

# Fine-Grained Visual Comparisons with Local Learning (Supplementary File)

Aron Yu and Kristen Grauman  
University of Texas at Austin

aron.yu@utexas.edu, grauman@cs.utexas.edu

## 1. Data Collection on Mechanical Turk

In order to obtain pairwise supervision data for our new dataset, we employed a crowd-sourcing strategy on Mechanical Turk (mTurk) by asking workers to perform relative comparison tasks. Before workers were allowed to work on our Human Intelligence Task (HIT), they must first complete a simple qualification test to learn about the visual attributes. Figure 1 shows the spectrum images used for the 4 attributes from Zap50K.



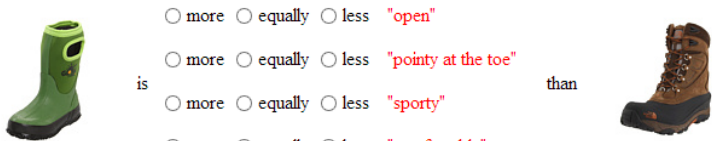
Figure 1: Shoe spectrums used to teach the workers about visual attributes.

The first stage of our tasks contained all 3,000 image pairs. We asked the workers to compare 10 pairs of unique images with respect to all 4 attributes. There were a total of 40 questions per HIT and we paid the workers 15 cents per HIT. The average completion time for these HITs was 5 minutes. The second stage of our tasks contained 4,612 fine-grained image pairs. Since these tasks were significantly more challenging, we asked the workers to compare 12 pairs of images with respect to only 1 attribute per pair. There were 12 questions per HIT and we paid the workers 18 cents per HIT. The average completion time for these HITs was 7 minutes. Screenshots of the HITs are shown in Figure 2.

## 2. Image Features

Different image features were used for each dataset. OSR scene images made use of 512-dimensional GIST descriptors while PubFig face images made use of a concatenation of 512-dimensional GIST descriptors and 30-dimensional LAB color

**Image Pair #7**



is  more  equally  less **"open"** than  high  mid  low


more  equally  less **"pointy at the toe"** with  high  mid  low confidence.

more  equally  less **"sporty"**  high  mid  low

more  equally  less **"comfortable"**  high  mid  low

(a) First Round

**Image Pair #4 (comfortable)**



**Shoe A** **Shoe B**

**Shoe A** is more "comfortable" than **Shoe B**.

**Shoe B** is more "comfortable" than **Shoe A**.

---

I'm **very confident** about my decision.

I'm **somewhat confident** about my decision.

I'm **not confident** about my decision.

---

Briefly explain your reasoning for this choice.

(b) Second Round

Figure 2: Sample questions from mTurk HITs.

histogram. Both sets of features were provided by the authors of Relative Attributes. Zap50K shoe images made use of a concatenation of 960-dimensional GIST descriptors and 30-dimensional LAB color histogram, which we extracted ourselves.

**3. Result Plots**

Due to space constraints, we could only include representative per-attribute plots in the paper. For completeness, here we show all of them. (None are new results; all outcomes are summarized in the main paper.) The cumulative accuracy curves for all Zap50K attributes are shown in Figure 3. The precision-recall curves for all OSR and PubFig attributes are shown in Figure 4 and 5.

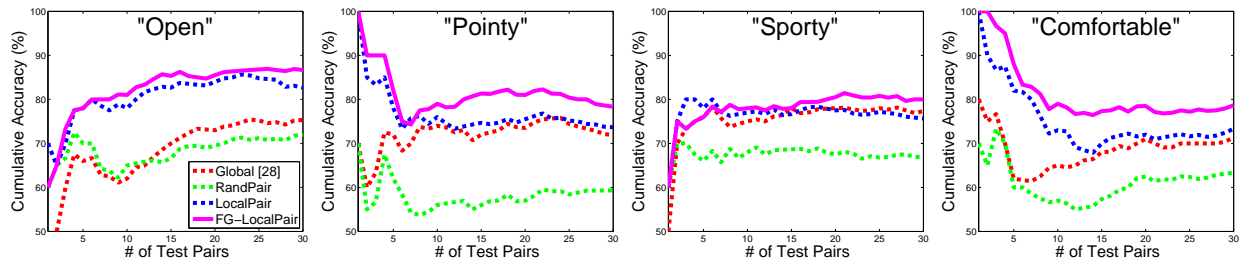


Figure 3: Cumulative accuracies for the 30 hardest pairs in Zap50K-1.

