Can domain adaptation make object recognition work for everyone?

Viraj Prabhu[*,1] **Ramprasaath R. Selvaraju**[†,2] **Judy Hoffman**[1] **Nikhil Naik**[3]
[1]Georgia Tech [2]Artera AI [3]Salesforce Research
{virajp,judy}@gatech.edu ram@artera.ai nnaik@salesforce.com

Abstract

Despite the rapid progress in deep visual recognition, modern computer vision datasets significantly overrepresent the developed world and models trained on such datasets underperform on images from unseen geographies. We investigate the effectiveness of unsupervised domain adaptation (UDA) of such models across geographies at closing this performance gap. To do so, we first curate two shifts from existing datasets to study the Geographical DA problem, and discover new challenges beyond data distribution shift: context shift, wherein object surroundings may change significantly across geographies, and subpopulation shift, wherein the intra-category distributions may shift. We demonstrate the inefficacy of standard DA methods at Geographical DA, highlighting the need for specialized geographical adaptation solutions to address the challenge of making object recognition work for everyone.

1. Introduction

As deep-learning based computer vision systems gain widespread adoption, it is crucial that they perform equitably across diverse geographical deployments. However, prior work [1] has found that in practice modern computer vision datasets significantly overrepresent the developed world and models trained on such datasets systematically underperform on images from the rest of the world [2] (see Fig. 1). Labeling images from target geographies is a natural solution but may be expensive and difficult to scale. Unsupervised domain adaptation [3–5] (UDA) has extensively studied the problem of adapting models trained on a labeled source to an unlabeled target domain. However, UDA typically considers specific kinds of shifts in data generating distributions (*e.g.* synthetic to real data [6], or clipart to sketch images [7]), rather than distribution shifts across space and time in the real world. In this work, we investigate the effectiveness of UDA techniques at the practical application of adapting trained object recognition models to novel geographies.

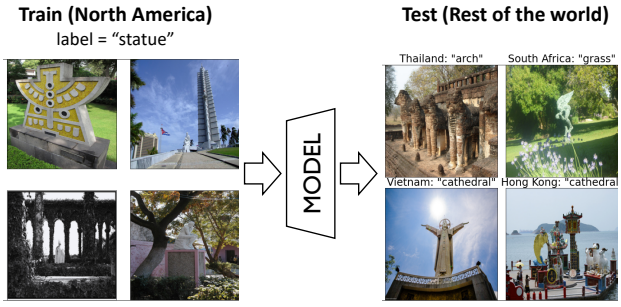


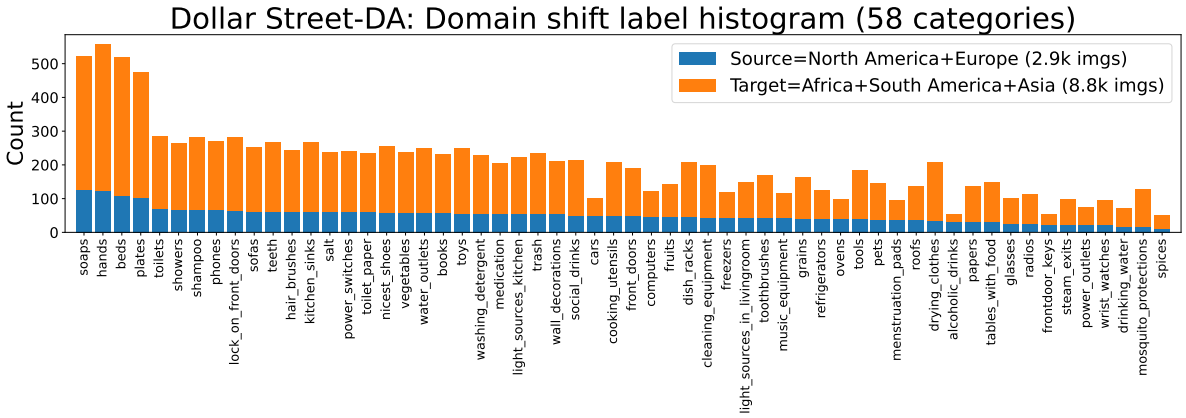
Figure 1. Modern computer vision datasets overrepresent the developed world [1]. This leads object recognition models trained on them (left) to underperform on images from novel geographies [2] (right – we show country of origin and model prediction above each image). In this work we investigate the effectiveness of domain adaptation [3] methods in bridging this performance gap.

