Upcycling Models under Domain and Category Shift — Supplementary Material

Sanqing Qu[1*], Tianpei Zou[1*], Florian Röhrbein[2], Cewu Lu[3], Guang Chen[1†],
Dacheng Tao[4,5], Changjun Jiang[1]

[1]Tongji University, [2]Chemnitz University of Technology,

[3]Shanghai Jiao Tong University, [4]JD Explore Academy, [5]The University of Sydney

1. Source Model Preparing

As aforementioned, in this paper, we focus on the K-way classification. For a given source domain $\mathcal{D}_s = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$ where $x_s^i \in \mathcal{X}_s$ and $y_s^i \in \mathcal{Y}_s \subset \mathbb{R}^K$, we adopt the same recipe as SHOT [7] and BMD [14] to prepare the source model. Specifically, the source model f_s parameterized by a deep neural network consists of two modules: the feature encoding module $g_s : \mathcal{X}_s \to \mathbb{R}^d$ and the classifier module $h_s : \mathbb{R}^d \to \mathbb{R}^K$, i.e., $f_s = h_s \circ g_s$. We optimize f_s with the following loss:

$$\mathcal{L}_{src} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} q_k \log \delta_k(f_s(x_s^i)) \tag{1}$$

where $\delta_k(f_s(x_s))$ denotes the softmax probability of source sample x_s belonging to the k-th category, q_k is the smoothed one-hot encoding of y_s, i.e., $q_k = (1-\alpha) * \mathbb{1}_{[k=y_s]} + \alpha/K$, and α is the smoothing parameter which is set to 0.1 for all benchmarks.

2. Experiments on Closed-set Adaptation

Existing methods designed for category shift, typically do not perform well for the vanilla closed-set domain adaptation scenario (CLDA). To examine the effectiveness and robustness of GLC, we further conduct experiments on Office-31, and Office-Home. All implementation details are the same as before, e.g., we adopt the ResNet-50 [5] as the backbone, the learning rate is set to 1e-3, and ρ is set to 0.75. The results are listed in Table 1 of this supplementary material. As shown in this Table, despite the fact that the GLC is not tailored for CLDA, we still attain comparable or even better performance compared to existing methods designated for CLDA, e.g., MDD [26]. Specifically, GLC obtains the overall accuracy of 88.1% and 70.4% on Office-31 and Office-Home, respectively. While MDD attains 88.9% and 68.1% on Office-31 and Office-Home, respectively. In a fairer comparison, GLC significantly outperforms UMAD [8], which is also model adaptation method designed for category-shift. Specifically, GLC outperforms UMAD by 6.4% and 7.3% on Office-31 and Office-Home, separately.

3. Experiments on Realistic Applications

So far, most of existing methods usually perform experiments only on standard computer science benchmarks. Here, we have further validated the effectiveness and superiority of GLC in realistic applications, including remote-sensing recognition, wild-animal classification, and single-cell RNA sequence identification. We present more details in the following.

3.1. Partial-set Model Adaptation on Remote Sensing Recognition

Remote sensing has great potential to manage global climate change, population movements, ecosystem transformations, and economic development. However, due to data protection regulations [24], it is difficult for researchers to obtain multi-scene, high-resolution satellite imagery. For example, there are strict data regulation policies in China for high-resolution remote sensing images in meteorological, oceanic, and environmental scenarios [25]. To validate the effectiveness of GLC on remote sensing, we conduct experiments on two existing large-scale datasets, the PatternNet [27] and the NWPU45 [1] dataset. PatternNet is one of the largest satellite image datasets collected from Google Earth imagery in the US. It contains 38 scene classes and 30,400 high-resolution ($0.2 \sim$ 6m per pixel) satellite images, such as airport, beach, dense residential, forest, etc. NWPU45 dataset consists of 45 scene classes and 31,500 satellite images covering more than 100 countries and regions around the world. Its spatial resolution varies from about 30 to 0.2 m per pixel. The heterogeneity of spatial resolution and geographic location poses a significant challenge to model adaptation. In this paper, we set the PatternNet as source dataset and the NWPU45 as target dataset to investigate the model adaptation from high-resolution satellite images to low-resolution satellite images. There

