Data-centric Machine Learning: Tackling social bias in computer vision datasets

Olga Russakovsky, Vikram V. Ramaswamy

PRINCETON UNIVERSITY

Many thanks to our sponsors:

NSF, Open Philanthropy, PRINCETON School of Engineering and Applied Science, Microsoft, Meta, Google Cloud

http://visualai.princeton.edu

@VisualAIlab

Work done jointly with:

Aaron B. Adcock, Jia Deng, Li Fei-Fei, Ruth Fong, Kyle Genova, Deepti Ghadiyaram, Kenji Hata, Yannis Karakozis, Leslie Kim, Sunnie S. Y. Kim, Anat Kleiman, Nicole Meister, Prem Nair, Sing Yu (Phoebe) Lin, Alexander Liu, Arvind Narayanan, Iroha Shirai, Klint Qinami, Laurens van der Maaten, Angelina Wang, Zeyu Wang, Kaiyu Yang, Jacqueline Yau and Ryan Zhang

Do the photographers **know** their data is being used?

Were the annotators **paid** a living wage?

Was the data downloaded from the web? Using what **queries**?

Who is allowed to use the data, and for what **purpose**?

Who is represented, and **how** are they represented?

How did search engine **bias** influence the dataset?

Did the subjects in the photos **consent**?

Who were the **researchers** collecting the data?

Who **annotated** the data, and what biases did that introduce?

Do the photographers **know** their data is being used?

Were the annotators **paid** a living wage?

Was the data downloaded from the web? Using what **queries**?

Who is allowed to use the data, and for what **purpose**?

Who is represented, and **how** are they represented?

How did search engine **bias** influence the dataset?

Did the subjects in the photos **consent**?

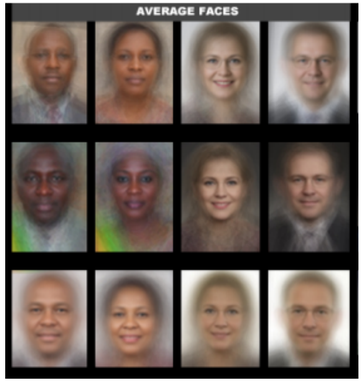
Who were the **researchers** collecting the data?

Who **annotated** the data, and what biases did that introduce?

Large scale ≠ equal representation

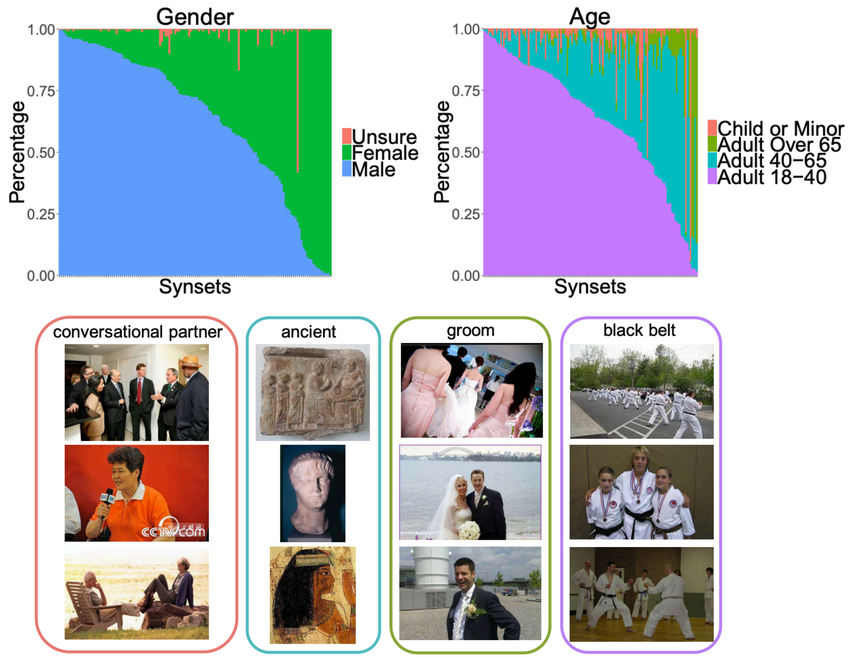
Imbalanced distribution of skin color in face recognition datasets

Dataset	Lighter
IJB-A	79.6%
Adience	86.2%



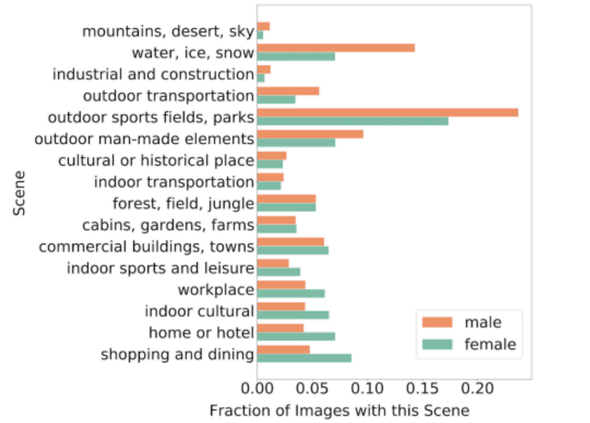
[Joy Buolamwini and Timnit Gebru. FAT*18. “GenderShades: Intersectional Accuracy Disparities in Commercial Gender…”]