Geographical domain adaptation presents two novel challenges beyond shifting data distributions: *context shift* and *subpopulation shift*. Context shift arises from a change in visual context for a given category across geographies (*e.g.* predominantly indoor v/s outdoor basketball courts). Subpopulation shift arises from a change in within-category data distributions (*e.g.* for a 'toothbrush' category, the relative proportion of electric v/s mechanical varieties may change across geographies). In our experiments, we demonstrate the inefficacy of conventional adaptation strategies in addressing these additional challenges.

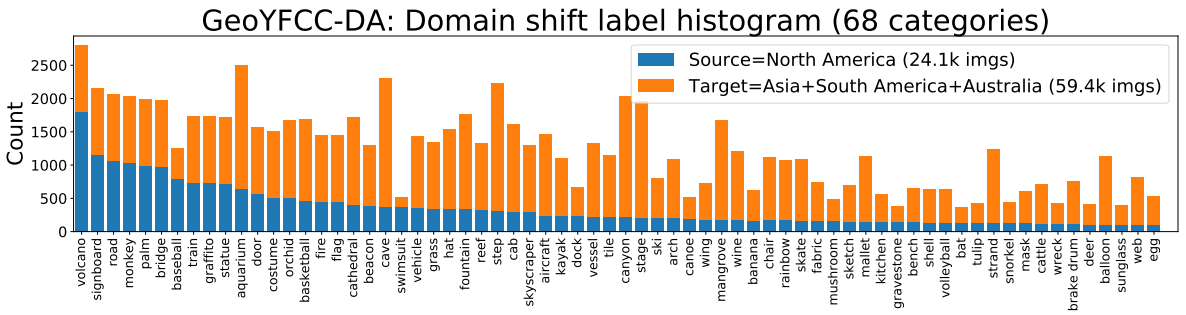
Some prior work has studied the problem of transferring deep visual models to new geographies. De Vries *et al.* [2] benchmark the drop in performance of publicly available vision API's on images from diverse geographies from the Dollar Street dataset [8], but do not propose a mitigation strategy. Recently, Dubey *et al.* [9] formulate this as a domain generalization problem and propose a solution that makes use of auxiliary target domain embeddings. We instead pose the problem as one of *domain adaptation* so as to leverage the full potential of unlabeled target data by allowing model updates on it. We make the following contributions:

1. To study the Geographical DA problem, we propose

[*]Work done partially as intern at Salesforce Research.
[†]Work done at Salesforce Research.



(a) Dollar Street-DA: {North America, Europe} → {Asia, Africa, South America}



(b) GeoYFCC-DA: {North America} → {Asia, Australia, South America}

Figure 2. Source (blue) and target (orange) label histograms for proposed Geographic DA shifts based on **(a)** Dollar Street [8] **(b)**: GeoYFCC [9].

two adaptation shifts curated from the Dollar Street [8] and GeoYFCC [9] datasets.

2. We validate the existence of context and subpopulation shift within these shifts, and experimentally verify that they pose significant challenges to model transfer.

3. We benchmark the performance of representative domain adaptation techniques from the literature on these shifts and find them to achieve limited success, illustrating the need for specialized adaptation solutions for Geographical DA.

2. Related Work

Unsupervised domain adaptation (UDA). UDA seeks to transfer a model trained on a labeled source to an unlabeled target domain, primarily via minimizing domain discrepancy statistics [10], domain-adversarial learning [4], or self-training [5]. We formulate the problem of transferring trained image classification models to images from unseen geographies as UDA, and study how standard DA methods fare in this setting.

Geographical transfer learning. Geographical transfer learning has received limited attention. Wang *et al.* [11] focus on cross-country adaptation of 3D object detectors, and propose a simple correction solution based on differences in average car sizes. Dubey *et al.* study the problem of geographical domain *generalization*, and propose an adaptive solution that uses auxiliary target domain embeddings but unlike our setting does not allow training on unlabeled target data. Concurrent work [12] extends the recently proposed WILDS [13] domain generalization benchmark to the unsupervised DA setting, and include one shift (FMoW [14]) for geographical adaptation of models trained for land use prediction from satellite imagery. In contrast, we study geographical adaptation of object recognition models trained on standard internet imagery from the Dollar Street and YFCC datasets, which poses unique challenges of context and subpopulation shift.

