GeoNet: Benchmarking Unsupervised Adaptation across Geographies

Supplementary Material

Tarun Kalluri   Wangdong Xu   Manmohan Chandraker
UC San Diego
https://tarun005.github.io/GeoNet

		GeoPlaces	
Train ↓ / Test →	USA	Asia	Africa
USA	56.35/85.15	36.27/63.27	32.20/51.97
Asia	21.03/44.81	49.63/78.45	26.77/47.90

Table S1. **Cross-Geography Drops on GeoPlaces** Top-1/Top-5 accuracies of Resnet-50 models across geographically different train and test domains, including a new test-set from Africa domain.

S1. Performance on additional geographies

In Table 2 in the main paper, we illustrated cross-domain drops across geographies for the case of USA↔Asia. We show that this phenomenon is not specific to these geographies, and similar cross-domain drop in accuracy can be observed in case of Africa as a new geographical domain. For this purpose, we follow a similar pipeline discussed in Section 3.1 of the main paper and collect images from Africa belonging to the 205 classes from Places-205, creating the test-set for Africa domain for GeoPlaces with 8358 images. For the case of GeoPlaces, we show in Tab. S1 that a model trained on USA obtains only 32.2% on test images from Africa with a significant drop of 24%, and a model trained on images from Asia only gets 26.77% top-1 accuracy on Africa test images with a drop of 23% compared to within-domain test accuracy. These results indicate that cross-domain transfer exhibits similar challenges across any geographically separated domains.

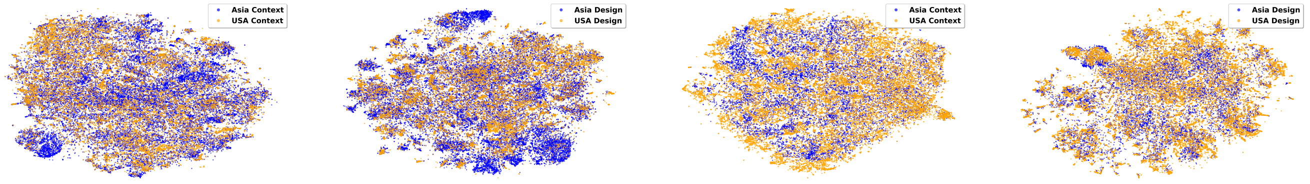
S2. Visualization of Context and Design Shifts

We provide deeper insight into the cross-domain shifts in contexts and designs induced by the geographies by visualizing their tSNE feature representations [4]. To this end, we first recall that we defined context of an image x as b_x representing the background regions in an image, and design f_x as the foreground objects (Section 3.4 in the main paper). However, we do not have box or mask annotation corresponding to the images in GeoNet, so it is not possible to directly infer the context and foreground in each image. Instead, we rely on a state-of-the-art object detector Mask-RCNN trained on COCO dataset [1] for this purpose. Specifically, we train a class-agnostic Mask-RCNN on the COCO dataset by mapping all the class labels to a single foreground class. We then identify all the masks detected by the network on our images, so that these masks then correspond to the foreground objects, while the other parts of the image corresponds to the background. To compute the feature representation of the foreground objects, we element-wise multiply the binary foreground mask with the deep feature map from the backbone Resnet-50, followed by a global pool. In other words, we use the binary foreground mask to select the area from the feature map corresponding to the foreground, and take an average of the locations to obtain a 2048-dimensional foreground feature vector per image. We similarly obtain a 2048-dimensional background vector by using the negation of the binary foreground mask as the background mask. Therefore, we end up with two feature representations per image pertaining to the foreground (design) and background (context) respectively. We repeat this for both domains USA and Asia from both the GeoPlaces and GeoImNet splits of our dataset. We then project this 2048 dimensional vector into a 2-dimensional vector using tSNE reduction and visualize the embeddings in Fig. S1.

Context Shift The pronounced distinction in the contexts between the two domains from GeoPlaces is highlighted in Fig. S1a, where we show minimum overlap between the features corresponding to the background regions in USA and Asia. Similar observations also hold for the case of GeoImNet in Fig. S1c. Since the background or the context plays a major role in identifying places or objects, this shift invariably results in drop in accuracy under cross-geography transfer.

