GeoNet: Benchmarking Unsupervised Adaptation across Geographies

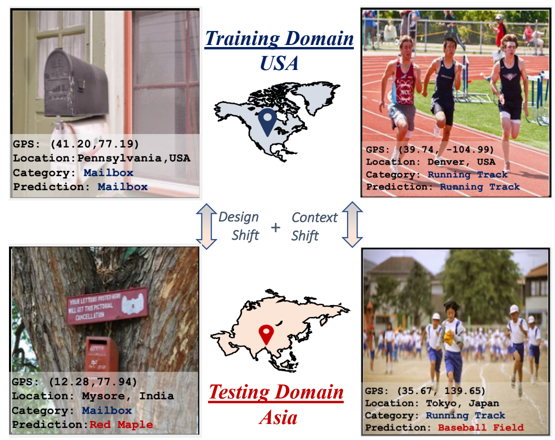
Tarun Kalluri Wangdong Xu Manmohan Chandraker
UC San Diego
{sskallur,w6xu,mkchandraker}@eng.ucsd.edu
https://tarun005.github.io/GeoNet

Abstract

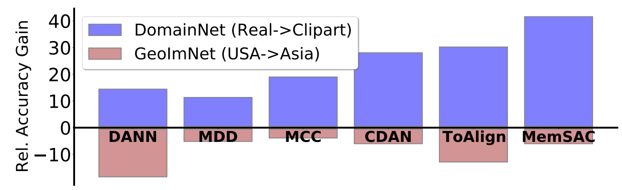
In recent years, several efforts have been aimed at improving the robustness of vision models to domains and environments unseen during training. An important practical problem pertains to models deployed in a new geography that is under-represented in the training dataset, posing a direct challenge to fair and inclusive computer vision. In this paper, we study the problem of geographic robustness and make three main contributions. First, we introduce a large-scale dataset ***GeoNet*** *for geographic adaptation containing benchmarks across diverse tasks like scene recognition (GeoPlaces), image classification (GeoImNet) and universal adaptation (GeoUniDA). Second, we investigate the nature of distribution shifts typical to the problem of geographic adaptation and hypothesize that the major source of domain shifts arise from significant variations in scene context (context shift), object design (design shift) and label distribution (prior shift) across geographies. Third, we conduct an extensive evaluation of several state-of-the-art unsupervised domain adaptation algorithms and architectures on GeoNet, showing that they do not suffice for geographical adaptation, and that large-scale pre-training using large vision models also does not lead to geographic robustness. Our dataset is publicly available at https://tarun005.github.io/GeoNet.*

1. Introduction

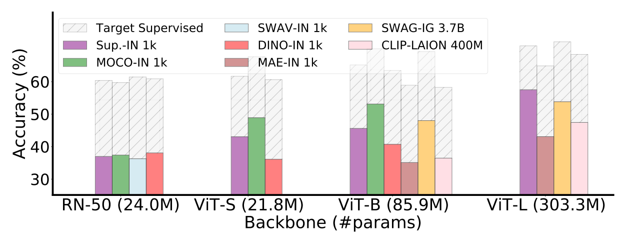
In recent years, domain adaptation has emerged as an effective technique to alleviate dataset bias [80] during training and improve transferability of vision models to sparsely labeled target domains [27, 36, 40, 42, 49–51, 68, 69, 87, 90]. While being greatly instrumental in driving research forward, methods and benchmark datasets developed for domain adaptation [56, 57, 64, 84] have been restricted to a narrow set of divergences between domains. However, the geographic origin of data remains a significant source of bias, attributable to several factors of variation between train and test data. Training on geographically biased datasets may cause a model



(a) **Geographic bias manifested in proposed GeoNet dataset**



(b) **Unsupervised domain adaptation does not suffice on GeoNet**



(c) **Large vision models exhibit cross-domain drops on GeoNet**

Figure 1. **Summary of our contributions**. (a): Training computer vision models on geographically biased datasets suffers from poor generalization to new geographies. We propose a new dataset called GeoNet to study this problem and take a closer look at the various types of domain shifts induced by geographic variations. (b) Prior unsupervised adaptation methods that efficiently handle other variations do not suffice for improving geographic transfer. (c) We highlight the limitations of modern convolutional and transformer architectures in addressing geographic bias, exemplified here by USA→Asia transfer on GeoImNet.

to learn the idiosyncrasies of their geographies, preventing generalization to novel domains with significantly different geographic and demographic composition. Besides robustness, this may have deep impact towards fair and inclusive computer vision, as most modern benchmark datasets like ImageNet [63] and COCO [47] suffer from a significant US or UK-centric bias in data [24, 73], with poor representation of images from various other geographies like Asia.

In this paper, we study the problem of geographic adaptation by introducing a new large-scale dataset called GeoNet, which constitutes three benchmarks – GeoPlaces for scene classification, GeoImNet for object recognition and GeoUniDA for universal domain adaptation. These benchmarks contain images from USA and Asia, which are two distinct geographical domains separated by various cultural, economic, demographic and climatic factors. We additionally provide rich metadata associated with each image, such as GPS location, captions and hashtags, to facilitate algorithms that leverage multimodal supervision.

GeoNet captures the multitude of novel challenges posed by varying image and label distributions across geographies. We analyze GeoNet through new sources of domain shift caused by geographic disparity, namely (i) *context shift*, where the appearance and composition of the background in images changes significantly across geographies, (ii) *design shift*, where the design and make of various objects changes across geographies, and (iii) *prior shift*, caused by different per-category distributions of images in both domains. We illustrate examples of performance drop caused by these factors in Fig. 1a, where models trained on images from USA fail to classify common categories such as *running track* and *mailbox* due to context and design shifts, respectively.

GeoNet is an order of magnitude larger than previous datasets for geographic adaptation [58, 61], allowing the training of modern deep domain adaptation methods. Importantly, it allows comparative analysis of new challenges posed by geographic shifts for algorithms developed on other popular adaptation benchmarks [56, 57, 64, 84]. Specifically, we evaluate the performance of several state-of-the-art unsupervised domain adaptation algorithms on GeoNet, and show their limitations in bridging domain gaps caused by geographic disparities. As illustrated in Fig. 1b for the case of DomainNet [56] vs. GeoNet, state-of-the-art models on DomainNet often lead to accuracies even worse than a source only baseline on GeoNet, resulting in negative *relative gain* in accuracy (defined as the gain obtained by an adaptation method over a source-only model as a percentage of gap between a source-only model and the target-supervised upper bound). Furthermore, we also conduct a study of modern architectures like vision transformers and various pre-training strategies, to conclude that larger models with supervised and self-supervised pre-training offer improvements in accuracy, which however are not sufficient to address the domain gap (Fig. 1c). This highlights that the new challenges introduced by geographic bias such as context and design shift are relatively under-explored, where our dataset may motivate further research towards this important problem.

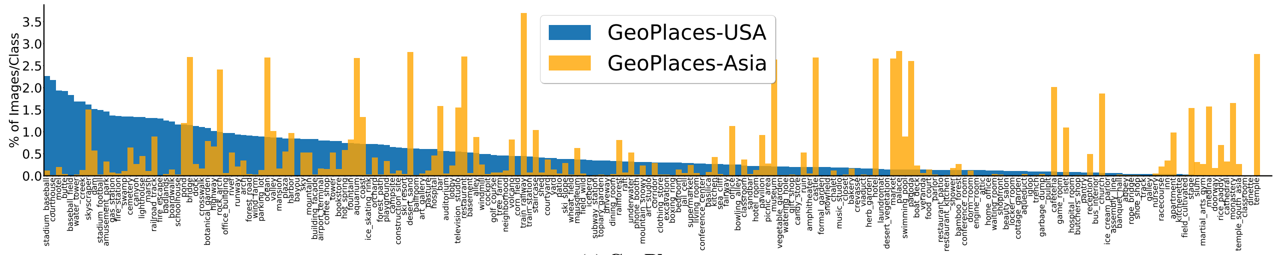
In summary, our contribution towards geographic domain adaptation is four-fold:

- A new large-scale dataset, GeoNet, with benchmarks for diverse tasks like scene classification and object recognition, with labeled images collected from geographically distant locations across hundreds of categories (Sec. 3).
- Analysis of domain shifts in geographic adaptation, which may be more complex and subtle than style or appearance variations (Sec. 3.4).
- Extensive benchmarking of unsupervised adaptation algorithms, highlighting their limitations in addressing geographic shifts (Sec. 4.2).
- Demonstration that large-scale pretraining and recent advances like vision transformers do not alleviate these geographic disparities (Sec. 4.3).

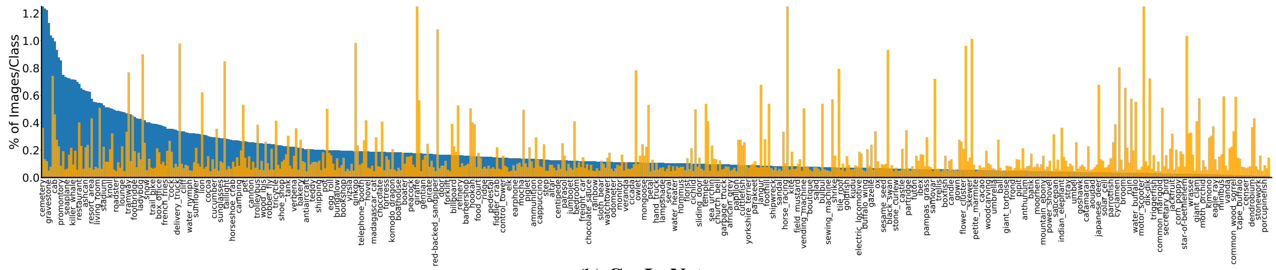
2. Related Works

Domain Adaptation Unsupervised domain adaptation enables training models on a labeled source domain along with unlabeled samples from a different target domain to improve the target domain accuracy. A large body of prior works aim to minimize some notion of divergence [4, 5] between the source and target distributions based on MMD [49, 51, 77, 78] adversarial [9, 13, 27, 50, 68, 81, 82, 93], generative [8, 36, 70], class-level [31, 44, 52, 55, 69, 89] or instance-level alignment [74, 85, 87] techniques. Clustering [23, 39, 41, 42, 54] and memory-augmentation approaches [40] have also been shown to be effective. However, most of these works are shown to improve performance using standard datasets such as Office-31 [64], visDA [57], OfficeHome [84] or DomainNet [56], where the distribution shifts typically arise from unimodal variations in style or appearance between source and target. While prior works also study semantic shift [6] and sub-population shift [10], we aim to address a more practical problem of geographic domain adaptation with more complex variations not covered by prior works.