Context and subpopulation shift. Singh *et al.* [15] study the problem of contextual biases learned by deep models based on frequently co-occuring categories. Some recent works [16, 17] study the problem of minimizing contextual biases when learning self-supervised representations from

Figure 3. Context shift for select categories across domains. **Left:** Dollar Street-DA **Right:** GeoYFCC-DA.

scene-level imagery. Recent work proposes the BREEDS benchmark [18] to study model robustness against subpopulation shift—the ability to generalize to novel data subclasses not seen during training. Cai *et al.* [19] study propose a input-consistency based label propagation algorithm to overcome subpopulation shift. To our knowledge, we are the first to study these challenges in the context of geographical DA.

3. Benchmarks and Challenges

We first present our two shifts for geographical domain adaptation curated from the Dollar Street and GeoYFCC datasets. We describe our curation process and analyze the characteristics of each geographical domain shift. We then describe and visualize the context and subpopulation shift present in these benchmarks.

3.1. Benchmarks

Dollar Street-DA. The Dollar Street dataset was collected as part of the GapMinder project with the aim of using "photos as data to kill country stereotypes". It contains photographs and videos of everyday objects from peoples' homes spanning 66 unique countries. We restrict our study to image data and download images belonging to 128 unique categories. We filter out categories that are scene-level or too broad ("agriculture lands", "play areas") or abstract / subjective ("most loved items", "things I dream of having"), as well as categories with less than 50 images, resulting in 62 categories. We further deduplicate the dataset and merge some highly similar categories (e.g., "plates of food" and "plates"), leaving us with images of 58 unique and distinct curated categories from 60 countries. We set up an adaptation problem from North America and Europe as source (2930 images) and Africa, South America, and Asia as target (8813 images). Fig. 2a presents a label histogram of each domain.

GeoYFCC-DA. The GeoYFCC dataset [9] contains 1.1 million images from 62 countries curated from the subset of YFCC100M [20] images with geotags that were then automatically labeled based on keyword matching of image tags against ImageNet-5K categories excluding those in ILSVRC12 [21]. We create an adaptation problem from countries in North America as source and countries in Asia, South America, and Australia as our target domain. Due to the automatic labeling pipeline, we notice a large amount of label noise in the dataset and take two measures to curate the dataset further: i) We train a ResNet50 [22] model on the source domain and measure heldout test accuracy, and only retain classes with $> 25\%$ accuracy. ii) We manually inspect 100 random qualitative examples from the source and target domains for the remaining categories and exclude categories with significant label noise. At the end of this process, we select 68 categories with 24.1k images in the source domain and 59.4k images in the target domain. See Fig. 2b for a label histogram of each domain.

3.2. Challenges: Context and subpopulation shift

Notation. Let $\mathcal{X}$ and $\mathcal{Y}$ denote input and ouput spaces, with the goal being to learn a convolutional neural network $h : \mathcal{X} \to \mathcal{Y}$ parameterized by Θ. In unsupervised DA we are given access to labeled source examples $(\mathbf{x}_{\mathcal{S}}, y_{\mathcal{S}}) \sim \mathcal{P}_{\mathcal{S}}(\mathcal{X}, \mathcal{Y})$, and unlabeled target examples $\mathbf{x}_{\mathcal{T}} \sim \mathcal{P}_{\mathcal{T}}(\mathcal{X})$, where $\mathcal{S}$ and $\mathcal{T}$ denote the source and target domains. The goal is to maximize model accuracy on the target domain, and we consider adaptation of models trained to perform K-way object recognition: the inputs $\mathbf{x}$ are images, and labels y are categorical variables $y \in \{1, 2, .., K\}$.