Skewed demographic distribution and cultural associations in ImageNet



[Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng and Olga Russakovsky. FAT*’20 "Towards Fairer Datasets: Filtering and Balancing the Distribution of the People Subtree…”]

Gender representation and associations captured in COCO



[Angelina Wang, A. Liu, R. Zhang, A. Kleiman, L. Kim, D. Zhao, I. Shirai, A. Narayanan, and O. Russakovsky. IJCV’22 "REVISE: A Tool for…”]

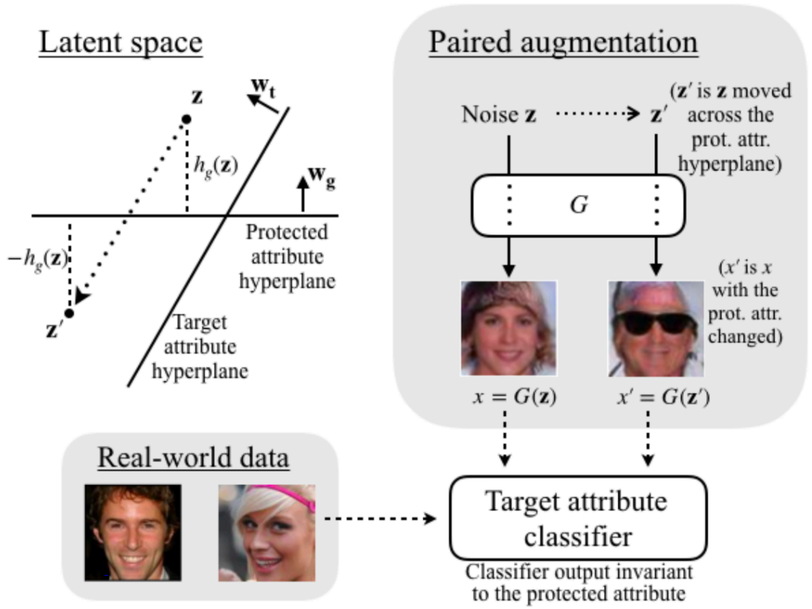
Algorithmically mitigating correlations in the data

Setting: Given a dataset that correlates target labels (e.g., whether the person is smiling, doing a particular activity, …) with protected attribute (e.g., the person's gender)

Goal: Develop a method for training visual classifiers that would ignore such correlations (if the correlations are deemed unnecessary and/or harmful)

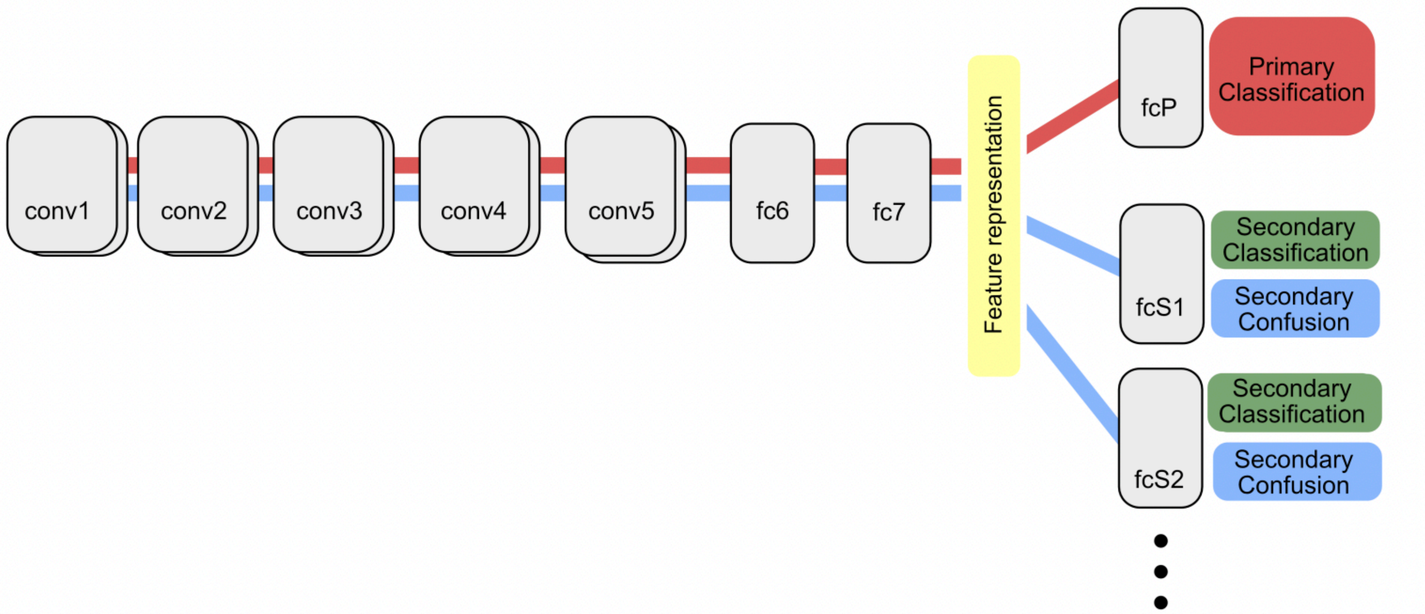
Model Name	Model	Test Inference	Bias (↓)	Accuracy (%, ↑) Color	Gray	Mean
Baseline	N-way softmax	$\arg\max_y \mathrm{P}(y\|x)$	0.074	89.0	88.0	88.5 ± 0.3
Oversampling	N-way softmax, resampled	$\arg\max_y \mathrm{P}(y\|x)$	0.066	89.2	89.1	89.1 ± 0.4
Adversarial	w/ uniform confusion [1, 46]	$\arg\max_y \mathrm{P}(y\|x)$	0.101	83.8	83.9	83.8 ± 1.1
	w/ ∇ reversal, proj. [51]	$\arg\max_y \mathrm{P}(y\|x)$	0.094	84.6	83.5	84.1 ± 1.0
DomainDiscrim	joint ND-way softmax	$\arg\max_y \sum_d \mathrm{P}_{\mathrm{tr}}(y,d\|x)$	0.844	88.3	86.4	87.3 ± 0.3
		$\arg\max_y \max_d \mathrm{P}_{\mathrm{te}}(y,d\|x)$	0.040	91.3	89.3	90.3 ± 0.5
		$\arg\max_y \sum_d \mathrm{P}_{\mathrm{te}}(y,d\|x)$	0.040	91.2	89.4	90.3 ± 0.5
	RBA [52]	$y = \mathcal{L}(\sum_d \mathrm{P}_{\mathrm{tr}}(y,d\|x))$	0.054	89.2	88.0	88.6 ± 0.4
DomainIndepend	N-way classifier per domain	$\arg\max_y \mathrm{P}_{\mathrm{te}}(y\|d^*,x)$	0.069	89.2	88.7	88.9 ± 0.4
		$\arg\max_y \sum_d s(y,d,x)$	**0.004**	**92.4**	**91.7**	$\mathbf{92.0 \pm 0.1}$