*Equal Contribution
†Corresponding author: guangchen@tongji.edu.cn

Table 1. Accuracy (%) comparison in CLDA scenario on Office-Home and Office-31. (Best in **bold**)

Methods	SF	OPDA	OSDA	PDA	CLDA	Office-Home													Office-31
						Ar2Cl	Ar2Pr	Ar2Re	Cl2Ar	Cl2Pr	Cl2Re	Pr2Ar	Pr2Cl	Pr2Re	Re2Ar	Re2Cl	Re2Pr	Avg	Avg
CDAN [10]	✗	✗	✗	✗	✓	49.0	69.3	74.5	54.4	66.0	68.4	55.6	48.3	75.9	68.4	55.4	80.5	63.8	86.6
MDD [26]	✗	✗	✗	✗	✓	**54.9**	73.7	77.8	60.0	71.4	71.8	61.2	53.6	78.1	72.5	**60.2**	82.3	68.1	**88.9**
UAN [23]	✗	✓	✗	✗	✗	45.0	63.6	71.2	51.4	58.2	63.2	52.6	40.9	71.0	63.3	48.2	75.4	58.7	84.4
CMU [3]	✗	✓	✗	✗	✗	42.8	65.6	74.3	58.1	63.1	67.4	54.2	41.2	73.8	66.9	48.0	78.7	61.2	79.9
DANCE [15]	✗	✓	✓	✗	✗	54.3	75.9	78.4	64.8	72.1	73.4	63.2	53.0	79.4	**73.0**	58.2	82.9	69.1	85.5
DCC [6]	✗	✓	✓	✓	✗	35.4	61.4	75.2	45.7	59.1	62.7	43.9	30.9	70.2	57.8	41.0	77.9	55.1	87.4
OVANet [16]	✗	✓	✓	✗	✗	34.5	55.8	67.1	40.9	52.8	56.9	35.4	26.2	61.8	53.8	35.4	70.8	49.3	70.4
Source-only	✓	-	-	-	-	44.8	67.4	74.2	53.0	63.3	65.1	53.7	40.5	73.5	65.6	46.3	78.3	60.5	78.8
UMAD [8]	✓	✓	✓	✗	✗	48.0	65.1	73.0	58.6	65.3	67.9	58.2	47.3	74.0	69.4	53.0	77.8	63.1	81.7
GLC	✓	✓	✓	✓	✗	51.2	**76.0**	**79.9**	**65.4**	**78.6**	**78.7**	**65.6**	**54.1**	**81.6**	70.9	58.4	**84.2**	**70.4**	88.1

are 21 overlapping scenes classes between PatternNet and NWPU45. Thus, we transfer the scene classes from the PatternNet to the 21 overlapping scene classes in the NWPU45 and compare the results with the original labels from the NWPU45 for performance evaluation.

An illustration of boxplot in Fig. 1b basically demonstrates that GLC effectively realizes model adaptation and achieves more accurate performance with less variance than existing methods. Quantitatively, GLC achieves $64.6 \pm 0.22\%$ overall accuracy with 5 different random seeds. In contrast to GLC, the baseline methods DCC, ETN, DANCE and BA3US obtains $55.2 \pm 0.80\%$, $54.9 \pm 0.77\%$, $62.0 \pm 1.02\%$ and $59.6 \pm 1.25\%$ overall accuracy, respectively. To verify the robustness, we further conduct an ablation experiment on decreasing the number of overlapping scenes between source and target domains. The number gets smaller, there is more probability of overlapping samples being categorized into other scenes. Despite this, the results in Fig. 1c show that GLC is still capable of addressing this challenge and even achieving better performance. Quantitatively, GLC obtains 64.5% average accuracy in four different target situations. In contrast, the baseline methods DCC, ETN, DANCE, and BA3US attains 55.4%, 55.0%, 52.4%, and 49.6% overall accuracy, respectively. We attribute this to our global one-vs-all clustering algorithm, which is able to discover non-existent scene categories and suppress model adaptation over these categories.