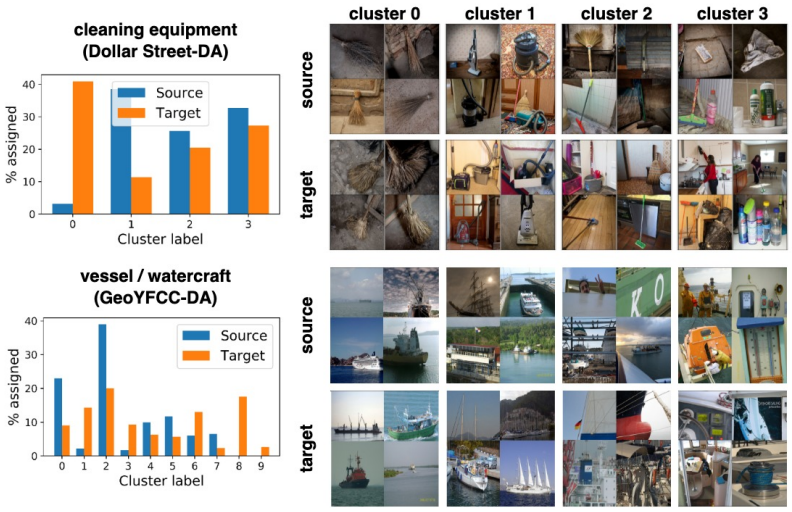


Figure 4. Subpopulation shift for select categories across domains. We plot normalized cluster assignments per-domain as approximate subpopulation distributions – blue denotes within-category subpopulation distribution on source and orange denotes target. As seen, subpopulation distributions shift significantly across domains. On the right we visualize random images from some of the discovered clusters, and verify that they generally correspond to distinct subpopulations.

Data and label distribution shift. As in conventional domain adaptation, geographical DA also presents data distribution shift ($\mathcal{P}_\mathcal{S}(\mathbf{x}) \neq \mathcal{P}_\mathcal{T}(\mathbf{x})$) as object appearances change across geographies (see Fig. 1), and label distribution shift ($\mathcal{P}_\mathcal{S}(y) \neq \mathcal{P}_\mathcal{T}(y)$), as task label distributions change across domains (see Fig. 2).

In addition, geographical adaptation presents two new challenges: context and subpopulation shift.

Context shift. We define context $c(\mathbf{x})$ for image $\mathbf{x}$ with label y as the *task-irrelevant* information in the image—this loosely corresponds to the background or surroundings of the object of interest. We define context shift as $\mathcal{P}_\mathcal{S}(c(\mathbf{x})|y) \neq \mathcal{P}_\mathcal{T}(c(\mathbf{x})|y)$, representing a change in object context across geographical domains.

In Fig. 3 we show qualitative examples of context shift within our proposed Dollar Street-DA and GeoYFCC-DA shifts for a few categories. For example, we find that in Dollar Street-DA, most "toothbrush" images in the source domain tend to be photographed inside bathrooms, whereas the surroundings in the target domains are *significantly* more diverse (*e.g.* walls and roofs). We see similar trends in the GeoYFCC-DA shift (*e.g.* indoor v/s outdoor basketball games). As deep neural networks are known to often employ "shortcut learning" [23] of potentially spurious features (*e.g.* object backgrounds) to make predictions, we hypothesize (and experimentally verify in Sec. 4.4) that such a context shift will present a challenge to visual recognition models deployed in new geographies.

Subpopulation shift. We define subpopulation shift as $\mathcal{P}_\mathcal{S}(\mathbf{x}|y) \neq \mathcal{P}_\mathcal{T}(\mathbf{x}|y)$, representing a change in within-category distribution across domains.

In Fig. 4, we show examples of subpopulation shift in the Dollar Street-DA and GeoYFCC-DA benchmarks. In the absence of subpopulation-level annotations, we use a simple strategy to obtain *approximate* annotations: we use a pretrained model (ResNet50 [22] trained on ImageNet) to extract features for source and target images of a given category, and perform agglomerative clustering on the combined set of embeddings. We then use the inferred cluster assignments as subpopulation annotations. We also plot the normalized within-class distribution of cluster assignments on the source and target domains, and measure the Wasserstein distance between the two as a measure of the degree of subpopulation shift.