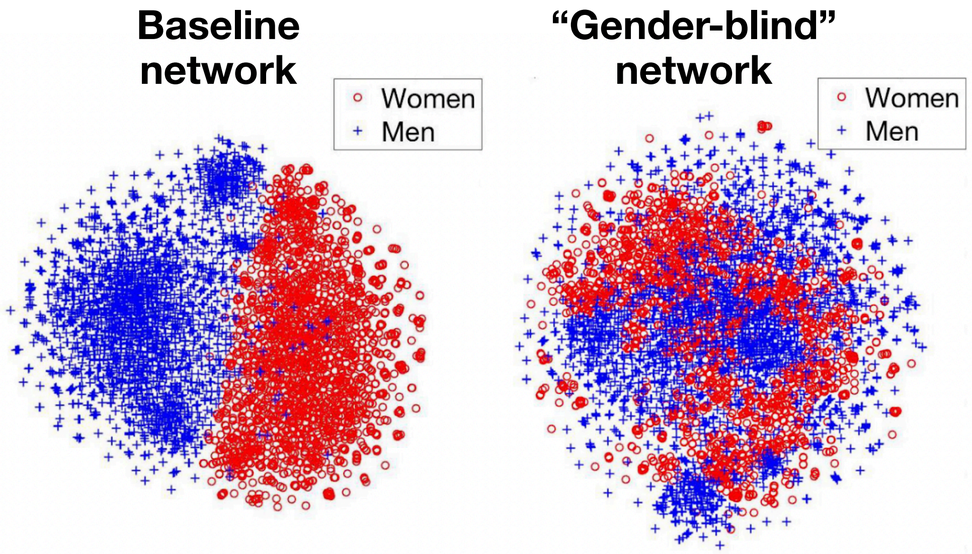
[Zeyu Wang, Klint Qinami, Yannis Karakozis, Kyle Genova, Prem Nair, Kenji Hata and Olga Russakovsky. CVPR'20 "Towards fairness in visual recognition: Effective strategies for bias mitigation"]



[Vikram V. Ramaswamy, Sunnie S. Y. Kim and Olga Russakovsky. CVPR'21 "Fair attribute classification through latent space debiasing"]

Algorithmically mitigating correlations in the data by learning a “de-biased” representation





[Mohsan Alvi et al. ECCV’18. “Turning a blind eye: Explicit Removal of Biases and Variation…”]
[David Madras et. al., ICML’18. “Learning Adversarially Fair and Transferable Representations”]
[Brian Zhang et al. AIES’18 “Mitigating Unwanted Biases with Adversarial Learning”]
[Ehsan Adeli, Qingyu Zhao, et al. WACV’21 “Representation Learning with Statistical Independence…”]

Algorithmically mitigating correlations in the data becomes increasing challenging in real datasets