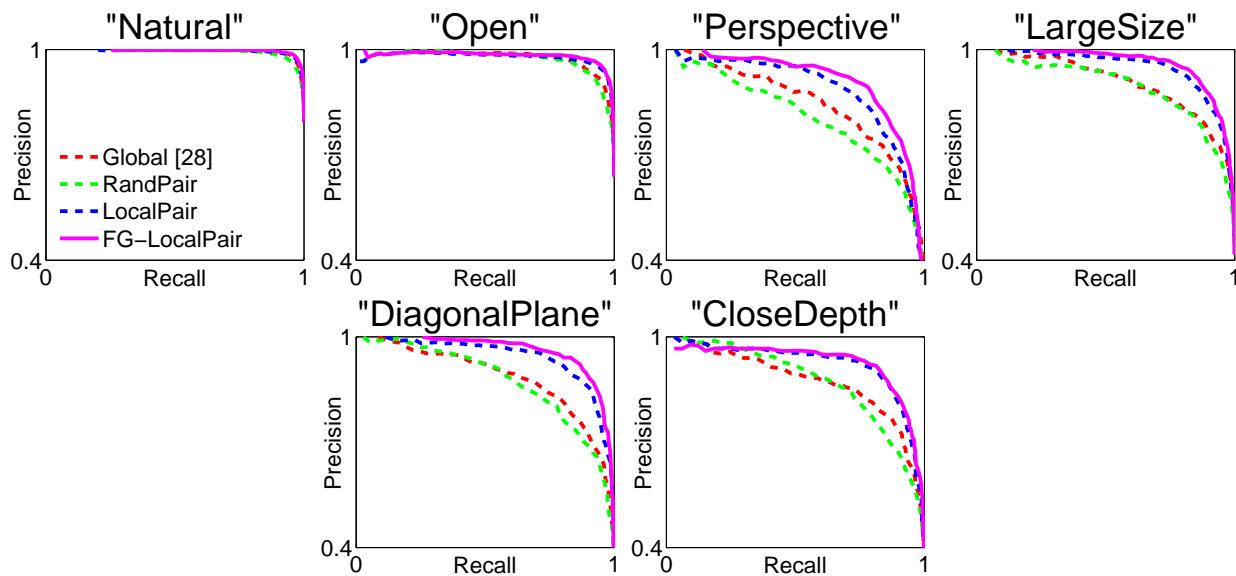


Figure 4: Precision-recall curves for OSR.

