PREDICTING OUT-OF-DOMAIN PERFORMANCE UNDER GEOGRAPHIC DISTRIBUTION SHIFTS

Haoran Zhang
Harvard University
haoran_zhang@g.harvard.edu

Konstantin Klemmer
Microsoft Research
koklemmer@microsoft.com

Esther Rolf
University of Colorado Boulder
esther.rolf@colorado.edu

David Alvarez-Melis
Harvard University
dam@seas.harvard.edu

ABSTRACT

In machine learning for geographic data, we often observe differences in data availability and distribution shifts across distinct geographic units, e.g., continents. This is a common challenge in remote sensing tasks, such as crop yield forecasting or flood mapping. In many of these scenarios, we have models trained on a data-rich region and apply domain adaptation to transfer predictive capabilities to the target region. However, the effectiveness of domain transfer can suffer from distribution shifts, posing critical challenges for model deployment. In this work, we show that, even in the absence of labels, certain domain distance measures, based on image and location embeddings, can serve as a proxy measure for transfer performance. We further highlight this capacity on a set of real-world geographic adaptation datasets, spatial splits for domains, and models for adaptation training.

1 INTRODUCTION

Machine learning on satellite images has been widely used for a range of applications relevant to climate change adaptation and mitigation, such as crop yield prediction (Ansarifar et al., 2021), disaster forecasting (Linardos et al., 2022), and pollution monitoring (Hu et al., 2017). However, the quantity, quality, and composition of satellite image data is not balanced across geographical regions due to many aspects, including population, environment, socioeconomic factors, etc. (Martin Sudmanns & Lang, 2020; Rolf et al., 2024). Hence, training models independently for each of the regions can lead to divergent performance given the discrepancies in the nature and availability of data. Instead, a common setting is to train models on data-rich regions to learn representations from satellite images and transfer predictive capabilities to the target data-poor regions via domain adaptation.

This approach is not straightforward as distribution shifts exist between geographic regions, with conditions in satellite images such as architectural styles, landscapes, and vegetation differing due to different climates, cultures, and environmental conditions (Federici et al., 2021). These shifts can pose significant challenges for machine learning models, as they may struggle to generalize effectively to target domains with substantially different ground conditions from source regions (Rolf et al., 2024). Therefore, it might not be ideal to blindly transfer models to out-of-distribution geographical regions for optimal performance.

We posit that distances between domain-specific data distributions can serve as a good indicator of the effectiveness of transfer between domains. Intuitively, a model is more likely to transfer adequately between domains that are distributionally similar than across those that are dissimilar. To explore the feasibility and effectiveness of using distance measures to predict out-of-distribution transfer performance for satellite imaging tasks, we design experiments to test and analyze the relationship between different distances and performance changes in domain adaptation.

2 METHODOLOGY

We seek a notion of distance measure that can serve as reliable predictors of domain adaptation performance for transferring a discriminative model between domains. We first introduce the distribution shift problem in satellite image data (and geographic data more generally). Then we discuss the two main steps of our methodology: **distance computation** between geographic domains and model training for **domain adaptation**.

Problem Formulation For satellite images, distribution shifts may appear when data used for training and evaluating models are collected at different locations, as visual markers in images can change due to human activity and environmental processes. In addition, climatic, social, and environmental factors can lead to shifts in marginal distribution on labels in the collected data.

In this work, we focus mainly on FMoW-Wilds, a classification dataset of land use and building function with RGB satellite image inputs. Figure 1 shows an example of distribution shifts in satellite image classification.

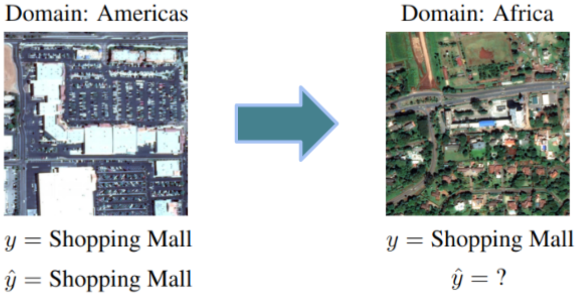


Figure 1: Example of geographic distribution shift in FMoW-wilds dataset: Two images from the class “Shopping Mall” appear vastly different depending on their geographic domain.

Distance Computation To obtain the distance between any domain pairs, we first compute the pairwise distance matrix for all the data points across the domains, and we then aggregate the distances through averaging or optimizing a cost objective over distribution. Given two domains from the same dataset, we propose 3 different sets of distances to compute:

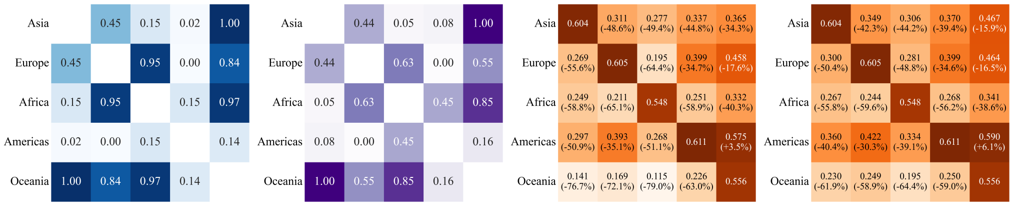
- **Average cosine distance:** angular distance between embedding vectors.
- **Wasserstein distance:** optimal transport based distance between embedding vectors.
- **Average arc distance:** geographic distance between image locations.

