Supplementary Material: Local Context-Aware Active Domain Adaptation

Tao Sun
Stony Brook University
tao@cs.stonybrook.edu

Cheng Lu
XPeng Motors
luc@xiaopeng.com

Haibin Ling
Stony Brook University
hling@cs.stonybrook.edu

In this supplementary material, we provide additional details, results and analyses on the experiments conducted in the main paper.

A. Dataset Details

Office-31 [10] contains 31 classes of 4,110 office environment related images. It has three domains: Amazon (A), DSLR (D) and Webcam (W). **Office-Home** is a similar dataset, containing 15,500 office images from 65 classes, split in four domains: Product (Pr), Clip Art (Cl), Artistic (Ar) and Real-World (Rw). **Office-Home RSUT** [12] is a subset of Office-Home created with the protocol of Reverse-unbalanced Source and Unbalanced Target to have a large label distribution shift. The major classes in the 'RS' fold become minor classes in the 'UT' fold, while the minor classes in the 'RS' fold become major classes in the 'UT' fold. **VisDA** [7] is a large-scale Synthetic-to-Real dataset of 12 objects. The training set contains 152,397 synthetic 2D renderings of 3D models and the validation set contains 55,388 real images. We use the training set as the source domain and the validation set as the target domain. **DomainNet** [6] consists of about 0.6 million images from 345 classes, distributed in six domains. Following [8, 3], we use five domains: Real (R), Clipart (C), Painting (P), Sketch (S), and Quickdraw (Q) for experiments.

B. Implementation Details

We implement all experiments with PyTorch 1.8. Results are run on servers with NVIDIA A5000/A6000 GPU. Following previous ADA works [1, 14, 13], we use ResNet-50 [2] pretrained on ImageNet [9] as the backbone network, a bottleneck layer (`Linear->BatchNorm1d`), and a classification head of one single `Linear` layer. The bottleneck feature dimension is 256. Training images are first resized to 256×256, and then randomly cropped to 224×224. Test images use center cropping instead. We adopt Adadelta optimizor with learning rate of 0.1 and a batch size of 32. On Office-Home and Office-31, we first train the models on only source data for 10 epochs, and then train on both source and target data with active domain adaptation for 30 epochs. At the epoch of 10, 12, 14, 16, 18, $B/5$ target data are selected for querying labels, where B is the labeling budget. On VisDA, we conduct source-only training for 1 epoch and ADA for 10 epochs. On DomainNet, we conduct source-only training for 10 epochs and ADA for another 10 epochs. Mean accuracies of 3 repeated experiments are reported. Code is available at https://github.com/tsun/LADA.

C. Additional Results and Analyses

Running time. Table A.1 reports running time in seconds with one A6000 GPU, including active sampling (AL) time averaged over 5 rounds (10%-budget) and model update (DA) time averaged over all training epochs. LAS consumes much less time than CLUE and slightly more than other AL methods. RAA/LAA is comparable to MCC and faster than CDAC.

Pseudo-label quality of LAA/RAA. We conduct experiments with different confidence thresholds τ and report the percentage of target samples in the anchor set with correct pseudo-labels. Shown in Fig. A.3, as τ reduces, the pseudo-label quality decreases. LAA has better pseudo-label quality than RAA. We also report results by fixing $\tau = 0.9$ and varying neighborhood size K in LAA (dashed lines). Overall, $\tau = 0.9$ used in the paper leads to a decent pseudo-label quality.

Comparison with other criteria on Office-31 Figure A.1 presents analyses on Office-31 similar to that on Office-Home in the Fig. 2 of the main paper. In the left figure, our LAS outperforms other active leaning criteria for labeling budgets ranging from 3% to 20%. When the labeling budget is small (*e.g.*, 3% or 5%), LAS boosts the accuracy by a large margin. Since in ADA the situation with a small labeling budget is more important, it shows the effectiveness of LAS. In the center figure with 5%-budget, the curve of LAS lies above others after 1% samples are selected. In the right figure, when combined with five different domain adaptation strategies, LAS consistently achieves the highest accuracies than three other active learning criteria that are previous arts.

