Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Appendix B.

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 8 of main text.
   (c) Did you discuss any potential negative societal impacts of your work? [No] In Section 1 of the main text, we describe the negative social impacts that modern vision models can have. Our paper hopes to shine light on these discrepancies, and explore what factors cause them to arise.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See the data card in Appendix B. Github link contains the code and data for reprehensibility.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] We evaluated pre-trained models. Computation usage is minimal.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes] See the data card in Appendix B.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See the data card in Appendix B.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] We provided additional annotations to the DollarStreet data. We did not collect any new photos.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] See Appendix A.1.3.

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [No]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [No]

A Appendix

A.1 Annotating Dollar Street with factor labels

A.1.1 DollarStreet Statistics

Region	Income Level low	medium	high
Africa	2141	1443	280
Asia	1362	8673	1424
Europe	0	1443	1455
The Americas	339	2093	1223

Table 3: Number of images for each region, income level pair in Dollar Street.

Table 3 shows the number of images in Dollar Street for each income and region pairing. We observe the distribution across images and regions is far from uniform, implying region and income distributions skew of counts are entangled. Consequently, we present both region and income comparisons where appropriate in our analysis.

A.1.2 Prototypical Image Selection

We define prototypical images for each class as those correctly classified by ResNet-50 model with the highest confidence. We use a ResNet-50 model pre-trained on ImageNet21k from Ridnik et al. [2021]. We select the ImageNet classes that overlap with Dollar Street labels, using the mapping as defined in [Goyal et al., 2022]. We use a soft-max over the sub-section of ImageNet classes that are in the mapping. We take the top predictions and confidence for these ImageNet classes and use the defined mapping from IN21k to Dollar Street in order to make DollarStreet class predictions. Out of the box, the model does not perform well on DollarStreet. Running a full pass over the dataset with Batch Norm in train mode, without any updates to the model weights, helps with the distribution shift from ImageNet to DollarStreet images, meaning overall accuracy is higher.

We select the three images that the model predicts successfully with the highest confidence. If such images do not exist, prototypical images are hand-selected. Table 4 shows the prototypical images used for five classes.

A.1.3 Annotation Setup

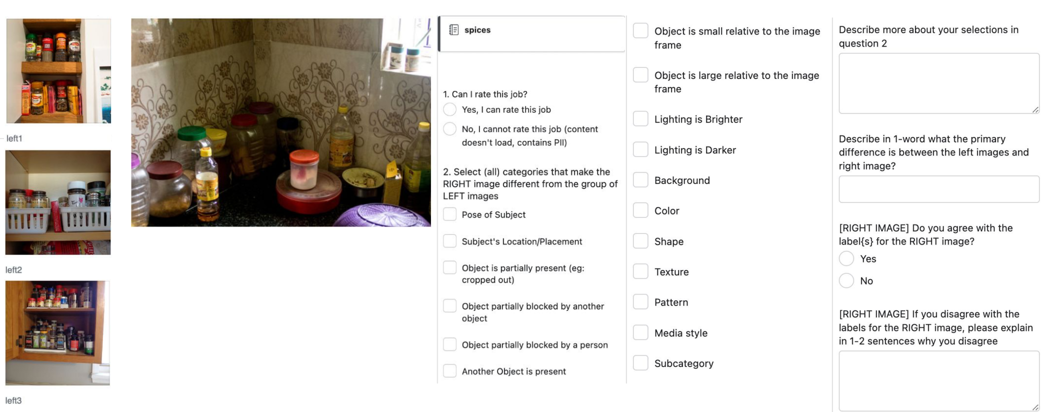


Figure 8: Example annotation task.

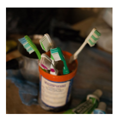
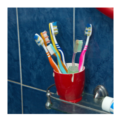
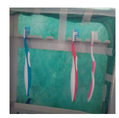
Class	Prototypical Images		
grains			
plates			
power outlets			
cleaning floors			
toothbrushes			

Table 4: Prototypical images used for five classes.

