Semantic Jitter: Dense Supervision for Visual Comparisons via Synthetic Images (Supplementary File)

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

Kristen Grauman
University of Texas at Austin
grauman@cs.utexas.edu

1. Fine-Grained Attribute Lexicon

We use the UT-Zap50K shoe dataset [5] to perform our lexical study. It contains 50,025 catalog shoe images along with a set of meta-data that are associated with each image. Our goal is to study how humans distinguish fine-grained differences in similar images. Specifically, we want to know what words humans use to describe fine-grained differences.

1.1. Experimental Design

We design our experiments in the form of "complete the sentence" questions and test them on the Amazon MTurk workers. We experiment with two kinds of designs: Design 1 compares two individual images while Design 2 compares one image against a group of six images. Given the meta-data which contains a category (i.e. slippers, boots) and subcategory (i.e. flats, ankle high) labels for each image, we combine these labels into a set of 21 unique category-subcategory pseudo-classes (i.e. slippers-flats, shoes-loaders). Using theses new pseudo-classes, we sample 4,000 supervision pairs (for each design) where 80% are comparing within the same pseudo-class and 20% are comparing within the same category. By focusing sampled pairs among items within a pseudo-class, we aim for a majority of the pairs to contain visually quite related items, thus forcing the human subjects to zero in on fine-grained differences.

For each question, the workers are asked to complete the sentence, "Shoe A is a *little more/less* ⟨insert word⟩ than Shoe B" using a single word ("Shoe B" is replaced by "Group B" for Design 2). They are instructed to identify **subtle differences** between the images and provide a short rationale to elaborate on their choices. Figure 2 shows a screenshot of a sample question.

1.2. Post-Processing

We post-process the fine-grained word suggestions through correcting for human variations (i.e. misspelling, word forms), merging of visual synonyms/antonyms, and evaluation of the rationales. For example, "casual" and "formal" are visual antonyms and workers used similar keywords in their rationales for "durable" and "rugged". In both cases, the frequency counts for the two words are combined. Over 1,000 MTurk workers participated in our study, yielding a total of 350+ distinct word suggestions[1]. In the end, we select the 10 most frequently appearing words as our fine-grained relative attribute lexicon for shoes: *comfort*, *casual*, *simple*, *sporty*, *colorful*, *durable*, *supportive*, *bold*, *sleek*, and *open*.

2. Generative Model Training

We train our attribute-conditioned image generator using a Conditional Variational Auto-Encoder (CVAE) [4]. The model requires a vector of real-valued attribute strengths for each training image. We detail the setup process for each dataset below.

2.1. Fashion Images of Shoes

We use a subset of 38,866 images from UT-Zap50K to train the generative shoe model. Using the meta-data once again, we select 40 attributes ranging from material types to toe styles (e.g. Material.Mesh, ToeStyle.Pointed, etc.) and assign binary pre-labels to them. In addition, we also use the 10 fine-grained relative attributes collected from our lexical study. We sample 500 supervision pairs for each attribute from the newly collected pairwise labels and train linear SVM rankers using RankSVM [2]. We then project all 1,000 images (used to train the ranker) onto the learned ranker to obtain their real-valued ranking scores, which we use as their pre-labels. While our focus is on the 10 relative attributes, the inclusion of additional attributes aids in overall learning of the generative model. However, we do not use any of those meta-data attributes for fine-grained relative attribute training as they are mostly binary in nature.

[1] We used only the words from Design 1 as the two designs produced very similar word suggestions.

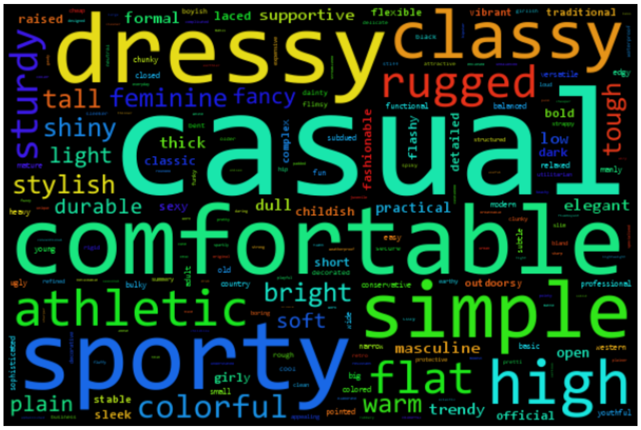


Figure 1: Word cloud depicting our crowd-mined data for a fine-grained relative attribute lexicon for shoes (before post-processing).

Finally, using these pre-labels from all 50 attributes, we train a linear classifier for each attribute. We apply the classifier on all 38,866 images and use their decision values as the real-valued attribute strength needed to train the generative model. All of this is a workaround, similar to the one used in [4], in order to supply the generative model with real-valued attribute strengths on its training data. If labeled binary attribute data were available for training the linear classifiers from the onset, that would be equally good if not better.