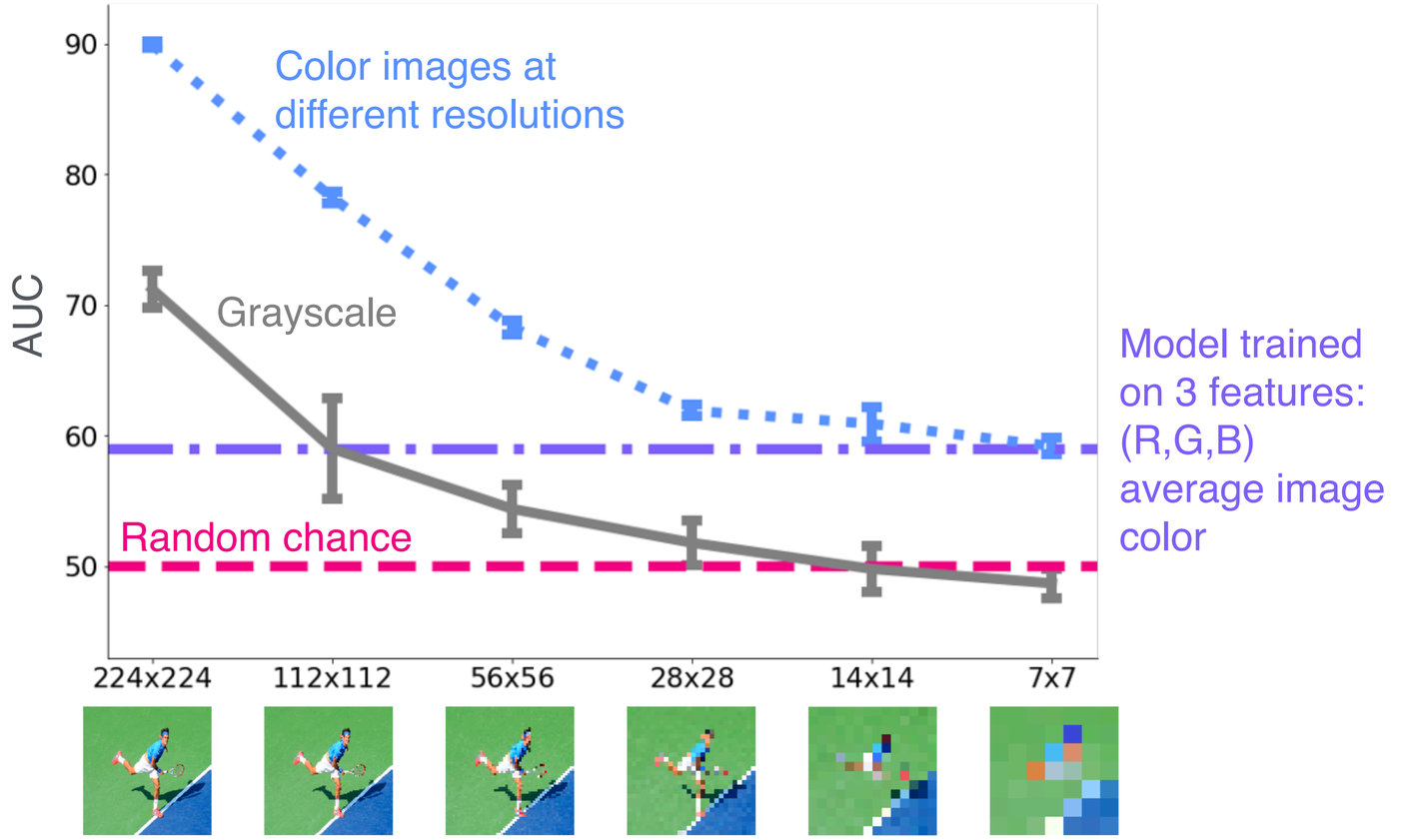
Goal: Understand how pervasive visual gender cues are in computer vision datasets

Method: Train a gender artifacts model (classifying if the image contains a person labeled “female” vs “male”), while obscuring or distorting parts of the image

If the model performs well despite distortion => gender cues are pervasive and difficult to remove

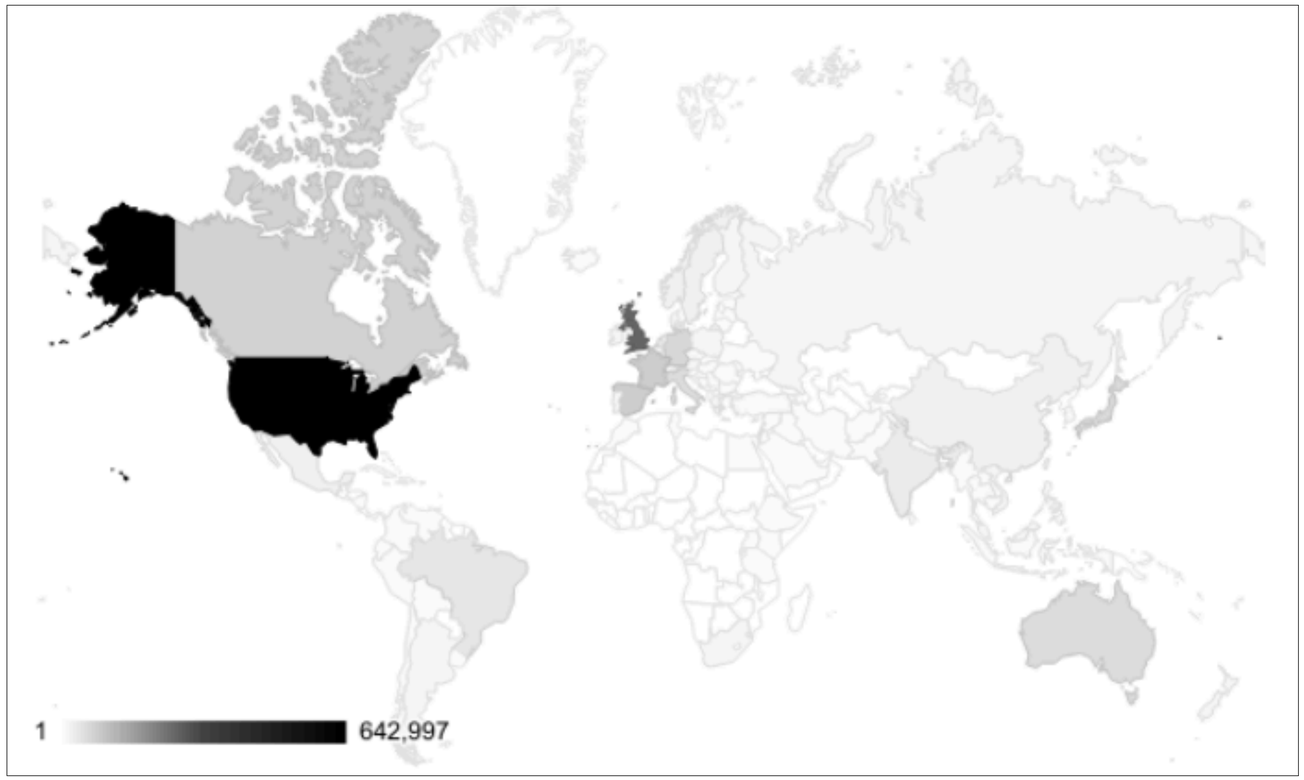
Images: COCO dataset [Lin et al. ECCV’14]