Specific definition and formulas of these distances are included in Appendix A.1. To make these distance measures comparable to each other, we apply normalization so that results for each type of distances have range 0 to 1.

Domain Adaptation In the step of domain adaptation, we fit pretrained image models on a source domain and evaluate the model performance on a target domain. Specifically, given a set $\mathbb{D}$ of k domains in the dataset $\{\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_k\}$, we fine-tune one image model $\mathcal{M}_i$ for each domain $\mathcal{D}_i$. Then for every model $\mathcal{M}_i$, we firstly evaluate and record its performance on the source domain $\mathcal{D}_i$ it is trained on. Next, we evaluate the performance under zeroshot and fewshot settings on each of the other domains except the source domain, i.e., domain $\mathcal{D}_j \in \mathbb{D} \backslash \mathcal{D}_i$. We then correlate transfer performance scores with domain distances.

3 EXPERIMENTS

We conduct several experiments to evaluate our central hypothesis. First, we introduce the different embeddings we use for computing distances and the datasets for domain adaptation. For model



(a) ResNet18 embeddings (b) SatCLIP embeddings (c) 0-shot (d) 10-shot

Figure 2: The left two figures are normalized pairwise Wasserstein distances of embeddings between continent domains in FMoW-wilds dataset. The right two figures are domain adapatation accuracy on FMoW-wilds dataset with DenseNet121. Rows represent source domains which image models are initially trained on. Columns represent target domains which image models are evaluated on. Percentages in the parentheses following the accuracy values are the relative drop in classification accuracy on the test set compared with performance of models trained on the target domain.

training details and other factors we explore such as splitting criteria for domains and models used for domain adaptation, see our Appendix B.

Embeddings. To generate embeddings, we apply pretrained models to the data, including image-based models and location-based models. First, we use pretrained image models, including ResNet18, ResNet50 (He et al., 2016) and ViT-Small (Dosovitskiy et al., 2021), to obtain vector representations of satellite images. Besides, geographic context can be relevant in many real-world modeling tasks, as similar images from different regions may have subtle differences yet belong to different classes. To obtain location-specific characteristics, we use Satellite Contrastive Location-Image Pretraining (SatCLIP), a location encoder model pretrained by matching a large dataset of satellite images with their geographic coordinates (Klemmer et al., 2025). With global geographic coverage, SatCLIP projects longitude and latitude pairs into rich representations that correspond to image features expected at a given location.

Experimental Datasets. We primarily focus on distance analysis and model performance on the FMoW-Wilds dataset, a classification dataset that contains satellite images of 62 different functional classes of built-environment and land use (Koh et al., 2021). We use a subset of the dataset that contains approximately 200 thousand data points, and we follow the same spatial splits in the original dataset and include 5 main geographical continents (Asia, Europe, Africa, Americas, and Oceania) as domains. The downstream task is to predict the respective class from image features (classification).

4 Results and Analysis

In this section, we present the main results of our study. A more comprehensive set of results and figures can be found in Appendix C.

Domain Distance. Figure 2a and Figure 2b show the Wasserstein pairwise distances across different pairs of continent domains in FMoW-wilds dataset. Compared with cosine distances, Wasserstein distances are larger where embedding distributions between domains have more discrepancies. Normalized distances from image embeddings are more concentrated towards the lower and upper bound ends, where SatCLIP embeddings yield smoother distance distributions. Arc distance, on the other hand, exhibits a completely different pattern as raw geographic distance.

Domain Adaptation. In Figure 2c and Figure 2d, we show primary results of domain adaptation performance of DenseNet121 model in FMoW-wilds dataset under zeroshot and fewshot settings. Entries on the diagonal show test results of models pretrained on the same domain. To make the accuracy values comparable across different models and different domain pairs, we further compute relative change in the performance compared with the models trained and evaluated on the target domain, given in parentheses.

4.1 Analysis

We now compare the relative drop in test performance against the corresponding distances between the two domains, accounting for different factors that may affect the relationship. We show our main findings in Figure 3 with analysis across different types of distance measures.

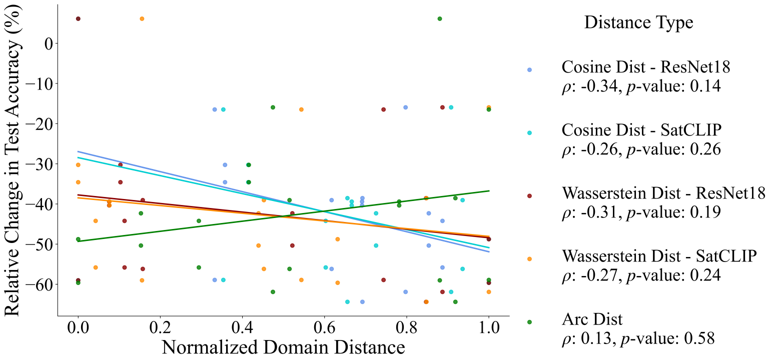


Figure 3: Comparison of domain distance and relative change of domain adaptation performance in FMoW-wilds dataset across different types of distance measures.

Both cosine and Wasserstein distances follow a downward trend, where larger distances between domains would lead to a greater drop in adaptation performance, and the correlation is moderate (with p-values between 0.14 and 0.27) across image and location embeddings. This suggests that cosine distance and Wasserstein distance can be predictive of the potential changes in domain adaptation performance. On the other hand, arc distance exhibits a weak upward trend suggesting an unstable and insignificant correlation. This is not surprising: satellite images taken at two far apart locations can still be very similar to each other, and raw geographic distance provides an incomplete signal.

