

# Just Noticeable Differences in Visual Attributes

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

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

## Abstract

We explore the problem of predicting “just noticeable differences” in a visual attribute. While some pairs of images have a clear ordering for an attribute (e.g., A is more sporty than B), for others the difference may be indistinguishable to human observers. However, existing relative attribute models are unequipped to infer partial orders on novel data. Attempting to map relative attribute ranks to equality predictions is non-trivial, particularly since the span of indistinguishable pairs in attribute space may vary in different parts of the feature space. We develop a Bayesian local learning strategy to infer when images are indistinguishable for a given attribute. On the UT-Zap50K shoes and LFW-10 faces datasets, we outperform a variety of alternative methods. In addition, we show the practical impact on fine-grained visual search.

## 1. Introduction

Imagine you are given a pile of images of Barack Obama, and you must sort them according to where he looks most to least serious. Can you do it? Surely there will be some obvious ones where he is more serious or less serious. There will even be image pairs where the distinction is quite subtle, yet still perceptible. However, you are likely to conclude that forcing a *total* order is meaningless: while the images exhibit different degrees of the attribute seriousness, at some point the differences become indistinguishable. It’s not that the pixel patterns in indistinguishable image pairs are literally the same—they just can’t be characterized consistently as anything other than “equally serious”.

Attributes are visual properties describable in words, capturing anything from material properties (*metallic*, *furry*), shapes (*flat*, *boxy*), expressions (*smiling*, *surprised*), to functions (*sittable*, *drinkable*). Since their introduction to the recognition community [7, 15, 16], attributes have inspired a number of useful applications in image search [13, 14, 15, 26], biometrics [4, 21], and language-based supervision for recognition [2, 16, 19, 25].

Existing attribute models come in one of two forms: categorical or relative. Whereas categorical attributes are



Figure 1: At what point is the strength of an attribute indistinguishable between two images? While existing relative attribute methods are restricted to inferring a total order, in reality there are images that look different but where the attribute is nonetheless perceived as “equally strong”. For example, in the fourth and fifth images of Obama, is the difference in *seriousness* noticeable enough to warrant a relative comparison?

suitable only for clear-cut predicates, such as *male* or *wooden*, relative attributes can represent “real-valued” properties that inherently exhibit a spectrum of strengths, such as *serious* or *sporty*. Typically one learns a relative attribute in the learning-to-rank setting; training data is ordered (e.g., we are told image A has it less than B), and a ranking function is optimized to preserve those orderings. Given a new image, the function returns a score conveying how strongly the attribute is present [1, 3, 5, 6, 14, 17, 18, 19, 22, 23, 27].

The problem is that existing models for relative attributes assume that all images are orderable. In particular, they assume that *at test time*, the system can and should always distinguish which image in a pair exhibits the attribute more. Yet, as our Obama example above illustrates, this assumption is incompatible with how humans actually perceive attributes. In fact, recent work reports that in a fine-grained domain like fashion, 40% of the time human judges asked to compare images for a relative attribute declare that no difference is perceptible [27]. Within a given attribute, sometimes we can perceive a comparison, sometimes we can’t. See Figure 1.

We argue that this situation calls for a model of *just noticeable difference* among attributes. Just noticeable difference (JND) is a concept from psychophysics. It refers to the amount a stimulus has to be changed in order for it to be detectable by human observers at least half the time. For example, JND is of interest in color perception (which light sources are perceived as the same color?) and image quality

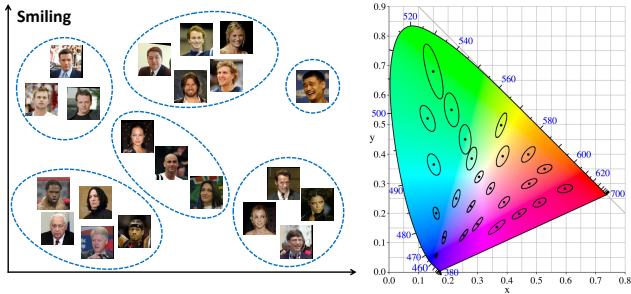


Figure 2: Analogous to the MacAdam ellipses in the CIE  $x,y$  color space (right) [8], relative attribute space is likely not uniform (left). That is, the regions within which attribute differences are indistinguishable may vary in size and orientation across the high-dimensional visual feature space. Here we see the faces within each “equally *smiling*” cluster exhibit varying qualities for differentiating smiles—such as age, gender, and visibility of the teeth—but are still difficult or impossible to order in terms of *smilingness*. As a result, simple metrics and thresholds on attribute differences are insufficient to detect just noticeable differences.

assessment (up to what level of compression do the images look ok?). JNDs are determined empirically through tests of human perception. For example, JND in color can be determined by gradually altering the light source just until the human subject detects that the color has changed [8].

