

Appendices

A Gender-based Metric: Contextual Representation

As one of our gender-based metrics from Section 5.1, we look into the contexts that each gender tends to be featured in, and with COCO as an example (Fig. 8), can see that images with females tend to be more indoors in scenes like **shopping** and **dining** and with object groups like **furniture**, **accessory**, and **appliance**. On the other hand, males tend to be in more outdoors scenes like **sports fields** and **water**, **ice**, **snow**, and with object groups like **sports** and **vehicle**. These trends reflect gender stereotypes in many societies and can propagate into the models. While there is work on algorithmically intervening to break these associations, there are often too many proxy features to robustly do so. Thus it is useful to intervene at the dataset creation stage.

B Validating Distance as a Proxy for Interaction

In Section 5.1, Instance Counts and Distances, we make the claim that we can use distance between a person and an object as a proxy for if the person, p , is actually interacting with the object, o , as opposed to just appearing in the same image with it. This allows us to get more meaningful insight as to how genders may be interacting with objects differently. The distance measure we define is $dist = \frac{\text{distance between } p \text{ and } o \text{ centers}}{\sqrt{\text{area}_p * \text{area}_o}}$, which is a relative measure within each object class because it makes the assumption that all people are the same size, and all instances of an object are the same size. To validate the claim we are making, we look at the SpatialSense dataset [67]; specifically, at 6 objects that we hope to be somewhat representative of the different ways people interact with objects: **ball**, **book**, **car**, **dog**, **guitar**, and **table**. These objects were picked over ones such as **wall** or **floor**, in which it is more ambiguous what counts as an interaction. We then hand-labeled the images where this object cooccurs with a human as “yes” or “no” based on whether the person of interest is interacting with the object or not. We pick the threshold by optimizing for mean per class accuracy, where every distance below it as classified as a “yes” interaction and every distance above it as a “no” interaction. The threshold is picked based on the same data that the accuracy is reported for.

As can be seen in Table 4, for all 6 categories the mean of the distances when someone is interacting with an object is lower than that of when someone is not. This matches our claim that distance, although imperfect, can serve as a proxy for interaction. From looking at the visualization of the distribution of the distances in Fig. 9, we can see that for certain objects like **ball** and **table**, which also have the lowest mean per class accuracy, there is more overlap between the distances for “yes” interactions and “no” interactions. Intuitively, this makes some sense, because a **ball** is an object that can be interacted with both from a

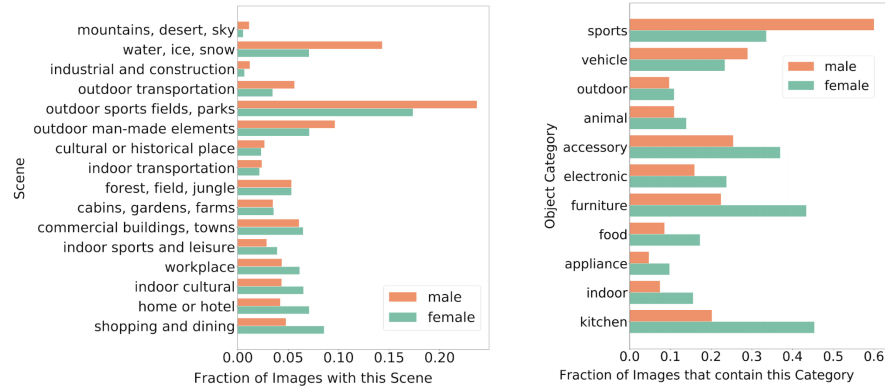


Fig. 8: Contextual information of images by gender, represented by fraction that are in a scene (left) and have an object from the category (right).

Table 4: Distances are classified as “yes” or “no” interaction based on a threshold optimized for mean per class accuracy. Visualization of the classification in Fig. 9. Distances for “yes” interactions are lower than “no” interactions in all cases, in line with our claim that smaller distances are more likely to signify an interaction.

