Flickr Africa: Examining Geo-Diversity in Large-Scale, Human-Centric Visual Data

Keziah Naggita[1*†] **Julienne LaChance**[2†] **Alice Xiang**[2]

[1]Toyota Technological Institute at Chicago   [2]Sony AI America
knaggita@ttic.edu, {Julienne.LaChance, Alice.Xiang}@sony.com

Abstract

Biases in large-scale image datasets are known to influence the performance of computer vision models as a function of geographic context. To investigate the limitations of standard Internet data collection methods in low- and middle-income countries, we analyze human-centric image geo-diversity on a massive scale using geotagged Flickr images associated with each nation in Africa. We report the quantity and content of available data with comparisons to population-matched nations in Europe as well as the distribution of data according to fine-grained intra-national wealth estimates. Furthermore, we present findings for an "othering" phenomenon as evidenced by a substantial number of images from Africa being taken by non-African photographers. The results of our study suggest that further work is required to capture image data representative of African people and, ultimately, improve the applicability of computer vision models in a global context.

Introduction

Data collection and processing are crucial to the machine learning (ML) pipeline and are the source of many biases in AI systems, which have been shown to largely stem from a lack of diverse representation in training datasets (Buolamwini and Gebru 2018). Currently, most large-scale computer vision datasets are collected via webscraping and subsequent data cleaning. For example, the ImageNet database ((Deng et al. 2009); 42607 citations per Google Scholar, accessed Sept. 14, 2022) is comprised of images sourced from search engines like Google and Flickr, while the COCO dataset ((Lin et al. 2014); 26751 citations per Google Scholar, accessed Sept. 14, 2022) is comprised of images sourced entirely from Flickr. Thus, biases inherent to Flickr influence the performance of models for visual tasks as diverse as object classification, pose estimation, instance segmentation, image captioning, and beyond. Some of these dataset biases have been explored in detail: for ImageNet and the Flickr-sourced Open Images dataset (Kuznetsova et al. 2018) it has been shown that data from India, China, and African and South-East Asian countries is vastly underrepresented despite their large populations (DeVries et al. 2019); while for COCO, data has been shown to be heavily skewed towards lighter-skinned and male individuals (Zhao,

Wang, and Russakovsky 2021). In particular, such biases impact the applicability of models in a global context. For instance, DeVries et al. (2019) manually sourced image data from 264 globally-distributed households and demonstrated how object recognition model performance drops when applied in lower-income nations. Motivated by the popularity of datasets sourced using Flickr data, we here analyze **1.5M** geotagged images in the Flickr database to deeply explore its representation of African people (see Figure 1).

In this paper, we aim to highlight the limitations of webscraping generic and human-centric image data from Africa for ML training purposes. We analyze image data for every African nation with direct comparisons to population-matched higher-GDP European nations and show that there is far less data available from Africa. We report the distribution of African geotagged image data as a function of fine-grained, intra-national wealth estimates (Chi et al. 2022) and assess data with respect to license restrictions, population size, nominal GDP, Internet usage, and official languages. Additionally, we collect crowdsourced annotations to explore image content, and provide evidence for an "othering" phenomenon as the majority of African geotagged images we analyzed were taken by foreigners, while the opposite trend is shown for select European nations. Such results highlight the importance of considering geodiversity metrics beyond ancestry/ethnicity of individuals within images and, moreover, how the mechanisms by which images are obtained can quantitatively and qualitatively affect how the image corpus represents the world (*e.g.* imposing a "Western gaze"). Overall, we find that Flickr provides a very limited and skewed representation of African countries which likely contributes to many of the biases in models trained on popular, large-scale image datasets.

Methodology

Data Collection: Flickr Africa *For each nation in Africa, we utilized Flickr queries to construct a dataset of images and associated metadata.* Using the FlickrAPI, we scraped images and associated metadata from Flickr between dates 2004-02-10 and 2022-02-10 (18 years) by querying by country name (*e.g.* "*Togo*") and the country name + "people" (*e.g.* "*Togo people*") for all 54 countries in Africa. Images without valid geotags were excluded, and city/country information were determined using open-source reverse geocode (Pen-

---

(a) Africa: Image data distribution colored by RWI.

(b) Africa: Total number of geotagged images.

(c) Madagascar: RWI group overlaid with image count per region.

