A. Limitations

While we can verify that exploring and exploiting minimum happy points are critical for ASFDA through a significant improvement, the existence of the MH points is challenging to prove due to the lacking of theoretical guarantees and interpretability of deep networks. Further, ASFDA relies on the well-trained source model that may be impaired by some causes, e.g., model transfer, model sharing, and source model training process. In these uncontrollable and unforeseen circumstances, the robustness of ASFDA methods would face serious challenges.

B. Implementation details

Model Details. We implement our experiments on the PyTorch platform. For a fair comparison, we adopt the backbone of ResNet-50 [4] for Office-Home and Office31, ResNet101 for VisDA-2017. Besides, we also add experiments of ResNet-50 on VisDA-2017, VGG16 [12] and Alexnet [7] on Office-Home and Office-31. We utilize the same network architecture as SHOT and conduct label smoothing as SHOT in the process of training the source model. For all datasets without a train-validation split, we view the all source data as a test set and train the optimal source model using the test set as validation. The maximization number of epochs for Office-31, Office-Home, and VisDA is set as 100, 50, and 10. Meanwhile, when we train the target model, we train 20 epochs for Office-31 and Office-Home, and 10 epochs for VisDA. We run three times and report the average results. In the process of training, we adopt minibatch SGD with momentum 0.9 and weight decay 1e-3. The learning rate is set to 1e-2 for Office-31 and Office-Home, and 1e-3 for VisDA-2017. Besides, we set \( \alpha = 3 \), \( \beta = 0.3 \), and \( o = 5 \) for all datasets. Experiments are conducted on a TITAN Xp.

Labeling budget. In active source free domain adaptation, a few samples are utilized to aid the adaptation process. Following the previous active domain adaptation works [3, 15], we mainly shows the results of 5% actively labeled target data. We also report the results of various selection ratios, e.g., 1%, 5%, ..., 10%, for analysis.

Baseline implementation. The active domain adaptation methods cannot be reproduced without the source data in ASFDA, so we report the best results shown in papers corresponding to our active source free domain adaptation, such as AADA [13], EADA [15], and so on. Similarly, because the ELPT [8] does not open the code, we report its original results. We now elaborate on our implementation of other baseline algorithms:

1. Base: we acquire the pseudo-labels of the target samples based on a deep clustering [9], and train the model with entropy loss, KL divergence and the standard cross-entropy loss. Here the pseudo-labels are not credible, so we apply the weight \( \beta = 0.3 \) and \( \alpha = 0.3 \) to all samples.

2. Random: select samples randomly.

3. Entropy [14]: select samples for which the model has highest predictive entropy.

4. BVSB (Best-Versus-Second-Best) [6]: select samples for which the the smallest difference between the top-2 predicted probabilities.

5. LC (Least Confidence) [5]: select samples for which the smallest of the maximum probability of model output.

6. CoreSet [11]: CoreSet formulates active sampling as a set-cover problem. We implement the CoreSet using the released code: https://github.com/JordanAsh/badge.

7. CTC (closest to its center): select samples which are the closet to their centers after the Kmeans clustering.

8. BADGE (batch active learning by diverse gradient embeddings) [2]: BADGE ensures diverse batches by running Kmeans++ [1] on ‘gradient embeddings’, which incorporates model uncertainty and diversity. We implement BADGE using the released code: https://github.com/JordanAsh/badge.

*denotes corresponding author.
Methods (3), (4), and (5) are based on model uncertainty, (6) and (7) are diversity-based, and (8) is a hybrid approach that combines uncertainty and diversity.

Regarding implementation details, MHPL selects active samples and learns selected samples with a proposed neighbor focal loss compared with the base. In contrast to other active strategies, MHPL explores more informative MH points and exploits these MH points with a proposed neighbor focal loss instead of the standard cross-entropy loss.