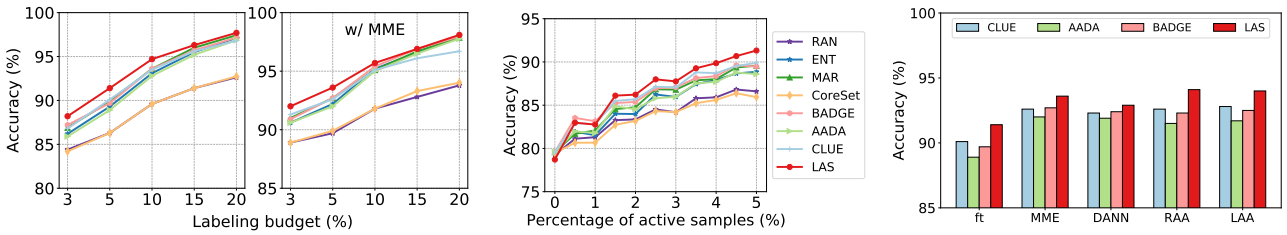
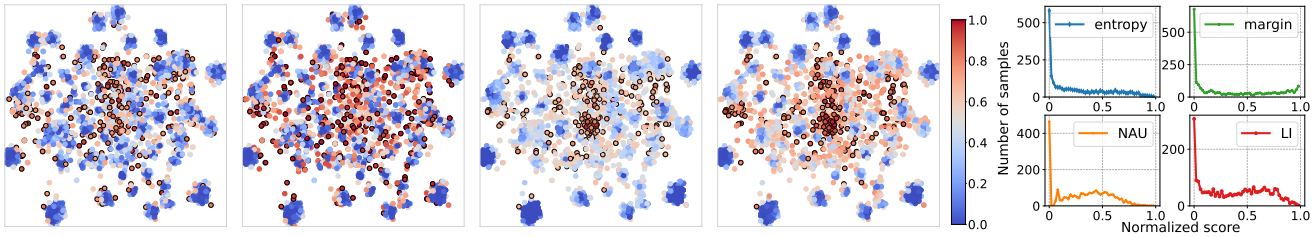


Figure A.1: Analysis on Office-31. (Left) varying labeling budget; (Center) accuracy curves with 5%-budget; (Right) combining AL criteria with different DA strategies.



(a) entropy (b) margin (c) NAU (d) LI (e) histogram

Figure A.2: (a-d) t-SNE visualization of target features on Office-Home Rw→Ar. Samples are colored according to their normalized uncertainty scores, where red indicates large values and blue indicates small values. The top 10% samples with highest scores are marked with black boarders. (e) Histogram of target samples by normalized scores.

Table A.1: Running time on **Office-Home Ar→Rw** and **VisDA** in seconds.

	AL					DA				
	BADGE	AADA	CLUE	MHPL	LAS	ft	MCC	CDAC	RAA	LAA
Ar→Rw	17.6±0.5	18.4±0.8	**24.0**±0.9	17.8±0.4	18.2±0.7	30.2±8.7	39.6±8.6	**67.4**±16.5	63.0±18.2	62.0±10.8
VisDA	80.4±2.4	57.0±1.7	**733.2**±56.6	70.0±2.6	105.6±17.9	1136.6±45.3	1623.4±18.4	**2662.0**±17.2	1449.8±15.8	1624.2±245.2

Table A.2: Comparison with Semi-Supervised Domain Adaptation methods on **Office-Home** using 10%-budget.

Task	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.
SSDA	ECACL	72.2	86.7	82.8	70.5	85.0	82.6	70.9	71.5	82.9	76.0	74.0	88.9	78.7
	CDAC	69.5	83.2	80.2	66.9	82.4	78.7	66.1	70.6	80.9	72.3	70.5	87.2	75.7
ADA	TQS	64.3	84.8	83.5	66.1	81.0	76.7	66.5	61.4	82.0	73.7	65.9	88.5	74.5
	CLUE	62.1	80.6	73.9	55.2	76.4	75.4	53.9	62.1	80.7	67.5	63.0	88.1	69.9
	S^3VAADA	67.8	83.9	82.9	67.0	81.4	79.5	65.8	65.9	82.4	74.8	68.6	87.9	75.7
	LAMDA	74.8	88.5	86.9	73.8	88.2	83.3	74.6	75.5	86.9	80.8	77.8	91.7	81.9
	LADA	**77.2**	**91.9**	**88.1**	**76.9**	**91.1**	**86.8**	**76.6**	**78.1**	**88.3**	**82.0**	**79.0**	**93.8**	**84.2**

Remarks on SSDA. Semi-supervised DA (SSDA) is closely related to Active DA. In both task, a few labeled target data and many unlabeled target data are available. Yet there are some differences. In SSDA, all labeled target data are provided for once at the beginning of training and fixed afterwards. While in ADA, labeled target data are actively selected. The active querying and model update interleave for several rounds during the training of ADA.