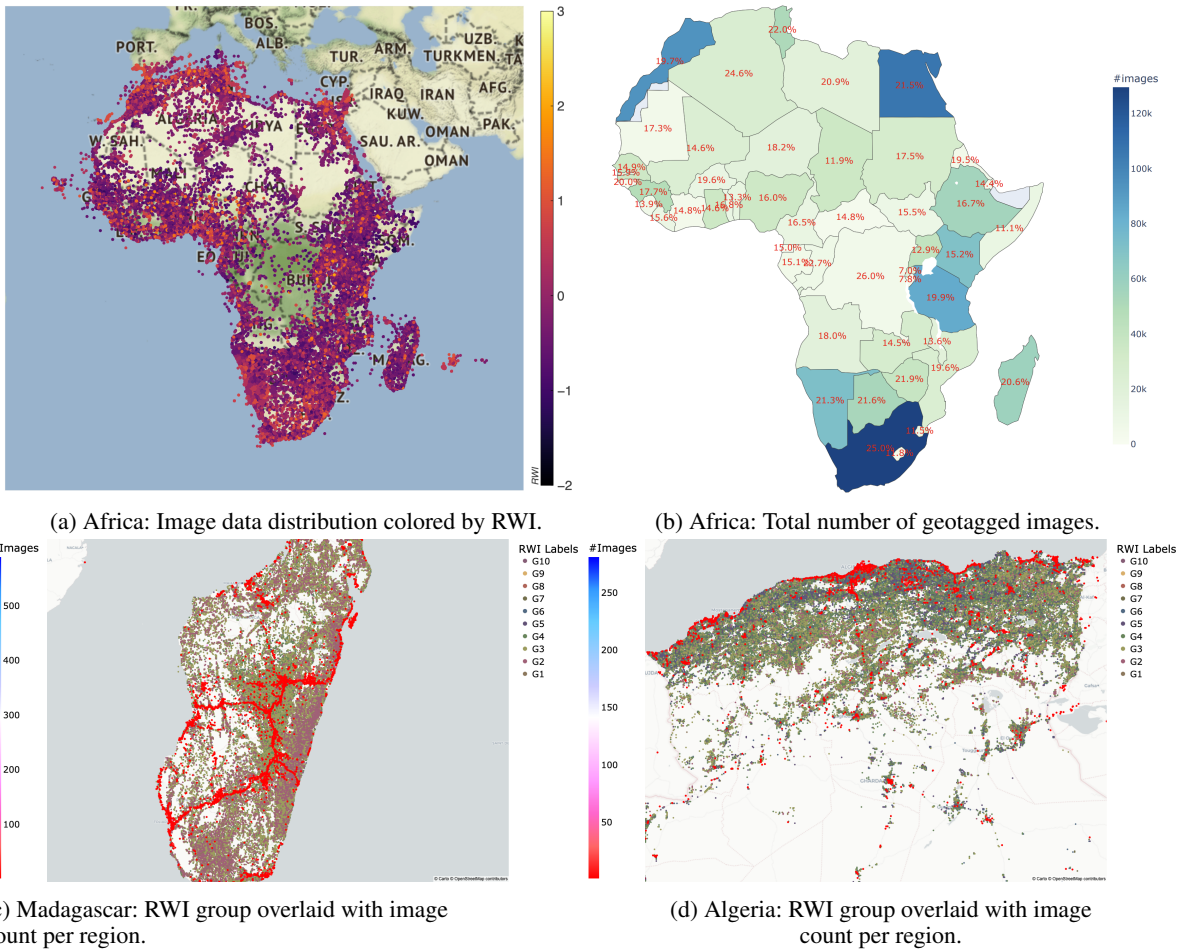(d) Algeria: RWI group overlaid with image count per region.

Figure 1: A collection of maps displaying relative wealth index (RWI) and geolocation of Flickr Africa images via country name query. Tolerance distance from geotag to nearest RWI-labeled point are: ((**a, b**) dist: $\leq 300km$; (**c, d**) $\leq 10km$). (**b**) Nations are colored according to total number of geotagged images and the percentages (rounded to one decimal place) is the percentage of geotagged images. South Africa had the highest number of geotagged images and Sao Tome and Principe had the smallest number of geotagged images while Cape Verde had the highest percentage of geotagged images and Rwanda had the lowest percentage of geotagged images.

man 2020). All data is available at https://doi.org/10.5281/zenodo.7133542

Population-matched European countries *The data collection process was repeated for four European nations.* In the interest of comparing data availability and content to higher-GDP European nations, we chose the following countries as a function of similar population size ((Wikipedia.org 2022b,d,a)): Switzerland and Sierra Leone (GDP: $841.97k$ vs. $4.27k$); Cyprus and Djibouti (GDP: $27.73k$ vs. $3.84k$); Finland and Central African Republic (CAF) (GDP: $297.62k$ vs. $2.65k$); and Slovenia and Lesotho (GDP: $63.65k$ vs. $2.56k$). For all 58 countries we collected data pertaining to percentage of internet users (Wikipedia.org 2022c), nominal GDP (Wikipedia.org 2022b), population size (Wikipedia.org 2022d,a) and official languages (Wikipedia.org 2022e).