Geographic Robustness Many prior works study biases of CNNs towards 3D poses [1, 95], textures [29], styles [35], natural variations [7, 60, 79] and adversarial inputs [35], but robustness of computer vision towards shift induced by geography is relatively under-explored. While algorithms for bridging geographic domain gaps have been proposed in [18, 41, 86], they are restricted to road scenes with limited number of classes. A major hindrance has been the lack of suitable benchmark datasets for geographic adaptation, so several datasets have been recently proposed to address this issue [24, 58, 61, 72]. Datasets based on dollar street images [61] highlight the geographic differences induced by income disparities between various countries, Ego4D [30] contains egocentric videos with actions from various geogra-



(a) **GeoPlaces**



(b) **GeoImNet**

Figure 2. **Class distribution in GeoNet** Percentage of images per class from USA and Asia domains shown for the GeoPlaces benchmark in (a) and GeoImNet benchmark in (b). The label distributions are long-tailed in both, and the dominant and tail classes are widely different across geographies in each setting indicating a strong prior shift. (Best viewed in color, zoom in to see the class names).

	Split	GeoPlaces	GeoImNet	GeoUniDA
USA	Train	178110	154908	100136
	Test	17234	16784	25034
Asia	Train	187426	68722	33912
	Test	26923	9636	8478
classes-shared		205	600	62
classes-private		-	-	138

Table 1. **Summary of GeoNet** Number of images in train and test splits in each of our benchmarks. While GeoPlaces and GeoImNet are developed for unsupervised adaptation, GeoUniDA is developed for universal domain adaptation across geographies.

phies, while researchers in [58] design an adaptation dataset with images from YFCC-100M [26] to analyze geographic shift. Adding to these efforts, we propose a much larger-scale dataset for geographic adaptation consisting of more diverse categories for place and object classification, across factors of variation beyond income disparities.

3. Dataset Creation and Analysis

We present the overall summary of various datasets in our benchmark in Tab. 1, including the number of images and categories from each of our settings. In this paper, we broadly consider US and Asia as the two domains, as these two geographies have considerable separation in terms of underlying cultural, environmental and economical factors, while also providing the appropriate level of abstraction and leaving enough data from each domain to perform meaningful analysis. Although Asia is less homogeneous than USA with greater within-domain variance, our adopted geographical granularity follows from the amount of data we could retrieve from different countries using concepts in GeoNet, where we observed general paucity in images from many low-resource countries on Flickr. We also note that the domain shifts caused by geographic disparities are not restricted to these regions, and use images from Africa to show similar observations of domain gaps in the supplementary.

3.1. GeoPlaces

We propose GeoPlaces to study geographic adaptation in scene classification, which involves predicting the semantic category of the place or location present in the image [96]. In contrast to object classification, it is necessary to accurately identify and understand various interactions and relationships between the objects and people in the scene to predict the appropriate scene category. In spite of rapid progress in datasets [88, 96] and methods [14] for this task, robustness of scene classification networks to unseen domains in general, and across geographies in particular, has received little attention, for which we propose a suitable benchmark.

Selecting Concepts and Images We use the 205 scene categories from Places-205 [96] to build GeoPlaces, as these semantic categories cover a wide range of real world scenes commonly encountered in most geographies. We build our GeoPlaces benchmark from the labeled Places-205 dataset [97]. We first collect the unique Flickr identifier (`Flickr-id`) associated with each image in the Places-205 dataset, and then use the publicly available Flickr API[1] to extract the GPS location of the image. Since only a fraction of images belong to Flickr and a further smaller fraction

[1] Flickr.com/services/api/explore/Flickr.photos.geo.getLocation

(a) **GeoPlaces** (b) **GeoImNet**

Figure 3. **Context Shift in GeoNet** A few examples showing the nature of context shifts across categories from GeoPlaces benchmark in (a), and GeoImNet benchmark in (a), arising due to a variety of differences between geographical disparity. For example, outdoor scenes (shopfront, marketplace) reflect the demographies across geographies, indoor-scenes (living rooms, cafeteria) reflect cultural and economic variations and wildlife images reflect the habitat and climatic variations.

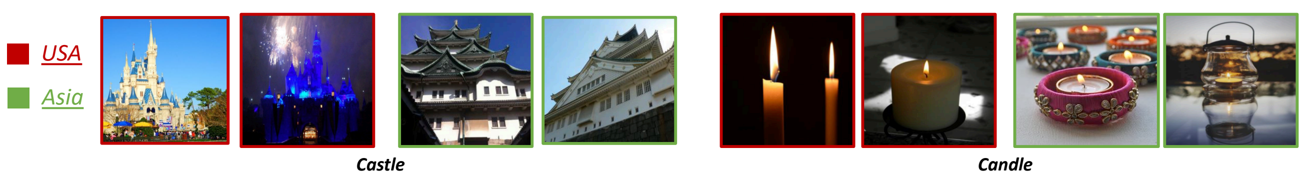


Figure 4. **Design Shift in GeoNet** We show examples illustrating the design shifts for the cases of *castle* from GeoPlaces and *candle* from GeoImNet. Note that differences in designs of castles as well as the variety of objects like candles found across geographies lead to design shifts between the domains.

contain valid geotags, we end up with around 400k images from 205 classes with associated geographical information. Of these, 190k images are from the US domain, and we use 178k of them for training and 17k for testing. In Asia domain however, we obtain only 27k images. To match the scale of images from both domains, we perform an additional step and manually collect more images as explained next.

Additional Data Due to the inherent US-centric bias of photo-sharing websites like Flickr, a major portion of images are US-based. In order to collect more images from the Asia domain, we directly scrape images from Flickr using the 205 category names from Places-205 as the *seed concepts*. As many Asian users often post descriptions and tags for pictures in languages other than English, we use translations of these seed concepts in English to 6 Asian languages, namely {Hindi, Korean, Japanese, Chinese, Russian, Hebrew}, and use these along with the original concepts, as the augmented or *expanded concepts*. Then, we search Flickr for images which match the criterion that (i) they are geotagged in Asia, and (ii) the tags associated with the image match with exactly one of the categories in the expanded concept list (which we assign as the label). We collect around 190k images this way, and use this as the training set. Since images collected from web tend to be nosier than human labeled ones, we use the manually labeled 27k images from Places-205 as the test set for Asia domain to ensure robust benchmarking.

3.2. GeoImNet

We propose the GeoImNet benchmark to investigate the domain shift due to geographical disparities on object classification. Different from existing object-level datasets for domain adaptation [56, 57, 64, 84], GeoImNet provides domain shifts induced by geographic disparities.

Dataset curation We collect images in the GeoImNet benchmark from the WebVision dataset [46], which itself is scraped from Flickr using queries generated from 5000 concepts in the Imagenet-5k dataset [22]. We then follow the same pipeline as explained above for GeoPlaces benchmark, and identify the GPS coordinates of each images using its `Flickr-id`.

Concept Selection Although the original dataset contains 5000 classes, many of these classes are indigenous to a particular geography. For example, *Bengal Tigers* are found in Indian subcontinent, and *Bald Eagle* is a North-American bird. Since unsupervised domain adaptation typically demands matching label spaces across source and target, we select 600 categories out of the original 5000 with at least 20 images in each domain from each category. We then assign roughly 15% of images from each domain into the test set and use the remaining as the training images.

Dataset filtering WebVision is *webly supervised* [16], which does not guarantee object-centric images or clean labels. Therefore, we remove all the images from the dataset which have more than one tag that match our selected concepts (the 600 chosen categories) to handle multi-labeled images. Furthermore, we manually quality-check all the test images and remove all the images with noisy labels. Finally, we perform de-duplication to remove images from the training set which are very similar to those in the test set. More insights into each step of our data collection and filtering process is provided in the supplementary material. The final label distribution for both US and Asia domains in both our benchmarks is shown in Fig. 2.

3.3. GeoUniDA

Universal Domain Adaptation (UniDA) [91] facilitates domain adaptation between source and target domains that have few private classes, in addition to shared classes which are common to both. While this is a realistic problem, prior works [45, 65, 67, 91] use benchmarks created from existing UDA datasets for evaluation. However, our proposed geographical adaptation setting gives us an unique opportunity to design benchmarks for UniDA such that the private categories from the source and the target are a natural reflection of the presence or absence of these categories in the respective geographical domains. In order to select the shared and private categories for our Geo-UniDA benchmark, we first start with the 1000 categories in the original Imagenet-1k dataset [63], and select top 200 categories each in the USA and Asia domains that have the most number of images from the WebVision dataset. Out of these, we use the 62 common classes as the shared categories, and the remaining 138 as the private classes in each domain.