As seen, this simple strategy discovers distinct clusters corresponding to semantically distinct subpopulations: *e.g.* for "cleaning equipment" on Dollar Street-DA we discover separate clusters roughly corresponding to brooms, vaccuum cleaners, mops, and miscellaneous cleaning items. Crucially, we find that the *intra-class distribution* of many categories changes significantly across geographies (*e.g.* brooms make up a significantly larger proportion of cleaning equipment in the target domain than in the source).

To quantitatively validate the subpopulation shift, we compute the *degree* of per-class subpopulation shift: for each class, we compute the normalized subpopulation distribution per-domain (as visualized in Fig. 4, left), measure the cross-domain Wasserstein

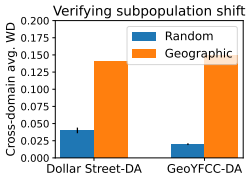


Figure 5. Verifying subpopulation shift.

Method	Dollar Street-DA	GeoYFCC-DA
source	54.66±0.62	42.88
target oracle*	67.73± 0.30	56.78
MMD [10]	55.77±0.75	43.53
DANN [4]	54.80±0.38	42.64
SENTRY [5]	55.73±0.34	42.58
SST	58.71±0.53	45.22

Table 1. Average accuracy on the target test set (20%) for the Dollar Street-DA and GeoYFCC-DA shifts. * denotes that the target oracle was trained on target data non-overlapping with the test set (80%) whereas DA methods were adapted without labels on the entire target dataset.

distance, and average across classes. Fig. 5 presents results for this measure for our proposed geographical shifts versus randomly constructed source and target domains of the same size, for the Dollar Street and GeoYFCC datasets. As seen, geographical shifts lead to *significantly* higher subpopulation shift.

4. Experiments

4.1. Setup

To account for label imbalance, we report per-class average accuracy as our metric. As described in Sec. 3, we consider adaptation on two shifts:

Dollar Street-DA: We consider images from North America and Europe as the source domain (2.9k) and images from Asia, Africa, and South America as the target domain (8.8k).

GeoYFCC-DA: We consider images from North America as the source domain (24.1k) and images from Asia, Australia, and South America as the target domain (59.4k).

On both shifts we create a 90%-10% train-test split on the source domain and report transfer performance on 20% heldout target data, but use the entire target dataset for unsupervised adaptation. On Dollar Street-DA, due to the relatively small size of the dataset, we report performance mean and standard deviation over three experimental runs.

4.2. Domain Adaptation Baselines

We benchmark 4 representative DA methods from the literature on our Geographical DA shifts: one domain discrepancy-based method, one domain adversarial method, and two self-training based methods.

1) MMD [10]: Aligns domains by computing mean source and target embeddings in a reproducing kernel hilbert space and minimizing their distance as a maximum mean discrepancy measure.

2) DANN [4]: Domain-adversarial neural networks adversarially learn a domain discriminator to distinguish source and target features against a feature encoder that is trained to fool the discriminator.

3) SENTRY [5]: SENTRY measures model predictive consistency across randomly augmented versions of each target image and selectively increases predictive entropy on highly consistent instances, while decreasing it on highly inconsistent ones. SENTRY also uses pseudolabel-based approximating class balancing on the target domain, and employs a slightly modified ResNet50 architecture [24].

4) Selective Self-training (SST): We implement a simplified self-training baseline that self-trains against predicted labels on target instances on which the model is atleast 90% confident, and also employs pseudolabel-based target class balancing and a modified ResNet50 architecture [24].

Finally, we also report performance for a *target oracle* that is trained in a supervised fashion on the target domain. The target oracle is meant to represent a performance *upper bound* in the absence of domain shift.