3.2. Open-set Model Adaptation on Wild Animal Classification

We next study a more challenging setting, the open-set model adaptation on wild animal classification. Having the ability to accurately classify wild animals is important for studying and protecting ecosystems [12], especially the ability to identify novel species [11]. However, it is almost impossible for a database to cover all animal species, and it is also typically difficult to collect and annotate a large number of wild animal images in practice. Thereby, it would be ideal if we develop an animal classification system based on the existing large number of virtual animal images on the Internet. In this article, we execute experiments on the I2AWA [28] benchmark to investigate open-set model adaptation from virtual to real-world. I2AWA consists of a virtual source domain dataset and a real-world target domain dataset with a total of 50 animal categories. We divide the first 30 into known categories in alphabetical order and the remaining 20 into unknown categories. The source domain dataset consists of 2,970 virtual animal images collected through the Google-Image search engine, while the target domain dataset comes from the AWA2 [22] dataset with a total of 37,322 images from the real world. Due to differences in image styles between virtual and real-world datasets, directly deploying a DNN model trained on the virtual images can lead to severe performance degradation. For a qualitative demonstration, we enumerate some animal images on target domain in Fig. 2a and apply the Grad-CAM heatmap [19] technique to compare the source model with the adapted target model by our GLC technique. From this, we can conclude that the source model typically fails to locate and extract key information for animal identification, while the upcycled model overcomes these failures well. For quantitative performance evaluation, we compare GLC with the methods dedicated to open-set domain adaptation (DANN [4], OSBP [17]), and the methods designed for universal domain adaptation (DANCE [15], OVANet [16]).

An inspection of the tSNE plots (Fig. 2b) indicates that our GLC algorithm effectively realizes known animal classification and unknown animal separation. This observation is further demonstrated by the quantitative metric in Fig. 2c. Specifically, GLC achieves $79.1 \pm 0.28\%$ overall H-Score. In contrast, the baseline methods DANN, OVANet, OSBP and DANCE obtains $70.1 \pm 0.85\%$, $70.8 \pm 0.58\%$, $72.2 \pm 1.61\%$, $74.5 \pm 0.32\%$ overall H-Score. As presented in Fig. 2d, compared to existing methods, GLC further provides significant savings in target-domain side computational resource overhead (about 48.4% training time re-