Figure 8 shows an example of the annotation task. Annotators select the factors distinguishing each image among sixteen factors such as pose, various forms of occlusion, size, style, type or breed. Annotators can select any number of distinctive factors for each image. We source 10 annotators through a third party vendor from South East Asia. In addition, we ask annotators to provide text descriptions to account for factors outside the sixteen we provide. We trained annotators with examples so that they were familiar with the task before annotating the target images. We had intermediate QA from the third party vendor monitoring annotations for quality. We also ask annotators whether they agree with the original class label for each image.

A.1.4 Label Agreement Annotation Setup

Country	Number of annotators
India	8
Nigeria	9
Brazil	13
United Arab Emirates	6
United States	8

Table 5: Annotator breakdown for label agreement task.

For our follow up annotations about label agreement, we sourced 44 annotators from 5 different countries, with the full demographics shown in Table 5. We asked one annotator per country about each image in question. In Table 9, we show example images from the three most disputed classes, along with alternative labels suggested by annotators. In Table 6, we show the classes with the highest and lowest levels of disagreement among annotators.

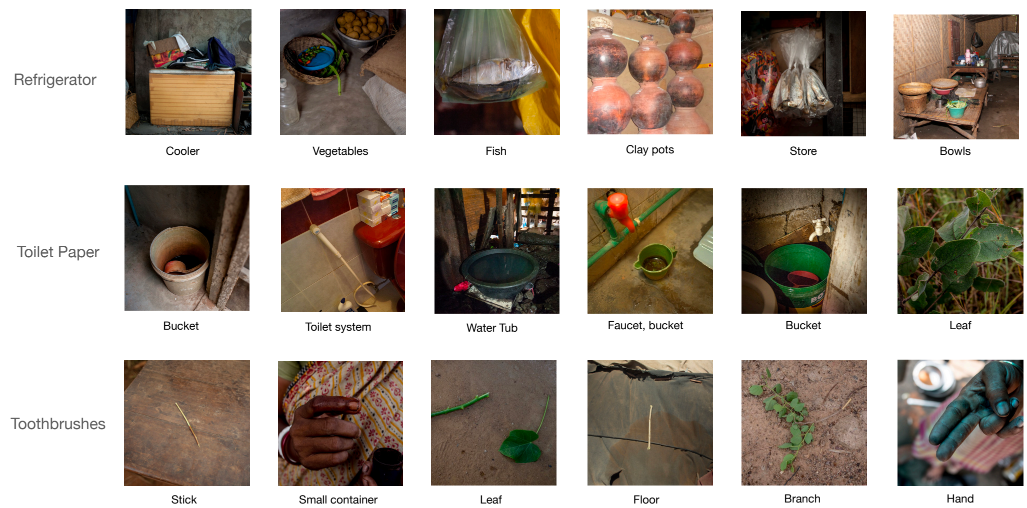


Figure 9: Randomly sampled example images and alternative labels given for the three classes with most disagreement. The original class label is shown on the left, and the alternative label given by the annotator shown below each image.

class	% disagreement	class	% disagreement
toilet paper	88.4	medication	10.0
refrigerators	83.5	fruit trees	11.8
toothbrushes	79.3	plates of food	12.3
sofas	77.8	trash	14.3
diapers	72.0	cleaning floors	14.5
armchairs	70.6	ceilings	15.0
showers	66.3	homes	15.0
kitchen sinks	64.5	books	15.0
wall clocks	63.2	cooking pots	19.8
radios	60.4	wheel barrows	20.0

Table 6: Top ten classes with the highest percentage of annotators who *disagreed* (left) and *agreed* (right) with the original class label

A.2 How do objects vary across incomes and geographies?