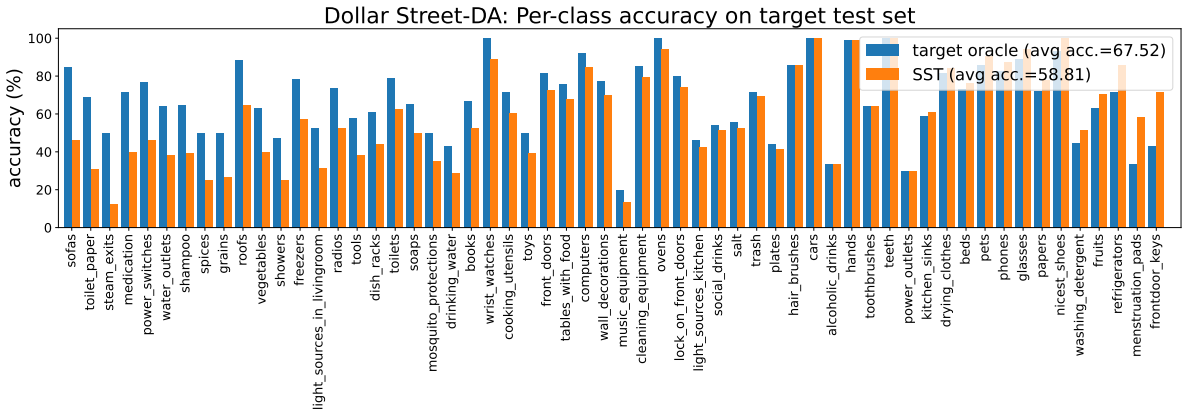
Implementation details. We use a ResNet50 [22] as our CNN architecture, and for SENTRY and SST use a modified few-shot variant [24] that replaces the last linear layer with a K-way (where K is the number of classes) fully-connected layer without bias. Input activations to this layer are L_2-normalized and its output is passed into a softmax layer with a temperature of 0.05. For optimization, we use Adam [25] with a learning rate of 0.001 and weight decay of $5e^{-4}$. We use a batch size of 128 across experiments. All DA methods are applied to a source model that is first trained using supervised training on the source domain for 50 epochs. We employ 100 epochs of adaptation on Dollar Street-DA and 40 epochs on GeoYFCC-DA. All methods additionally minimize a supervised cross-entropy loss on source labels during adaptation (with a loss weight of 1.0 and 0.1 on Dollar Street-DA and GeoYFCC-DA respectively). To combat label distribution shift, we follow Tan *et al.* [26] and use class-balanced sampling on the source domain across experiments.

4.3. Results

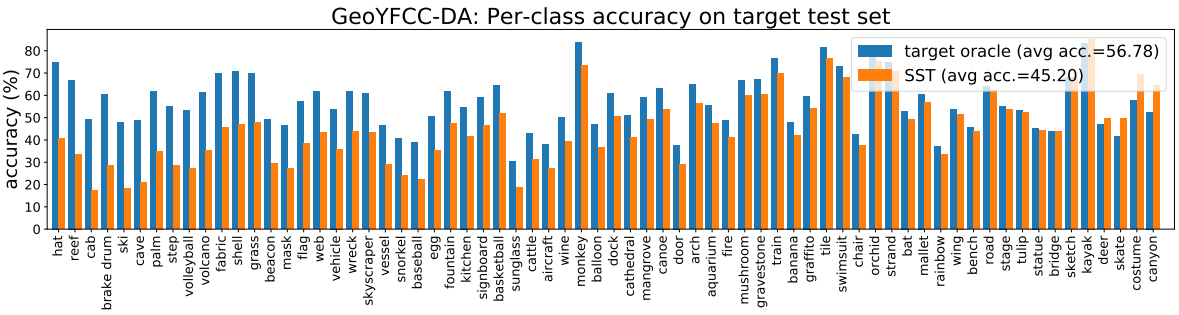
Table 1 presents average accuracy on the target test set for the Dollar Street-DA and GeoYFCC-DA shifts. We observe:

▷ **Geographical shifts lead to significant performance drops (Row 1 v/s 2).** As seen, the target oracle achieves 67.7% and 56.78% whereas the source model achieves a significantly lower performance of 54.66% (-13.1%) and 42.88% (-13.9%). Clearly, geographical variations pose a significant challenge to transfer.

▷ **DA methods offer limited improvements (Rows 3-6).** All the benchmarked methods achieve limited success at geographical adaptation, sometimes performing no better than the source model. Surprisingly, we find the simple SST method to achieve the best (albeit small) improvement over the source model (+4.1% / +2.3%), but still well short of the target oracle (-9% / -11.6%). Altogether, these results



(a) Dollar Street-DA



(b) GeoYFCC-DA

Figure 6. Per-category target test accuracy for a model trained on **blue:** target train set and **orange:** source train set and adapted to the target domain via the SST method. Categories are ordered in decreasing order of accuracy drop.

indicate the need for specialized adaptation solutions for the Geographical DA problem.