Gender labels: binarized, socially-perceived inferred gender expression [Zhao et al. EMNLP’17], [Zhao et al. ICCV'21]



[Nicole Meister, Dora Zhao, Angelina Wang, Vikram V. Ramaswamy, Ruth Fong, Olga Russakovsky. ICCV’23. “Gender artifacts in visual datasets.”]

Algorithmically mitigating bias in the data becomes impossible when data is simply missing



[Shreya Shankar et al. NeurIPS Workshop'17
"No classification without representation…"]

[Terrance DeVries et al. CVPR Workshop'19
"Does object recognition work for everyone"]

GeoDE: **Geo**graphically **D**iverse **E**valuation Dataset

Goal: Rethink data collection approaches and create a geographically diverse dataset

Approach: Partnered with Appen to solicit photographs from people around the world

Result: 61,940 images from 6 different regions and 40 different objects across the world

Sing Yu (Phoebe) Lin

Dora Zhao

Aaron B. Adcock

Laurens van der Maaten

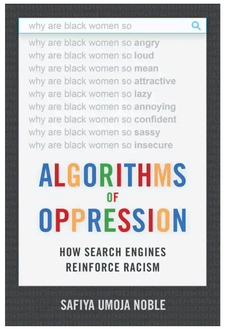
Deepti Ghadiyaram

[Vikram V. Ramaswamy, S. Y. Lin, D. Zhao, A. B. Adcock, L. van der Maaten, D. Ghadiyaram and O. Russakovsky. arXiv'23 "GeoDE: a Geographically Diverse…"]

GeoDE is collected using crowd-sourcing

Issues with web-scraping of images:

- Stereotypes, underrepresentation
- Geographic bias: most images are from Western Europe and North America
- Consent: who created the image? Do we have their consent to use their work to train an AI model?
- Privacy: who is pictured within the image? Do they know that their likeness is being used to train an AI model?



[Timnit Gebru et al. CACM'21 "Datasheets for datasets."]
[Eun Jo and Timnit Gebru. FAT*20 "Lessons from archives:…"]
[Morgan Scheuerman et al. CSCW'21 "Do Datasets Have Politics?"]
[Amandalynne Paullada et al. Patterns'21 "Data and its (Dis)contents…"]
[Abeba Birhane et al. arxiv'21 "Multimodal Datasets: Misogyny,…"]
[Abeba Birhane et al. arxiv'21 "The Values Encoded in Machine…"]
[Vinay Prabhu and Abeba Birhane. WACV'21 "Large datasets: a…"]
[Emily Denton et al. Big Data & Society'21 "On the Genealogy of…"]
[Shreya Shankar et al. NeurIPS Workshop'17 "No classification…"]
[Terrance DeVries et al. CVPR Workshop'19 "Does object…"]
[Bernard Koch et al. NeurIPS D&B track'21 "Reduced, Reused and…"]
[Margot Hanley et al. NeurIPS workshop'21 "An Ethical Highlighter…"]
[Milagros Miceli et al. FAccT'21 "Documenting Computer Vision…"]
[Ben Hutchinson et al. FAccT'21 "Towards Accountability for…"]

GeoDE is collected using crowd-sourcing

Advantages of crowd-sourcing:

- Target specific distributions of data
- Have consent; can ensure no recognizable people or PII in images
- No selection bias: images are not created to generate excitement or novel content

Disadvantages of crowd-sourcing:

- Cost!