3.4. Analysis of Distribution Shifts

We denote the source dataset using $D_s{=}\{X_s, Y_s\}$, and assume that $X_s{\sim}P_s(x)$ and $(X_s, Y_s){\sim}P_s(x, y)$ where $P_s(x)$ and $P_s(x, y)$ are the image marginal and image-label joint distribution respectively. Target dataset $D_t = \{X_s, Y_s\}$ and target distributions $P_t(x)$ and $P_t(x, y)$ are defined similarly, and the domain discrepency assumption states that $P_s(x, y) \neq P_t(x, y)$. In order to formulate domain shift across geographies, we define f_x as the part of image referring to the foreground objects (corresponds to the salient objects in a scene) and b_x to be the rest of the image corresponding to the background regions (corresponding to the surrounding regions or context). For example, for the task of classifying *living room* in Fig. 3a from GeoPlaces, common objects like sofa and table are foreground, while floor, roof and walls are backgrounds. We make a simplifying assumption that an image is completely explainable using its foreground and background and replace the class-conditional distribution of the images $P(x|y)$ with the joint class-conditional $P(b_x, f_x|y)$. Further, we also assume that given a class label, the background is conditionally independent of the foreground. Then,

$$\begin{aligned} P(x,y) &= P(x|y)\cdot P(y) \\ &= P(b_x, f_x|y)\cdot P(y) \\ &= P(b_x|y)\cdot P(f_x|b_x,y)\cdot P(y) \\ \implies P(x,y) &= \underbrace{P(b_x|y)}_{\text{context}}\cdot\underbrace{P(f_x|y)}_{\text{design}}\cdot\underbrace{P(y)}_{\text{prior}} \end{aligned} \tag{1}$$

We define the class-conditional background distribution $P(b_x|y)$ as context, class-conditional object distribution $P(f_x|y)$ as design and the label distribution $P(y)$ as prior. Note that standard covariate shift assumption [4] assumes uniform domain discrepency across all the images ($P_s(x){\neq}P_t(x)$), which does not hold for geographic adaptation due to the diverse source of variations. We analyze each of these from a geographic adaptation perspective next.

Context Shift We define context shift to be the changes in the context around an object or scene given by $P_s(b_x|y) \neq P_t(b_x|y)$. Deep learning models are generally sensitive to object contexts and backgrounds, and learn spurious correlations that impede their ability to recognize objects and scenes in novel contexts [19, 20, 62, 75]. In geographic adaptation, context shift can be caused by differences in cultural or economic factors across geographies, and few examples illustrating context shift from GeoPlaces and GeoImNet are shown in Fig. 3. While prior works already introduce context shift for domain adaptation [58], a key difference lies in their modeling assumption that the context is irrelevant while training, while in our case context might play a key role in improving scene classification on GeoPlaces.

Design Shift We define "design" shift as the change in object structure, shape and appearance, where the foreground objects belonging to the same semantic category look different across geographies, given by $P_s(f_x|y) \neq P_t(f_x|y)$. Few examples are shown in Fig. 4, where categories like *castle* from GeoPlaces and *candle* from GeoImNet datasets look widely different due to high intra-class variance, although they belong to the same semantic category. It is important to note that context and design shifts might also occur within a domain or within a geography. However, it is easier to account for intra-domain variations on labeled source datasets than ensuring robustness to new and unlabeled geographies.

Prior Shift The label distributions across the domains in our benchmarks widely differ due to natural prominence or rarity of the classes according to the geography, as shown in Fig. 2, where the head classes of one domain might be tail classes in another. This leads to a prior shift where $P_s(y) \neq P_t(y)$. For example, categories like *railway station, outdoor markets, monasteries* are common in Asia while *baseball stadiums*

		GeoPlaces	
Train ↓ / Test →	USA	Asia	Drop (%)
USA	56.35/85.15	36.27/63.27	-20.08/-21.88
Asia	21.03/44.81	49.63/78.45	-28.60/-33.64
		GeoImNet	
Train ↓ / Test →	USA	Asia	Drop (%)
USA	56.35/77.95	36.98/63.42	-19.37/-14.53
Asia	40.43/64.60	60.37/80.22	-19.94/-15.62

Table 2. Top-1/Top-5 accuracies of Resnet-50 models across geographically different train and test domains. Note the significant drop in accuracies caused by the geographical domain shifts in each setting.

	Original			Balanced		
	USA	Asia	Δ	USA	Asia	Δ
GeoPlaces	56.35	36.27	20.08%	55.52	42.6	12.92%
GeoImNet	56.35	36.98	19.37%	52.72	37.3	15.42%

Table 3. USA $\rightarrow$ Asia comparison between GeoNet and its label-balanced version. Non-trivial gaps between the geographies still exist even after accounting for prior shift between the domains.

are more common in USA. Prior works examining prior shift or label shift across domains [2, 3, 28, 48, 92] generally assume that the class conditionals remain the same, which is not true in the case of geographic adaptation due to context and design shifts as illustrated above.

4. Experiments

4.1. Domain Shifts in Proposed Datasets

We illustrate the severity of domain differences across geographies using the drop in accuracy caused by cross-geography transfer in Tab. 2. Specifically, we train a Resnet-50 [34] model using images only from one domain, and compute the accuracies on both within-domain and cross-domain test sets. Since a lot of categories in GeoNet are close (example, *train station* vs. *subway station*), we use both top-1 and top-5 accuracies to report the performance. We observe a significant drop in accuracy caused by direct transfer of models across domains which can be attributed to the geographic bias in the training data. For example, a model trained on GeoPlaces benchmark on US images gives 56.35% Top-1 accuracy on US images, but only 36.27% on images from Asia with a notable drop of 20%. On the GeoImNet benchmark, within-domain testing on images collected from USA gives 56.35% top-1 accuracy while cross-domain testing on Asia images gives only 36.98% with a drop of 19.37%. The 36.98% accuracy is also much inferior to the supervised accuracy on the Asia domain (60.37%) which can be considered as the target upper bound.

Meta-category wise error analysis for GeoImNet We relate the drop in performances across geographies to the proposed notions of domain discrepency in geographic adaptation like context and domain shifts in Fig. 5. Specifically, since the concepts in GeoImNet are sourced from ILSVRC, we leverage the wordnet hierarchy to group our 600 classes into 9 meta-labels. We then average the accuracy within each meta-class from USA→Asia domain transfer, and plot

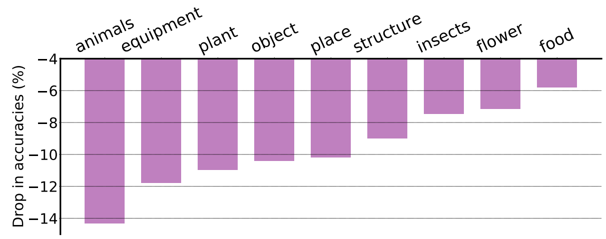


Figure 5. Drop in accuracies for each meta-category in GeoImNet. Groups that showcase context and design shifts suffer a larger drop in accuracy.

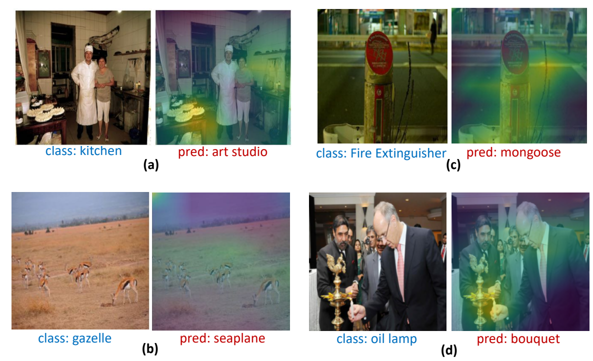


Figure 6. GradCAM visualization of predictions of a USA-trained model on Asia images show that prominent context and design shifts across geography hurts accuracy. (a) is from GeoPlaces, (b,c,d) are from GeoImNet.

the difference in accuracy across domains per meta-label in Fig. 5. We note that categories in the meta-label "animals" have minimum design-shift across domains, but suffer from context shift due to shifts in weather and habitats across geographies leading to significant drop in accuracy. On the other hand, many categories in "equipment" and "object"(like *candle, broom, sewing machine*) have prominent design shifts (Fig. 4) leading to notable performance drop. Finally, categories in "food" (like *bottled water, ice-cream*) have minimum change in both design and context and hence suffer the least fall in accuracy across domains.

GradCAM visualization of the failure cases We present few examples in Fig. 6 of predictions made on Asia test images by a model trained on USA, along with their Grad-CAM visualizations. As shown, when the model focuses on the context and background, it fails to generalize to new scenes from target geographies with notable shifts in context (*kitchen* classified as *art studio*). Even in cases when the model accurately focuses on the foreground object, it sometimes leads to incorrect predictions due to design shifts between geographies, where *oil lamp* is accurately localized, but predicted as *bouquet*.

Separating the prior shift To further delineate prior shift from context and design shifts, we curate a balanced subset out of GeoNet such that each category has about 200-300 images, and drop categories which have fewer images (about $3/4^{th}$ of the categories remain). From Tab. 3, the drop in accuracy after addressing the prior shift is 12.9% on GeoPlaces

Method	GeoPlaces				GeoImNet			
	USA → Asia		Asia → USA		USA → Asia		Asia → USA	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Source Only	**36.27**	**63.27**	**21.03**	**44.81**	**36.98**	**63.43**	**40.43**	**64.6**
DANN [27]	29.58	55.23	16.59	35.32	32.88	57.77	38.42	62.90
CDAN [50]	30.48	55.94	17.01	36.26	35.94	60.21	39.88	63.74
MCC [38]	30.09	55.85	17.17	<u>36.85</u>	35.71	60.48	39.86	64.00
SAFN [90]	32.50	57.93	14.34	35.68	32.40	58.43	36.26	61.58
MDD [94]	34.18	59.10	<u>17.81</u>	36.44	36.26	62.13	40.15	63.91
MCD [69]	33.49	59.41	16.57	34.74	25.60	48.45	36.69	60.68
ToAlign [87]	29.86	56.16	16.32	33.58	32.13	58.64	37.98	63.17
MemSAC [40]	<u>34.68</u>	<u>60.52</u>	15.75	32.83	<u>36.71</u>	<u>63.16</u>	<u>40.34</u>	<u>64.40</u>
Tgt. Supervised	49.63	78.45	56.35	85.15	60.37	80.22	56.35	77.95

Table 4. **UDA on GeoNet** Top-1 and Top-5 accuracies of various unsupervised adaptation methods on GeoNet. Most of the methods fail to sufficiently handle cross-geography transfer on both GeoPlaces and GeoImNet benchmarks and often give lower accuracies even compared to a baseline model trained only using source data calling attention to the need for novel methods that can handle domain shifts beyond style and appearance.