4.4. Analyzing failure modes

We now analyze the performance of our best-performing SST model and contrast it with the target oracle.

Per-class performance. In Fig. 6 we show per-category accuracies on the target test set for i) **blue:** a *target oracle* model trained on the target train set and ii) **orange:** the best performing SST model which was trained on the source train set and adapted to the entire target domain. We order categories in descending order of accuracy drop. As seen, the SST model performance lags behind the target oracle on most categories. We further analyze failure modes arising from context shift and subpopulation shift:

Context shift. In Fig. 7 we show examples of model errors on the target domain arising from context shift. We also visualize GradCAM [27] explanations along side each image, and find that, in most cases, the model makes erroneous predictions on target images with contexts that are uncommon in the source domain (see Fig. 3 for source examples) while fixating on spurious background features.

Subpopulation shift. To verify that subpopulation shift is a failure mode, we measure per-class average subpopulation accuracy, using the approximate subpopulation-level annotations described in Sec. 3.2. We then measure the correlation between the drop in per-class average subpopulation accuracy for the SST model compared to the target oracle (that does not experience subpopulation shift), against the degree of subpopulation shift measured via the per-category average of cross-domain Wasserstein distance between subpopulation distributions. On the 40 classes with the highest subpopulation shift in Dollar Street-DA, we observe a Pearson correlation coefficient of 0.44, and 0.39 on GeoYFCC-DA. This indicates the tendency of the adapted model to underperform to a larger degree—when compared to the target oracle—on categories with high subpopulation shift.

5. Limitations & Conclusion

Our work has some important limitations: we do not consider semantic drift, where the meaning of a category itself may change across geographies *e.g.* a "chair" in one country might be considered as a "sofa" in another. We also restrict our study to adaptation across continent-level shifts, but analyzing shifts across a more fine-grained (*e.g.* country) level is also potentially valuable. Moreover, variation in

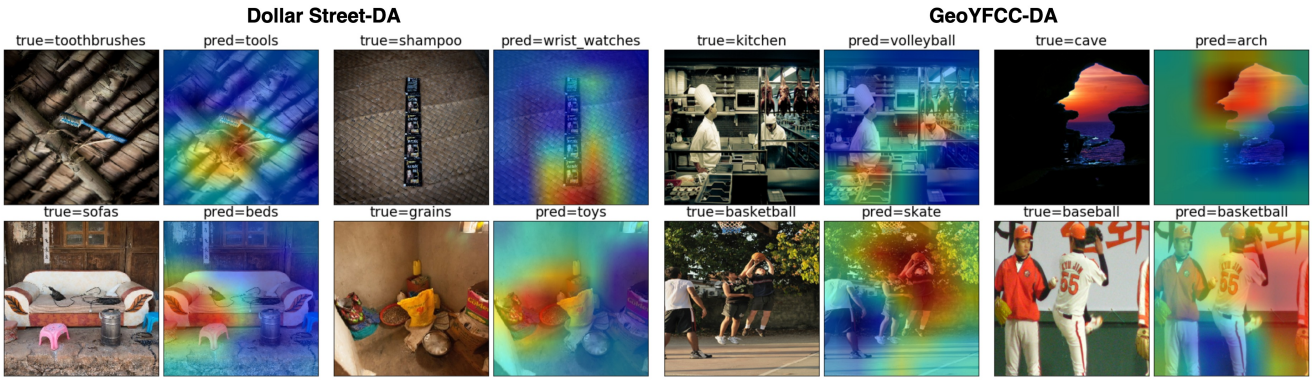


Figure 7. Context shift as a failure mode for Geographical DA. We visualize incorrect predictions from the best-performing SST model on the target test set alongside visual explanations generated with GradCAM [27] for Dollar Street-DA (left) and GeoYFCC-DA (right). As seen, the model frequently attends to spurious background features and makes incorrect predictions.

visual appearance within a geography can sometimes be larger than that across geographies, due to other confounding factors like income and demographics—we do not study these differences here. Finally, in the absence of human annotations, we are restricted to using inferred subpopulation annotations for our analysis.