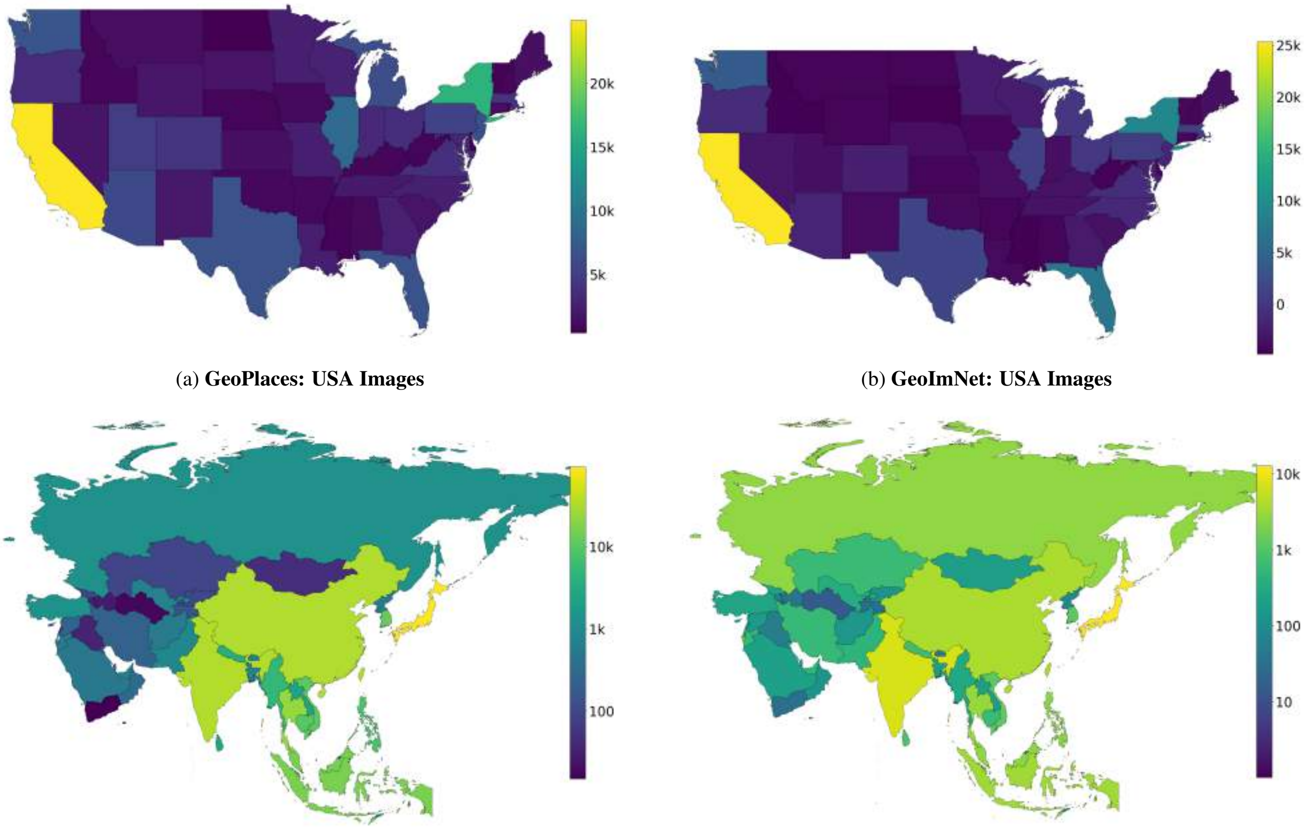
Design Shift The tSNE features of the foreground regions is shown in Fig. S1b for the case of GeoPlaces and in Fig. S1d for GeoImNet. Minimum overlap is observed between the features corresponding to the foreground, or design of the objects, in each case indicating the presence of notable design shift between the domains.

We also note that datasets like COCO are predominantly US-biased, so the use of COCO in analyzing distribution shifts on Asia images is not completely fair. To this end, manually annotating images with finer-grained foreground



(a) **Context Shift in GeoPlaces** (b) **Design Shift in GeoPlaces** (c) **Context Shift in GeoImNet** (d) **Design Shift in GeoImNet**

Figure S1. **tSNE Visualizations of context and design shifts in GeoNet**. As shown, there is a notable separation between the context and design features between USA (in orange) and Asia (in blue) in both GeoPlaces and GeoImNet.