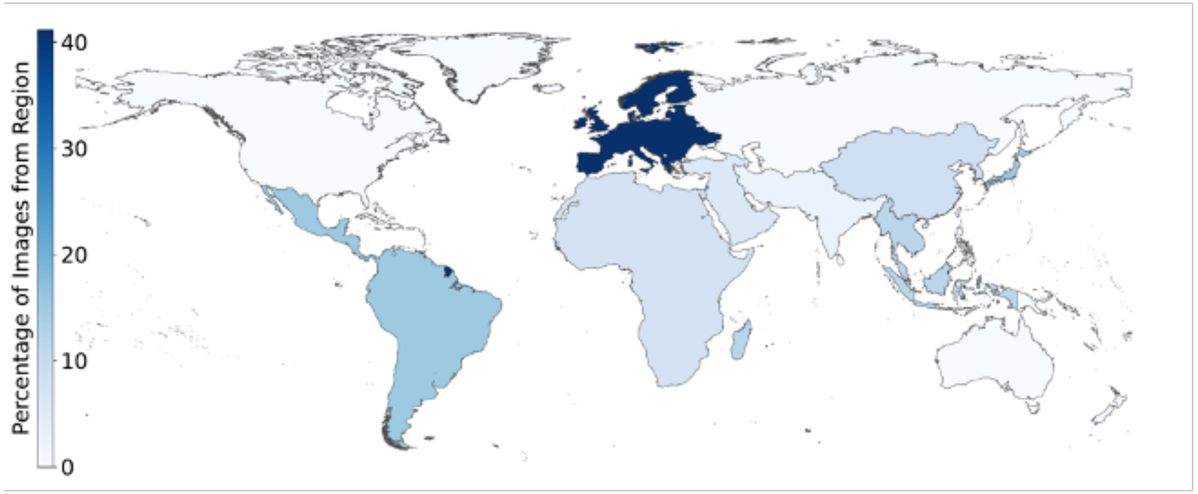
[Vikram V. Ramaswamy, S. Y. Lin, D. Zhao, A. B. Adcock, L. van der Maaten, D. Ghadiyaram and O. Russakovsky. arXiv'23 "GeoDE: a Geographically Diverse…"]

Takeaway 1: Distribution of images

We can ensure specific distribution of images:

GeoYFCC

[Abhimanyu Dubey et al. CVPR 2021 “Adaptive Methods for Real-World Domain Generalization”]



GeoDE

(ours)

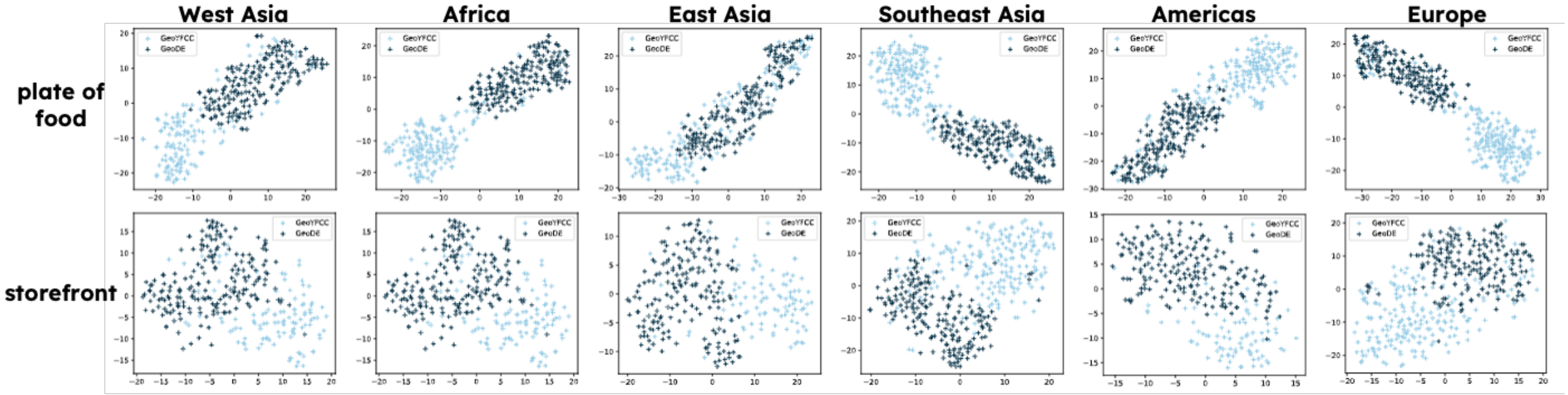
[Vikram V. Ramaswamy, S. Y. Lin, D. Zhao, A. B. Adcock, L. van der Maaten, D. Ghadiyaram and O. Russakovsky. arXiv’23 ”GeoDE: a Geographically Diverse…”]

Takeaway 2: Crowd-sourced images look different

[Vikram V. Ramaswamy, S. Y. Lin, D. Zhao, A. B. Adcock, L. van der Maaten, D. Ghadiyaram and O. Russakovsky. arXiv'23 "GeoDE: a Geographically Diverse…"]

