In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

Thinking Outside the Pool: Active Training Image Creation for Relative Attributes (Supplementary File)

Aron Yu
University of Texas at Austin
aron.yu@utexas.edu

Kristen Grauman
UT Austin and Facebook AI Research
grauman@cs.utexas.edu

1. Active Training Individual Plots

In Figure 6 of the main text, we present the gain curves of our active training experiment over the Real baseline, averaged over all attributes. Here, we show the individual gain curves for each attribute from both datasets. Figure 1 and 2 represent the shoes and face attributes, respectively. Our approach learns the fastest for almost every attribute.

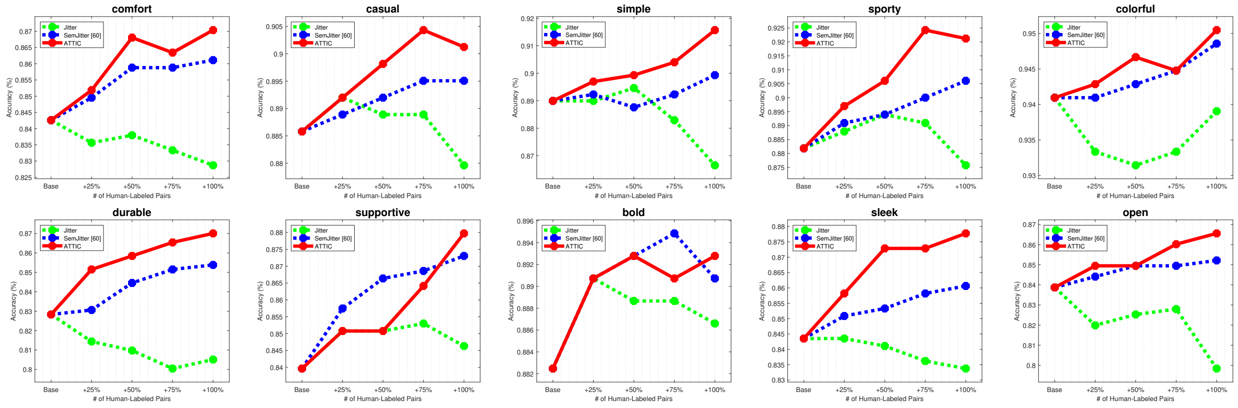
comfort, casual, simple, sporty, colorful, durable, supportive, bold, sleek, open

Figure 1: Individual active learning curves for the shoes attributes.

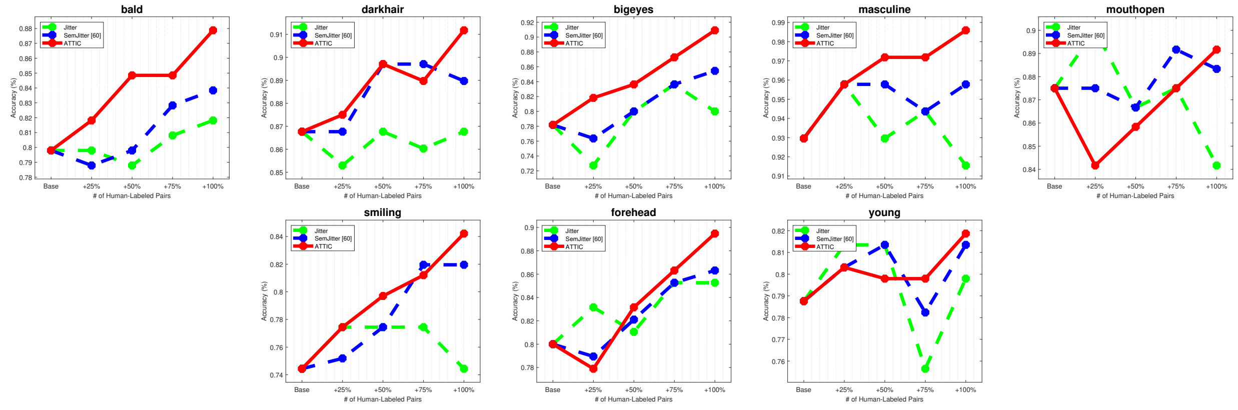
bald, darkhair, bigeyes, masculine, mouthopen, smiling, forehead, young

Figure 2: Individual active learning curves for the face attributes.

2. Shoes Results on Dataset Train/Test Split

The results in the main paper compare our method to the relevant baselines on UT-Zappos50K. Here in Table 1 we show the results reported in [4] alongside our results for the same Zappos train/test split, for completeness. Note that as in [4], for an apples-to-apples comparison, all methods are applied to the same 64 × 64 images. These results use the method exactly as described and applied in the main paper for the "Auto" scenario.

Our method outperforms all the existing methods for the majority of the attributes. Semantic Jitter [4] outperforms ours for *sporty* in the first test set and *open* in the second test set, indicating that those attributes were similarly well-served by that method's heuristic choice for generated images. However, our automated method overall has the advantage.

		Open	**Sporty**	**Comfort**
Zap50K-1	RelAttr [1]	88.33	89.33	91.33
	FG-LP [3]	90.67	91.33	93.67
	DeepSTN [2]	93.00	93.67	94.33
	SemJitter [4]	95.00	**96.33**	95.00
	ATTIC (ours)	**95.67**	96.00	**95.67**
Zap50K-2	RelAttr [1]	60.36	65.65	62.82
	FG-LP [3]	69.36	66.39	63.84
	DeepSTN [2]	70.73	67.49	66.09
	SemJitter [4]	**72.18**	68.70	67.72
	ATTIC (ours)	71.68	**69.62**	**68.64**

Table 1: Extension to the result table from [4] that includes our results for the same Zappos splits. All methods are trained and tested on $\mathbf{64 \times 64}$ images for an apples-to-apples comparison.