(a) **GeoPlaces: USA Images** (b) **GeoImNet: USA Images**

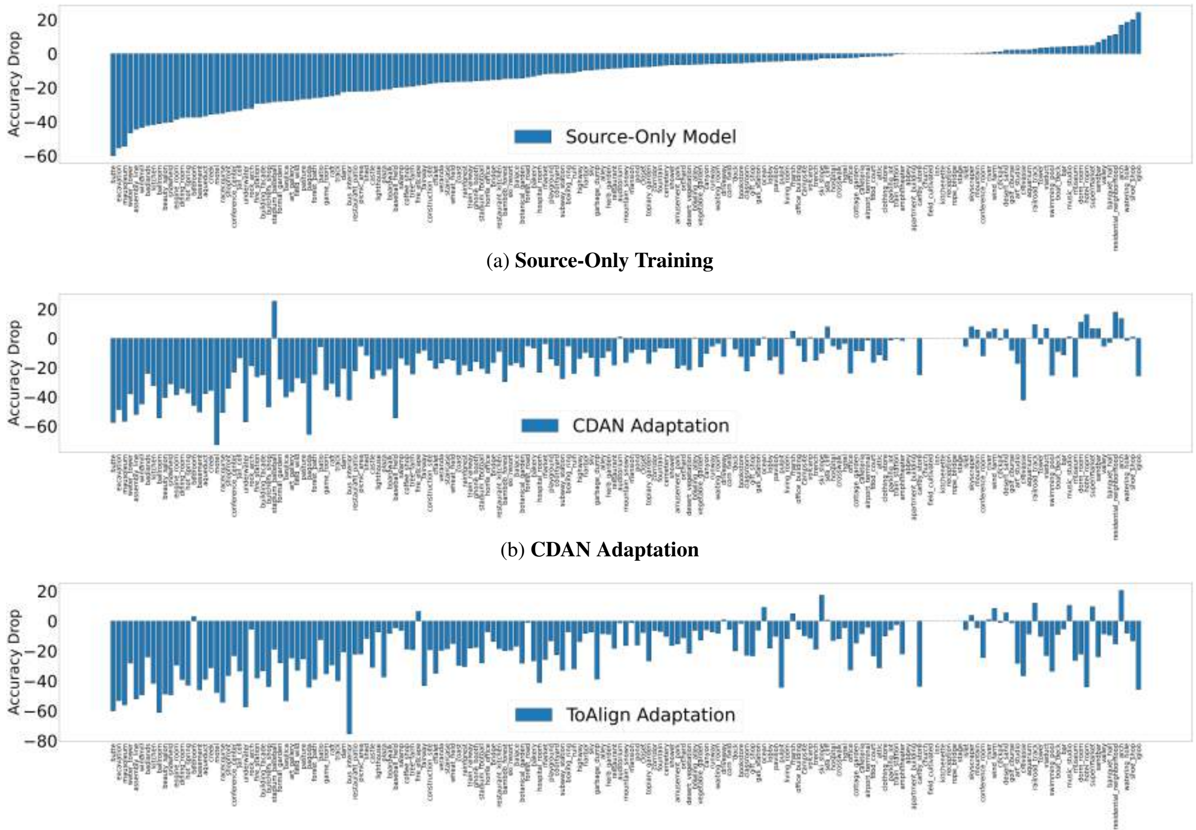
(c) **GeoPlaces: Asia Images** (d) **GeoImNet: Asia Images**

Figure S2. **Geographical Distribution of images from USA and Asia domains**. We show the images per geographical sub-region in both domains on GeoNet. As shown, in Asia, a majority of images are from Japan, India, Korea, China and Taiwan while in USA, a majority of images are from populous regions like California and New York. Note that the color-bar scale is linear for USA and log-scale for Asia.

and context labels in both geographies would yield more accurate analysis, which is left as a future work.