Takeaway 2: Crowd-sourced images look different

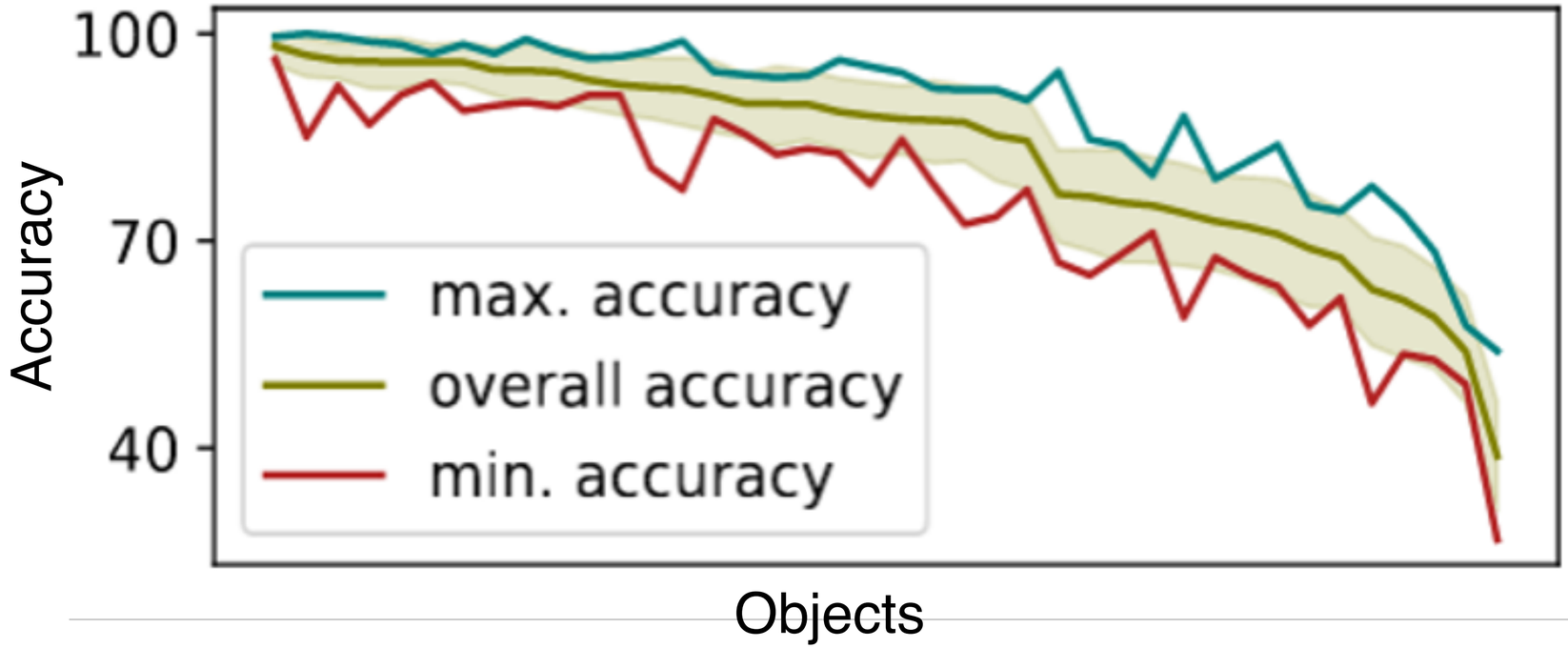
Feature representations between crowd-sourced and web-scraped images are different



[Vikram V. Ramaswamy, S. Y. Lin, D. Zhao, A. B. Adcock, L. van der Maaten, D. Ghadiyaram and O. Russakovsky. arXiv'23 "GeoDE: a Geographically Diverse…"]

Takeaway 3: GeoDE can find gaps in large models

Example: CLIP [Alec Radford et al. ICML'21. "Learning transferable visual models from natural language supervision."]



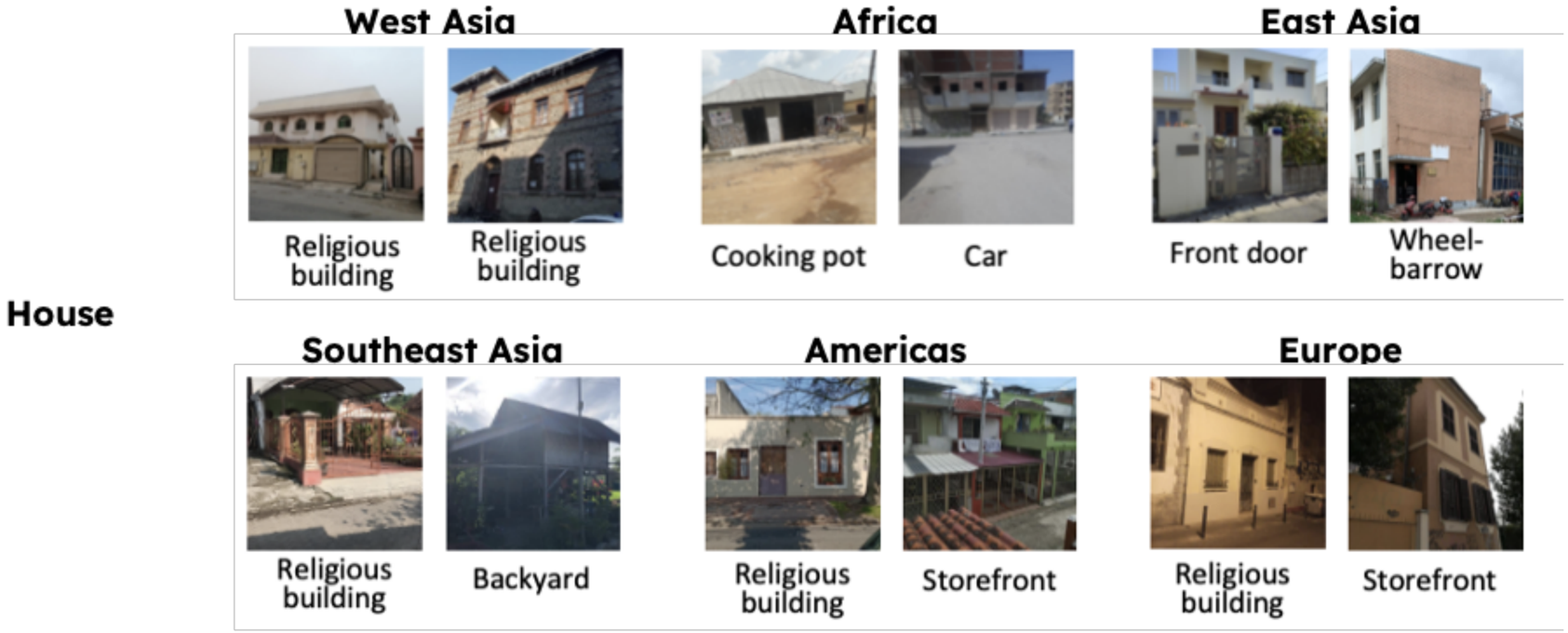
Discrepancy among regions: 31/40

objects have at least one region whose accuracy falls outside the 95% confidence interval