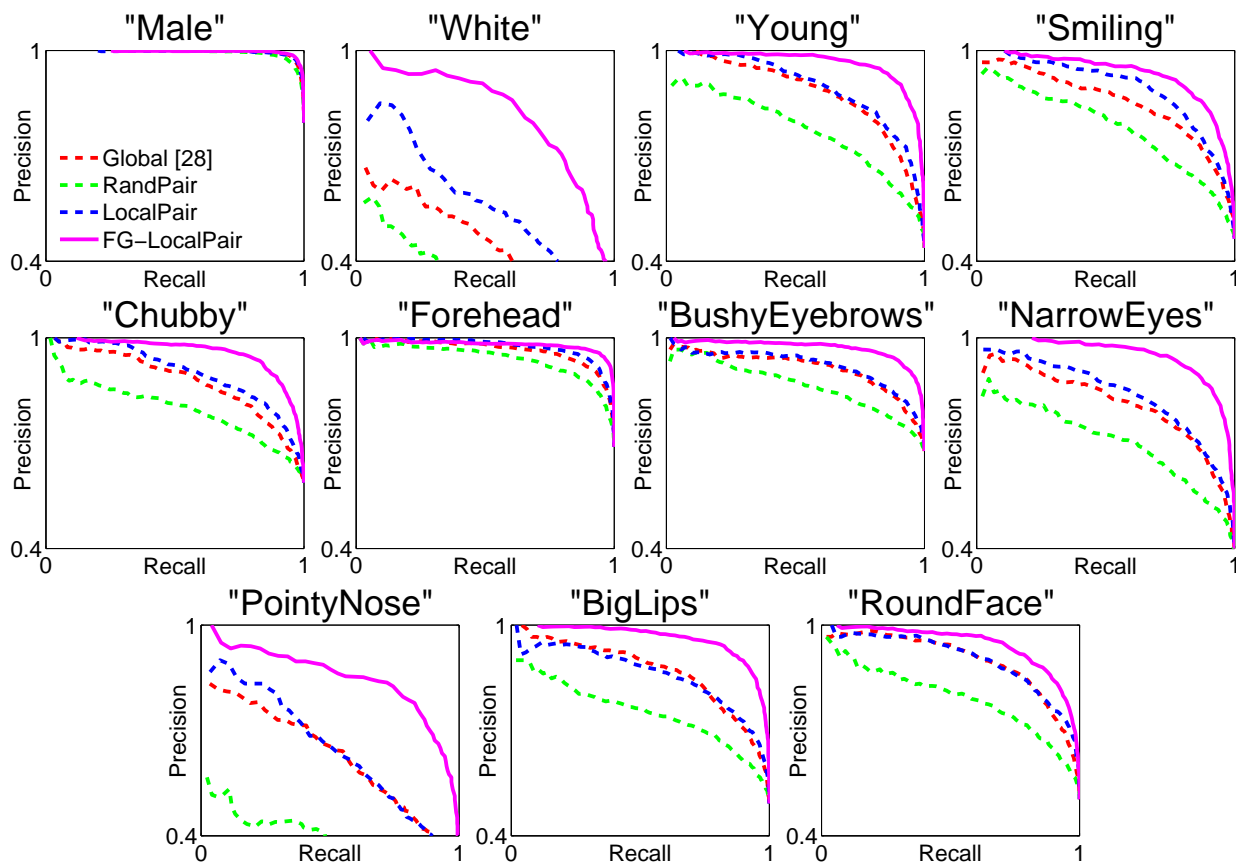


Figure 5: Precision-recall curves for PubFig.

## 4. Neighborhood Size

As mentioned in the paper, the optimal setting for  $K$  depends on the specific attribute, query, and even training pairs. To understand exactly how  $K$  affects our results, we performed experiments on our proposed method (FG-LocalPair) over a spectrum of  $K$  values using the exact same setup procedures as in the paper. The results on all 3 datasets are shown below.  $K = All$  represents the Global baseline. As we can see, the optimal  $K$  value falls somewhere in-between  $K = 50$  and 200 for most of the attributes. We note that even the accuracies for  $K = 1$  can be better than Global’s due to presence of almost identical pairs in the training set (more prevalent in OSR and PubFig).

$K$	Open	Pointy	Sporty	Comfort
1	83.43	84.33	87.87	87.23
10	90.10	89.67	93.43	92.57
20	89.87	90.13	<b>94.07</b>	<b>92.70</b>
50	89.83	<b>91.33</b>	93.00	92.60
80	90.20	91.03	92.57	92.33
100	<b>90.67</b>	90.83	92.67	92.37
200	90.50	90.50	92.90	91.37
300	90.37	89.70	92.37	91.83
400	90.10	89.93	92.17	91.43
600	89.67	89.80	92.27	91.30
800	89.13	90.13	92.00	90.63
1000	88.70	89.40	91.57	90.70
<i>All</i>	87.77	89.37	91.20	89.93

Table 1:  $K$ -sweep accuracies for **Zap50K**.

$K$	Natrl	Open	Persp.	LgSize	Diag	ClsDepth
1	94.20	92.23	88.47	88.90	90.53	85.97
10	95.37	93.53	90.13	90.63	91.83	87.33
20	95.33	93.50	90.33	90.77	92.33	88.27
50	95.63	93.87	90.10	91.13	92.57	89.37
80	<b>95.77</b>	<b>94.17</b>	90.37	<b>91.30</b>	<b>92.70</b>	90.03
100	95.70	94.10	<b>90.43</b>	91.10	92.43	90.47
200	<b>95.77</b>	<b>94.17</b>	90.13	90.63	91.87	90.57
300	95.73	94.00	89.97	90.57	91.67	<b>90.67</b>
400	95.70	93.77	89.60	90.63	91.37	90.40
600	95.30	94.00	89.60	90.47	91.17	90.50
800	95.33	93.87	89.60	90.03	90.87	90.07
1000	95.37	93.83	89.37	89.87	90.80	89.67
<i>All</i>	95.03	90.77	86.73	86.23	86.50	87.53

Table 2:  $K$ -sweep accuracies for **OSR**.