C. Additional experimental results

Full experimental results under different networks. To further prove the effectiveness of MHPL, we conduct experiments on Office-31 and Office-Home with different backbones: VGG and Alexnet. We find two essential observations from the results in Table 1 and 2. (1) Even a small set of labeled target data may bring larger performance gains, which proves the practicality and effectiveness of ASFDA. On average, when only 5% of the Office-home target data is used for annotation, MHPL is 11.3% and 9.2% higher than the base with Alexnet and VGG, respectively. Similarly, the accuracy of MHPL is 11.6% and 4.8% higher than the base on Office-31 with Alexnet and VGG, respectively. (2) Our method MHPL could fully explore and exploit MH points which are crucial for ASFDA. As shown in Table 1 and 2, the performance gains brought from MHPL are significantly higher than other active strategies in all tasks with VGG and Alexnet, especially in several challenging tasks.
Ablation studies on hyperparameters \( q, o \) and one-shot querying. 

(a) Neighbor number of NAU  
(b) Neighbor number of NDR  
(c) Pr→Cl  
(d) Re→Ar

Figure 1. Ablation studies on hyperparameters \( q, o \) and one-shot querying.

Figure 2. Feature visualization for the source model with 5% actively labeled target data on Ar→Cl task. Different colors in (a) represent different classes of pseudo-labels by clustering. Blue blocks include easily-adaptive source-similar samples with label-clean neighbors that can be learned well by SFDA methods. Red blocks include the hard-adaptive source-dissimilar samples with label-chaotic neighbors. In (b), (c), (d), (e), and (f), the dark green indicates that the pseudo-label is consistent with the true label, and light blue indicates the opposite. The red stars indicate the selected samples based on different criteria, respectively.

**Ablation study on number of neighbors** \( q \). To study the sensitivity to a crucial hyperparameter \( q \), we run ablation experiments on three tasks, Pr→Cl, Re→Ar, and Re→Cl, in ResNet-50. As shown in Fig. 1 (a), utilizing 5% labeled target data, results with \( q = 5/10/20/30/40 \) show that our method is not sensitive to the choice of a reasonable \( q \) on three tasks.

**Ablation study on number of neighbors** \( o \) in NDR. To study the sensitivity to a hyperparameter \( o \) in neighbor diversity relaxation, we run ablation experiments on Ar→Cl and Pr→Cl in ResNet-50. As shown in Fig. 1 (b), when \( o = 0 \), the NDR is not being executed, and the model is not performing well. The model’s accuracy is significantly improved with the increase of \( o \) value. When the value of \( o \) is larger, the accuracy gradually tends to be stable. We set \( o = 5 \) on all datasets.

**Ablation study on one-shot querying.** To further verify the effect of one-shot querying, we conduct the experiments on Pr→Cl and Re→Ar to compare the effect with the samples selected at different epochs of model training. The results in Fig. 1 (c) and (d) verify that the samples chosen by the source model are source-dissimilar and effective.

**Feature visualization.** As shown in Fig. 2, compared with BVSB [6], Coreset [11], and BADGE [2], the samples selected by our MHPL mostly fall into source-dissimilar and label-chaotic regions. Even the samples chosen by Entropy [14] are also mostly label-chaotic, but they are not diverse. Therefore, the samples selected by our MHPL meet all three conditions of minimum happy points: neighbor-chaotic, individual-different, and source-dissimilar.

**Ablation on focal loss.** As shown in Fig. 3, with different active sample ratios in Cl→Pr and Pr→Cl, the results from focal loss are always higher than those from CE loss.

**More ratios.** Since LC [5] is the best approach in the active strategy, we compare LC with our MHPL in different ratios of active samples on the Pr→Cl and Re→Cl tasks. For the two tasks shown in Fig. 4, our MHPL is always far
superior to LC in different sample selection ratios.

Ablation study on hyperparameters $\alpha$. We utilize 5% and 10% active labeled target samples in the Cl→Pr task to analyze the hyperparameter of $\alpha$. As shown in Fig. 5 (a) and (b), the accuracy with $\alpha$Pu is significantly higher than that of cross-entropy loss. We set $\alpha = 3$ on all datasets.

Ablation study on hyperparameters $\beta$. We utilize 5% and 10% active labeled target samples in the Ar→Cl task to analyze the hyperparameter of $\beta$. When $\beta = 0$, the performance is worse as the $D_U^1$ is not utilized. With the increase of $\beta$, the effect of the model first increases and then decreases, indicating that the model is supposed not to focus more on $D_U^1$. We set $\beta = 0.3$ for all datasets.

References