Existing SSDA methods can be directly applied in ADA. In the paper, we have compared different active query methods when using MME [11] and CDAC [4] as the model adaptation methods. Additional results on Office-Home with CDAC is provided in Tab. A.4. Generally, the proposed LAS can select more informative samples than other active selection criteria.

Nevertheless, it may be sub-optimal to simply combine active query with existing SSDA methods. A unified ADA solution that considers both active query and model adaptation would be better effective. Tables A.2, A.3 present comparison results with two state-of-the-art SSDA methods, ECACL [5] and CDAC [4]. The comparison results are taken from [3]. It should be noted that although ADA can select more informative labeled samples, the performance is also affected by the way to utilize unlabeled data.

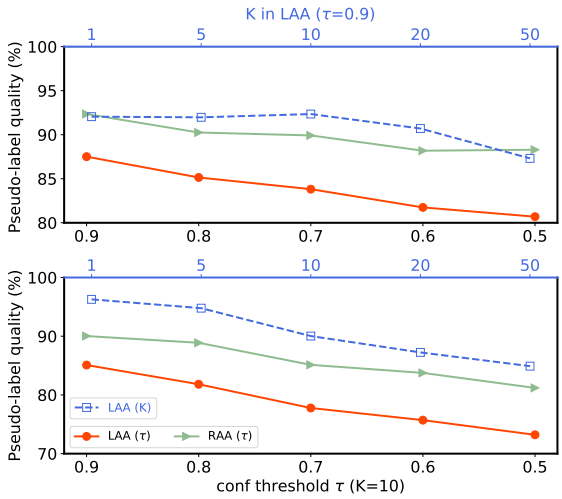


Figure A.3: Pseudo-label quality averaged over Office-Home Cl→{Ar, Pr, Rw} (upper) and Office-Home RSUT C→{P, R} (lower) using 10%-budget.

Table A.3: Comparison with SSDA methods on **Office-Home RSUT** using 10%-budget.

Task	Method	C→P	C→R	P→C	P→R	R→C	R→P	Avg.
SSDA	ECACL	78.6	68.6	59.5	77.1	61.9	82.0	71.3
	CDAC	73.0	58.7	55.8	73.3	50.3	77.3	64.7
ADA	TQS	69.4	65.7	53.0	76.3	53.1	81.1	66.4
	CLUE	69.7	65.9	57.1	73.4	59.5	82.7	68.1
	S^{3}VAADA	73.0	63.0	50.7	69.6	52.6	78.3	64.5
	LAMDA	81.2	75.7	**64.1**	81.6	65.1	87.2	75.8
	LADA	**83.2**	**77.2**	63.8	**83.0**	**65.4**	**88.1**	**76.8**

From the tables, ECACL and CDAC surpass three early ADA methods. The state-of-the-art LAMDA method selects target data to approximate the entire target distribution, and addresses the issue of label distribution mismatch between source and target domains. It obtains better performances than SSDA arts. Our proposed LADA (LAS w/ LAA) selects locally-representative samples, and progressively expand the labeled data with confident samples in a class-balanced manner. LADA outperforms LAMDA by +2.3% on Office-Home and +1.0% on Office-Home RSUT. When replacing LAA with CDAC in LADA, the performance drops, as we show in the paper.

Uncertainty measures in LAS. Figure A.2 visualizes the target features on Office-Home Rw→Ar. Similar to the plots in Fig. 3 of the paper, entropy and margin include some outliers in the top 10% samples (see circles with black boarders in the bottom part of Figs. A.2a, A.2b). For NAU in Fig. A.2c, target data have small normalized scores. Target data with high normalized LI scores form several small clusters in Fig. A.2d. From the histogram in Fig. A.2e, a maximal can be observed around 0.6-0.7 for LI. These phenomenons are similar to Office-31 W→A in the paper.