Relative Wealth Estimates *Fine-grained relative wealth estimates were associated with each geotagged image.* We utilize the relative wealth index (RWI) data collected from Low- and Middle-Income Countries (LMICs) by Facebook's Data for Good project ((Chi et al. 2022)); however, this dataset excludes: {*Somalia, Seychelles, Sao Tome and Principe, Sudan,* and *South Sudan*}. RWI scores are normalized by nation, so the data should only be utilized for intra-national wealth comparisons. Using k-nearest neighbour, we computed the nearest RWI-labeled geographic location of each image and excluded geotagged images beyond $300km$ to the nearest RWI-labelled location.

Manual Content Annotation *Crowdsourced annotations were collected for six additional image features.* We used Amazon Mechanical Turk (AMT) to collect annotations describing image contents. Each task (HIT) had 21 images, with six binary questions per image which required the an-

notator label the image according to: indoor vs. outdoor setting, public vs. private setting, nature vs. manmade setting, the presence of people, real vs. synthetic image type, and offensive vs. inoffensive content. We compensated workers at a rate of $15/hour and utilized gold standard images (1 per 20 images) to assess annotator performance.

Limitations of Our Approach   We acknowledge three notable limitations of our method. First, we recognize that geolocation data (longitude, latitude) is inherently unreliable. Values may be modified or removed by the user or otherwise not reflect the location of capture, and reverse geolocation methods are computationally expensive and often fail, particularly with geographic locations close to region borders. This motivates our use of both geotags and country name tags for cross-validation of location, though this restricts us to fewer data samples overall. Additionally, some forms of geodiversity are difficult or impossible to determine from visual inspection alone, such as an individual's gender, ethnicity, or religion. Finally, we were limited to obtaining data using only two queries, namely, by country name or country name + "people". We anticipate future work exploring a wider variety of query terms, both in English and local languages; here, no correlation was determined between dominant national languages and geotagged image availability.

Ethical Considerations   We note that although the Flickr images analyzed here are all publicly viewable, we show that most have the Flickr default license of "All Rights Reserved". Thus, we have opted to provide image URLs in lieu of images for direct download to avoid duplication of protected content, particularly in the event that a Flickr user chooses to remove or modify the permissions of an image. We acknowledge the weaknesses of this method in terms of consent, as public Flickr images are typically not taken by those in the images (as pointed out in (Birhane and Prabhu 2021)) and Flickr users may wish to avoid the utilization of their images for research purposes. Given that our objective is to critique large-scale image dataset curation strategies which do not respect image licenses (*e.g.* the methodology for generating the COCO dataset), we deemed it justifiable to perform basic analyses on protected images and to build awareness regarding widespread license violations in standard AI training pipelines.

Results and Discussion

Data Availability and Geographic Distribution   There are very few geotagged images from Africa (see Figure 1b) with drastically fewer than European nations of comparable size (see Table 1); *e.g.* Switzerland has $18\times$ as many geotagged images as Sierra Leone. The number of geotagged images positively increased with population size (correlation: 0.412 & 0.538), internet usage (0.474 & 0.385), and GDP (0.599 & 0.748) respective of country-name and country-name+people query; this was found to be statistically significant. Official language was not found to have a meaningful correlation to the number of geotagged images (*p-value* $= 0.2021$ & *p-value* $= 0.846$) after coding by: 1- (English is the only official language), 2- English is among the two official languages, 3- English among greater than three official languages, 4- English not among less than three official languages, and 5- English not among greater than three official languages.

Tags and Licenses   Analysis of the tags found on the images revealed that the most frequent tags were the name of the place where the image was taken, including "Africa", in addition to image contents. The least frequent tags were usually those in foreign languages or tags containing typos or consisting of multiple concatenated words. Such tags are difficult to interpret, let alone utilize for Flickr image queries in the interest of constructing diverse datasets.

Additionally, the vast majority of images with query by "[country name]" and "[country name] people" respectively are licensed as "All Rights Reserved" (80.46%, 81.99%), indicating the Flickr default setting when images are uploaded to the platform. Thus, those constructing datasets using Flickr Africa data must be aware that most images are unavailable for open public use. This further limits ethical access to geographically diverse data.