$K$	Male	White	Young	Smiling	Chubby	Forehead	Eyebrow	Eye	Nose	Lip	Face
1	88.03	76.40	82.33	83.70	81.43	88.47	85.93	86.63	82.33	85.60	77.23
10	92.40	82.27	88.27	85.80	84.87	91.90	90.23	90.97	87.53	89.67	83.00
20	<b>92.80</b>	84.53	89.83	86.20	86.00	92.97	<b>91.10</b>	92.30	88.40	90.47	84.13
50	92.17	86.67	91.27	86.80	87.10	93.40	90.57	<b>92.47</b>	<b>89.80</b>	<b>90.70</b>	85.77
80	91.77	87.17	91.73	87.23	87.10	93.50	89.90	91.43	89.20	90.13	86.60
100	91.77	87.43	<b>91.87</b>	87.00	<b>87.37</b>	94.00	89.83	91.40	89.07	90.43	86.70
200	90.77	<b>88.30</b>	91.27	<b>87.50</b>	86.13	94.27	89.37	90.20	88.83	89.37	88.80
300	89.83	88.10	90.87	87.20	86.40	<b>94.30</b>	89.47	89.87	88.53	88.60	<b>88.80</b>
400	88.87	88.20	90.53	86.73	85.83	<b>94.33</b>	88.10	89.03	88.57	88.17	<b>89.60</b>
600	88.33	87.87	89.83	86.40	85.00	93.97	87.60	87.47	88.13	87.47	89.47
800	87.37	87.43	89.07	86.43	83.97	93.80	86.80	86.60	87.63	86.40	88.87
1000	86.93	87.13	88.53	86.17	83.27	93.67	86.13	86.07	86.87	85.67	88.50
<i>All</i>	81.80	76.97	83.20	79.90	76.27	87.60	79.87	81.67	77.40	79.17	82.33

Table 3:  $K$ -sweep accuracies for **PubFig**.

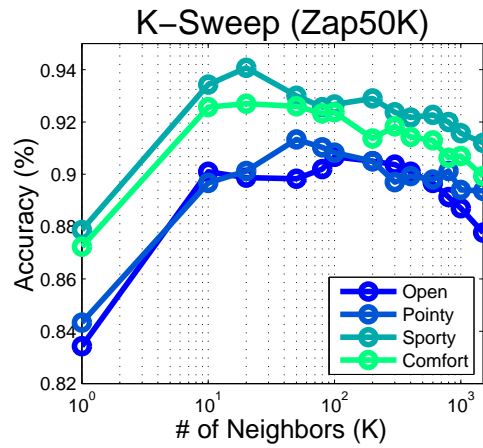


Figure 6: *K*-sweep accuracy curves for **Zap50K**.

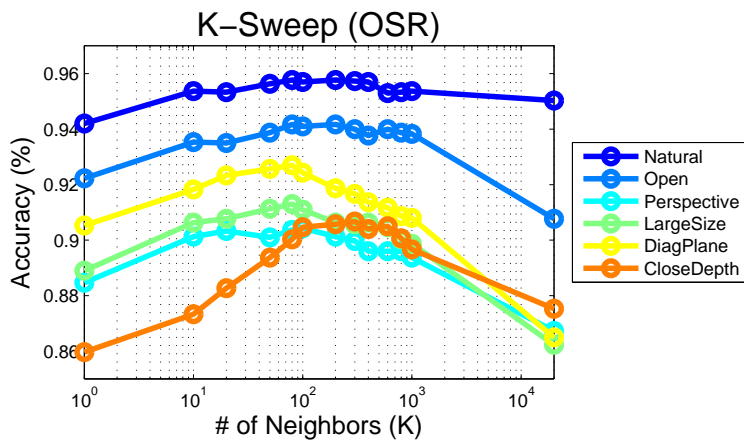


Figure 7: *K*-sweep accuracy curves for **OSR**.

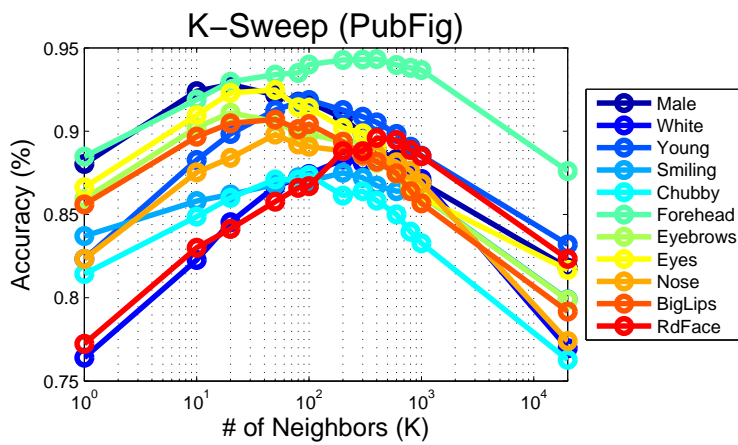


Figure 8: *K*-sweep accuracy curves for **PubFig**.