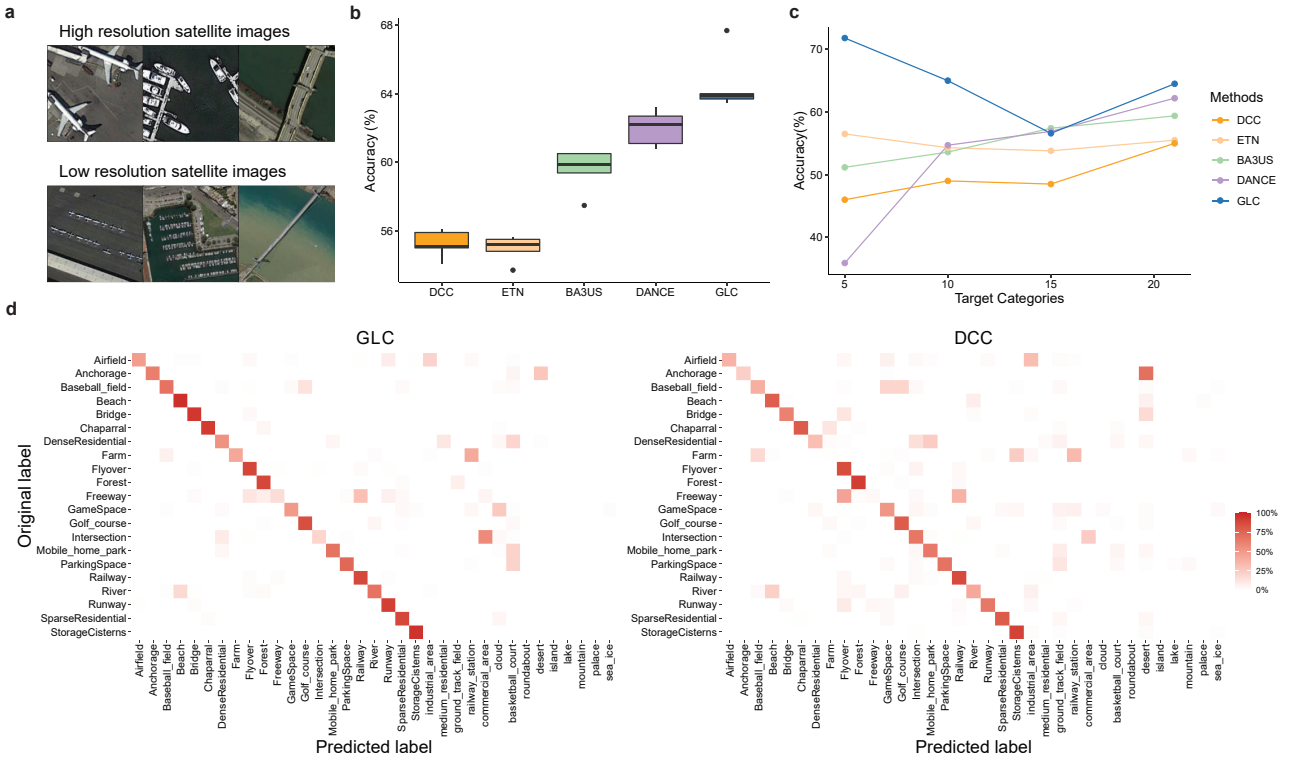

Figure 1. **Analysis of partial-set model adaptation on remote sensing recognition from high-resolution satellite images to low-resolution satellite images.** **a**, Example satellite images of different spatial resolutions on three scene classes, airfield, anchorage, and bridge (from left to right). **b**, Accuracy rate of DCC, ETN, BA3US, DANCE, and GLC. Each boxplot ranges from the upper and lower quartiles with the median as the horizontal line and whiskers extend to 1.5 times the interquartile range. **c**, Accuracy rate of DCC, ETN, BA3US, DANCE, and GLC when applying on different target domain scenarios. **d**, Confusion matrix visualization of GLC and DCC. Clearer diagonal structure indicates better overall accuracy.

duction in this case). This is due to the fact that our GLC merely fine-tunes the source model to realize adaptation, while existing source data-dependent methods need to train the target models from scratch.

To visually assess the separation between the known and unknown classes, we present the uncertainty density distribution in the Fig. 2e. The higher the uncertainty, the more the model treated the input animal image as an unknown species. The results show that while OVANet and DANCE are able to achieve promising classification of known classes of animals, they have troubles in unknown animal separation. In contrast, our GLC draws a better trade-off between known animal classification and unknown animal identification.

We further examine the robustness of GLC in different open-set situations, e.g., varying target domain unknown categories and source domain known categories. As illustrated in Fig. 2f, we can find that GLC maintains a promising H-score compared to existing methods. Specifically, GLC achieves 78.3% overall H-Score in four different target domain unknown categories situations, while DANN, OSBP, OVANet, and DANCE obtains 69.3%, 72.9%, 70.8%, 75.4% average H-Score, respectively. Similarly, when source domain known categories varies, GLC arrives 83.4% overall H-Score in four different situations, still significantly outperforming existing methods.

3.3. Open-partial-set Model Adaptation on Single-cell Identification

We finally consider the most challenging scenario, open-partial-set adaptation, where both source and target domains contain private categories. Here, we implement experiments on single-cell identification. It has great potential in the studies of cell heterogeneity, developmental dynamics, and cell communications [21]. Currently, there are two main types of single-cell sequencing technologies, namely scRNA-seq and scATAC-seq. However, it has been noted that the extreme scarcity of scATAC-seq data tends to limit its ability for cell type identification. In contrast, large amounts of well-annotated scRNA-seq datasets have been curated as cell atlases. It motivates us to upcycle models trained on the scRNA-seq datasets and adapt them to the scATAC-seq datasets. Nevertheless, the cell types contained in different atlas data are generally inconsistent,