We show the most dissimilar classes across incomes and regions by comparing the Jensen-Shannon Distance of the factor annotation distributions in Tables 7 and 8.

class	income bucket	differentiating factors
roofs	`low` vs. `high`	subcategory, pose, smaller
ceilings	`low` vs. `high`	pose, subcategory, texture
diapers	`low` vs. `high`	color, shape, texture
radios	`low` vs. `high`	color, shape, subcategory
floors	`low` vs. `high`	texture, pose, pattern
sofas	`low` vs. `high`	color, texture, multiple objects
kitchen sinks	`low` vs. `high`	shape, pose, background
toilet paper	`low` vs. `high`	color, pose, background
wardrobes	`low` vs. `high`	background, pattern, color
mosquito protections	`low` vs. `high`	color, subcategory, pattern

Table 7: Classes with most stark differences in factor distributions by Jensen-Shannon Distance (JSD) across incomes.

class	regions	distinctive factors
chickens	`Asia` vs. `Europe`	partial view, color, shape
chickens	`Europe` vs. `Africa`	pose, partial view, color
diapers	`The Americas` vs. `Africa`	pose, color, partial view
pet foods	`Asia` vs. `The Americas`	color, texture, pattern
pet foods	`Asia` vs. `Europe`	pattern, subcategory, color
ceilings	`Europe` vs. `Africa`	pose, subcategory, texture
roofs	`Europe` vs. `Africa`	subcategory, pose, texture
car keys	`Asia` vs. `Europe`	pattern, partial view, subcategory
make up	`Europe` vs. `Africa`	background, subcategory, pattern
goats	`Asia` vs. `Africa`	pattern, color, subcategory

Table 8: Classes with most stark differences in factor distributions by Jensen-Shannon Distance (JSD) across regions

We show the most similar classes across incomes and regions using the same procedure of comparing Jensen-Shannon Distance of the factor annotation distributions in Tables 9 and 10.

class	income buckets	distinctive factors
vegetable plots	`low` vs. `high`	multiple objects, background, color
phones	`medium` vs. `high`	background, pose, multiple objects
pens	`medium` vs. `high`	color, background, pattern
bikes	`low` vs. `high`	background, subcategory, smaller
armchairs	`medium` vs. `high`	color, background, pose
latest furniture bought	`medium` vs. `high`	subcategory, background, color
child rooms	`medium` vs. `high`	pose, pattern, color
wall clocks	`medium` vs. `high`	color, pose, shape
cooking utensils	`medium` vs. `high`	pose, shape, pattern

Table 9: Classes most similar in factor distributions by Jensen Shannon Distance across incomes

class	regions	distinctive factors
vegetable plots	The Americas vs. Africa	pose, background, pattern
phones	Asia vs. Europe	pose, background, color
pens	Europe vs. Africa	pose, color, pattern
wheel barrows	Europe vs. The Americas	color, pose, background
ceilings	Asia vs. Africa	subcategory, pattern, texture
pets	Asia vs. Europe	background, pattern, subcategory
stoves	Asia vs. Africa	subcategory, color, pattern
menstruation pads	Asia vs. The Americas	pose, subcategory, pattern
tvs	Europe vs. The Americas	partial view, subcategory, background
everyday shoes	Europe vs. The Americas	color, partial view, shape

Table 10: Classes most similar in factor distributions by Jensen Shannon Distance (JSD) across regions

A.3 Evaluation Setup

CLIP Prompt Engineering We use CLIP in a zero shot setting, where we prompt the model using the set of Dollar Street classes (e.g. *medication, plates of food*) for each image to generate predictions. We generate the text prompts for CLIP by combining the 80 prompt templates used in the original CLIP paper with each Dollar Street class name, substituting _ for spaces. We consider an image correctly predicted if the top 5 classes predicted by CLIP is associated with the photo. *Note: Most photos in DollarStreet have only one label, but a small subset of (638) images containing multiple class labels (e.g. (cups, plates, dish racks) and (child rooms, kids bed, beds)).*