3. Real+ Baseline

Here we show Real+, which is the same as the Real baseline reported in the main paper, but enhanced to use the additional labeled images used to train the image generator. The rankers are trained with ordered pairs, and all methods tested in the experiments reported in the main paper use the same number n of ordered pairs. However, the image generator does additionally access attribute-labeled training images (not ordered pairs) to learn how to perform attribute-conditioned generation. For Real+ we attempt to convert those binary attribute labels used by the generator into ordered pairs to bolster training of the ranker for Real. To that end, we use the real-valued attribute strength from the output of an attribute classifier as the pseudo-ground truth labels to form ordered pairs. For training, we sample an equal number of pseudo pairs using these estimated attribute strength in addition to the existing real pairs.

Tables 2 and 3 show the results, alongside the Real baseline results copied from the main paper. We see that overall the image generator training images have a net neutral effect on the baseline's results. This is an indication that both Real and Real+ suffers from the same sparsity issue, as the images are taken from similar pool of real images. The addition of similarly distributed (real) images lacks the fine-grained details needed to train a stronger model.

	Comfort	**Casual**	**Simple**	**Sporty**	**Colorful**	**Durable**	**Supportive**	**Bold**	**Sleek**	**Open**	**Mean**
Real [2]	84.26	88.58	88.99	88.18	94.10	82.83	83.96	88.25	84.35	83.87	86.74
Real+	81.71	87.96	87.12	87.58	91.05	82.60	84.41	87.63	85.82	83.87	85.98

Table 2: Real+ results for the shoes dataset.

	Bald	**DarkHair**	**BigEyes**	**Masculine**	**MouthOpen**	**Smiling**	**Forehead**	**Young**	**Mean**
Real [2]	79.80	86.77	78.18	92.96	87.50	74.44	80.00	78.76	82.30
Real+	81.82	86.03	80.00	92.96	86.67	75.94	81.21	79.28	82.99

Table 3: Real+ results for the face dataset.

4. Collecting Active Labels from Human Annotators on MTurk

In Figure 3, we show the interface we used to collect labels from human annotators on MTurk for the actively synthesized training images. In addition to the relative decision, we also instruct the workers to reflect their level of confidence with their decision. Image pairs with low overall confidence and/or low agreement among workers are pruned and not used in training.

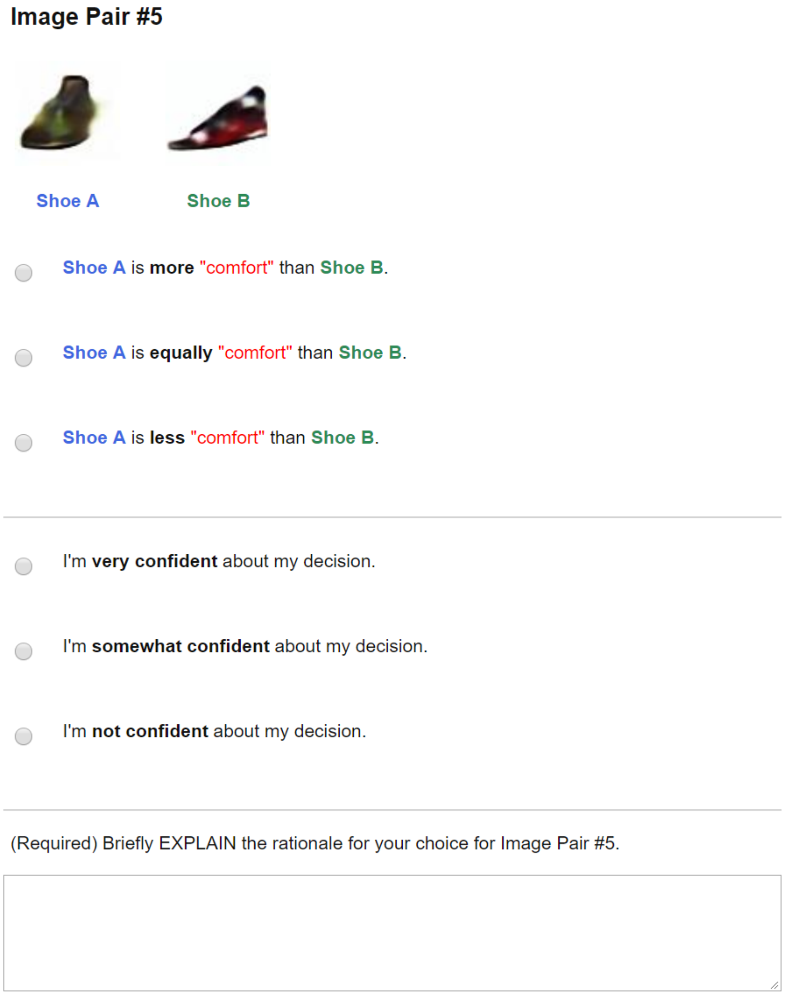

Image Pair #5

Shoe A   Shoe B

○ Shoe A is **more** "comfort" than Shoe B.

○ Shoe A is **equally** "comfort" than Shoe B.

○ Shoe A is **less** "comfort" than Shoe B.

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

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

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

(Required) Briefly EXPLAIN the rationale for your choice for Image Pair #5.

Figure 3: Example of a single task within a HIT.

References

[1] D. Parikh and K. Grauman. Relative Attributes. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2011.

[2] K. Singh and Y. J. Lee. End-to-end localization and ranking for relative attributes. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2016.

[3] A. Yu and K. Grauman. Fine-Grained Visual Comparisons with Local Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[4] A. Yu and K. Grauman. Semantic Jitter: Dense Supervision for Visual Comparisons via Synthetic Images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.