S3. Geographic Distribution of Images

While we broadly categorize Asia and USA to be the two major geographical domains, not all sub-regions in these geographies have equal representation. We show the geographic distribution over respective geographies in Fig. S2, by leveraging the per-image GPS metadata provided in GeoNet. For images from Asia from Fig. S2c for GeoPlaces and Fig. S2d for GeoImNet, we observe a large fraction of images from Japan, India, Korea, China and Taiwan, while some countries are more sparsely represented. Likewise, in USA in Fig. S2a and Fig. S2b, we observe a significant share of images from California, New York and Florida than other regions. These distributions reflect the larger user demographic biases in photo-sharing websites like Flickr from

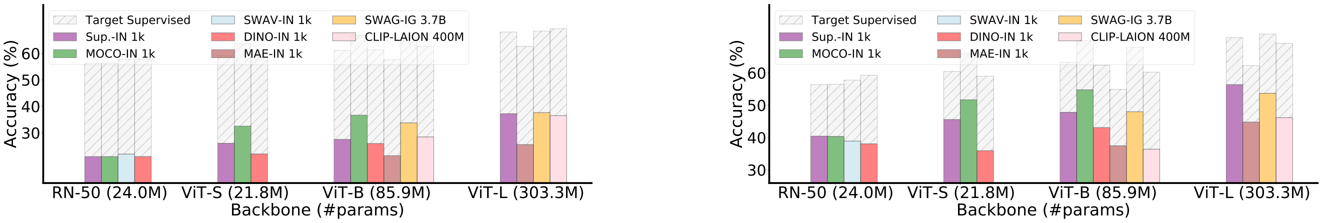


(a) **Source-Only Training**

(b) **CDAN Adaptation**

(c) **ToAlign Adaptation**

Figure S3. **Per-class accuracy drops** on USA→Asia transfer for a plain source-only model as well as post-adaptation using CDAN [2] and ToAlign [5] adaptation methods. Note that the trend of per-class accuracy drops is the same before and after the adaptation indicating the limited benefit offered by existing state-of-the-art adaptation methods in bridging geographical shifts.



(a) **Asia→USA on GeoPlaces**

(b) **Asia→USA on GeoImNet**

Figure S4. **Large-Scale pre-training on GeoNet** We show that most architectures and pre-training strategies exhibit significant cross-domain drops when fine-tuned on geographically biased datasets. Shown for Asia→USA on GeoPlaces in Fig. S4a and GeoImNet in Fig. S4b, refer main paper for other transfer settings.

where all our images have been taken from.

S4. Error Analysis of Unsupervised Adaptation

While we show in the main paper (Table 3) that existing unsupervised adaptation approaches yield limited benefit for geographical adaptation, we conduct a deeper analysis into the per-class accuracy post-adaptation in Fig. S3 for the case of USA→Asia on GeoPlaces. Specifically, we first take a model trained only on USA images, and compute the drop in per-class accuracy suffered by direct cross-domain transfer on Asia test images. We show this in Fig. S3a, where classes like *mausoleum*, *assembly line* and *kitchen* suffer the largest drops in accuracy. Next, we carry the same analysis using a model trained with CDAN [2] adaptation method. From Fig. S3b, we observe that the trends in per-class accuracy drops are mostly similar with or without using CDAN adaptation, indicating that the benefit achieved using an adaptation method is negligible on all the categories. Similar observations also hold for the case of adaptation using ToAlign [5], underlining the fact that existing state-of-the-art adaptation methods cannot handle geographic shifts across most categories.

S5. Data De-duplication

Since a lot of users tend to upload multiple pictures of the same scene on sites like Flickr, we carry a data de-duplication exercise so that there are no such duplicate copies of same images in train and test sets which would unfairly improve within-domain accuracy. We first group all the images in the train and test sets which belong to the same geographical location, by discretizing the GPS coordinates within one degree. Then, within each group, we first resize the images to 32x32x3, and compute a histogram of the images along the RGB channels. We also flatten the image and compute the euclidean distance between all pairs of images within the same group and remove all images from the training set which are "similar" to images in test set, where two images are similar if they belong to the same GPS group, and have RGB histogram, euclidean distance lower than preset thresholds.

S6. Large-scale pre-training on GeoNet

In Fig. S4, we show the effect of large-scale pretraining on the transfer setting Asia→USA from GeoPlaces(Fig. S4a) and GeoImNet(Fig. S4b). We make similar observations as the transfer setting from USA→Asia in the main paper. Specifically, we show that transformers outperform Resnets, pre-training using billion-scale datasets like SWAG [3] outperforms ImageNet-pretraining and all models still have significant gap with the target supervised accuracy indicating the limitations of these models in bridging cross-geography domain shifts.