Method	closed-set	open-set	H-Score	Target Sup.
UniDA [91]	27.64	43.93	33.93	70.70%
DANCE [66]	38.54	78.73	51.75	
OVANet [67]	36.54	66.89	47.26	

Table 5. **Universal domain adaptation methods on GeoUniDA.** *closed-set* and *open-set* refer to the closed set and open set accuracies, and *H-Score* is the harmonic-mean of the two. Note the significant gap that still exists with target supervised accuracy on closed-set labels with the best adaptation method DANCE [66].

and 15.4% on GeoImNet, compared to 20.08% and 19.37% on the original datasets, showing that non-trivial accuracy drops caused by context and design shifts still exist even after accounting for label imbalance between the domains.

4.2. Benchmarking Domain Adaptation

We study the effectiveness of prior unsupervised adaptation algorithms in bridging novel notions of domain gaps like context shift and design shift on GeoNet. We review various standard as well as current state-of-the-art domain adaptation methods to examine their geographical robustness.

Architecture and training details We follow the standard protocol established in prior works [40, 50, 69] and use an ImageNet pre-trained Resnet-50 [34] as the feature extractor backbone and a randomly intialized classifier layer. We use a batch size of 32 and SGD with a learning rate of 0.01 for the classifier head and 0.001 for the already pretrained backbone. We report the top-1 and top-5 accuracy numbers using the test splits from each benchmarks. We perform comparisons between traditional adversarial methods (DANN [27], CDAN [50]), class-aware adaptation methods (MCC [38], MDD [94]), non-adversarial methods (SAFN [90], MCD [69]) as well as recent state-of-the-art (ToAlign [87], MemSAC [40]). We train prior works using their publicly available code and adopt all hyper-parameters as recommended in the respective papers.

Existing UDA methods do not suffice on GeoNet We show the Top-1 and Top-5 accuracies of all the transfer settings from GeoNet in Tab. 4. A key observation is that most of the domain adaptation approaches are no better, or sometimes even worse, than the baseline model trained only using source domain data, indicating their limitations for geographic domain adaptation. For example, on GeoPlaces, training using data from USA achieves a top-1 accuracy of 36.27% on test data from Asia test images, while the best adaptation method (MemSAC) obtains lesser accuracy of 34.7%, indicating negative transfer. Likewise, on GeoImNet, a USA-trained source model achieves 36.98% on test images from Asia which is comparable to the best adaptation accuracy of 36.71%. To further illustrate this, we define relative accuracy gain as the improvement in accuracy obtained by a method over a source-only model as a percentage of gap between a source-only model and the target-supervised upper bound (which is 100% if the method achieves the target supervised upper bound). From Fig. 1b, it is notable that the same adaptation methods that yield significantly high relative accuracy gains on DomainNet [56] yield negative relative accuracy gains on GeoNet, highlighting the unique the nature of distribution shifts in real-world settings like geographic adaptation that challenge existing methods. These observations also suggest that future research should focus on context-aware and object-centric representations in addition to domain invariant features to improve cross-domain transfer amidst context and design shifts.

Universal domain adaptation on Geo-UniDA We run SOTA universal domain adaptation methods (You et.al. [91], DANCE [66] and OvaNET [67]) on the Geo-UniDA benchmark of GeoNet. Following prior works [67], we adopt the H-score metric which is a harmonic mean of closed-set and open-set accuracies giving equal importance to closed set transfer as well as open set accuracy. In Tab. 5, we

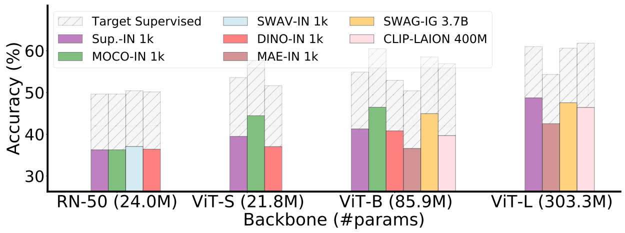


Figure 7. We show that most architectures and pre-training strategies exhibit significant cross-domain drops when fine-tuned on geographically biased datasets. Shown for USA→Asia on GeoPlaces, refer Fig. 1c for the plot on GeoImNet and supplementary material for other transfer settings.

show that DANCE [66] outperforms both You et.al. [91] and OVANet [67] on the Geo-UniDA benchmark. We also show that a significant gap still exists between target supervised accuracy when trained using supervision (70.7%) and best adaptation accuracy (38.5%) on our benchmark, highlighting the limitations of existing methods to efficiently address universal adaptation in a geographic context.

4.3. Large-scale pre-training and architectures

It is common to use large scale self-supervised [11, 12, 15, 17, 32, 33] and weakly-supervised [37, 53, 76] pre-trained models as starting points in various downstream applications. While recent works explored role of pre-training on domain robustness [43], we are interested in the extent to which large scale pre-training effectively preserved robustness when fine-tuned on geographically under-represented datasets. We investigate the performance of a variety of methods on GeoNet in terms of backbone architectures, pre-training strategies and supervision.

Experimental setup Our backbone architectures include Resnet50 [34] as well as the small (ViT-S), base (ViT-B) and large (ViT-L) vision transformers [25]. In terms of supervision, in addition to the standard supervised pre-training on ImageNet-1k, we also consider self-supervised methods MoCo-V3 [17], SwAV [11], DINO [12], MAE [32] trained on ImageNet-1k, the weakly supervised SWAG [76] trained on 3.6B uncurated instagram images and CLIP [59] trained on 400M image-language pairs [71]. We denote {Backbone-Supervision-Data} for different model choices (for example, Resnet50-sup-IN1k indicates a Resnet50 pre-trained on supervised data from ImageNet-1k).

For evaluating geographic robustness of these models, we first take the pre-trained model and fine-tune it on training data from a "source" geography, then evaluate the performance on test data from the "target" geography. We show the results using USA as the source and Asia as the target from the GeoPlaces benchmark in Fig. 7, and GeoImNet benchmark in Fig. 1c. For reference, we also report accuracy after fine-tuning on labeled data from the target geography for each {Backbone-Supervision-Data} pair (denoted as target-supervised), which serves as an upper bound for the transfer performance.

Large-scale pretraining is not geographically robust From Fig. 7, we make a few observations. Firstly, comparison between Resnet50 and ViT-S which have roughly the same number of parameters suggests the superiority of the vision transformer architectures over CNNs. For example, ViT-S-sup-IN1k is better than Resnet50-sup-IN1k, and ViT-S-moco-IN1k is better than Resnet50-moco-IN1k, indicating that global reasoning using self-attention layers in vision transformers benefits context-dependent tasks like GeoPlaces. Next, comparing different pre-training strategies, we observe that MoCo gives best accuracy on ViT-S and ViT-B, while supervised pre-training outperforms other approaches on large models like ViT-L. However, the gap between target supervised accuracy and the best adaptation accuracy achieved using either Resnet50 or any of the vision transformers is still high, highlighting the need for better transfer strategies. In terms of data, weakly-supervised pre-training using billion-scale dataset IG3.6B (ViT-B-swag-3B) shows significant improvements over self-supervised training methods like MAE (ViT-B-mae-IN1k) and DINO (ViT-B-dino-IN1k). But despite training on massive-scale data, ViT-L-swag-3B and ViT-L-clip-400M are still inferior to the target supervised accuracies, revealing the limitations of current pre-training strategies towards robust cross-geography transfer after fine-tuning. While the success of large-scale pre-training strategies are well-documented on popular datasets like ImageNet, our results indicate that similar benefits might not be observed when application domains significantly differ from pre-training or fine-tuning datasets [21].

5. Conclusion

We introduce a new dataset called GeoNet for the problem of geographic adaptation with benchmarks covering the tasks of scene and object classification. In contrast to existing datasets for domain adaptation [56,57,64,84], our dataset with images collected from different locations contains domain shifts captured by natural variations due to geographies, cultures and weather conditions from across the world, which is a novel and understudied direction in domain adaptation. Through GeoNet, we analyze the sources of domain shift caused by changes in geographies such as context and design shift. We conduct extensive benchmarking on GeoNet and highlight the limitations of current domain adaptation methods as well as large-scale pretraining methods towards geographical robustness. Finally, in spite of geographical diversity in GeoNet, we note a possible limitation of indirect bias towards USA as the user-base on photo-sharing sites like Flickr is dominated by the US. Creating datasets that are a more natural reflection of cultures and trends from diverse geographies and devising learning algorithms robust to those variations is an exciting proposition for the future.

Acknowledgements We thank NSF CAREER 1751365, Google AI Award for Inclusion Research and National Research Platform for hardware access.