ImageNet21k as a shared taxonomy For models outside of CLIP, we use ImageNet21k to ground our models in a shared taxonomy. Following Goyal et al. [2022], we map the ImageNet21k labels to DollarStreet classes. We consider the image correctly classified if any of the top 5 ImageNet21k classes predicted by the model are mapped to any of the DollarStreet classes associated with the photo. We note that the mapping is not 1:1, and multiple classes in DollarStreet have multiple classes in ImageNet 21k that map to the single class. All of the models used for evaluation excluding CLIP and SEER are trained on ImageNet 21k. SEER is pre-trained in a self-supervised manner, and the model is fine-tuned on the 108 classes in ImageNet 21k that overlap with DollarStreet prior to evaluation. For ImageNet-21k pretraining, we use models from Ridnik et al. [2021].

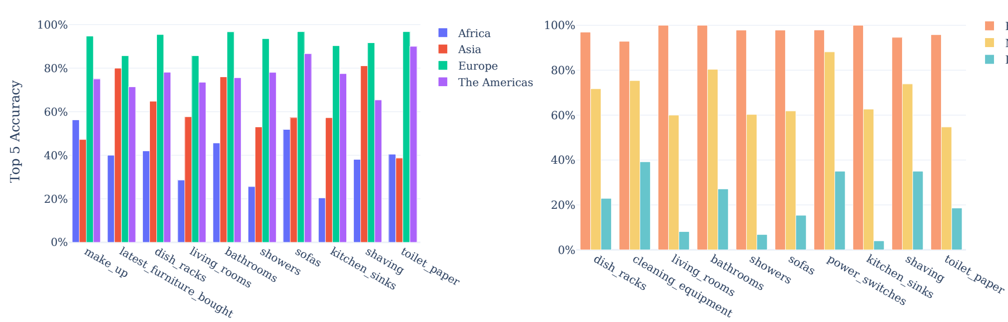


Figure 10: 10 classes with biggest performance discrepancy over regions (left) and income bucket (right).

Class level performance disparities Figure 10 shows the top 10 classes with the biggest performance disparity between groups for regions and incomes. We define the largest performance discrepancy as the maximum difference in accuracy between any two regions (or income buckets). At a class level, we find that the discrepancy in accuracy can be stark - over 50% for the classes with the widest gap. For both incomes and geographies, we find that the differences mostly pertain to items in kitchens (*dish racks, kitchen sinks*) and items in bathrooms (*showers, shaving, toilet paper, bathrooms*).

A.4 Explaining model performance disparities with factor labels

As part of our analysis of model performance disparities, we investigate the impact of pretraining class balance and image quality. In Table 11, we show the Pearson correlation coefficients and p-values between each model's top-5 accuracy and the Image DPI, a measure of image resolution. In Table 12, we show the Pearson correlation coefficients and p-values between each model's top-5 accuracy and the ImageNet-21K class count. We excluded CLIP from this analysis as CLIP was trained on a proprietary dataset.

Model	Correlation, Top 5 Accuracy and Image DPI
ViT	-0.019 (p = 0.035)
ResNet50	-0.023 (p = 0.008)
MLPMixer	-0.026 (p = 0.002)
BeIT	-0.003 (p = 0.72)
SEER	-0.016 (p = 0.057)
CLIP	-0.035 (p = 0.00005)

Table 11: Pearson Correlation coefficients and p-values between each model's top-5 accuracy and image quality, as measured by DPI.

Model	Correlation, Top-5 Accuracy and Class Count
ViT	0.126 ($p < 0.0001$)
ResNet50	0.142 ($p < 0.0001$)
MLPMixer	0.135 ($p < 0.0001$)
BeIT	0.222 ($p < 0.0001$)
SEER	0.103 ($p < 0.0001$)

Table 12: Pearson Correlation coefficients and p-values between each model's top-5 accuracy and ImageNet-21K class counts. CLIP is not included, as it was trained on a proprietary dataset.