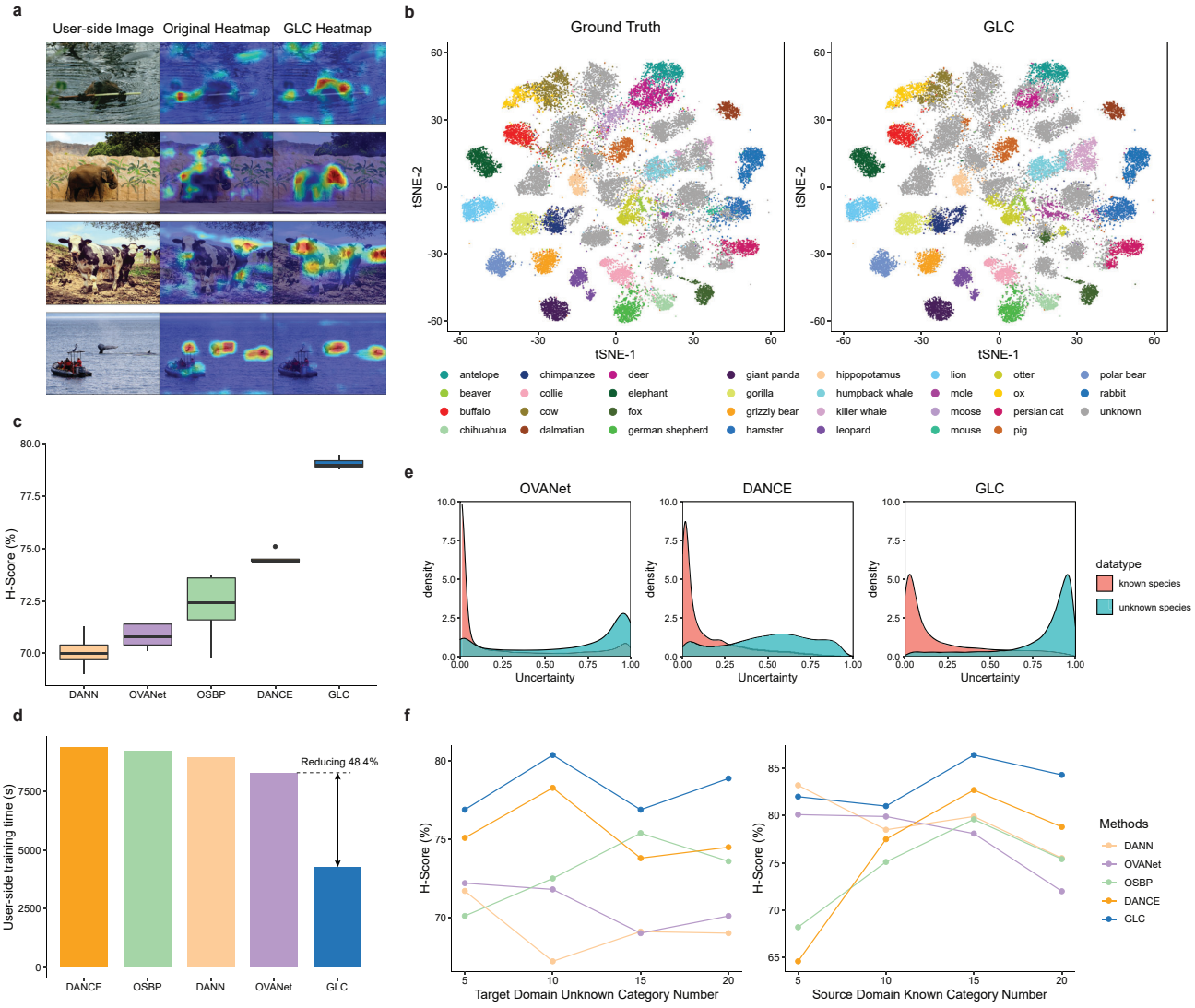

Figure 2. **Analysis of open-set model adaptation on wild animal classification from virtual images to real-world images. a**. Example wild animal images of the I2AWA dataset, and comparison of the feature heatmap images between the pre-trained source model and the adapted target model. **b**. tSNE visualization of ground truth labels and GLC predicted labels. Each tSNE subplot is generated from the same extracted features by our GLC. We apply gray to denote those animal species that are not presented in the source domain. **c** H-Score rate of DANN, OVANet, OSBP, DANCE, and GLC. Each boxplot ranges from the upper and lower quartiles with the median as the horizontal line and whiskers extend to 1.5 times the interquartile range. **d**, Target domain training time of DANN, OVANet, OSBP, DANCE, and GLC. **e**, Uncertainty distribution of OVANet, DANCE, and our GLC for known and unknown animal categories. **f**, H-Score rate of DANN, OVANet, OSBP, DANCE, and GLC when the number of target domain unknown animal categories and source domain known animal categories varying.