It is important to reiterate here that distances based on SatCLIP embeddings represent semantic features obtained from satellite imagery, while only requiring geographic coordinates as inputs. Compared with image embeddings, which require a full pass through all the image data, this becomes a much more convenient method to compute and obtain results for performance analysis. Furthermore, SatCLIP embeddings can be obtained for any location over landmass globally.

Lastly, we expand our analysis to potential confounding factors affecting our hypothesized relationship, including datasets, geographic domains, and models used for domain adaptation tasks. The results of ablation studies are included in Appendix C.3. Despite the changes in different components of a domain adaptation task, we observe the common relationship that domain pairs with larger distances tend to exhibit greater performance drops when adapting from one domain to another.

5 Conclusion

Our analysis demonstrates that distance measures between two geographic domains can serve as a predictor for performance change in adaptation from source domain to target domain, and larger domain distances typically imply decreasing domain adaptation performance. This finding holds across different datasets, different definitions of geographic domains (continents and biomes), and different predictive models. In the future, we plan to leverage this finding for selecting optimal datasets from all available source domains to achieve maximize adaptation performance on the target domain.

Despite the contributions and promising future steps, there are some limitations in our experiments. We use the average values of pairwise distance matrix for cosine distance and arc distance, but they are point estimates of the distances between data, which cannot accurately describe the distribution of distances across two domains. Therefore, other choices of estimates that account for the distribution should be considered. The spatial splits for domains are at relatively large scale (e.g., continents or biomes), which could make our correlation findings less effective. Moreover, we should take into account of a larger variety of datasets and models in each experiment setting to ensure the robustness and reliability of our conclusion.

REFERENCES

Javad Ansarifar, Lizhi Wang, and Sotirios V. Archontoulis. An interaction regression model for crop yield prediction. *Scientific Reports*, 11(1):17754, 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-97221-7. URL https://doi.org/10.1038/s41598-021-97221-7.

Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *CVPR*, 2018.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

Marco Federici, Ryota Tomioka, and Patrick Forré. An information-theoretic approach to distribution shifts. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 17628–17641. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/93661c10ed346f9692f4d512319799b3-Paper.pdf.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Ke Hu, Ashfaqur Rahman, Hari Bhrugubanda, and Vijay Sivaraman. Hazeest: Machine learning based metropolitan air pollution estimation from fixed and mobile sensors. *IEEE Sensors Journal*, 17(11):3517–3525, 2017. doi: 10.1109/JSEN.2017.2690975.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, 2017. doi: 10.1109/CVPR.2017.243.

Konstantin Klemmer, Esther Rolf, Caleb Robinson, Lester Mackey, and Marc Rußwurm. SatCLIP: Global, general-purpose location embeddings with satellite imagery. *AAAI*, 2025.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*, 2021.

Vasileios Linardos, Maria Drakaki, Panagiotis Tzionas, and Yannis L. Karnavas. Machine learning in disaster management: Recent developments in methods and applications. *Machine Learning and Knowledge Extraction*, 4(2):446–473, 2022. ISSN 2504-4990. doi: 10.3390/make4020020. URL https://www.mdpi.com/2504-4990/4/2/20.

Hannah Augustin Martin Sudmanns, Dirk Tiede and Stefan Lang. Assessing global sentinel-2 coverage dynamics and data availability for operational earth observation (eo) applications using the eo-compass. *International Journal of Digital Earth*, 13(7):768–784, 2020. doi: 10.1080/17538947.2019.1572799. URL https://doi.org/10.1080/17538947.2019.1572799.

National Geographic Society. The five major types of biomes. https://education.nationalgeographic.org/resource/five-major-types-biomes, 2024. Accessed: 2024-11-18.

Esther Rolf, Jonathan Proctor, Tamma Carleton, Ian Bolliger, Vaishaal Shankar, Miyabi Ishihara, Benjamin Recht, and Solomon Hsiang. A generalizable and accessible approach to machine learning with global satellite imagery. *Nature Communications*, 12, 2021. doi: 10.1038/s41467-021-24638-z.

Esther Rolf, Konstantin Klemmer, Caleb Robinson, and Hannah Kerner. Position: mission critical - satellite data is a distinct modality in machine learning. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

A Methodology

A.1 Distance Computation

A.1.1 Average Cosine Embedding Distance

Cosine distance measures the differences in the embedding space of geographical data. Let $u_{x_{ik}}$ and $v_{x_{jl}}$ be the embeddings of data points x_{ik} from domain $\mathcal{D}_i$ and x_{jl} from domain $\mathcal{D}_j$. The cosine embedding distance is defined as

$$c(u_{x_{ik}}, v_{x_{jl}}) = 1 - \frac{u_{x_{ik}} \cdot v_{x_{jl}}}{\|u_{x_{ik}}\|_2 \|v_{x_{jl}}\|_2}$$

We take the embeddings of input data and compute cosine distances across each pair of the embeddings from different domains. To further obtain the distance between two domains, we take the average over the matrix of cosine embedding distances.