Object	# Labeled Examples	Mean Per Class Accuracy (%)	“Yes” Distance mean \pm std	“No” Distance mean \pm std	Threshold
ball	107	67	6.16 \pm 2.64	8.54 \pm 4.15	7.63
book	27	78	2.45 \pm 1.99	4.84 \pm 2.24	3.88
car	135	71	2.94 \pm 3.20	4.59 \pm 2.97	2.74
dog	58	71	1.08 \pm 1.12	2.07 \pm 1.79	0.60
guitar	40	88	0.90 \pm 1.77	2.13 \pm 1.21	1.61
table	76	67	1.88 \pm 1.19	3.28 \pm 2.45	2.47

distance and from direct contact, and for **table** in the labeled examples, people were often seated at a table but not directly interacting with it.

C Pairwise Queries

In Section 4.2, another claim we make is that pairwise queries of the form “[Desired Object] and [Suggested Query Term]” could allow dataset collectors to augment their dataset with the types of images they want. One of the examples we gave is that if one notices the images of **airplane** in their dataset are overrepresented in the larger sizes, our tool would recommend they make the query “**airplane** and **surfboard**” to augment their dataset, because based on the distribution of training samples, this combination is more likely than other kinds of queries to lead to images of smaller airplanes.

However, there are a few concerns with this approach. For one, certain queries might not return any search results. This is especially the case when the suggested query term is a scene category, such as **indoor cultural**, in which the query “**pizza** and **indoor cultural**” might not be very fruitful. To deal with this, we

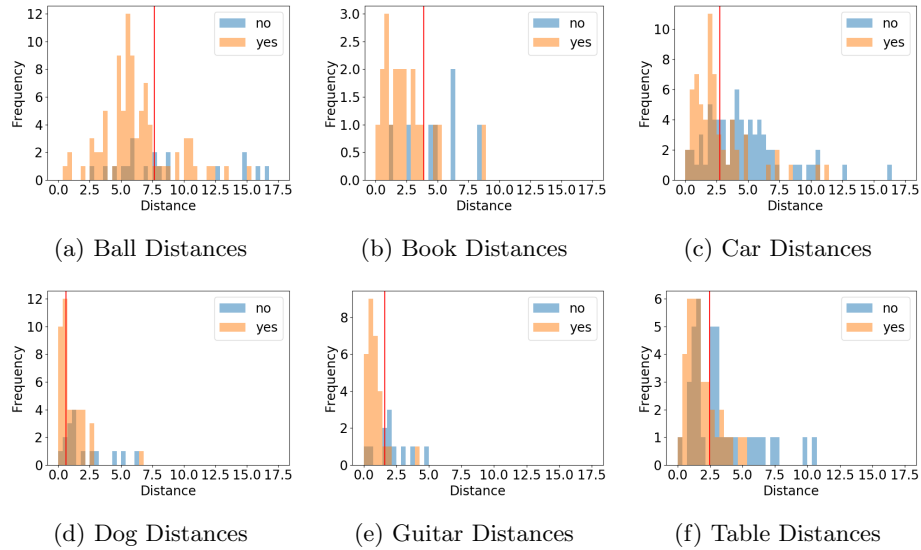


Fig. 9: Distances for the objects that were hand-labeled, orange if there is an interaction, and blue if there is not. The red vertical line is the threshold along which everything below is classified as “yes”, and everything above is classified as “no”.

can substitute the scene category, `indoor cultural`, for more specific scenes in that category, like `classroom` and `conference`, so that the query becomes something like “`pizza` and `classroom`”. When the suggested query term involves an object, there is another approach we can take. In datasets like PASCAL VOC [19], the set of queries used to collect the dataset is given. For example, to get pictures of `boat`, they also queried for `barge`, `ferry`, and `canoe`. Thus, in addition to querying, for example, “`airplane` and `boat`”, one could also query for “`airplane` and `ferry`”, “`airplane` and `barge`”, etc.