S7. Sample Images

We show few sample images from selected classes across both USA and Asia domains in GeoPlaces benchmark in Fig. S5, Fig. S6 and GeoImNet benchmark in Fig. S7, Fig. S8.

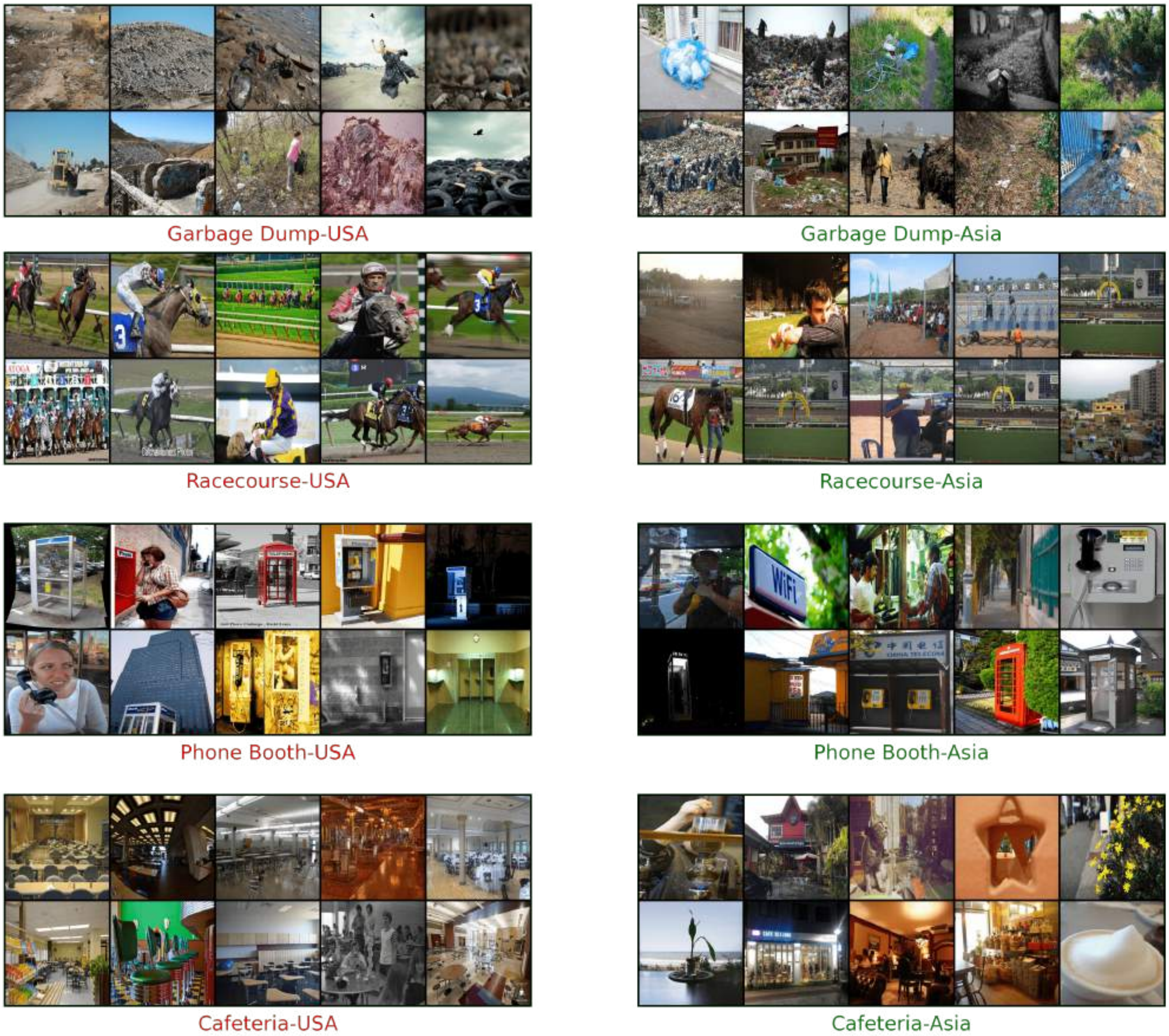


Figure S5. Sample images showing the domain gap between USA (left) and Asia (right) domains for classes garbage dump, race course, phone booth and cafetaria from GeoPlaces.