Geodiversity by RWI   To assess the impact of wealth on the availability of geotagged image data, we examine image counts by RWI values binned into 10 percentile groups ($G1$-$G10$). For most nations, the majority of image data comes from the middle RWI regions ($G4$, $G5$, $G6$ and $G7$) and the least from low RWI regions ($G1$, $G2$ and $G3$). However, this is not always the case, *e.g.* Madagascar (Figure 1c) and Algeria (Figure 1d) from which data is commonly sourced from low-income areas (along main roads close to national parks) or high-income areas (in major cities), respectively. RWI has potential as a mechanism for constructing geodiverse datasets in future work.

Local vs. Non-Local Representation   For the African geotagged images we randomly selected for the crowdsourcing task, images were far more likely to be taken by foreigners whereas the opposite trend was observed for high-GDP European nations, according to comparisons of geotags and user-reported location. For Sierra Leone, +169% of images were captured by foreigners, compared to Switzerland (-31%). The same trend applies to Djibouti and Cyprus (+335% and -49%) and CAF and Finland (+272% and -49%). Thus, images sourced from Africa may be more predisposed to bias resulting from an "othering" phenomenon, and less representative of African cultural viewpoints.

Image Content   By utilizing crowdsourced annotations, we examine image content data from each matched African/European nation pair. The AMT results revealed that query-by-name images from both African countries and European countries were predominantly "real" (93.47% and 91.94%), "inoffensive" (88.51% and 89.41%), "outdoor" (77.89% and 79.28%), "public" (90.27% and 90.19%), and "nature" (63.68% and 62.96%) images. Nations varied more in terms of total number of images containing people (Sierra Leone 68.25%, Lesotho 59.76%, CAF 52.13%, Djibouti 54.86%, Switzerland 68.45%, Finland 60.76%, Slovenia 71.81%, and Cyprus 67.28%). Given that image content was fairly similar across most attributes annotated, and there exist far fewer geotagged images from Africa, we anticipate

Table 1: Geotagged image counts for matched African and European nations (query-by-name).

| European Countries				African Countries			
Country	population	#geotagged	%geotagged	Country	population	#geotagged	%geotagged
Switzerland	8.75M	129,517	24.170699	Sierra Leone	8.30M	7,303	13.902532
Finland	5.55M	119,901	22.941545	CAF	5.60M	2,954	14.843475
Slovenia	2.11M	86,630	23.313706	Lesotho	2.10M	5,121	11.831708
Cyprus	918.10k	77,826	20.948900	Djibouti	976.11k	5,179	14.374532
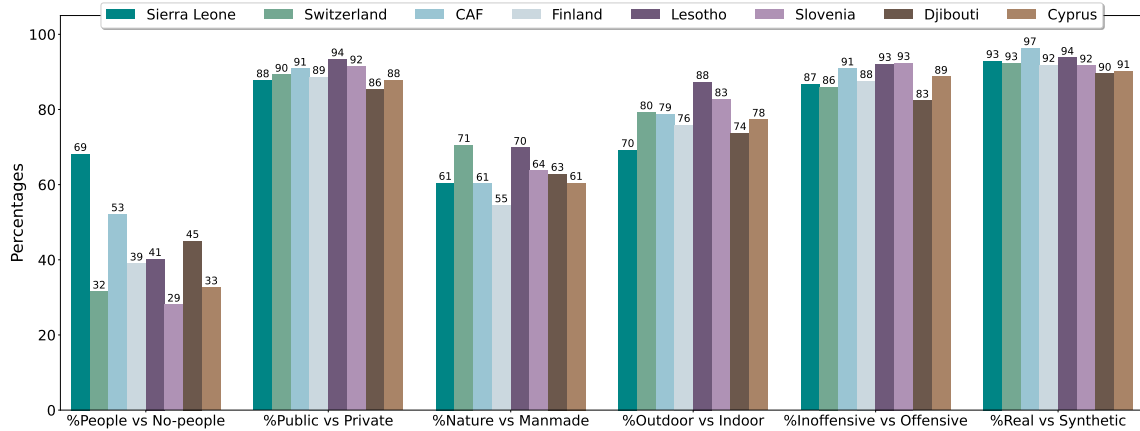
Figure 2: In general, images from African nations and population-matched European nations did not contain people (e.g., 33% of Cyprus geotagged images contained people and 67% did not) and were captured in outdoor (e.g., 78% of Cyprus images were captured outdoor and 22% indoor), public (e.g., 88% of Cyprus images were public and 12% private), and natural settings (e.g., 61% of Cyprus images were captured in nature and 39% in man-made settings). The majority of images were real photographs (as opposed to synthetically-generated images, or pictures of pictures) and did not contain content considered to be offensive by annotators.

insufficient African data availability for certain computer vision tasks. For example, the lower prevalence of images captured in "private" and "indoor" settings indicates *e.g.* household object image data inaccessibility, which thereby impacts downstream object recognition system models.