A.1.2 Wasserstein Distance

The Kantorovich formulation of the Optimal Transport (OT) problem is defined as

$$\min_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} c(x,y) d\pi(x,y),$$

where $c(\cdot,\cdot)$ is a cost function, and $\Pi(\mu,\nu)$ represents the set of couplings satisfying

$$\begin{aligned}\Pi(\mu,\nu) = \{\pi \in \mathcal{P}(\mathcal{X}\times\mathcal{X}) : &\int_{\mathcal{X}} \pi(x,y)dy = \mu(x), \\ &\int_{\mathcal{X}} \pi(x,y)dx = \nu(y)\}\end{aligned}$$

We use cosine distance as the cost metric, and the formulation then quantifies the transformation cost between distributions in the embedding space, while accounting for distribution geometry and underlying data structure.

For computational complexity concerns, we choose to use the entropic regularized version of the Wasserstein distance, the Sinkhorn distance, formulated with an additional penalty term in the optimization problem

$$\min_{\pi \in \Pi(\mu,\nu)} \int c(x,y)d\pi(x,y) + \epsilon \int \pi(x,y) \log \pi(x,y) dxdy,$$

where ϵ is the regularization parameter.

In implementation, we make use of the ot.solve function in the Python Optimal Transport (POT) package to efficiently solve the above optimization objective.

A.1.3 Average Arc Distance

Arc distance represents the geodetic distance, or the geographical distance measured in the length of the shortest arc between two locations along the earth surface in miles. This can be computed with longitude-latitude coordinates of a pair of locations. Similar to cosine embedding distance, we take the average of arc distances between data points across two domains as the distance between the domain pair.

To keep the range of arc distance consistent with the other types of distances, we divide all raw results of arc distances by 12436, which is the distance between the north pole and south pole, as well as the maximum possible arc distance for any two places. Hence, the rescaled arc distance has the range of 0 to 1.

B Experiment Details

B.1 More Experiment Settings

In addition to embeddings from different types of models, we explore whether the hypothesis holds for different criteria to split domains from a dataset and different models for domain adaptation tasks.

Domain Split Criteria Different criteria for splitting geographical domains can lead to data clusters with distinct structures, and thus it is important to check if our observations only pertain to continent domains. Firstly, we split domains by continent regions, including Asia, Europe, Africa, Americas, and Oceania, which is consistent with the domain categories in FMoW-wilds dataset. We also examine the biome splits, where biome describes a large geographical area that has particular traits of vegetation, climate, and ecosystem. The five major biome categories include aquatic, grassland, forest, desert, and tundra (National Geographic Society, 2024).

Model Training We train multiple image models for the domain adaptation schemes of satellite images to ensure that changes in task performance across domains are not solely attributed to the choice of a specific model. In our work, we include ResNet18 and DenseNet121 (Huang et al., 2017) for training.

B.2 Experiment and Model Training Details

B.2.1 Model Encodings

Image based models To ensure the variety of image models used for embedding generation and distance computation, we include ResNet18, ResNet50, and ViT-Small, covering both convolutional neural network and vision transformer architectures.

ResNet18 and ResNet50 are widely used Convolutional Neural Network (CNN) architectures in the field of deep learning (He et al., 2016). ResNet18 and ResNet50 have the same backbone of Residual Networks, where ResNet50 (with $\sim$25.6 million parameters) contains more layers and parameters than ResNet18 (with $\sim$11 million parameters).

Vision Transformer (ViT) is a more recent family of pretrained models that use transformer architecture and patch-based representations on image tasks (Dosovitskiy et al., 2021). In this work, We choose ViT-Small (with $\sim$22 million parameters) on a similar scale of parameters to other models.

All the image models used are initially loaded with the pretrained weights on ImageNet dataset.

Location based models SatCLIP has multiple pretrained checkpoints trained with different vision encoders and spatial resolution L, where L describes the number of Legendre polynomials used for spherical harmonics location encoding. To account for the potential differences of geospatial features between small and large scale of resolution, we use two SatCLIP pretrained models trained with ResNet50 image encoder to generate embeddings, one with $L = 10$ covering low resolution and the other with $L = 40$ yielding high-resolution and more fine-grained embeddings.

B.2.2 Datasets

FMoW-wilds FMoW-wilds dataset is a part of the Wilds benchmarks (Koh et al., 2021) and adapted from the original Functional Map of World (FMoW) dataset (Christie et al., 2018). FMoW dataset contains over 1 million satellite images with abundant metadata for classification tasks of 62 different functional purposes of buildings and land use. FMoW-wilds further adapts the dataset with the primary focus of distribution shifts, and data is split into domains defined by years and geographical regions.

Note that FMoW-wilds targets the domain generalization problems across both temporal and spatial components, whereas our work only keeps the spatial part, i.e., building domains based on their geographical locations. Specifically, we combine the data from year 2016 and 2017, and we split the data into 5 main geographical domains, following the original region splits in the dataset: Asia, Europe, Africa, Americas, and Oceania.

Table 1: Data size of train, validation, and test splits for each domain in FMoW-wilds dataset.

Domains	Train	Validation	Test	Total
Asia	33736	4839	4963	43538
Europe	39567	5813	5858	51238
Africa	18624	2657	2593	23874
Americas	55361	7749	8024	71134
Oceania	4040	578	666	5284

In addition to the satellite image and land use category, each data is accompanied by metadata, including longitude and latitude of the location in the image, the timestamp when the satellite image was taken, the country code, etc.

Specific sizes of each partition by domains and train, validation and test splits can be found in Table 1.

Forest Cover In addition to FMoW-wilds dataset, we obtain a forest cover dataset that is included as part of the MOSAIKS benchmark. The dataset includes the percentage of vegetation with height greater than 5 meters for satellite images taken globally with resolution of approximately 30m by 30m (Rolf et al., 2021). The downstream task is to predict forest cover percentage from image features (regression), and we report the performance based on coefficient of determination.