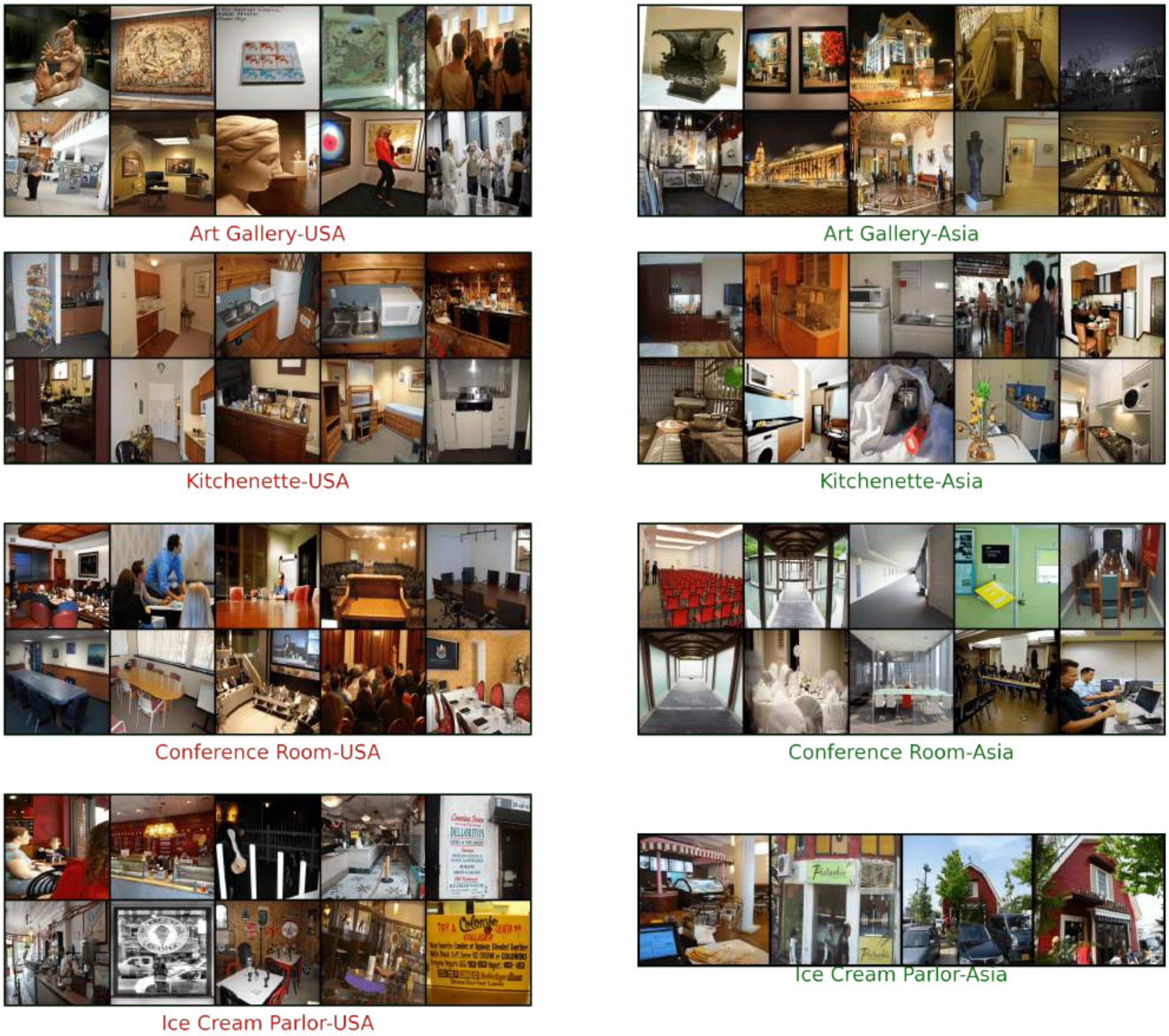


Figure S6. Sample images showing the domain gap between USA (left) and Asia (right) domains for classes `art gallery`, `kitchenette`, `conference room` and `ice-cream parlor` from GeoPlaces.