2.2. Human Faces

We use a subset of 11,154 images from LFW [1] to train the generative face model. Following [4], we use the 73 dimensional attribute strength provided in [3] to train the generative face model.

3. Nearest Neighbors

In Figure 3, we provide additional qualitative examples of the neighboring pairs given actual test pairs, expanding upon Figure 4 in the main paper. Notice that for the face images, the synthetic image pairs exhibit fine-grained differences while preserving the underlying identity, something that is valuable for learning but hard to obtain using real image pairs.

4. Attribute Lexicon

Figure 1 shows a word cloud of the raw results collected from the Turkers. The frequency of word usage corresponds directly with the size of the words.

References

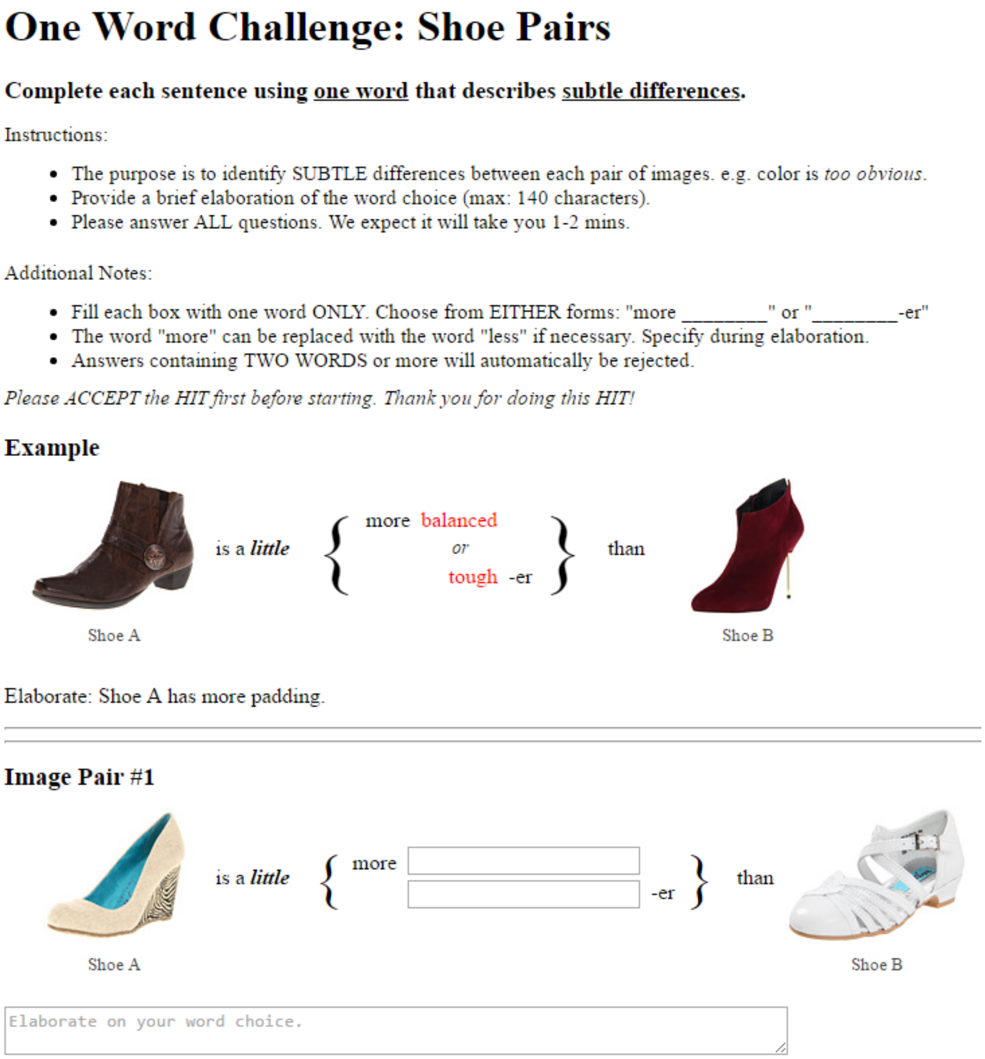
[1] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[2] T. Joachims. Optimizing search engines using clickthrough data. In *SIGKDD*, 2002.

[3] N. Kumar, A. Berg, P. Belhumeur, and S. Nayar. Attribute and Simile Classifiers for Face Verification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2009.

[4] X. Yan, J. Yang, K. Sohn, and H. Lee. Attribute2Image: Conditional image generation from visual attributes. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2016.

[5] 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.



One Word Challenge: Shoe Pairs

Complete each sentence using one word that describes subtle differences.

Instructions:

- The purpose is to identify SUBTLE differences between each pair of images. e.g. color is *too obvious*.
- Provide a brief elaboration of the word choice (max: 140 characters).
- Please answer ALL questions. We expect it will take you 1-2 mins.