In total, there are 398484 samples in the training set and 99622 samples in the test set. For validation purposes, we randomly select 20% of the training data and hold them out as a validation set.

B.2.3 Domain Split Criteria

To obtain different sets of domains with the same dataset, we make use of the shapefiles provided from United States Geological Survey (USGS) which describe the domain multipolygons in terms of longitude and latitude coordinates. Then with the location information in metadata, we partition the dataset and group data into region domains.

Continent The continent domains include Asia, Europe, Africa, Americas, and Oceania. Note that Americas consist of both North America and South America. We drop location points that do not belong to these main continents (e.g., places in Antarctica or in the ocean), since they only constitute a very small portion of the data.

Biome The biome domains include aquatic, grassland, forest, desert, and tundra (National Geographic Society, 2024). Since the main tasks included this work primarily focus on objects on land, such as buildings and forests, we do not include aquatic biome in the domain list. For some of the datasets used, the size of data for tundra biome is very small, because the biome is mainly located in the Arctic regions. We then exclude tundra biome in such cases.

B.2.4 Model Training and Hyperparameters

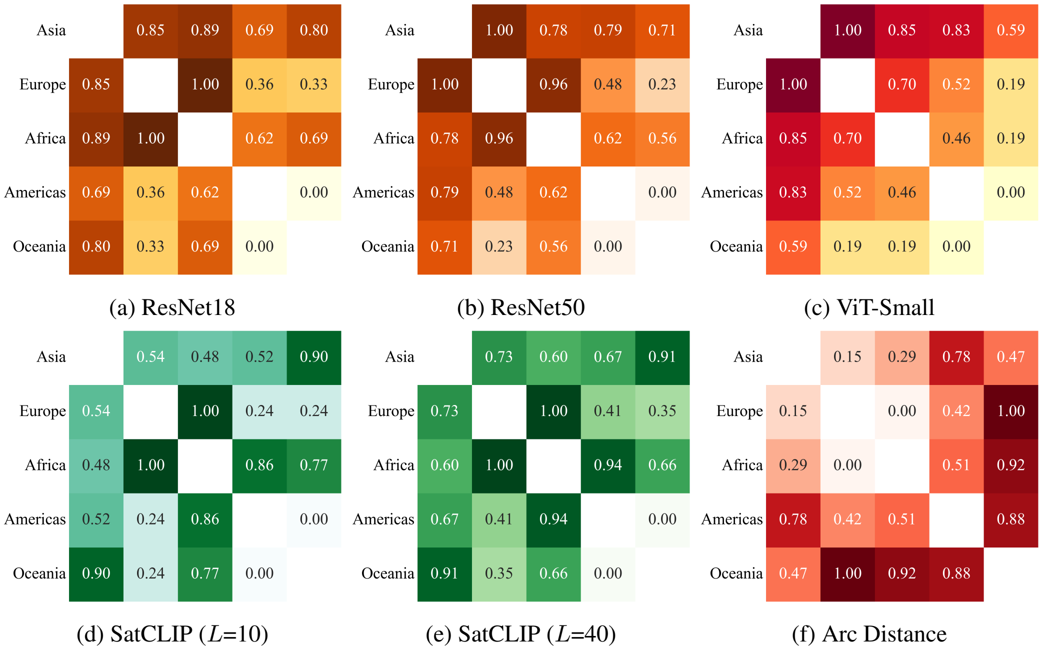
Training on source domain After loaded with pretrained weights on ImageNet, the final linear layers of all models are replaced with classification heads with proper number of classes. During training, we use batch size of 64 and Adam optimizer with learning rate 10^{-4}. A maximum number of epochs is set to 50 with early stopping implemented.

Adapting to target domain With image models trained on source domains, we further fine-tune and evaluate them on target domains in zeroshot and fewshot settings. For zeroshot, we directly evaluate the performance of image models on test set of target domain. In fewshot cases, we fit models on a subset of training data in the target domain with restrictions on the number of samples randomly selected per class. Since randomness is involved with subsampling examples, we run the pipelines 3 times and report the average as the transfer performance. During training, we use batch size of 32 and Adam optimizer with learning rate 10^{-4}. We set maximum number of training epochs to 50 and use early stopping. For evaluation on target domain, we always load model weights with the best validation performance.

C RESULTS

C.1 DOMAIN DISTANCE

We show normalized pairwise cosine and arc distances in FMoW-wilds dataset in Figure 4, and we present normalized Wasserstein distances with entropic regularization $\epsilon = 0.1$, $\epsilon = 0.01$, and $\epsilon = 0.001$ in Figure 5.



(a) ResNet18 (b) ResNet50 (c) ViT-Small

(d) SatCLIP (L=10) (e) SatCLIP (L=40) (f) Arc Distance

Figure 4: Normalized pairwise distances between continent domains in FMoW-wilds dataset. (a)-(e) Average cosine distances of embeddings from different pretrained models. (f) Arc distance.

