} \tag{3}$$

where $\mathcal{L}_{kt}$ is the supervision of the high-confidence known sample labels and $\mathcal{L}_{alg}$ is the constraint of the cluster alignment in our paper, to achieve superior knowledge transfer on known classes and generalization on unknown classes. $\widetilde{\boldsymbol{z}}_i$ represents the soft-distribution output by the main network C on the explored class space. n_1 and n_2 represent the number of known and unknown samples, respectively.

method	SF	bcycle	bus	car	mcycle	train	truck	OS*	UNK	HOS
Cluster	✗	73.7	79.5	66.0	83.5	84.2	0.9	**65.0**	44.0	52.0
SHOT	✓	23.1	76.6	58.5	89.9	82.4	1.4	55.3	37.2	44.5
AaD	✓	49.5	63.1	32.3	60.1	69.9	27.9	50.5	90.0	64.7
Ours(w/o s)	✓	54.3	38.7	44.7	75.0	66.6	50.5	55.0	**99.6**	<u>70.8</u>
Ours(w/ s)	✓	54.9	50.3	41.3	72.0	69.1	52.7	<u>56.7</u>	**99.6**	**72.3**

Table 1. Accuracy (%) on VisDA for OS-SFDA. OS* represents the average accuracy of the per known class (bicycle, bus, car, motorcycle, train, truck), while UNK is the unknown class accuracy. HOS represents the harmonic mean between OS* and UNK.

Methods	FINCH	DenMune	**Ours(w/o s)**	**Ours(w/ s)**
ACC	33.0	25.0	58.0	58.3
K	4	15	12	12

Table 2. Comparing ACC(%) and the inferred target class number K on VisDA with baselines.

C. Results Analysis on VisDA

Dataset VisDA. VisDA [3] is a large-scale dataset for the synthetic-to-real domain adaptation task with two domains and 12 classes. Its source domain has 152k synthetic images, while the target domain has 55k real-world images. We choose 6 classes (bicycle, bus, car, motorcycle, train, truck) as the known class, while the remaining 6 classes are considered the unknown class.

Results Analysis. As shown in Tab. 1, our method (w/ s)'s accuracy (OS*, UNK, HOS) is much higher than the traditional method SHOT [4]. Compared with the state-of-the-art method AaD [5], our method (w/ s) achieves 6.2%, 9.6%, and 7.6% relative improvements over AaD for OS*,

UNK, and HOS, respectively. Furthermore, as shown in Tab. 2, our method (w/ s) outperforms all no-parameter clustering methods (FINCH, DenMune) for ACC and obtains the more accurate inferred target class number K. The above results demonstrate the superiority and effectiveness of our method (w/ s). Meanwhile, our method (w/o s) outperforms all compared methods, indicating that our method (w/o s) is also efficient and superior without the supervision of reliable known knowledge.

D. Time Complexity Analysis

Our Proposed Method. In addition to the pre-trained model, we introduce an additional unknown diffuser to solve the open-set source-free domain adaptation task (OS-SFDA). Compared with the pre-trained model, we pay more attention to the time complexity consumed by the unknown diffuser, whose time complexity for each part and overall are as follows:

Model Encoding Process. The time complexity of the unknown diffuser to encode data is $\mathcal{O}(n_b dK)$. n_b and d represent the number of training target samples in each batch and the dimension of the feature output by the feature extractor f in the pre-trained model, respectively. The value of the target class number K changes dynamically with the exploration process during training, which is initialized by the known class number K_n.

Optimization in the target class space and target domain. For OS-SFDA, our method utilizes the unknown diffuser to explore the target class space. The exploration process is divided into two parts: cluster distribution optimization and class space exploration, the time complexity of which are $\mathcal{O}(n_b dK + n_b K)$ and $\mathcal{O}(K^2 d + n_b K)$, respectively. Therefore, the total time complexity for optimization (exploration) in the target class space is $\mathcal{O}(n_b dK + n_b K + K^2 d)$ at each iteration. Then, based on the explored wider class space, we perform the optimization in the target domain; that is, we realize the known knowledge transfer and unknown generalize in the target domain based on the wider class space. We obtain reliable known knowledge through contrastive learning and its time complexity can be represented as $\mathcal{O}(n_b n_b^- K_n)$, where we select n_b^- negative samples ($n_b^- \ll n_b$) and the time complexity is approximately written as $\mathcal{O}(n_b K_n)$. After that, as shown in Eq. (3), reliable known knowledge and cluster alignment on the explored class space are used as supervision information, to perform known knowledge transfer and unknown knowledge generalization, whose time complexities are $\mathcal{O}(n_h K_n)$ and $\mathcal{O}(n_b dK + n_b K)$, respectively. n_h ($n_h < n_b$) is the number of the high-confidence known samples. Therefore, the total time complexity in the optimization process is $\mathcal{O}(n_b dK + n_b K + n_b K_n + n_h K_n)$.

Overall Time Complexity. Ultimately, the overall time complexity of our proposed method is $\mathcal{O}(n_b dK + n_b K + K^2 d + n_b K_n + n_h K_n)$, which is approximately written as $\mathcal{O}(n_b dK + n_b K + K^2 d + n_b K_n)$.

Existing Methods. For the basic network architecture, our method uses the same pre-trained model M with the existing traditional method SHOT [4] and advanced method AaD [5], which utilize the backbone of Resnet-50 on Office-Home and Office-31 while that of Resnet-101 on VisDA. Except for the same pre-trained model, SHOT and AaD utilize corresponding strategies to transfer knowledge in the source-free and open-set setting, whose time complexities are $\mathcal{O}(n_b d)$ and $\mathcal{O}(n_b d + n_b^2 K_n)$, respectively. Since the final K obtained by exploration and optimization is about twice as large as K_n, the time complexity consumed by our method can be regarded as the same magnitude order with the one consumed by the existing methods.

In summary, through the above time complexity analysis, our method is consistent with SHOT and superior to AaD for time complexity. It shows that our method does not sacrifice time complexity to obtain a better adaptation effect in OS-SFDA.

References

[1] Jason Chang and John W Fisher III. Parallel sampling of dp mixture models using sub-cluster splits. *Proc. Adv. Neural Inform. Process. Syst.*, 26, 2013. 1

[2] Jason Chang et al. *Sampling in computer vision and Bayesian nonparametric mixtures*. PhD thesis, Massachusetts Institute of Technology, 2014. 1

[3] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017. 2

[4] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *Proc. Int. Conf. Mach. Learn.*, pages 6028–6039. PMLR, 2020. 2, 3

[5] Shiqi Yang, Shangling Jui, Joost van de Weijer, et al. Attracting and dispersing: A simple approach for source-free domain adaptation. *Proc. Adv. Neural Inform. Process. Syst.*, 35:5802–5815, 2022. 2, 3