References

[1] Michael A Alcorn, Qi Li, Zhitao Gong, Chengfei Wang, Long Mai, Wei-Shinn Ku, and Anh Nguyen. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4845–4854, 2019. 2

[2] Amr Alexandari, Anshul Kundaje, and Avanti Shrikumar. Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation. In *International Conference on Machine Learning*, pages 222–232. PMLR, 2020. 6

[3] Kamyar Azizzadenesheli, Anqi Liu, Fanny Yang, and Animashree Anandkumar. Regularized learning for domain adaptation under label shifts. *arXiv preprint arXiv:1903.09734*, 2019. 6

[4] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010. 2, 5

[5] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19:137–144, 2006. 2

[6] Ryan Y Benmalek, Sabhya Chhabria, Pedro O Pinheiro, Claire Cardie, and Serge Belongie. Learning to adapt to semantic shift. 2021. 2

[7] Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020. 2

[8] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3722–3731, 2017. 2

[9] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *Advances in neural information processing systems*, pages 343–351, 2016. 2

[10] Tianle Cai, Ruiqi Gao, Jason Lee, and Qi Lei. A theory of label propagation for subpopulation shift. In *International Conference on Machine Learning*, pages 1170–1182. PMLR, 2021. 2

[11] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020. 8

[12] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 8

[13] Chaoqi Chen, Weiping Xie, Wenbing Huang, Yu Rong, Xinghao Ding, Yue Huang, Tingyang Xu, and Junzhou Huang. Progressive feature alignment for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 627–636, 2019. 2

[14] Gongwei Chen, Xinhang Song, Bohan Wang, and Shuqiang Jiang. See more for scene: Pairwise consistency learning for scene classification. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 4987–4999. Curran Associates, Inc., 2021. 3

[15] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 8

[16] Xinlei Chen and Abhinav Gupta. Webly supervised learning of convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1431–1439, 2015. 5

[17] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. pages 9640–9649, 2021. 8

[18] Yi-Hsin Chen, Wei-Yu Chen, Yu-Ting Chen, Bo-Cheng Tsai, Yu-Chiang Frank Wang, and Min Sun. No more discrimination: Cross city adaptation of road scene segmenters. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1992–2001, 2017. 2

[19] Jinwoo Choi, Chen Gao, Joseph CE Messou, and Jia-Bin Huang. Why can't i dance in the mall? learning to mitigate scene bias in action recognition. *Advances in Neural Information Processing Systems*, 32, 2019. 5

[20] Myung Jin Choi, Antonio Torralba, and Alan S Willsky. Context models and out-of-context objects. *Pattern Recognition Letters*, 33(7):853–862, 2012. 5

[21] Elijah Cole, Xuan Yang, Kimberly Wilber, Oisin Mac Aodha, and Serge J. Belongie. When does contrastive visual representation learning work? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 1–10. IEEE, 2022. 8

[22] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 4

[23] Zhijie Deng, Yucen Luo, and Jun Zhu. Cluster alignment with a teacher for unsupervised domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9944–9953, 2019. 2

[24] Terrance DeVries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. Does object recognition work for everyone? *CoRR*, abs/1906.02659, 2019. 2

[25] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 8

[26] Abhimanyu Dubey, Vignesh Ramanathan, Alex Pentland, and Dhruv Mahajan. Adaptive methods for real-world domain

generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14340–14349, 2021. 3

[27] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 1, 2, 7

[28] Saurabh Garg, Yifan Wu, Sivaraman Balakrishnan, and Zachary Lipton. A unified view of label shift estimation. *Advances in Neural Information Processing Systems*, 33:3290–3300, 2020. 6

[29] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018. 2

[30] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 2

[31] Xiang Gu, Jian Sun, and Zongben Xu. Spherical space domain adaptation with robust pseudo-label loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9101–9110, 2020. 2

[32] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 8

[33] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 8

[34] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6, 7, 8

[35] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021. 2

[36] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. PMLR, 2018. 1, 2

[37] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 8

[38] Ying Jin, Ximei Wang, Mingsheng Long, and Jianmin Wang. Minimum class confusion for versatile domain adaptation. In *European Conference on Computer Vision*, pages 464–480. Springer, 2020. 7

[39] Tarun Kalluri and Manmohan Chandraker. Cluster-to-adapt: Few shot domain adaptation for semantic segmentation across disjoint labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4121–4131, 2022. 2

[40] Tarun Kalluri, Astuti Sharma, and Manmohan Chandraker. Memsac: Memory augmented sample consistency for large scale domain adaptation. *arXiv preprint arXiv:2207.12389*, 2022. 1, 2, 7

[41] Tarun Kalluri, Girish Varma, Manmohan Chandraker, and CV Jawahar. Universal semi-supervised semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5259–5270, 2019. 2

[42] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4893–4902, 2019. 1, 2

[43] Donghyun Kim, Kaihong Wang, Stan Sclaroff, and Kate Saenko. A broad study of pre-training for domain generalization and adaptation. *arXiv preprint arXiv:2203.11819*, 2022. 8

[44] Abhishek Kumar, Prasanna Sattigeri, Kahini Wadhawan, Leonid Karlinsky, Rogerio Feris, Bill Freeman, and Gregory Wornell. Co-regularized alignment for unsupervised domain adaptation. In *Advances in Neural Information Processing Systems*, pages 9345–9356, 2018. 2

[45] Jogendra Nath Kundu, Suvaansh Bhambri, Akshay Kulkarni, Hiran Sarkar, Varun Jampani, and R Venkatesh Babu. Subsidiary prototype alignment for universal domain adaptation. *arXiv preprint arXiv:2210.15909*, 2022. 5

[46] Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862*, 2017. 4

[47] 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. 2, 13

[48] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, pages 3122–3130. PMLR, 2018. 6

[49] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015. 1, 2

[50] 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. 1, 2, 7, 14, 15

[51] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, pages 2208–2217. PMLR, 2017. 1, 2

[52] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2507–2516, 2019. 2

[53] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, pages 181–196, 2018. 8

[54] Changhwa Park, Jonghyun Lee, Jaeyoon Yoo, Minhoe Hur, and Sungroh Yoon. Joint contrastive learning for unsupervised domain adaptation. *arXiv preprint arXiv:2006.10297*, 2020. 2

[55] Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. Multi-adversarial domain adaptation. *arXiv preprint arXiv:1809.02176*, 2018. 2

[56] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1406–1415, 2019. 1, 2, 4, 7, 8

[57] 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, 2, 4, 8

[58] Viraj Prabhu, Ramprasaath R Selvaraju, Judy Hoffman, and Nikhil Naik. Can domain adaptation make object recognition work for everyone? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3981–3988, 2022. 2, 3, 5

[59] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 8

[60] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5389–5400. PMLR, 09–15 Jun 2019. 2

[61] William A Gaviria Rojas, Sudnya Diamos, Keertan Ranjan Kini, David Kanter, Vijay Janapa Reddi, and Cody Coleman. The dollar street dataset: Images representing the geographic and socioeconomic diversity of the world. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 2

[62] Amir Rosenfeld, Richard Zemel, and John K Tsotsos. The elephant in the room. *arXiv preprint arXiv:1808.03305*, 2018. 5

[63] 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. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 2, 5

[64] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010. 1, 2, 4, 8

[65] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, and Kate Saenko. Universal domain adaptation through self supervision. *arXiv preprint arXiv:2002.07953*, 2020. 5

[66] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, and Kate Saenko. Universal domain adaptation through self-supervision. 2020. 7, 8

[67] Kuniaki Saito and Kate Saenko. Ovanet: One-vs-all network for universal domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9000–9009, 2021. 5, 7, 8

[68] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Adversarial dropout regularization. *arXiv preprint arXiv:1711.01575*, 2017. 1, 2

[69] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3723–3732, 2018. 1, 2, 7

[70] Swami Sankaranarayanan, Yogesh Balaji, Carlos D Castillo, and Rama Chellappa. Generate to adapt: Aligning domains using generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8503–8512, 2018. 2

[71] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 8

[72] D. Sculley, Eric Breck, Igor Ivanov, James Atwood, Miha Skalic, Pallavi Baljekar, Pavel Ostyakov, Roman Solovyev, Weimin Wang, and Yoni Halpern. *The Inclusive Images Competition*. 2019. 2

[73] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv preprint arXiv:1711.08536*, 2017. 2

[74] Astuti Sharma, Tarun Kalluri, and Manmohan Chandraker. Instance level affinity-based transfer for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5361–5371, 2021. 2

[75] Krishna Kumar Singh, Dhruv Mahajan, Kristen Grauman, Yong Jae Lee, Matt Feiszli, and Deepti Ghadiyaram. Don't judge an object by its context: learning to overcome contextual bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11070–11078, 2020. 5

[76] 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. 8, 15

[77] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer, 2016. 2

[78] Shuhan Tan, Xingchao Peng, and Kate Saenko. Class-imbalanced domain adaptation: an empirical odyssey. In *European Conference on Computer Vision*, pages 585–602. Springer, 2020. 2

[79] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33, 2020. 2

[80] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011. 1

[81] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE international conference on computer vision*, pages 4068–4076, 2015. 2

[82] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017. 2

[83] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 13

[84] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5018–5027, 2017. 1, 2, 4, 8

[85] Rui Wang, Zuxuan Wu, Zejia Weng, Jingjing Chen, Guo-Jun Qi, and Yu-Gang Jiang. Cross-domain contrastive learning for unsupervised domain adaptation. *arXiv preprint arXiv:2106.05528*, 2021. 2

[86] Yan Wang, Xiangyu Chen, Yurong You, Li Erran Li, Bharath Hariharan, Mark Campbell, Kilian Q Weinberger, and Wei-Lun Chao. Train in germany, test in the usa: Making 3d object detectors generalize. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11713–11723, 2020. 2

[87] Guoqiang Wei, Cuiling Lan, Wenjun Zeng, Zhizheng Zhang, and Zhibo Chen. Toalign: Task-oriented alignment for unsupervised domain adaptation. In *NeurIPS*, 2021. 1, 2, 7, 14, 15

[88] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 3

[89] Shaoan Xie, Zibin Zheng, Liang Chen, and Chuan Chen. Learning semantic representations for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 5423–5432, 2018. 2

[90] Ruijia Xu, Guanbin Li, Jihan Yang, and Liang Lin. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1426–1435, 2019. 1, 7

[91] Kaichao You, Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Universal domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2720–2729, 2019. 5, 7, 8