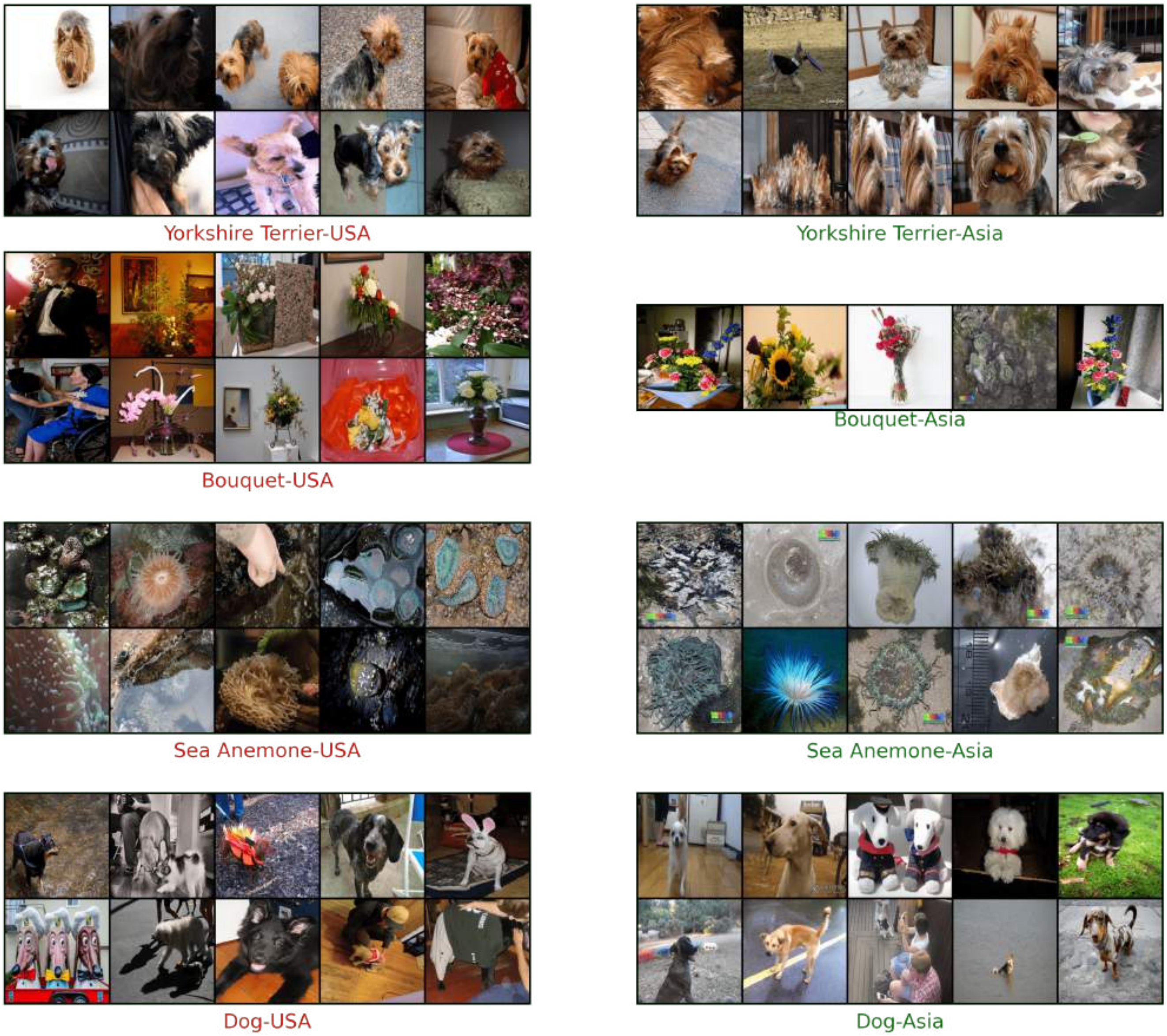


Figure S7. Sample images showing the domain gap between USA (left) and Asia (right) domains for classes `Yorkshire Terrier`, `bouquet`, `sea anemone` and `dog` from GeoImNet.