Another concern is there might be a distribution difference between the correlation observed in the data and the correlation in images returned for queries. For example, just because `cat` and `dog` cooccur at a certain rate in the dataset, does not necessarily mean they cooccur at this same rate in search engine images. However, our query recommendation rests on the assumptions that datasets are constructed by querying a search engine, and that objects cooccur at roughly the same relative rates in the dataset as they do in query returns; for example, that because `train` cooccurring with `boat` in our dataset tends to be more likely to be small, in images returned from queries, `train` is also likely to be smaller if `boat` is in the image. We make an assumption that for an image that contains a `train` and `boat`, the query “`train` and `boat`” would recover these kinds of images back, but it could be the case that the actual query used to find this image was “`coastal transit`.” If we had access to the actual query used to find each image, the conditional probability could then be calculated over the queries themselves rather than the object or scene cooccurrences. It is

because we don’t have these original queries that we use cooccurrences to serve as a proxy for recovering them.

To gain some confidence in our use of these pairwise queries in place of the original queries, we show qualitative examples of the results when searching on Flickr for images that contain the tags of the object(s) searched. We show the results of querying for (1) just the object (2) the object and query term that we would hope leads to more of the object in a smaller size, and (3) the object and query term that we would hope leads to more of the object in a bigger size. In Figs. 10 and 11 we show the results of images sorted by relevance under the Creative Commons license. We can see that when we perform these pairwise queries, we do indeed have some level of control over the size of the object in the resulting images. For example, “**pizza** and **classroom**” and “**pizza** and **conference**” queries (scenes swapped in for **indoor cultural**) return smaller pizzas than the “**pizza** and **broccoli**” query, which tends to feature bigger pizzas that take up the whole image. This could of course create other representation issues such as a surplus of **pizza** and **broccoli** images, so it could be important to use more than one of the recommended queries our tool surfaces. Although this is an imperfect method, it is still a useful tactic we can use without having access to the actual queries used to create the dataset.⁶

⁶ We also looked into using reverse image searches to recover the query, but the “best guess labels” returned from these searches were not particularly useful, erring on both the side of being much too vague, such as returning “sea” for any scene with water, or too specific, with the exact name and brand of one of the objects.

Object of Interest	train	
Query for Smaller Train	train + boat	
Query for Bigger Train	train + backpack	
Object of Interest	pizza	
Query for Smaller Pizza	pizza + classroom or conference	
Query for Bigger Pizza	pizza + broccoli	

Fig. 10: Screenshots of top results from performing queries on Flickr that satisfy the tags mentioned. For **train**, when it is queried with **boat**, the **train** itself is more likely to be farther away, and thus smaller. When queried with **backpack**, the image is more likely to show travelers right next to, or even inside of, a **train**, and thus show it more in the foreground. The same idea applies for **pizza** where it's imaged from further in the background when paired with an **indoor cultural** scene, and up close with **broccoli**.

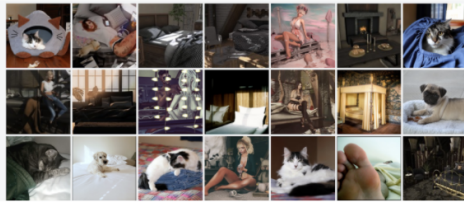
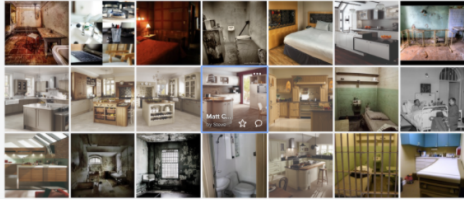



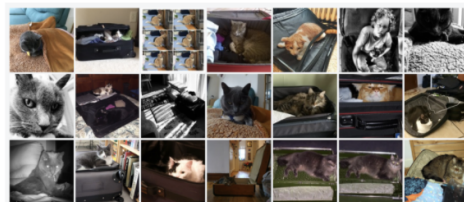
Object of Interest	bed	
	bed + sink	
	bed + cat	
Object of Interest	cat	
	cat + sheep	
	cat + suitcase	

Fig. 11: Screenshots of top results from performing queries on Flickr that satisfy the tags mentioned. For **bed**, **sink** provides a context that makes it more likely to be imaged further away, whereas **cat** brings **bed** to the forefront. The same is the case when the object of interest is now **cat**, where a pairwise query with **sheep** makes it more likely to be further, and **suitcase** to be closer.