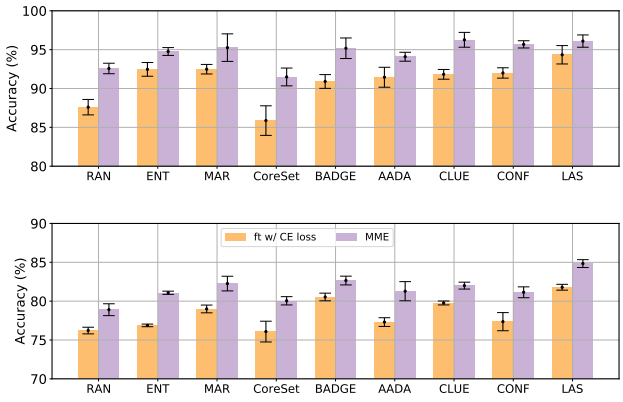


Figure A.4: Visualization of standard deviations on Office-31 A→W (upper) and W→A (lower) using 5%-budget.

Training with a joint labeled set. In the implementation of some early ADA works [1, 14], the queried labeled data are added to the source labeled data, and training mini-batches are sampled from this joint labeled set. The objective is

$$\mathcal{L} = \mathbb{E}_{(x,y)\sim\mathcal{D}_s\cup\mathcal{D}_{tl}}\ell_{\text{ce}}(h(x),y) \tag{A.1}$$

where ℓ_{ce} is the cross entropy loss. Differently, recent works [13, 3] and ours adopt

$$\mathcal{L} = \mathbb{E}_{(x,y)\sim\mathcal{D}_s}\ell_{\text{ce}}(h(x),y)+\mathbb{E}_{(x,y)\sim\mathcal{D}_{tl}}\ell_{\text{ce}}(h(x),y) \tag{A.2}$$

Comparing Eq. A.1 with Eq. A.2, the advantage of training with a joint labeled set is that it only needs to back-propagate through one batch of data, thus reducing the memory and computation usage. The disadvantage is that the labeled data set is dominated by the source data. When there is a large domain gap (*e.g.* when label distribution shift exists), the performance may be hurt.

Nevertheless, to better demonstrate the effectiveness of LAS, Table A.4 lists the results using fine-tuning with a joint labeled set. Accuracies are slightly lower than their counterparts in Table 1 of the main paper. LAS still achieves the best scores among all AL methods.

Visualization of standard deviations. Figure A.4 plots the standard deviations on two Office-31 tasks over 3 repeated experiments. Performances are relatively stable to different random initializations. Of all active selection methods, our proposed LAS obtains the highest average accuracies.

Visualization of LAS. To visualize how LAS selects target samples, we present t-SNE plots of target features on Office-Home Pr→Ar and Ar→Pr in Fig. A.5 and Fig. A.6, respectively. We choose the 10th epoch, where 10% target data are selected as candidates based on LI-scores, of which 1% target data from cluster centroids are selected for querying labels. Candidate, selected and remaining target samples are marked with squares, stars and points, respectively. The top 20 candidates and queried images are also

Table A.4: Accuracies (%) on **Office-Home** with 5% labeled target samples. (†Training mini-batches are sampled from a joint labeled set.)

AL method	DA 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.
RAN	ft w/ CE loss†	60.5	78.4	79.0	60.2	74.4	72.7	61.5	56.2	77.6	69.9	59.6	82.8	69.4
ENT		62.8	81.5	82.7	64.1	78.1	75.7	63.4	57.5	81.2	72.8	62.5	87.3	72.5
MAR		64.0	81.8	82.7	64.1	79.0	74.9	64.6	59.9	80.7	73.2	64.6	87.8	73.1
CoreSet		58.5	77.1	79.3	60.8	72.4	71.8	60.9	54.9	77.3	70.9	58.8	81.3	68.7
BADGE		65.5	83.6	82.1	63.1	79.8	75.3	64.9	61.0	80.8	73.1	65.1	87.1	73.5
AADA		61.8	82.0	82.1	62.3	77.7	76.0	63.1	59.4	81.8	72.9	62.4	87.2	72.4
CLUE		65.3	81.8	81.7	62.6	78.5	74.8	63.9	61.4	79.9	72.9	63.1	87.6	72.8
CONF		63.4	81.9	82.9	63.8	78.2	75.8	64.2	60.2	81.6	73.3	63.2	87.4	73.0
MHPL		65.6	82.1	82.9	65.3	79.1	74.6	64.7	61.4	81.6	73.3	63.7	88.1	73.5
LAS		67.2	84.3	83.1	65.1	80.9	77.0	65.3	62.5	81.4	73.8	66.7	89.0	74.7
RAN	CDAC	61.6	78.8	80.1	67.7	80.2	77.6	68.7	61.9	79.7	74.1	63.0	85.2	73.2
ENT		62.9	81.9	83.4	69.0	82.0	80.0	70.3	63.3	84.2	75.6	67.7	87.1	75.6
MAR		65.6	83.8	83.3	69.0	83.7	81.0	70.2	65.7	84.6	75.9	67.0	88.1	76.5
CoreSet		58.9	77.7	79.6	67.1	77.9	77.2	67.4	58.6	81.6	73.6	63.4	83.4	72.2
BADGE		63.3	80.4	81.3	69.6	83.0	78.8	70.4	62.7	83.6	76.1	67.4	88.0	75.4
AADA		61.8	81.8	82.8	69.6	83.2	80.4	70.7	63.5	84.3	76.2	66.2	87.2	75.6
CLUE		65.2	83.6	82.3	68.8	84.4	79.8	69.7	64.9	83.6	75.2	68.0	87.5	76.1
CONF		62.6	82.9	83.8	70.6	83.6	79.7	70.0	64.6	84.3	76.4	66.8	88.1	76.1
MHPL		65.5	82.4	82.7	70.8	84.1	81.7	70.5	66.2	84.3	76.9	68.7	87.8	76.8
LAS		67.4	85.4	83.1	71.0	85.0	81.7	72.1	67.8	85.1	77.4	70.4	89.5	78.0
LAS	RAA	**71.2**	**88.1**	**85.3**	**73.2**	**87.8**	**83.8**	**72.6**	**72.2**	**86.6**	79.2	74.4	**91.7**	**80.5**
LAS	LAA	**71.2**	87.4	84.6	72.1	87.0	83.6	71.5	71.6	85.3	**79.3**	**75.5**	90.4	80.0