[92] Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *International conference on machine learning*, pages 819–827. PMLR, 2013. 6

[93] Weichen Zhang, Wanli Ouyang, Wen Li, and Dong Xu. Collaborative and adversarial network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3801–3809, 2018. 2

[94] Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael I Jordan. Bridging theory and algorithm for domain adaptation. *arXiv preprint arXiv:1904.05801*, 2019. 7

[95] Bingchen Zhao, Shaozuo Yu, Wufei Ma, Mingxin Yu, Shenxiao Mei, Angtian Wang, Ju He, Alan Yuille, and Adam Kortylewski. Ood-cv: A benchmark for robustness to individual nuisances in real-world out-of-distribution shifts. In *ICML 2022 Shift Happens Workshop*, 2022. 2

[96] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 3

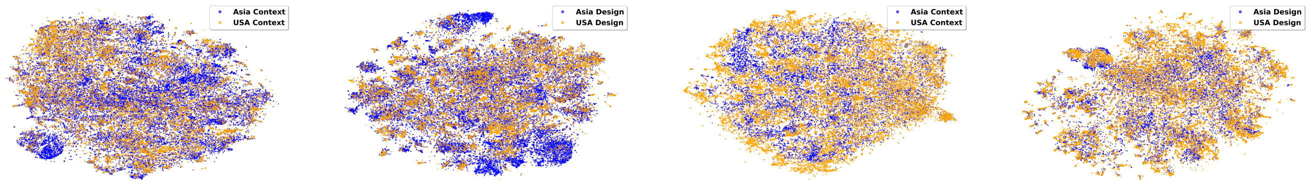
[97] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. *Advances in neural information processing systems*, 27, 2014. 3

A. 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. 6 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.

B. Visualization of Context and Design Shifts

We provide deeper insight into the cross-domain shifts in contexts and designs induced by the geographies by vi-



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

Figure 8. **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.

	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 6. **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.

sualizing their tSNE feature representations [83]. 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 [47] 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 elementwise 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. 8.

Context Shift The pronounced distinction in the contexts between the two domains from GeoPlaces is highlighted in Fig. 8a, 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. 8c. 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. 8b for the case of GeoPlaces and in Fig. 8d 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.

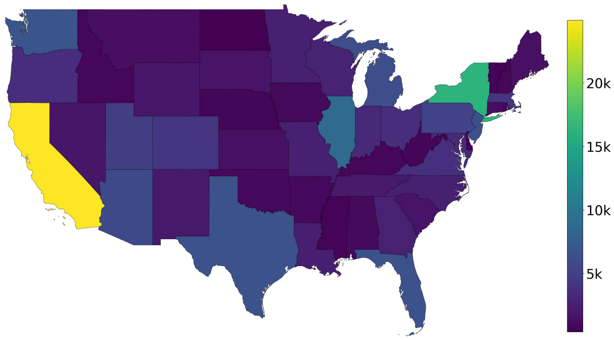
We also note that datasets like COCO are predominantly US-biased, so the use of COCO in analyzing distribution shifts on Asia images is not completely fair. To this end, manually annotating images with finer-grained foreground and context labels in both geographies would yield more accurate analysis, which is left as a future work.

C. Geographic Distribution of Images

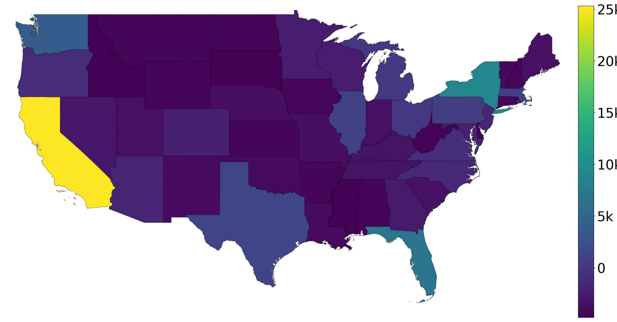
While we broadly categorize Asia and USA to be the two major geographical domains, not all sub-regions in these geographies have equal representation. We show the geographic distribution over respective geographies in Fig. 9, by leveraging the per-image GPS metadata provided in GeoNet. For images from Asia from Fig. 9c for GeoPlaces and Fig. 9d 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. 9a and Fig. 9b, 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 where all our images have been taken from.

D. Error Analysis of Unsupervised Adaptation

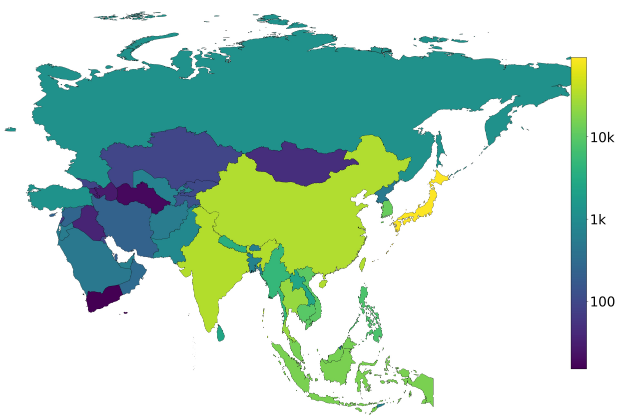
While we show in the main paper (Table 3) that existing unsupervised adaptation approaches yield limited benefit for geographical adaptation, we conduct a deeper analysis into the per-class accuracy post-adaptation in Fig. 10 for the case of USA→Asia on GeoPlaces. Specifically, we first take a model trained only on USA images, and compute the drop in



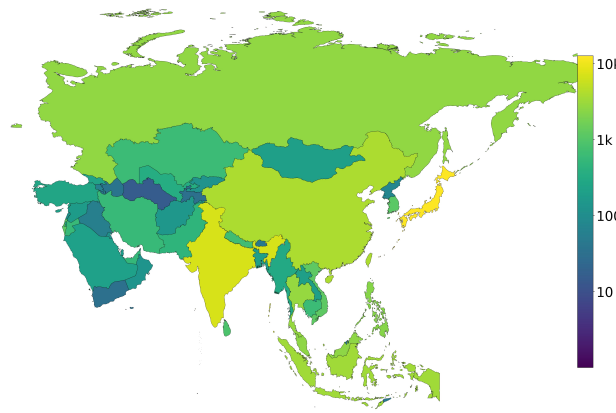
(a) **GeoPlaces: USA Images**



(b) **GeoImNet: USA Images**



(c) **GeoPlaces: Asia Images**



(d) **GeoImNet: Asia Images**

Figure 9. **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.

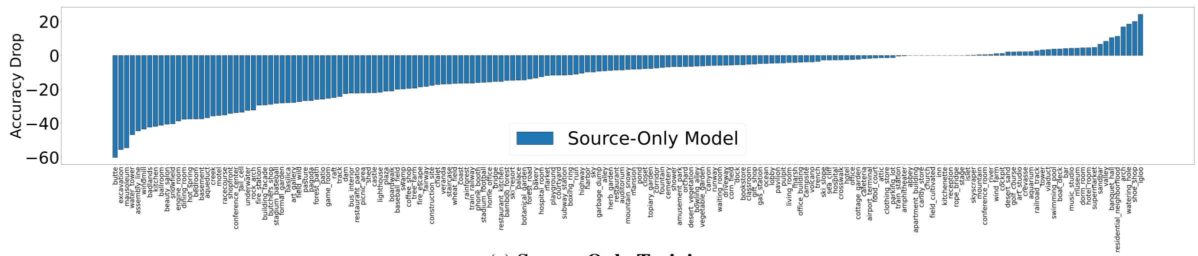
per-class accuracy suffered by direct cross-domain transfer on Asia test images. We show this in Fig. 10a, where classes like *mausoleum*, *assembly line* and *kitchen* suffer the largest drops in accuracy. Next, we carry the same analysis using a model trained with CDAN [50] adaptation method. From Fig. 10b, 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 [87], underlining the fact that existing state-of-the-art adaptation methods cannot handle geographic shifts across most categories.

E. Data De-duplication

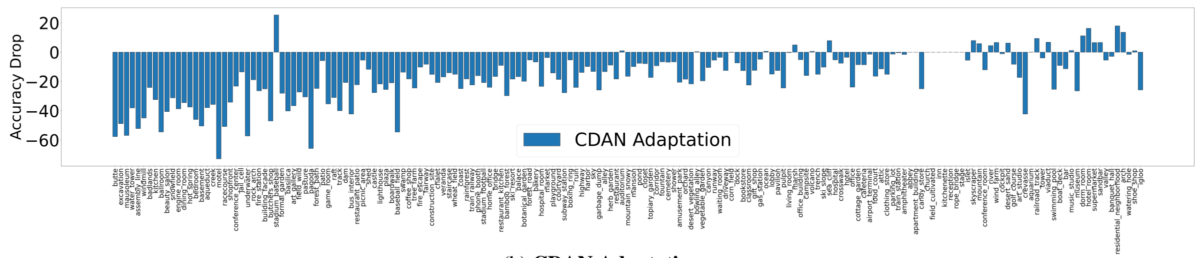
Since a lot of users tend to upload multiple pictures of the same scene on sites like Flickr, we carry a data de-duplication exercise so that there are no such duplicate copies of same images in train and test sets which would unfairly improve within-domain accuracy. We first group all the images in 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.