Factors most associated with misclassifications differ considerably across regions and incomes. We find for the `high` income bucket, objects marked as *smaller* are most associated with mistakes, appearing +2.8x more among mistakes. On the other hand, *texture* which is not among the top five factors among mistakes in the `high` income bucket is associated with mistakes in the `medium` and `low` income buckets. *Texture* is +0.6x and +1.7x more likely to appear among mistakes in the `medium` and `low` income buckets respectively. We also find in the `low` income bucket, factors such as *occlusion* and *darker lighting* to be associated with model mistakes, appearing +1.2x and +0.9x more so among mistakes in the `low` income bucket. This suggests specific factors such as *texture*, *occlusion*, and *darker lighting* are associated with the disparity in performance we observe across incomes.

Further discussion of actors associated with mistakes across regions. We also measured the factors associated with model mistakes across regions in Figure 6 in Section 5.3 of the main text. In `Asia` we observe the factors most associated with mistakes are similar to those associated with mistakes overall. However, we find distinctive factors are associated with mistakes across each of the other regions. In the Americas, we find *smaller objects* (+1.2x more likely to appear among mistakes), followed by images with *multiple objects* (+0.3x). Similarly in `Europe`, *smaller objects* and *multiple objects* are most associated with mistakes appearing +2.8x and +0.7x more so among mistakes respectively. In `Africa` however, we find instead *texture* (+1.6x) most associated with mistakes, followed by *occlusion* (+0.9x) and *darker lighting* (+0.8x). This suggests the disparity due to lower performance in regions such as `Africa` are associated with distinct factors related to *texture*, *occlusion*, and *darker lighting*.

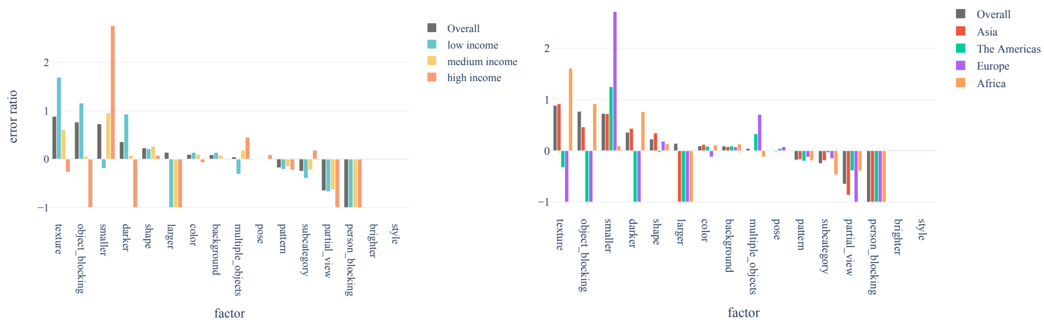


Figure 11: Shows the full error ratios for each factor per income bucket (left) and region (right). An error ratio higher than zero indicates the factor is more associated with model mistakes; less than zero indicates the factor is less likely to appear among a model's mistakes.

Statistical significance of error ratios for top factors. To confirm the top factors associated with model mistakes measured by our error ratio are statistically significant. We conduct a Chi-Squared test comparing the overall distribution of counts of the top factors to their distribution of counts among misclassifications. We find a statistically significant difference with a Chi-Squared statistic of 21.7 (p-value =0.0002).

Class	Income	Factors associated with mistakes
sofas	`low`	pattern (+0.5x), background (+0.3x), pose (+0.2x)
toilet paper	`low`	texture (+3.3x), shape (+2.7x), color (+0.8x)
living rooms	`low`	background (+0.8x), pose (+0.0x), color (-0.1x)
kitchen sinks	`low`	color (+0.5x), background (+0.3x), pose (+0.2x)
showers	`low`	background (+0.9x), pose (+0.3x), pattern (-0.5x)

Table 13: Class-specific vulnerabilities surfaced by our factor labels. We show vulnerabilities for the classes with lowest income performance.The values in parenthesis indicate how much more likely a factor is to appear for misclassified samples.