displayed under the t-SNE plots. As can be seen, the candidates (*i.e.*, samples with large LI-scores) generally lie in the regions where model predictions are inconsistent. It is also difficult to distinguish their semantic labels visually, especially for Pr→Ar, indicating that these images are hard cases. There are some highly similar images in the candidates. After the second step of diverse selection, images selected for querying labels become much more diverse.

References

[1] Bo Fu, Zhangjie Cao, Jianmin Wang, and Mingsheng Long. Transferable query selection for active domain adaptation. In *CVPR*, pages 7272–7281, 2021. 1, 3

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *ICCV*, pages 770–778, 2016. 1

[3] Sehyun Hwang, Sohyun Lee, Sungyeon Kim, Jungseul Ok, and Suha Kwak. Combating label distribution shift for active domain adaptation. In *ECCV*, 2022. 1, 2, 3

[4] Jichang Li, Guanbin Li, Yemin Shi, and Yizhou Yu. Cross-domain adaptive clustering for semi-supervised domain adaptation. In *CVPR*, pages 2505–2514, 2021. 2

[5] Kai Li, Chang Liu, Handong Zhao, Yulun Zhang, and Yun Fu. Ecacl: A holistic framework for semi-supervised domain adaptation. In *ICCV*, pages 8578–8587, 2021. 2

[6] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *ICCV*, pages 1406–1415, 2019. 1

[7] 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. 1

[8] Viraj Prabhu, Arjun Chandrasekaran, Kate Saenko, and Judy Hoffman. Active domain adaptation via clustering uncertainty-weighted embeddings. In *ICCV*, pages 8505–8514, 2021. 1

[9] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 1

[10] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *ECCV*, pages 213–226, 2010. 1

[11] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *ICCV*, pages 8050–8058, 2019. 2

[12] Shuhan Tan, Xingchao Peng, and Kate Saenko. Class-imbalanced domain adaptation: an empirical odyssey. In *ECCV Workshops*, pages 585–602, 2020. 1

[13] Binhui Xie, Longhui Yuan, Shuang Li, Chi Harold Liu, Xinjing Cheng, and Guoren Wang. Active learning for domain adaptation: An energy-based approach. In *AAAI*, 2021. 1, 3

[14] Ming Xie, Yuxi Li, Yabiao Wang, Zekun Luo, Zhenye Gan, Zhongyi Sun, Mingmin Chi, Chengjie Wang, and Pei Wang. Learning distinctive margin toward active domain adaptation. In *CVPR*, 2022. 1, 3