F. Large-scale pre-training on GeoNet

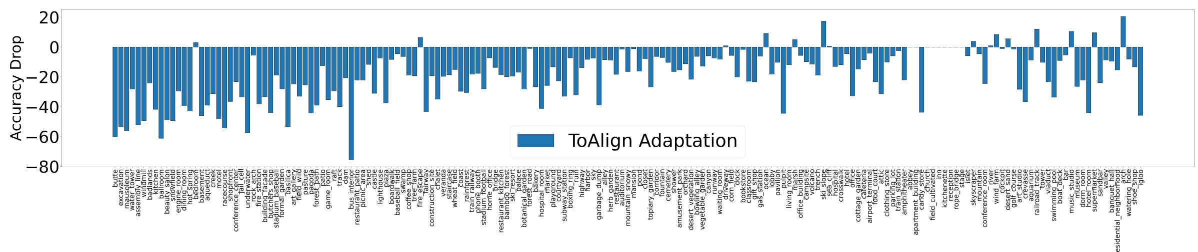
In Fig. 11, we show the effect of large-scale pretraining on the transfer setting Asia→USA from GeoPlaces(Fig. 11a) and GeoImNet(Fig. 11b). We make similar observations as the transfer setting from USA→Asia in the main paper. Specifically, we show that transformers outperform Resnets,



(a) **Source-Only Training**

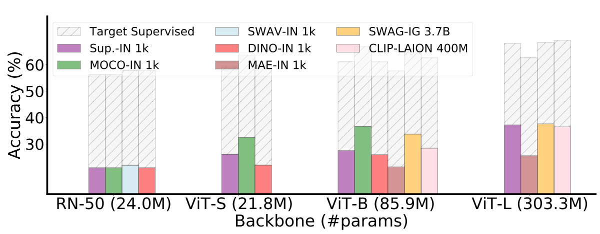


(b) **CDAN Adaptation**

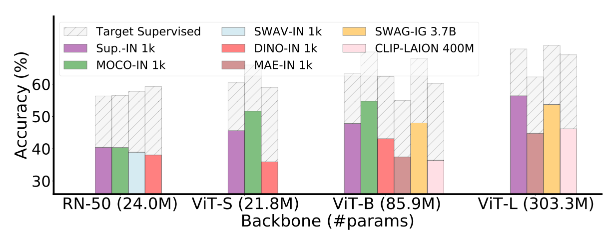


(c) **ToAlign Adaptation**

Figure 10. **Per-class accuracy drops** on USA→Asia transfer for a plain source-only model as well as post-adaptation using CDAN [50] and ToAlign [87] 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 11. **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. 11a and GeoImNet in Fig. 11b, refer main paper for other transfer settings.

pre-training using billion-scale datasets like SWAG [76] 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.

G. Effect of label-cleaning on GeoImNet

Before the current version of GeoImNet with 600 classes, we created a slightly larger, albeit more noisy 700 class version. We then observed that although these concepts have been selected from ImageNet, there were many ambiguous classes (like `fancy dress`, `frozen yogurt`, `prey`, `flash`, `walking stick`) etc. So, we removed 100 such classes with ambiguous concept meanings, and created a newer version with 600 classes, which is eventually used in benchmarking and release. In Tab. 7, we show the cross-domain accuracies with the older version. We observed that while the cross-domain drops remain the same,

Source ↓ / Target →	USA		Asia		Drop(%)	
GeoImNet-Before Filtering						
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
USA	46.63	67.85	29.69	51.43	-16.94	-16.42
Asia	31.55	52.28	52.93	72.96	-21.38	-20.68
GeoImNet-After Filtering						
	USA		Asia		Drop(%)	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
USA	56.35	77.95	36.98	63.42	-19.37	-14.53
Asia	40.43	64.60	60.37	80.22	-19.94	-15.62

Table 7. Top-1/Top-5 accuracy of models across geographically different train and test domains with a more noisier 700-class version of GeoImNet.

the absolute accuracy themselves are much higher using a cleaner version of the dataset.

H. Sample Images

We show few sample images from selected classes across both USA and Asia domains in GeoPlaces benchmark in Fig. 12, Fig. 13 and GeoImNet benchmark in Fig. 14, Fig. 15.

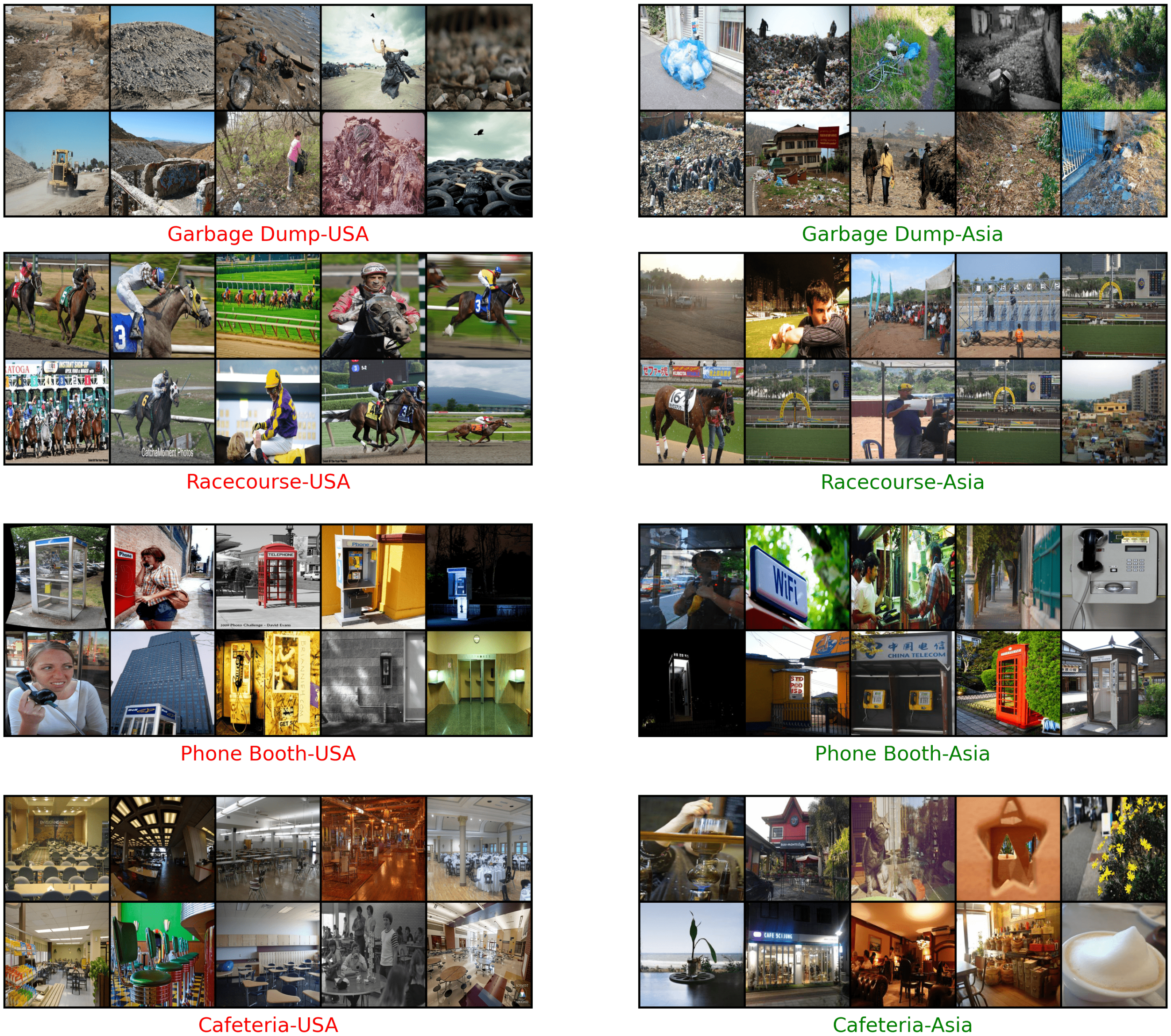


Figure 12. Sample images showing the domain gap between USA (left) and Asia (right) domains for classes `garbage dump`, `race course`, `phone booth` and `cafetaria` from GeoPlaces.

