Compare and Contrast: Learning Prominent Visual Differences (Supplementary)

Steven Chen Kristen Grauman
The University of Texas at Austin

This supplementary document contains the following:

- Additional dataset annotator agreement statistics, as referenced by Section 5.1.
- Additional detail on the binary attribute dominance baseline, as referenced by Section 5.2.
- Additional prominence prediction examples, as referenced by Section 5.3.
- Additional example of image search results, as referenced by Section 5.4.
- Description generation offline results using the SVM relative attribute ranker, as referenced by Section 5.5.

1. Dataset Annotator Agreement

Figure 1 shows the frequency of each attribute appearing as the ground truth most prominent difference for both Zap50K and LFW10. The statistics show that prominence occurs diversely across the attribute vocabularies. For both vocabularies of ten attributes, no single attribute was chosen as prominent more than 19% of the time.

2. Binary Attribute Dominance Baseline

To ensure a fair baseline, we follow the approach of Turakhia and Parikh [4] as closely as possible, collecting dominance annotations to train the dominance baseline model, and building binary attribute classifiers to produce input features for the dominance model. First, we directly convert our vocabulary of relative attributes into binary attributes, e.g., *sportiness* becomes *is sporty* or *is not sporty*, *fanciness* becomes *is fancy* or *is not fancy*, etc.

We collect binary attribute ground truth for each single image and attribute in our datasets, asking annotators whether the image contains or does not contain each attribute. We show each attribute and image to five different Mechanical Turk annotators, and take the majority presence/absence vote as the binary attribute ground truth. We use this ground truth to train M binary attribute SVM classifiers, one for each attribute.

Next, we collect dominance annotations at the category level, using the same interface and parameters as Turakhia and Parikh [4]. For each category, and for each possible combination of attribute pair, we ask annotators to choose which attribute pops out more. Dominance ground truth, as defined by [4], is the number of annotators that selected the attribute when it appeared as one of the options for that category. We follow the approach of Turakhia and Parikh [4] for training, projecting the category-level dominance ground truth to each training image in the split. We represent the images by their Platt scaled [3] binary attribute SVM classifier outputs.

Note that the method of [4] does not predict prominent differences. Nonetheless, in order to provide a comparison with our approach, we add a mapping from attribute dominance predictions to estimated prominent differences. In particular, to predict the most prominent difference given a novel image pair, we first compute binary attribute dominance values for each image in the pair, then select the attribute with the highest dominance value among both images as the predicted prominent difference. This method selects the attribute that sticks out as most dominant from either of the single images in the input pair.

3. Results

3.1. Prominence Prediction

Figure 2 shows additional qualitative examples of prominence predictions made by our approach on both Zap50K [5] and LFW10 [2], including both success cases and failure cases.

3.2. Image Search

Figure 3 shows an additional example of top search results returned by our algorithm and the WhittleSearch [1] baseline.

3.3. Description Generation

Figure 4 displays description generation offline results using the SVM relative attribute ranker scores as input features. Our approach significantly outperforms all baselines. These results are similar to the description generation re-

sults in the main work using the CNN ranker, and were omitted due to space constraints.

References

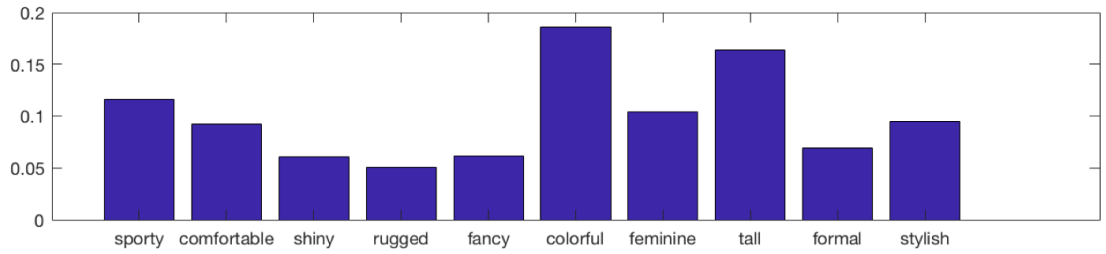
[1] A. Kovashka, D. Parikh, and K. Grauman. Whittlesearch: Image search with relative attribute feedback. In *CVPR*, 2012.

[2] S. Maji and G. Shakhnarovich. Part and attribute discovery from relative annotations. *IJCV*, 108(1):82–96, 2014.

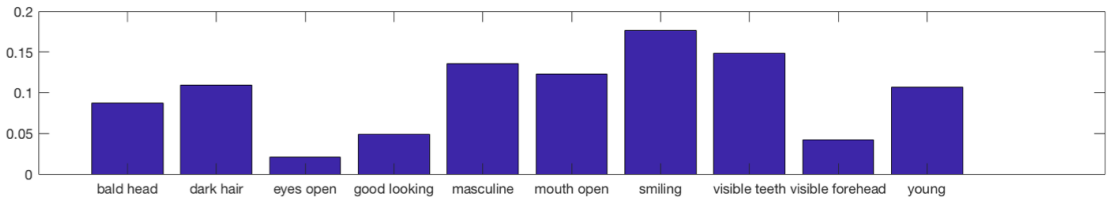
[3] J. C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 1999.