Additional Notes:

- Fill each box with one word ONLY. Choose from EITHER forms: "more ________" or "________-er"
- The word "more" can be replaced with the word "less" if necessary. Specify during elaboration.
- Answers containing TWO WORDS or more will automatically be rejected.

Please ACCEPT the HIT first before starting. Thank you for doing this HIT!

Example

Shoe A is a ***little*** { more balanced *or* tough -er } than Shoe B

Elaborate: Shoe A has more padding.

Image Pair #1

Shoe A is a ***little*** { more ________ ________ -er } than Shoe B

Elaborate on your word choice.

Figure 2: Screenshot of our lexical experiment on MTurk in Design 1.

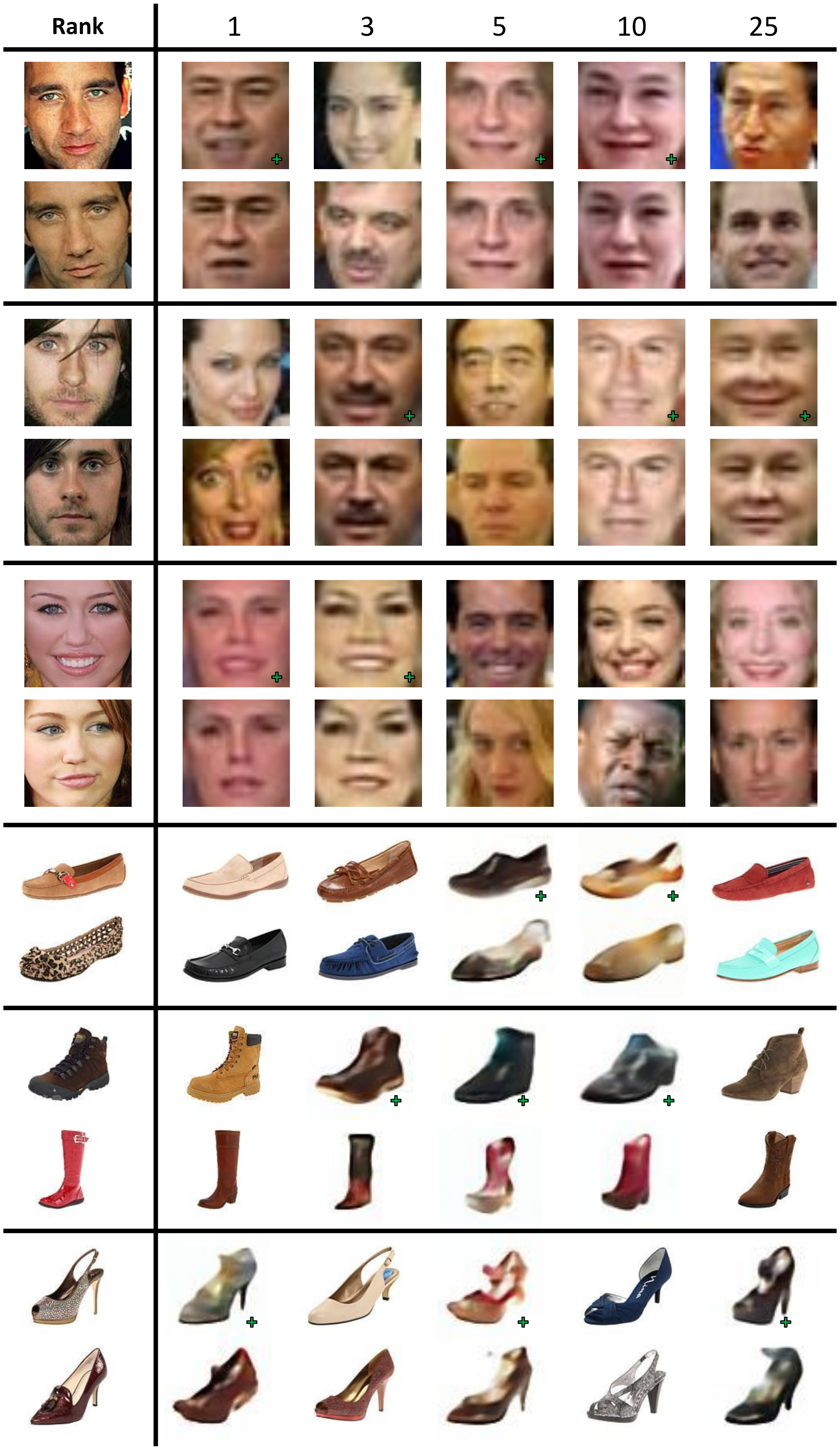


Figure 3: Examples of nearest neighbor image pairs given novel test pairs (left). A green plus sign denotes a synthetic image pair.