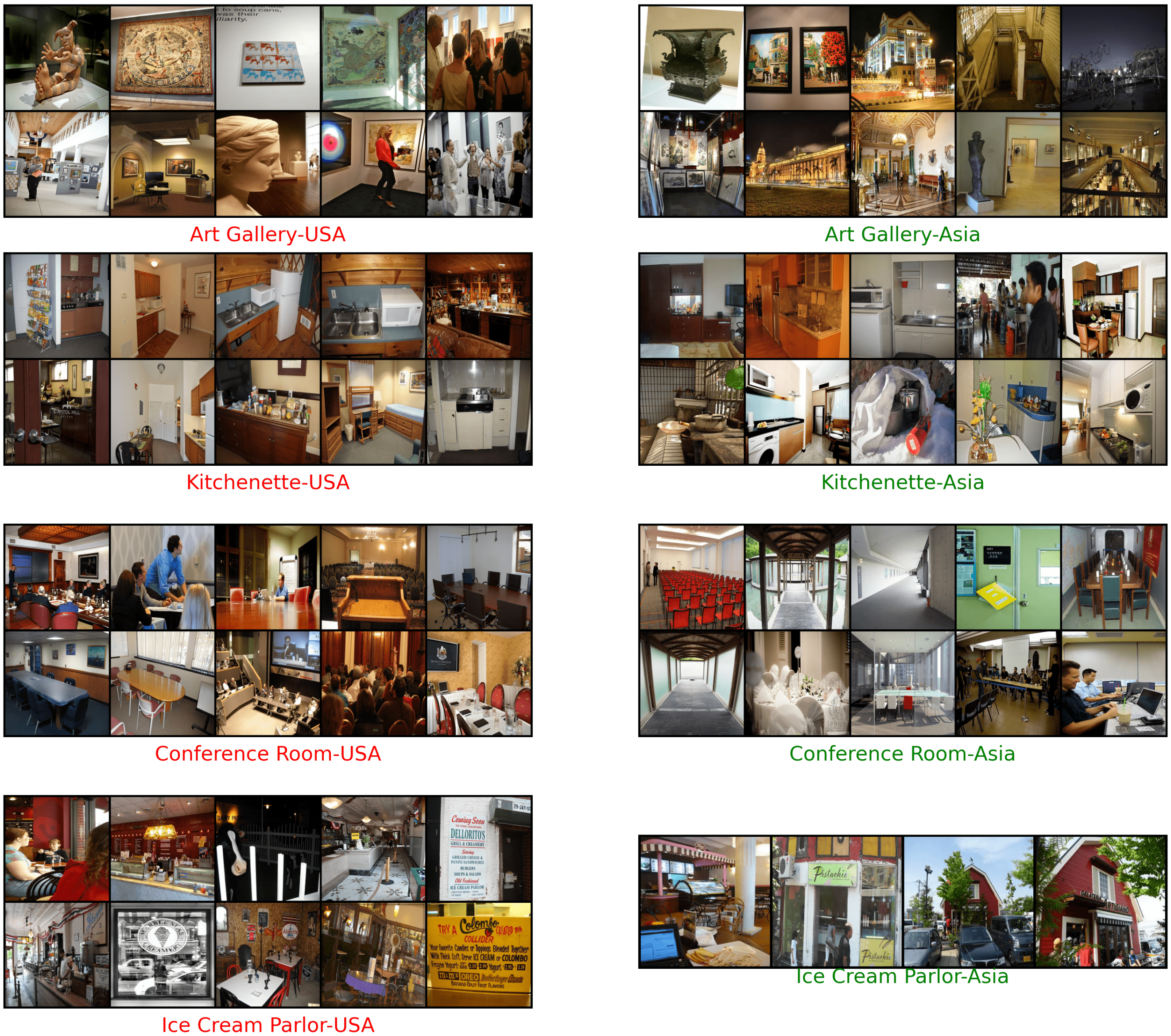


Figure 13. Sample images showing the domain gap between USA (left) and Asia (right) domains for classes `art gallery`, `kitchenette`, `conference room` and `ice-cream parlor` from GeoPlaces.

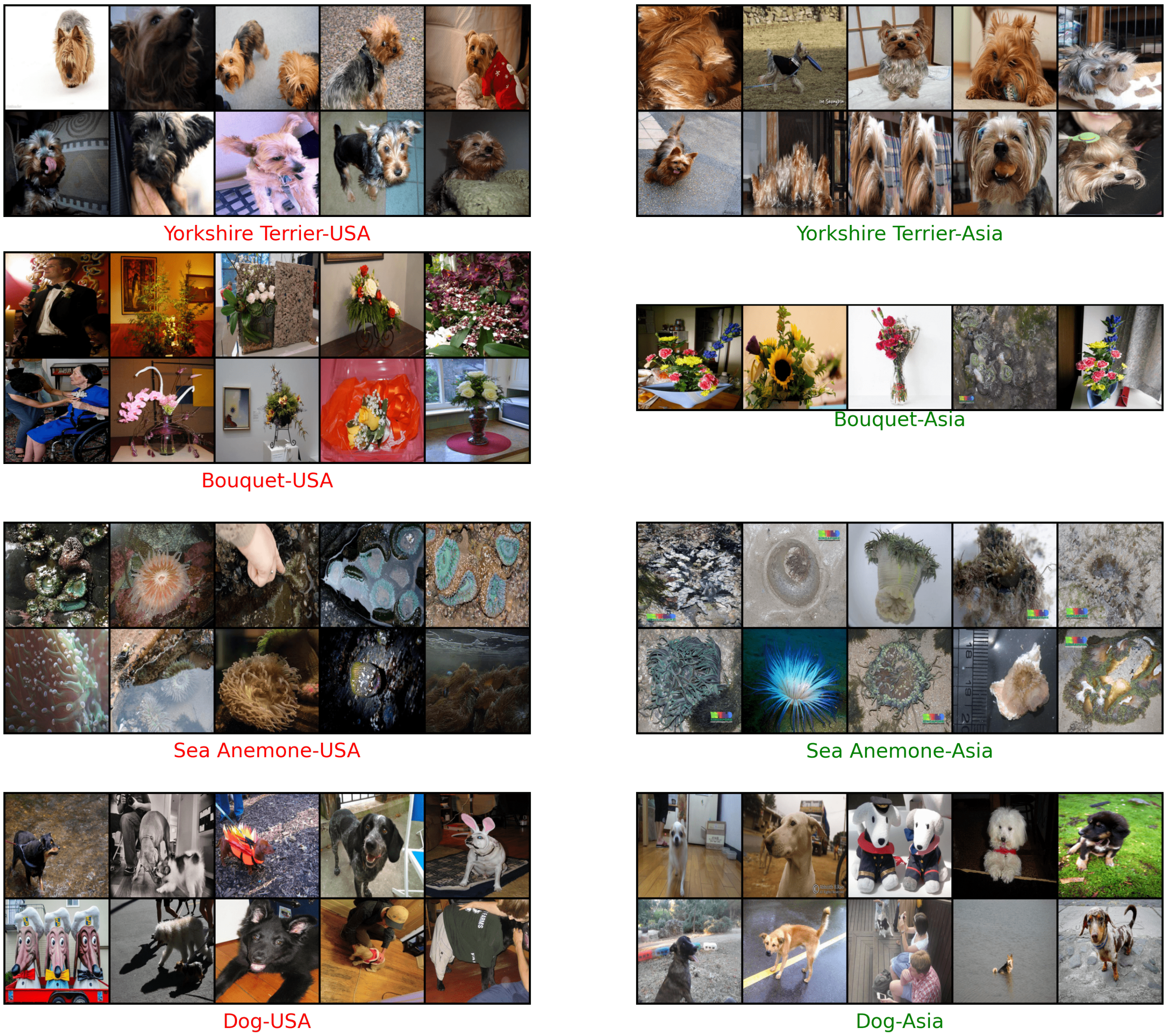


Figure 14. Sample images showing the domain gap between USA (left) and Asia (right) domains for classes `Yorkshire Terrier`, `bouquet`, `sea anemone` and `dog` from GeoImNet.

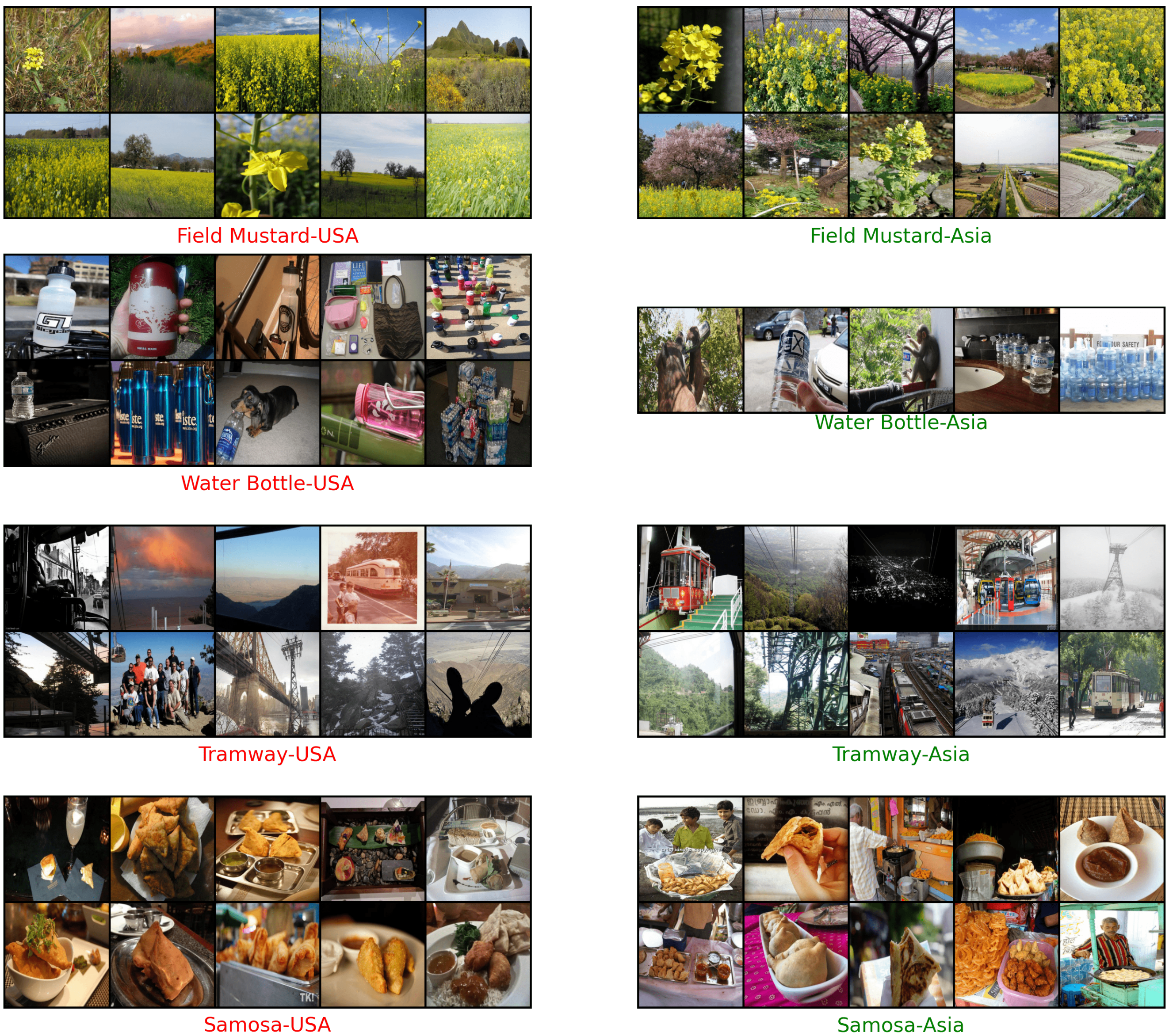


Figure 15. Sample images showing the domain gap between USA (left) and Asia (right) domains for classes `Field Mustard`, `Water Bottle`, `Tramway` and `Samosa` from GeoImNet.