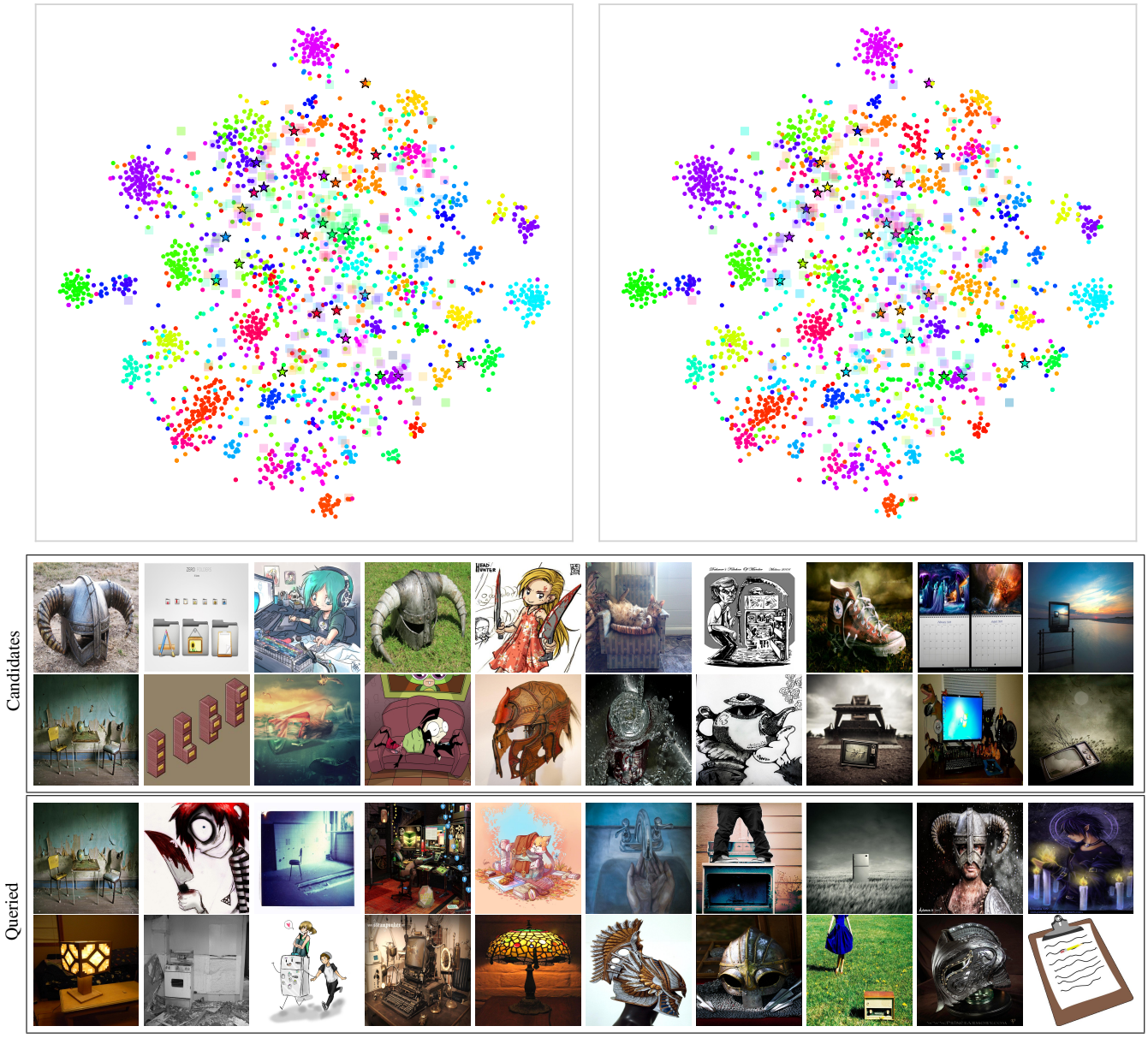


Figure A.5: Visualization of LAS sampling on Office-Home Pr→Ar. The first row presents t-SNE plots. Squares denote candidate target samples based on LI-scores; stars denote selected target samples for querying labels; and points denote the rest target samples. Each marker is colored according to its (left) ground-truth label and (right) pseudo label from the current model. The last two rows plot top 20 candidate samples and queried samples with largest LI-scores, respectively.

Figure A.6: Visualization of LAS sampling on Office-Home Ar→Pr. The first row presents t-SNE plots. Squares denote candidate target samples based on LI-scores; stars denote selected target samples for querying labels; and points denote the rest target samples. Each marker is colored according to its (left) ground-truth label and (right) pseudo label from the current model. The last two rows plot top 20 candidate samples and queried samples with largest LI-scores, respectively.