To summarize, we studied the problem of domain adaptation of trained object recognition models to new geographies, and investigated the effectiveness of off-the-shelf adaptation methods. We proposed two shifts to study this problem and demonstrated the existence of two unique challenges: cross-domain context shift and subpopulation-shift. We found existing DA methods to offer limited success at Geographical DA, suggesting the need for future work to develop specialized adaptation solutions for this important but understudied problem.

References

[1] S. Shankar, Y. Halpern, E. Breck, J. Atwood, J. 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. 1

[2] T. De Vries, I. Misra, C. Wang, and L. Van der Maaten, "Does object recognition work for everyone?," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 52–59, 2019. 1

[3] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *European conference on computer vision*, pp. 213–226, Springer, 2010. 1

[4] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*, pp. 1180–1189, PMLR, 2015. 1, 2, 5

[5] V. Prabhu, S. Khare, D. Kartik, and J. Hoffman, "Sentry: Selective entropy optimization via committee consistency for unsupervised domain adaptation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8558–8567, 2021. 1, 2, 5

[6] X. Peng, B. Usman, N. Kaushik, D. Wang, J. Hoffman, and K. Saenko, "Visda: A synthetic-to-real benchmark for visual domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2021–2026, 2018. 1

[7] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang, "Moment matching for multi-source domain adaptation," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1406–1415, 2019. 1

[8] https://www.gapminder.org/dollar-street. 1, 2

[9] A. Dubey, V. Ramanathan, A. Pentland, and D. Mahajan, "Adaptive methods for real-world domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14340–14349, 2021. 1, 2, 3

[10] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *International conference on machine learning*, pp. 97–105, PMLR, 2015. 2, 5

[11] Y. Wang, X. Chen, Y. You, L. E. Li, B. Hariharan, M. Campbell, K. Q. Weinberger, and W.-L. 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*, pp. 11713–11723, 2020. 2

[12] S. Sagawa, P. W. Koh, T. Lee, I. Gao, S. M. Xie, K. Shen, A. Kumar, W. Hu, M. Yasunaga, H. Marklund, *et al.*, "Extending the wilds benchmark for unsupervised adaptation," in *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021. 2

[13] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, *et al.*, "Wilds: A benchmark of in-the-wild distribution shifts," in *International Conference on Machine Learning*, pp. 5637–5664, PMLR, 2021. 2

[14] G. Christie, N. Fendley, J. Wilson, and R. Mukherjee, "Functional map of the world," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6172–6180, 2018. 2

[15] K. K. Singh, D. Mahajan, K. Grauman, Y. J. Lee, M. Feiszli, and D. 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*, pp. 11070–11078, 2020. 2

[16] R. R. Selvaraju, K. Desai, J. Johnson, and N. Naik, "Casting your model: Learning to localize improves self-supervised representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11058–11067, 2021. 2

[17] S. Mo, H. Kang, K. Sohn, C.-L. Li, and J. Shin, "Object-aware contrastive learning for debiased scene representation," *Advances in Neural Information Processing Systems*, vol. 34, 2021. 2

[18] S. Santurkar, D. Tsipras, and A. Madry, "Breeds: Benchmarks for subpopulation shift," in *International Conference on Learning Representations*, 2020. 3

[19] T. Cai, R. Gao, J. Lee, and Q. Lei, "A theory of label propagation for subpopulation shift," in *International Conference on Machine Learning*, pp. 1170–1182, PMLR, 2021. 3

[20] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "Yfcc100m: The new data in multimedia research," *Communications of the ACM*, vol. 59, no. 2, pp. 64–73, 2016. 3

[21] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015. 3

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016. 3, 4, 5

[23] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," *Nature Machine Intelligence*, vol. 2, no. 11, pp. 665–673, 2020. 4

[24] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. F. Wang, and J.-B. Huang, "A closer look at few-shot classification," in *International Conference on Learning Representations*, 2018. 5

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014. 5

[26] S. Tan, X. Peng, and K. Saenko, "Class-imbalanced domain adaptation: an empirical odyssey," in *European Conference on Computer Vision*, pp. 585–602, Springer, 2020. 5

[27] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017. 6, 7