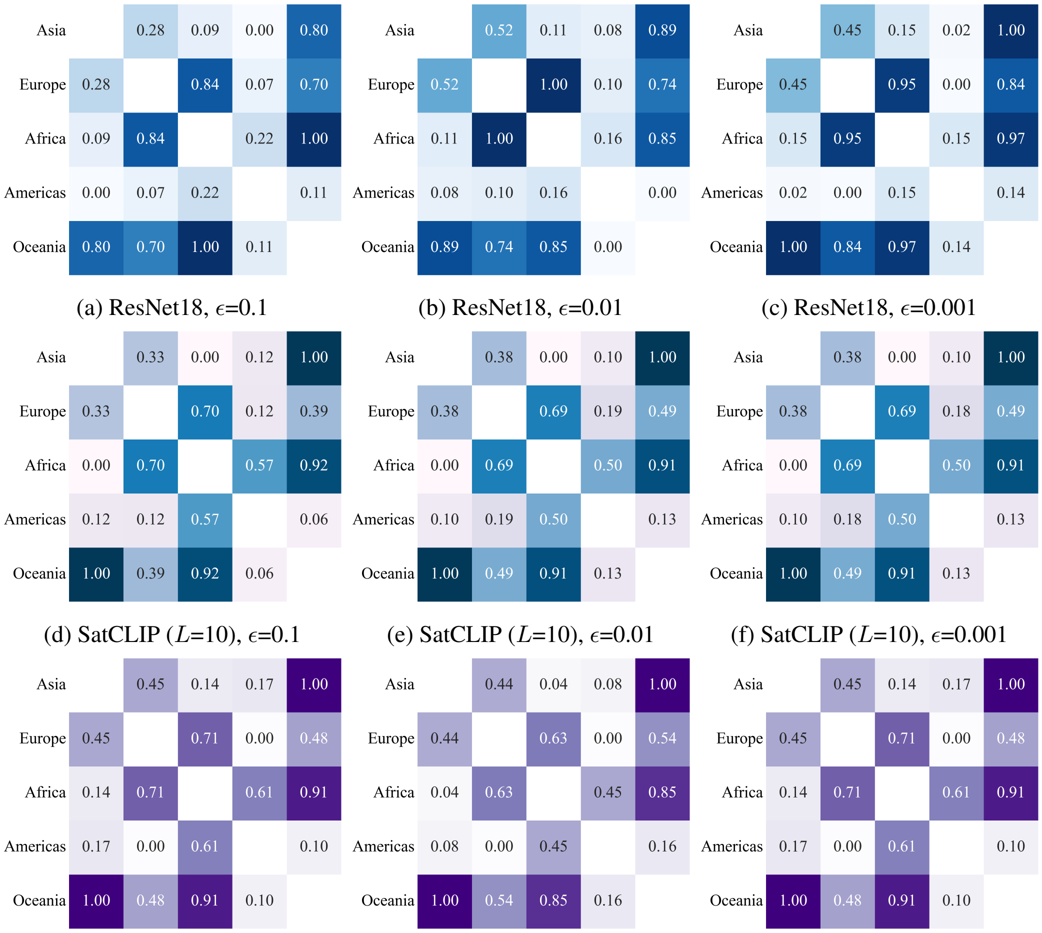
C.2 DOMAIN ADAPTATION

We show results of domain adaptation performance in FMoW-wilds dataset in Figure 6, including both ResNet18 and DenseNet121 under zeroshot and fewshot settings. To make the accuracy values comparable across different models and different domain pairs, we compute relative change in performance compared with the models trained and evaluated on the target domain:

$$\Delta = \frac{\text{acc}(\mathcal{D}_s \Rightarrow \mathcal{D}_t) - \text{acc}(\mathcal{D}_t)}{\text{acc}(\mathcal{D}_t)},$$

where $\text{acc}(\mathcal{D}_s \Rightarrow \mathcal{D}_t)$ represents the adaptation performance from source domain $\mathcal{D}_s$ to target domain $\mathcal{D}_t$, and $\text{acc}(\mathcal{D}_t)$ is the test accuracy of the same model fully trained on $\mathcal{D}_t$ and evaluated on test set for $\mathcal{D}_t$.

The domain adaptation tasks from Oceania to other continents are far more difficult with the greatest relative drop in performance, while the tasks transferring from other continents to Oceania are comparatively easier. We believe the main reason is the large gap of dataset size between Oceania and other continents. Besides, DenseNet121 achieves even better performance (6.13% increase) through adapting from Americas to Oceania in fewshot settings than DenseNet121 directly trained on the full dataset of Oceania. This further supports the prospects that having models pretrained on data-rich source domains and adapted to data-poor target domains. However, being trained with data from other continents such as Asia (15.95% drop), Europe (16.49% drop), Africa (38.56% drop)