Factors most associated with largest discrepancies for classes across income buckets. We show the three factors most associated with model mistakes for the classes across income buckets with largest performance gap in Table 13. Trends are similar to those shown in the main paper for the largest disparity per region.

Additional analysis of vulnerabilities by country In Table 14 we show the most vulnerable factor by country along with its error ratio for CLIP with a ViT encoder.

Additional analysis on the effect of architecture and training procedure We extend our evaluation of CLIP models to include a ResNet50 encoder, to enable more consistency between the architectures of our CLIP and supervised models. Results per income are given in 15.

A.5 Texture debiasing experimental details

To measure the effect of reducing texture bias from Geirhos et al., we create a mapping from Dollar Street classes to ImageNet-1k similar to Rojas et al. [2022]. We initialize the mapping by matching the embedding similarity of each class name to its nearest neighbors from ImageNet-1k using a pre-trained Spacy language model `eng-large` `https://spacy.io/usage/linguistic-features#vectors-similarity`. We then manually correct any

Country	Most vulnerable factor	Error Ratio
Bangladesh	shape	1.77
Bolivia	color	0.48
Brazil	multiple_objects	3.25
Burkina Faso	texture	1.56
Burundi	color	0.42
Cambodia	texture	0.7
Cameroon	background	0.23
China	smaller	3.52
Colombia	color	0.34
Egypt	pattern	0.47
France	pose	0.88
Haiti	shape	0.26
India	texture	1.2
Indonesia	smaller	2.15
Cote d'Ivoire	texture	2.17
Jordan	background	0.29
Kazakhstan	background	0.82
Kenya	background	0.51
South Korea	shape	1.71
Latvia	smaller	7.88
Lebanon	background	0.82
Liberia	color	0.49
Malawi	texture	2.27
South Africa	color	0.73
Mexico	pose	0.18
Myanmar	texture	1.84
Nepal	texture	1.65
Netherlands	pose	0.94
Nigeria	texture	1.29
Pakistan	background	1.05
Palestine	background	0.49
Papua New Guinea	background	0.29
Peru	pose	0.91
Philippines	texture	1.98
Romania	background	0.43
Russia	pose	0.42
Rwanda	texture	3.72
Somalia	background	0.58
Sri Lanka	background	2.65
Sweden	pose	0.28
Thailand	subcategory	0.16
Tunisia	background	0.4
United States	smaller	3.49
Ukraine	pose	0.38
United Kingdom	pose	0.39
Vietnam	background	0.3
Zimbabwe	background	0.55

Table 14: The most vulnerable factor for a CLIP ViT per country.

issues in this mapping to produce ImageNet-1k mappings for approximately half of the Dollar Street classes. Note for all other analysis we use the ImageNet-21k mapping from Goyal et al. [2022].

Model	high	middle	low
BEiTPretrained21k	64.9	60.4	51.7
MLPMixerPretrained21k	88.3	79.9	61.9
SeerPretrained	89.6	81.7	64.3
ViTPretrained21k	88.3	79.9	61.3
ResNet50Pretrained21k	86.5	77.4	58.4
CLIP ViTB/32	92.6	83.4	66.9
CLIP ResNet101	91.2	82.9	68.2
CLIP ResNet50	95.7	88.8	73.4

Table 15: Comparison of model architectures across incomes. Overall, we find that while architecture and learning objective are important factors for fairness considerations, there are consistent and similar vulnerabilities across models.

B Data Card

We provide a data card for our annotations, following the guidance of Pushkarna et al. [2022].

DollarStreet Factor Annotations

We provide annotations for Dollar Street images with distinctive factor labels such as pose, background, and color to explain performance disparities in models.