Conclusion and Future Work

Geographical context shapes data, and data shapes the performance of models trained using such data. The key findings from our Flickr Africa data analysis have the potential to be highly impactful in both (1) exposing new limitations of current large-scale image data collection methodologies, and (2) exposing unique data challenges to Africa, including the lack of data crucial to specific domains (*e.g.* a researcher cannot source sufficient, representative household object data if very few images are taken within indoor/private scenes). Notably, we reported on the extreme lack of data availability when compared to wealthy European nations; for instance when querying by country name, Switzerland had 18x the geotagged image data as Sierra Leone, an African nation of similar population size (8.75M vs. 8.30M, respectively), while Sao Tome and Principle only had (776, 116) geotagged images in total (depending on query). Moreover, data may be even less accessible according to use case,

given that most of the Flickr Africa data has a restrictive use license, and certain image content attributes were found to appear less frequently (*e.g.* private and indoor settings). Nationally, higher quantities of geotagged image data was found to positively correlate with population size, GDP, and Internet usage, but no significant correlation was discovered based on dominant national languages. Additionally, we interrogate where African image data comes from: generally from middle-wealth regions as measured intra-nationally by RWI, though this differs by nation; and with images mainly taken by foreigners, though the opposite trend is identified in wealthy European nations.

Looking forward, we encourage new scholarship centering novel methods for sourcing geodiverse datasets and measuring new forms of geodiversity specific to Africa, such as analyses of tribal diversity as opposed to the more commonly studied diversity by race/ethnicity. We openly provide our large-scale dataset to enable future researchers to utilize and augment Flickr Africa for model evaluations across a wide domain of computer vision tasks; likewise, more rigorous bias identification methods (*e.g.* (Wang et al. 2022)) may uncover still more limitations. Finally, we would be interested to explore the extent to which privacy and consent are respected in Africa.

Acknowledgments

References

Birhane, A.; and Prabhu, V. U. 2021. Large Image Datasets: A Pyrrhic Win for Computer Vision? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 1537–1547.

Buolamwini, J.; and Gebru, T. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Friedler, S. A.; and Wilson, C., eds., *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, 77–91. PMLR.

Chi, G.; Fang, H.; Chatterjee, S.; and Blumenstock, J. E. 2022. Microestimates of wealth for all low- and middle-income countries. *Proceedings of the National Academy of Sciences*, 119(3): e2113658119.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

DeVries, T.; Misra, I.; Wang, C.; and van der Maaten, L. 2019. Does Object Recognition Work for Everyone? In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, 52–59. Computer Vision Foundation / IEEE.

Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J. R. R.; Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Malloci, M.; Duerig, T.; and Ferrari, V. 2018. The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *CoRR*, abs/1811.00982.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In Fleet, D.; Pajdla, T.; Schiele, B.; and Tuytelaars, T., eds., *Computer Vision – ECCV 2014*, 740–755. Cham: Springer International Publishing. ISBN 978-3-319-10602-1.

Penman, R. 2020. reverse-geocode 1.4.1. https://pypi.org/project/reverse-geocode/#description. Accessed: 2022-06-15.

Wang, A.; Liu, A.; Zhang, R.; Kleiman, A.; Kim, L.; Zhao, D.; Shirai, I.; Narayanan, A.; and Russakovsky, O. 2022. REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets. *Int. J. Comput. Vision*, 130(7): 1790–1810.

Wikipedia.org. 2022a. List of African countries by population. https://en.wikipedia.org/wiki/List_of_African_countries_by_population. Accessed: 2022-08-27.

Wikipedia.org. 2022b. List of countries by GDP (nominal) - IMF. https://en.wikipedia.org/wiki/List_of_countries_by_GDP_(nominal). Accessed: 2022-08-27.

Wikipedia.org. 2022c. List of countries by number of Internet users. https://en.wikipedia.org/wiki/List_of_countries_by_number_of_Internet_users. Accessed: 2022-08-27.

Wikipedia.org. 2022d. List of European countries by population. https://en.wikipedia.org/wiki/List_of_European_countries_by_population. Accessed: 2022-08-27.

Wikipedia.org. 2022e. List of official languages by country and territory. https://en.wikipedia.org/wiki/List_of_official_languages_by_country_and_territory. Accessed: 2022-08-27.

Zhao, D.; Wang, A.; and Russakovsky, O. 2021. Understanding and Evaluating Racial Biases in Image Captioning. *CoRR*, abs/2106.08503.