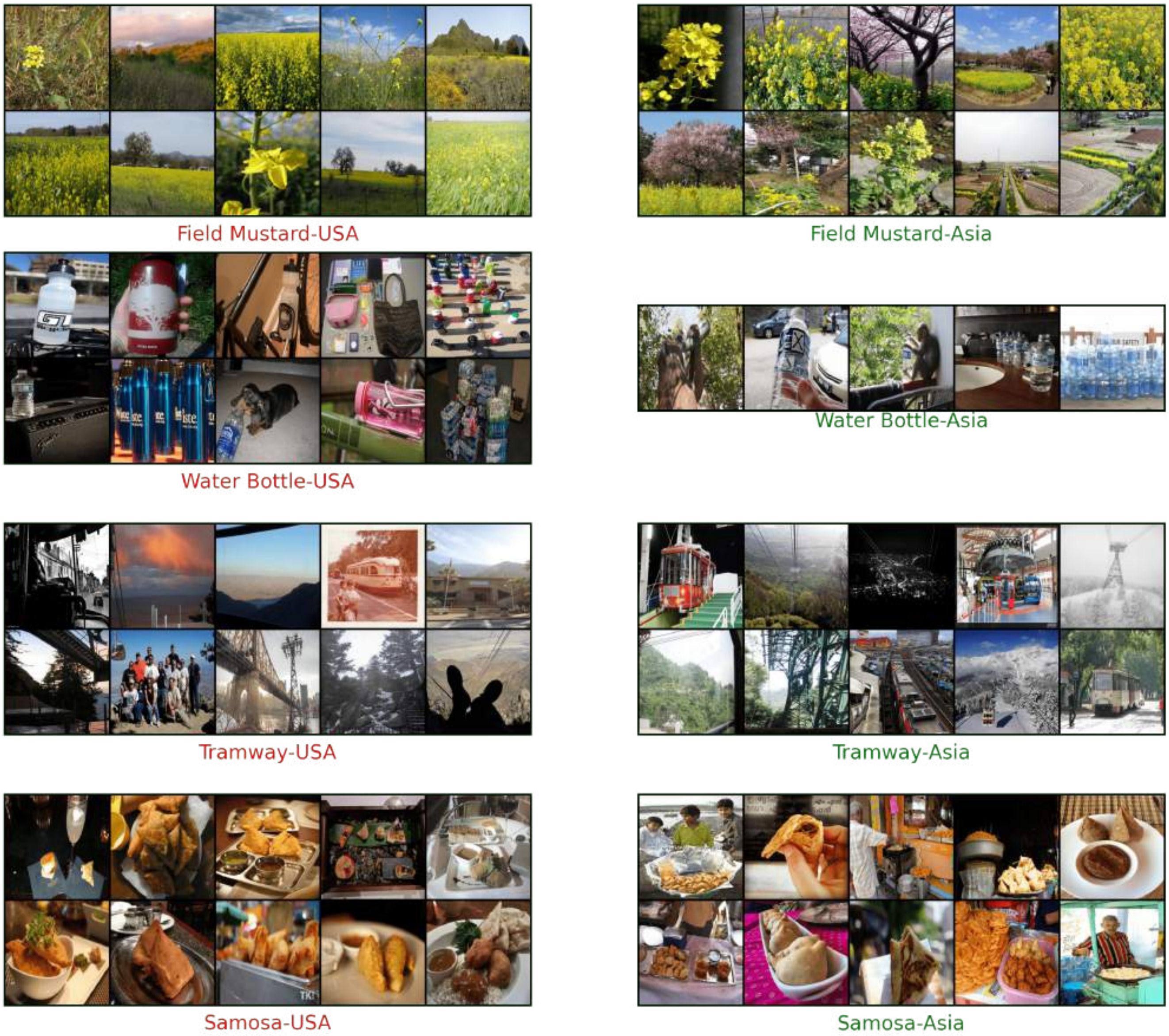


Figure S8. Sample images showing the domain gap between USA (left) and Asia (right) domains for classes Field Mustard, Water Bottle, Tramway and Samosa from GeoImNet.

References

[1] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. S1

[2] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pages 1640–1650, 2018. S3

[3] Mannat Singh, Laura Gustafson, Aaron Adcock, Vinicius de Freitas Reis, Bugra Gedik, Raj Prateek Kosaraju, Dhruv Mahajan, Ross Girshick, Piotr Dollár, and Laurens van der Maaten. Revisiting Weakly Supervised Pre-Training of Visual Perception Models. In *CVPR*, 2022. S4

[4] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. S1

[5] Guoqiang Wei, Cuiling Lan, Wenjun Zeng, Zhizheng Zhang, and Zhibo Chen. Toalign: Task-oriented alignment for unsupervised domain adaptation. In *NeurIPS*, 2021. S3, S4