(a) ResNet18, ϵ=0.1 (b) ResNet18, ϵ=0.01 (c) ResNet18, ϵ=0.001

(d) SatCLIP (L=10), ϵ=0.1 (e) SatCLIP (L=10), ϵ=0.01 (f) SatCLIP (L=10), ϵ=0.001

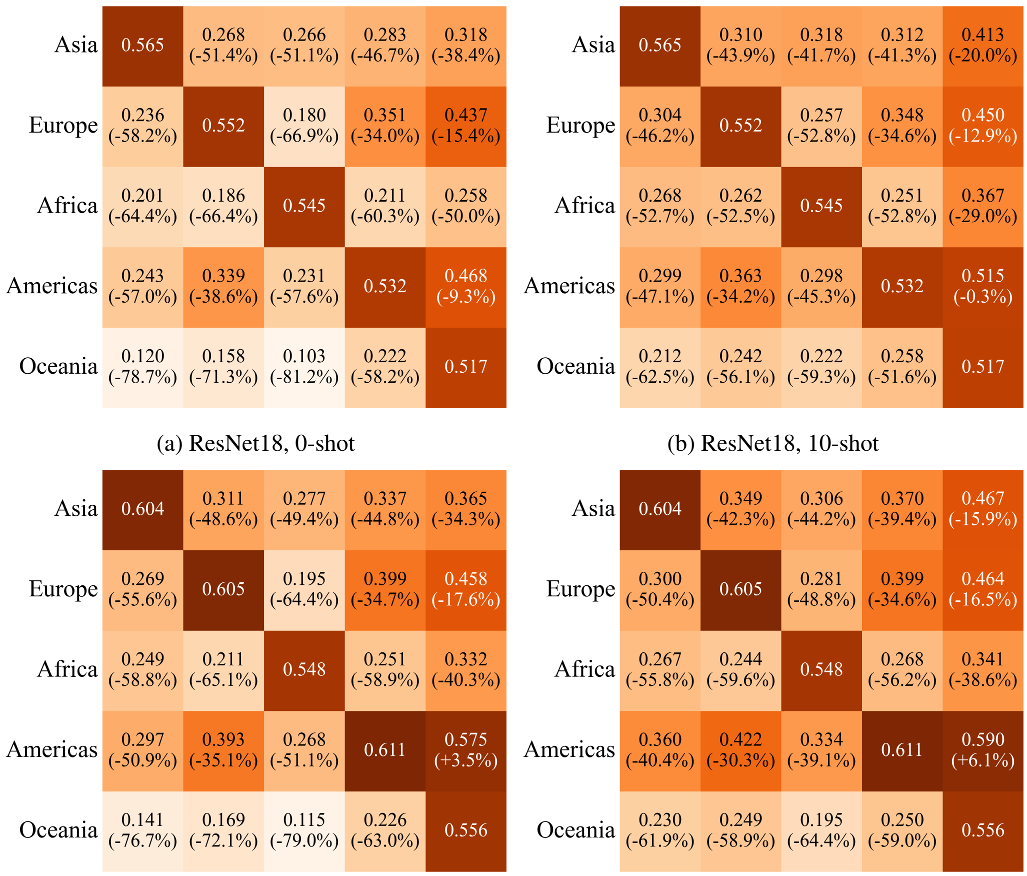
(g) SatCLIP (L=40), ϵ=0.1 (h) SatCLIP (L=40), ϵ=0.01 (i) SatCLIP (L=40), ϵ=0.001

Figure 5: Normalized pairwise Wasserstein distances between continent domains in FMoW-wilds dataset using embeddings from different pretrained models and weights of entropic regularization for optimal transport solver.

does not contribute positively despite having much more training data on Oceania. We notice that the class distribution of Americas data has sufficient examples in classes from the support of data distribution of Oceania, especially the majority classes of satellite images taken in Oceania. Hence, we believe it is also important that source domains on which models are mainly trained should have support closer to the distribution in target domains, and this further motivates leveraging domain distances to design and construct a good source dataset or domain to achieve optimal generalization performance.

C.3 Ablation Studies

We further conduct ablation studies to analyze potential confounding factors affecting our results. Figure 7 shows the trends under two different datasets, FMoW-wilds and MOSAIKS-Forest, which have very different target tasks, land use classification and forest coverage regression. In Figure 8, we explore the relationship using FMoW-wilds dataset with a different definition of geographic domain. As explained in section B.2.3, we choose to partition data based on their geographical locations into continents or biomes, and we apply domain adaptation and analyze the results. Figure 9 reports the relationship from the perspective of using different models for domain adaptation task.



Asia	0.565	0.268 (-51.4%)	0.266 (-51.1%)	0.283 (-46.7%)	0.318 (-38.4%)
Europe	0.236 (-58.2%)	0.552	0.180 (-66.9%)	0.351 (-34.0%)	0.437 (-15.4%)
Africa	0.201 (-64.4%)	0.186 (-66.4%)	0.545	0.211 (-60.3%)	0.258 (-50.0%)
Americas	0.243 (-57.0%)	0.339 (-38.6%)	0.231 (-57.6%)	0.532	0.468 (-9.3%)
Oceania	0.120 (-78.7%)	0.158 (-71.3%)	0.103 (-81.2%)	0.222 (-58.2%)	0.517

(a) ResNet18, 0-shot

Asia	0.565	0.310 (-43.9%)	0.318 (-41.7%)	0.312 (-41.3%)	0.413 (-20.0%)
Europe	0.304 (-46.2%)	0.552	0.257 (-52.8%)	0.348 (-34.6%)	0.450 (-12.9%)
Africa	0.268 (-52.7%)	0.262 (-52.5%)	0.545	0.251 (-52.8%)	0.367 (-29.0%)
Americas	0.299 (-47.1%)	0.363 (-34.2%)	0.298 (-45.3%)	0.532	0.515 (-0.3%)
Oceania	0.212 (-62.5%)	0.242 (-56.1%)	0.222 (-59.3%)	0.258 (-51.6%)	0.517

(b) ResNet18, 10-shot

Asia	0.604	0.311 (-48.6%)	0.277 (-49.4%)	0.337 (-44.8%)	0.365 (-34.3%)
Europe	0.269 (-55.6%)	0.605	0.195 (-64.4%)	0.399 (-34.7%)	0.458 (-17.6%)
Africa	0.249 (-58.8%)	0.211 (-65.1%)	0.548	0.251 (-58.9%)	0.332 (-40.3%)
Americas	0.297 (-50.9%)	0.393 (-35.1%)	0.268 (-51.1%)	0.611	0.575 (+3.5%)
Oceania	0.141 (-76.7%)	0.169 (-72.1%)	0.115 (-79.0%)	0.226 (-63.0%)	0.556