[4] N. Turakhia and D. Parikh. Attribute dominance: What pops out? In *ICCV*, 2013.

[5] A. Yu and K. Grauman. Fine-Grained Visual Comparisons with Local Learning. In *CVPR*, 2014.



(a) Ground truth attribute frequency for Zap50K [5].



(b) Ground truth attribute frequency for LFW10 [2].

Figure 1: Frequency of each attribute appearing as the ground truth prominent difference for Zap50K [5] and LFW10 [2].

(a) **shiny** (>),
feminine, colorful

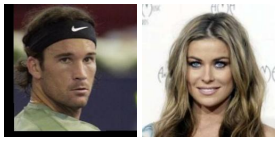
(b) **rugged** (<),
tall, feminine

(c) **tall** (<),
colorful, sporty

(d) **colorful** (>),
sporty, comfortable

(e) **formal** (>),
comfortable, shiny

(f) **tall** (<),
comfortable, sporty



(g) **masculine** (>),
smiling, visible teeth

(h) **masculine** (<),
mouth open, visible teeth

(i) **smiling** (>),
visible teeth, masculine

(j) **sporty** (>)
GT: feminine

(k) **visible teeth** (>)
GT: mouth open

(l) **dark hair** (>)
GT: bald

Figure 2: Success and failure cases of prominence predictions made by our approach. Success cases shown above line, with predicted most prominent attribute shown in bold, followed by next two most confident attributes. Failure cases shown below, with our prediction in bold, followed by the ground truth prominent difference.

Figure 3: Sample image search result. We show the target image along with the top eight ranked images produced by the baseline WhittleSearch [1] and our prominence approach, after two iterations of search. Our approach brings more relevant images: in this case, colorful and flat sneakers, whereas the baseline returns many unrelated images.

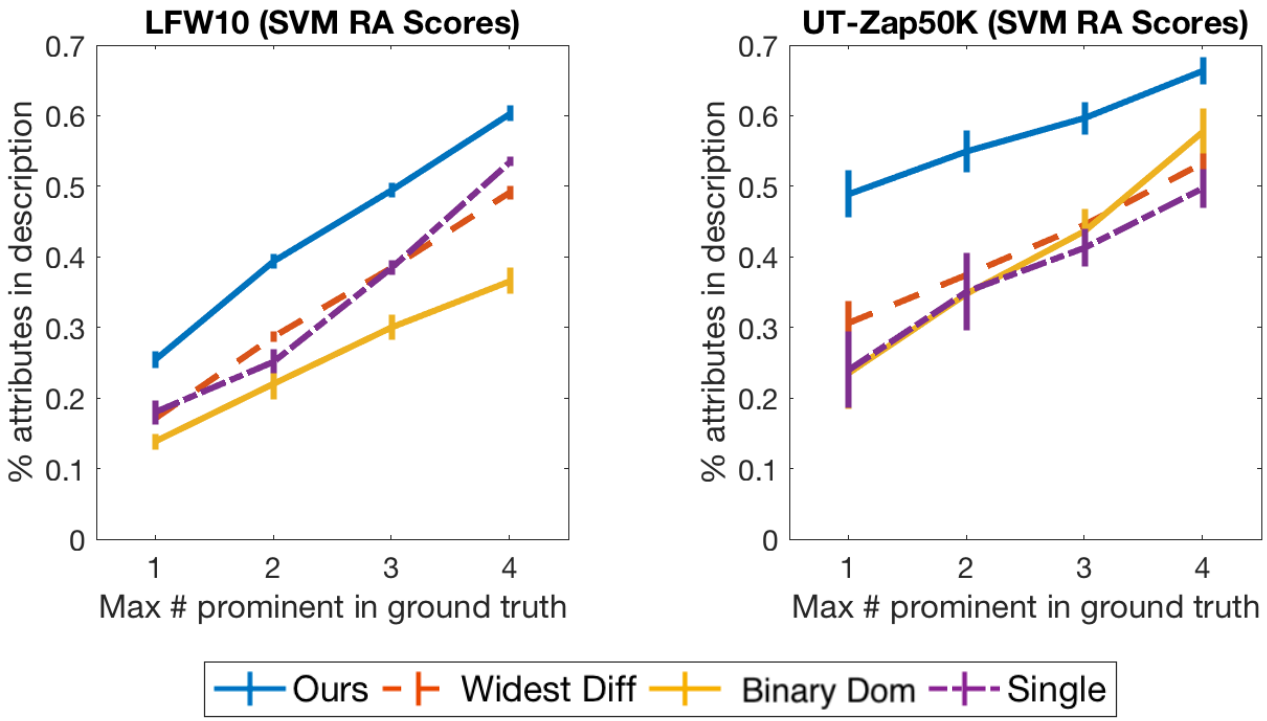


Figure 4: Description generation offline results using the SVM relative attribute ranker. Results are similar to the description generation results in the main paper using the CNN ranker, and were omitted due to space constraints.