[Vikram V. Ramaswamy, S. Y. Lin, D. Zhao, A. B. Adcock, L. van der Maaten, D. Ghadiyaram and O. Russakovsky. arXiv'23 "GeoDE: a Geographically Diverse…"]

Takeaway 3: GeoDE can find gaps in large models

Example: CLIP [Alec Radford et al. ICML'21. "Learning transferable visual models from natural language supervision."]



Some stereotypes emerge: houses are predicted to be religions buildings, especially from West Asia and Southeast Asia.

[Vikram V. Ramaswamy, S. Y. Lin, D. Zhao, A. B. Adcock, L. van der Maaten, D. Ghadiyaram and O. Russakovsky. arXiv'23 "GeoDE: a Geographically Diverse…"]

Takeaway 4: Training with GeoDE improves model performance

Adding GeoDE to ImageNet improves performance on DollarStreet

[William A. Gaviria Rojas et al. NeurIPS D&B'22. "The Dollar Street Dataset: ImagesRepresenting the Geographic…".]

	ImageNet	+GeoDE		ImageNet	+GeoDE		ImageNet	+GeoDE
bicycle	92	**95**	hand soap	49	**65**	medicine	**80**	78
chair	86	**88**	house	88	**91**	plate of food	84	**96**
cleaning equip.	19	**36**	light fixture	36	**63**	stove	**89**	85
cooking pot	49	**61**	light switch	77	**79**	toy	50	**58**

[Vikram V. Ramaswamy, S. Y. Lin, D. Zhao, A. B. Adcock, L. van der Maaten, D. Ghadiyaram and O. Russakovsky. arXiv'23 "GeoDE: a Geographically Diverse…"]

Conclusions about GeoDE:

- Constructed GeoDE, a crowd-sourced dataset
- Crowd-sourcing leads to very different features spaces from web-scraping.
- GeoDE helps measure performance discrepancies across different regions
- Training with GeoDE can improve model performance


Sing Yu (Phoebe) Lin


Dora Zhao


Aaron B. Adcock


Laurens van der Maaten


Deepti Ghadiyaram

[Vikram V. Ramaswamy, S. Y. Lin, D. Zhao, A. B. Adcock, L. van der Maaten, D. Ghadiyaram, O. Russakovsky. arXiv'23 "GeoDE: a geographically diverse…"]

Conclusions about GeoDE:

- Constructed GeoDE, a crowd-sourced dataset
- Crowd-sourcing leads to very different features spaces from web-scraping.
- GeoDE helps measure performance discrepancies across different regions
- Training with GeoDE can improve model performance

Conclusions overall:

- Algorithmic bias mitigation solutions can be effective, but only up to a limit
- Web-scraping is great for scale, but need to diversify our data collection methods


Sing Yu (Phoebe) Lin


Dora Zhao


Aaron B. Adcock


Laurens van der Maaten


Deepti Ghadiyaram


Nicole Meister


Angelina Wang


Ruth Fong

Jia Deng, Li Fei-Fei, Kyle Genova, Kenji Hata, Yannis Karakozis, Leslie Kim, Anat Kleiman, Sunnie S. Y. Kim, Prem Nair, Alexander Liu, Arvind Narayanan, Iroha Shirai, Klint Qinami, Zeyu Wang, Kaiyu Yang, Jacqueline Yau, Ryan Zhang

[Vikram V. Ramaswamy, S. Y. Lin, D. Zhao, A. B. Adcock, L. van der Maaten, D. Ghadiyaram, O. Russakovsky. arXiv'23 "GeoDE: a geographically diverse…"]

[Angelina Wang, A. Liu, R. Zhang, A. Kleiman, L. Kim, D. Zhao, I. Shirai, A. Narayanan, O. Russakovsky. IJCV'22 "REVISE: A tool for revealing and…"]

[Kaiyu Yang, K. Qinami, L. Fei-Fei, J. Deng, O. Russakovsky. FAT*'20 "Towards…"]

[Zeyu Wang, K. Qinami, Y. Karakozis, K. Genova, P. Nair, K. Hata, O. Russakovsky. CVPR'20 "Towards fairness in visual recognition: effective strategies for…"]

[Nicole Meister, Dora Zhao, A. Wang, V. V. Ramaswamy, R. Fong, O. Russakovsky. ICCV'23. "Gender artifacts in visual…"]

[Vikram V. Ramaswamy, S. S. Y. Kim, O. Russakovsky. CVPR'21 "Fair attribute…"]

[Kaiyu Yang, J. Yau, L. Fei-Fei, J. Deng, O. Russakovsky. ICML'22 "A study of…"]

[Angelina Wang, O. Russakovsky. ICML'21 "Directional bias amplification"]

[Dora Zhao, A. Wang, O. Russakovsky. ICCV'21 "Understanding and evaluating…"]

[Angelina Wang, V. V. Ramaswamy, O. Russakovsky. FAccT'22 "Towards…"]