which poses substantial challenges for model adaptation across atlases. In this article, we apply our GLC to two mouse cell atlases, the Tabula Muris atlas [18] for scRNA-seq data and the Cusanovich atlas [2] for scATAC-seq data. The Tabula Muris atlas consists of 73 cell types totaling 96,404 cells from 20 organs with two protocols profiling transcriptomics, while the Cusanovich atlas consists of 29 cell types totaling 81,173 cells from 13 tissues. There are 19 cell types common between the Tabula Muris atlas and the Cusanovich atlas. For performance evaluation, as in the wild animal experiments above, we utilize the harmonic mean accuracy H-Score of the known cell types and the unknown cell types as the quantitative metric. We compare our GLC with recently developed and applied methods for scRNA-seq and scATAC-seq integration, including the scJoint [9] and the Seurat v.3 [20].

We illustrate the tSNE plots in Fig. 3a to compare with the ground truth labels annotated in the Cusanovich atlas [2]. The tSNE plots are generated by applying the singular value of the term frequency-inverse document frequency

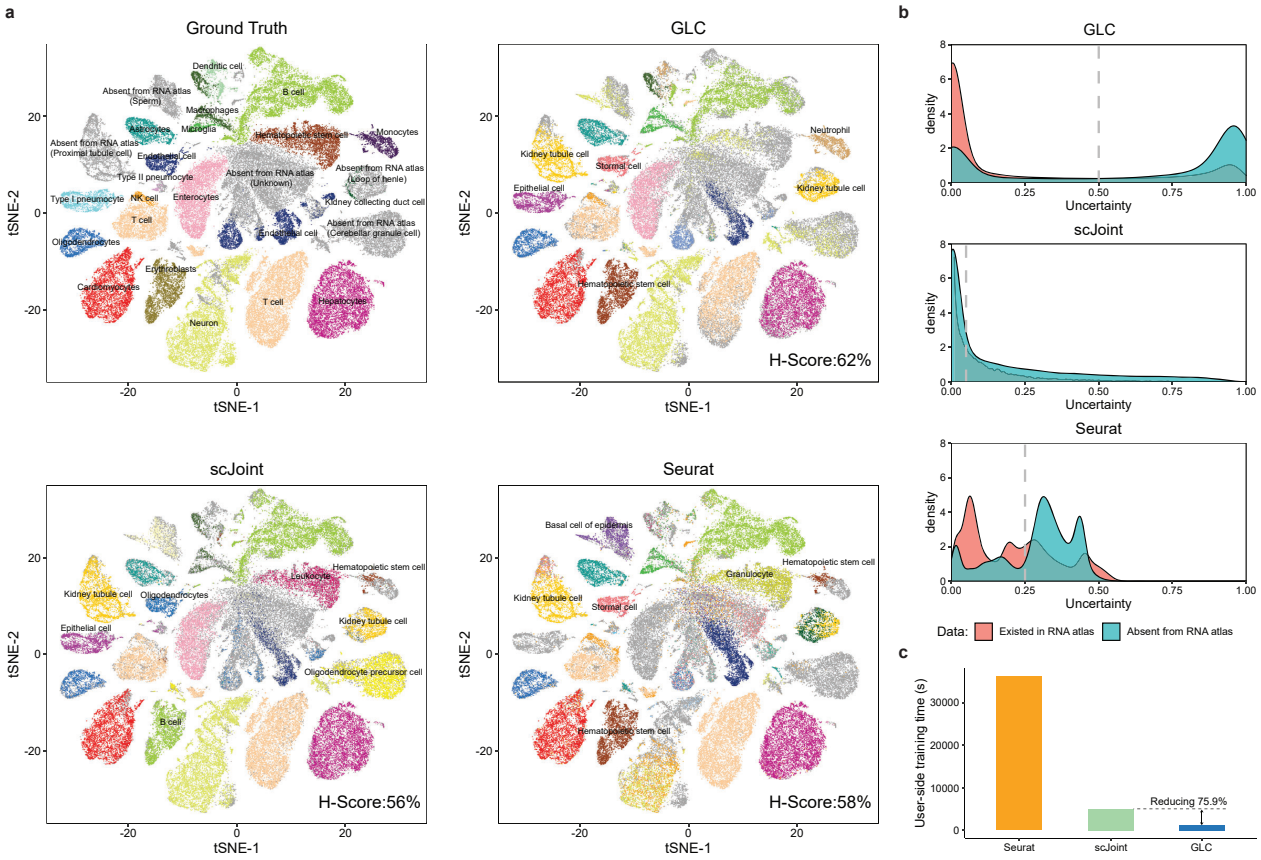

Figure 3. **Analysis of open-partial model adaptation on single-cell identification from scRNA-seq atlas to scATAC-seq atlas. a**, tSNE visualization of ground truth, GLC, scJoint, and Seurat predicted labels. We apply gray to indicate those cell types that are not presented in the source domain scRNA-seq atlas. **b**, Uncertainty distribution of GLC, scJoint, and Seurat for cell types that are existed and absent from the scRNA-seq atlas. The higher the uncertainty, the more the model tends to assign cell types to unknown that are absent from the scRNA-seq atlas. The dashed line denotes the decision boundaries for known and unknown cell types prediction. **c**, Target-domain training time of Seurat, scJoint and our GLC.