Data is available at:
https://github.com/facebookresearch/dollarstreet_factors
Data visualizer is available at:
https://dollarstreetfactors.metademolab.com/

Overview	
Publisher	Meta
Authors	Laura Gustafson, Megan Richards, Melissa Hall, Caner Hazirbas, Diane Bouchacourt, Mark Ibrahim
Contact	dollarstreet-factors@meta.com
Funding & Funding Type	Fundamental AI Research
License	CC BY-NC 4.0
Applications	
Dataset Purpose	Evaluate computer vision models robustness to common factors to help pinpoint where geographical and economical performance discrepancies arise.
Key Application	*Computer Vision, Robustness, Fairness*
Primary Motivations	We can use the factors to identify model vulnerabilities that contribute to these discrepancies. Pinpointing the vulnerabilities will help guide research into developing fairer models.
Intended Audience	Vision researchers aiming to analyse their trained vision models.
Suitable Use Case	Evaluation of Computer Vision models and analysis as to a model's strengths and weaknesses.
Data Type	
Primary Data Type	Annotations for existing DollarStreet dataset of images
Primary Annotation Type	Annotations are manually gathered from expert annotators. Annotations are booleans for each of the factors, along with single word and paragraph responses detailing the annotator's logic.
Data SnapShot	Dataset contains • Annotations for **14k** images • Each image is annotated with **16** factors
Data Sources	Annotations were manually gathered. Annotations are for images from the existing public DollarStreet dataset. https://www.gapminder.org/dollar-street Images are licensed under CC-BY.

DollarStreet Factor Annotations	
Annotation format	Each item in the annotation file will contain: 1. Image information: • `url`: Public url of image • `full_image_id`: Unique ID of image • `household_id`: Unique ID of household who took the photo • `class`: Image classification class 2. Group information: • `region`: Region where the image is from. Options are *The Americas, Europe, Africa, Asia*. Derived from `country`. • `income_bucket`: Income bucket of household who took the image. Options are *high income, middle income, low income*. Derived from `income` • `country`: String of country name where image is taken. • `lat`: Latitude of country • `lng`: Longitude of country • `income`: Integer of income of household. TODO metric 3. Summary: • `one_word`: One word describing how the image differs from the prototypical images for it's class. • `justification`: String description for the annotators' justification of their one word summary. • `agree_right`: Boolean describing whether the annotator agreed with the class label • `why_disagree`: If `agree_right` is `False` this will contain a string explanation as to why the annotator disagreed with the class label 4. Factors (Boolean): • `multiple_objects` • `background` • `color` • `brighter` • `darker` • `style` • `larger` • `smaller` • `object_blocking` • `person_blocking` • `partial_view` • `pattern` • `pose` • `shape` • `subcategory` • `location` • `texture`

B.1 Interactive factor dashboard

We show screenshots of our interactive dashboard for exploring the factor labels across regions in Figures 12 and 13. The dashboard allows for interactive queries by region, income, factor label. Each query yields sample images, which you can interactively explore annotations for as shown in 13. We hope this tool will allow researchers to easily explore factor labels associated with images across axes such as regions or incomes to spur further research into reliable vision systems.

Figure 12: Interactive dashboard for Dollar Street factor annotations with an income and factor label query (for texture).

Figure 13: Interactive dashboard for Dollar Street factor annotations illustrating an example of the annotations.

B.2 Sample images

Country				
The Americas				
Africa				

Table 17: Examples of diaper images. Our factors surfaced that images of diapers in Dollar Street between regions differed most among *pose*, *color*, *partial view*.

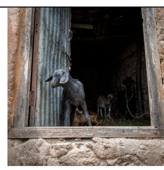
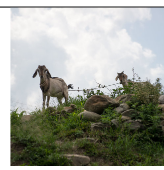
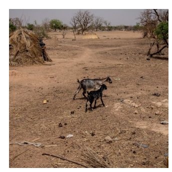
Country				
Asia				
Africa				

Table 18: Examples of goat images. Our factors surfaced that images of goats in Dollar Street between regions differed most among *pattern*, *color*, *subcategory*

In Tables 18 and 17 we show example images from classes and regions that were found to have some the starkest difference in factors, as measured by JSD.