(c) DenseNet121, 0-shot

Asia	0.604	0.349 (-42.3%)	0.306 (-44.2%)	0.370 (-39.4%)	0.467 (-15.9%)
Europe	0.300 (-50.4%)	0.605	0.281 (-48.8%)	0.399 (-34.6%)	0.464 (-16.5%)
Africa	0.267 (-55.8%)	0.244 (-59.6%)	0.548	0.268 (-56.2%)	0.341 (-38.6%)
Americas	0.360 (-40.4%)	0.422 (-30.3%)	0.334 (-39.1%)	0.611	0.590 (+6.1%)
Oceania	0.230 (-61.9%)	0.249 (-58.9%)	0.195 (-64.4%)	0.250 (-59.0%)	0.556

(d) DenseNet121, 10-shot

Figure 6: Domain Adaptation Accuracy for ResNet18 and DenseNet121 on FMoW-wilds dataset. Rows represent source domains which image models are initially trained on. Columns represent target domains which image models are finally evaluated on. Percentages in the parentheses following the accuracy values are the relative drop in classification accuracy on the test set compared with performance of models trained on the target domain.

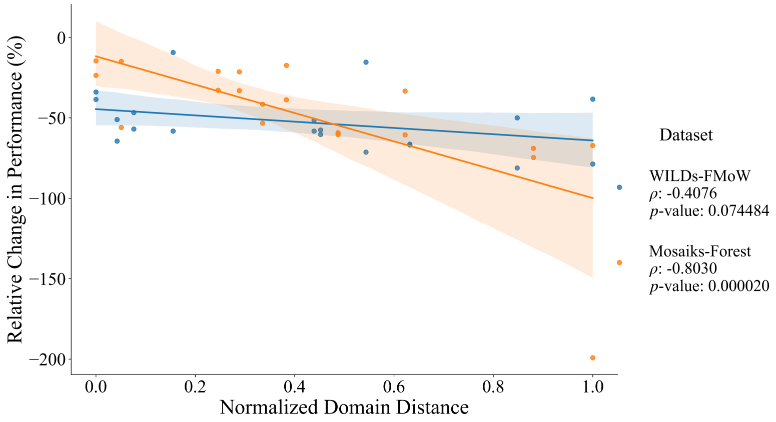


Figure 7: Comparison of Wasserstein domain distance and relative change of domain adaptation performance across different datasets. Performance results are reported on 0-shot adaptation accuracy of ResNet18.

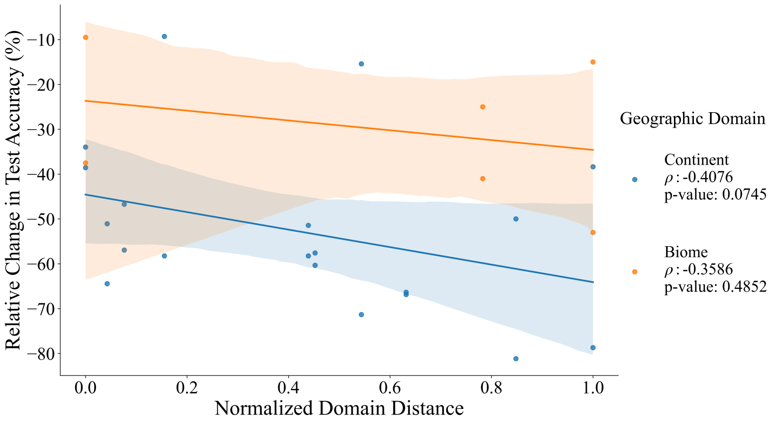


Figure 8: Comparison of Wasserstein domain distance and relative change of domain adaptation performance in FMoW-wilds dataset across different domain split criteria. Performance results are reported on 0-shot adaptation accuracy of ResNet18.

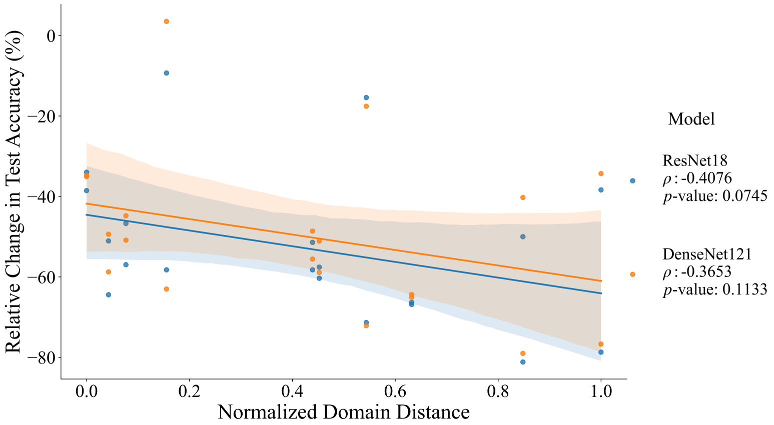


Figure 9: Comparison of Wasserstein domain distance and relative change of domain adaptation performance in FMoW-wilds dataset across different models for adaptation task. Performance results are reported on 0-shot adaptation accuracy of ResNet18 and DenseNet121.