(TD-IDF) transformation of scATAC-seq peak matrix as in the Cusanovich atlas [2]. It observes that GLC achieves a better trade-off between known cell types identification and unknown cell types separation than the other methods. This observation is further quantitatively demonstrated by the H-Score metric. Specifically, GLC obtains 62% overall H-Score compared with 58% for Seurat and 56% for scJoint. As presented in Fig. 3c, not only is there a significant performance improvement, but our GLC also brings significant savings in target domain computational resource overhead (about 75.9% training time reduction).

In addition to tSNE plots, we also present the uncertainty density distribution in Fig. 3b, where the higher the uncertainty, the more the model tends to group the cell into the unknown cell types group. To find the best trade-off point, a global decision boundary search was performed for all methods. The decision boundary for GLC is 0.5 compared with 0.25 for Seurat and 0.05 for scJoint. It further indicates that our GLC attains an optimal trade-off in cell types identification to the other methods.

4. Discussion

During the past decades, deep neural networks (DNNs) have achieved remarkable success in various applications and fields. However, DNNs are typically restricted to the training data domain. If the test data is collected in another modality or from other types of instruments, we will typically suffer from a significant performance degradation [13]. This phenomenon is likely to worsen when training and testing data do not share the same ground-truth class space. Although DNNs can be adapted to different application scenarios with additional supervised learning, this paradigm asks for annotation of large-scale target domain data. It would require significant resources and experts in real-world applications, such as clinical staff for medical imaging diagnosis and genetic scientists for single-cell sequence analysis, making it extremely expensive and impossible for most scenarios. In this paper, we find that it is

possible to productively upcycle existing pre-trained models and adapt them to new scenarios. Numerous empirical evidences on standard computer science benchmarks and simulated realistic applications basically demonstrate that GLC is a promising, simple, and general solution for a variety of real-world application tasks, including single-cell sequence analysis, remote sensing recognition, and other such domain-dependent problems.

References

[1] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017. 1

[2] Darren A Cusanovich, Andrew J Hill, Delasa Aghamirzaie, Riza M Daza, Hannah A Pliner, Joel B Berletch, Galina N Filippova, Xingfan Huang, Lena Christiansen, William S DeWitt, et al. A single-cell atlas of in vivo mammalian chromatin accessibility. *Cell*, 174(5):1309–1324, 2018. 4, 5

[3] Bo Fu, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. Learning to detect open classes for universal domain adaptation. In *ECCV*, 2020. 2

[4] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 2016. 2

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[6] Guangrui Li, Guoliang Kang, Yi Zhu, Yunchao Wei, and Yi Yang. Domain consensus clustering for universal domain adaptation. In *CVPR*, 2021. 2

[7] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *ICML*, 2020. 1

[8] Jian Liang, Dapeng Hu, Jiashi Feng, and Ran He. Umad: Universal model adaptation under domain and category shift. *arXiv preprint arXiv:2112.08553*, 2021. 1, 2

[9] Yingxin Lin, Tung-Yu Wu, Sheng Wan, Jean YH Yang, Wing H Wong, and YX Wang. scjoint integrates atlas-scale single-cell rna-seq and atac-seq data with transfer learning. *Nature Biotechnology*, pages 1–8, 2022. 4

[10] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *NeurIPS*, 2018. 2

[11] Zhongqi Miao, Ziwei Liu, Kaitlyn M Gaynor, Meredith S Palmer, Stella X Yu, and Wayne M Getz. Iterative human and automated identification of wildlife images. *Nature Machine Intelligence*, 3(10):885–895, 2021. 2

[12] Mohammad Sadegh Norouzzadeh, Anh Nguyen, Margaret Kosmala, Alexandra Swanson, Meredith S Palmer, Craig Packer, and Jeff Clune. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences*, 115(25):E5716–E5725, 2018. 2

[13] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE TKDE*, 2009. 5

[14] Sanqing Qu, Guang Chen, Jing Zhang, Zhijun Li, Wei He, and Dacheng Tao. Bmd: A general class-balanced multicentric dynamic prototype strategy for source-free domain adaptation. In *ECCV*. Springer, 2022. 1

[15] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, and Kate Saenko. Universal domain adaptation through self supervision. In *NeurIPS*, 2020. 2

[16] Kuniaki Saito and Kate Saenko. Ovanet: One-vs-all network for universal domain adaptation. In *ICCV*, 2021. 2

[17] Kuniaki Saito, Shohei Yamamoto, Yoshitaka Ushiku, and Tatsuya Harada. Open set domain adaptation by backpropagation. In *ECCV*, 2018. 2

[18] Nicholas Schaum, Jim Karkanias, Norma F Neff, Andrew P May, Stephen R Quake, Tony Wyss-Coray, Spyros Darmanis, Joshua Batson, Olga Botvinnik, Michelle B Chen, et al. Single-cell transcriptomics of 20 mouse organs creates a tabula muris: The tabula muris consortium. *Nature*, 562(7727):367, 2018. 4

[19] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017. 2

[20] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck III, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902, 2019. 4

[21] Tim Stuart and Rahul Satija. Integrative single-cell analysis. *Nature reviews genetics*, 20(5):257–272, 2019. 3

[22] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. In *CVPR*, 2017. 2

[23] Kaichao You, Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Universal domain adaptation. In *CVPR*, 2019. 2

[24] Bing Zhang, Yuanfeng Wu, Boya Zhao, Jocelyn Chanussot, Danfeng Hong, Jing Yao, and Lianru Gao. Progress and challenges in intelligent remote sensing satellite systems. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2022. 1

[25] Lianchong Zhang, Guoqing Li, Chi Zhang, Huanyin Yue, and Xiaohan Liao. Approach and practice: integrating earth observation resources for data sharing in china geoss. *International Journal of Digital Earth*, 12(12):1441–1456, 2019. 1

[26] Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging theory and algorithm for domain adaptation. In *ICML*, 2019. 1, 2

[27] Weixun Zhou, Shawn Newsam, Congmin Li, and Zhenfeng Shao. Patternnet: A benchmark dataset for performance evaluation of remote sensing image retrieval. *ISPRS journal of photogrammetry and remote sensing*, 145:197–209, 2018. 1

[28] Junbao Zhuo, Shuhui Wang, Shuhao Cui, and Qingming Huang. Unsupervised open domain recognition by semantic discrepancy minimization. In *CVPR*, 2019. 2