Why is it challenging to develop a computational model of JND for relative attributes? At a glance, one might think it amounts to learning an optimal threshold on the difference of predicted attribute strengths. However, this begs the question of how one might properly and densely sample real images of a complex attribute (like seriousness) to gradually walk along the spectrum, so as to discover the right threshold with human input. More importantly, an attribute space need not be *uniform*. That is, depending on where we look in the feature space, the magnitude of attribute difference required to register a perceptible change may vary. Therefore, the simplistic “global threshold” idea falls short. Analogous issues also arise in color spaces, e.g., the famous MacAdam ellipses spanning indistinguishable colors in the CIE  $x,y$  color space vary markedly in their size and orientation depending on where in the feature space one looks (leading to the crafting of color spaces like CIE Lab that are more uniform). See Figure 2.

We propose a solution to infer when two images are indistinguishable for a given attribute. Following the non-uniformity intuition above—which says the decision function will likely vary depending on where in the feature space one looks—we develop a Bayesian approach that relies on *local* statistics of orderability. Our approach leverages both a low-level visual descriptor space, within which image pair proximity is learned, as well as a mid-level visual attribute space, within which attribute distinguishability is represented. To our knowledge, our framework offers the first attempt to unify a notion of “equality” (i.e., unnoticeable differences) into relative attributes during in-

ference. Whereas past ranking models have attempted to integrate equality into *training*, none attempt to distinguish between orderable and un-orderable pairs at test time.

We apply our method on two challenging datasets with fine-grained relative attributes, the UT Zappos 50K collection of catalog images of shoes and the Labeled Faces in the Wild (LFW) collection of human faces. The results show our approach’s superior performance compared to various baselines for detecting noticeable differences. Furthermore, we demonstrate how attribute JND has potential benefits for an image search application.

## 2. Related Work

**Comparing images by their attributes** Relative attributes are most commonly represented with learned ranking functions [1, 2, 3, 5, 6, 14, 17, 18, 19, 22, 23, 27]. Pairwise supervision is used for training: a set of pairs ordered according to the attribute is obtained from human annotators, and a ranking function that preserves those orderings is learned. Given a novel pair of images, the ranker indicates which image has the attribute more. In a similar spirit, regression [4] and paired-difference classification [9] have also been employed. While some implementations (including [19]) augment the training pool with “equal” pairs to facilitate learning, notably no existing work attempts to discern distinguishable from indistinguishable pairs at test time—our main goal. In Sec. 3 we discuss technical reasons why other common learning paradigms (e.g., ordinal regression) are not an easy solution to the problem.

**Fine-grained and unrankable attributes** Of all prior work in relative image ranking, those that come closest to our goal are our fine-grained relative attribute work [27] and the facial attractiveness ranking method of [3]. The former uses local learning to tackle attribute comparisons that are visually subtle, e.g., deciding which of two athletic shoes is more *sporty*. Like the methods cited above, this method also assumes all images are distinguishable at test time. In contrast, our method specifically deals with the boundary where “subtle” and “indistinguishable” meet.

In [3], the authors train a hierarchy of SVM classifiers to recursively push a image into buckets of more/less attractive faces. The leaf nodes contain images “unrankable” by the human subject, which can be seen as indistinguishability for the specific attribute of human attractiveness. Nonetheless, the proposed method is not applicable to our problem. It learns a ranking model specific to a single human subject, whereas we learn a subject-independent model. Furthermore, the training procedure [3] has limited scalability, since the subject must rank *all* training images into a partial order; the results focus on training sets of 24 images for this reason. In our domains of interest, where thousands or more training instances are standard, getting a reliable global partial order on all images remains an open challenge.

**Variability in how attributes are perceived** Differences in human perception are another source of ambiguity in attribute prediction, especially for subjective properties. Recent work deals with this by learning personalized models [1, 3, 12]. In contrast, we are interested in modeling attributes where there is consensus about comparisons, only they are subtle. Rather than personalize a model towards an observer, we want to discover the (implicit) map of where the consensus for JND boundaries in attributes exists. The attribute calibration method of [24] post-processes attribute classifier outputs so they can be fused for multi-attribute search. Our method is also conscious that differences in attribute outputs taken at “face value” can be misleading, but our goal and approach are entirely different.

**Choosing between relative and binary attributes** The “spoken attributes” [22] method learns to generate a human-like description for an image by intelligently selecting whether to use binary or relative attributes. The insight is that even when a person *can* distinguish an attribute, he may choose not to say so, depending on the context. For example, if one face is clearly smiling more than the other, but neither is smiling much, it is unusual for a human describing the image to say “the person on the left is smiling more than the one on the right.” The work is not concerned with detecting JND. It assumes a relative comparison is always possible, just not always worth mentioning.

### 3. Approach

Given a pair of images and specified attribute, our goal is to decide whether or not the attribute’s strength is distinguishable between the two. We develop a Bayesian prediction approach based on local learning. Our approach first constructs a predicted relative attribute space using sparse human-provided supervision about image comparisons (Sec. 3.1). Then, on top of that model, we combine a likelihood computed in the predicted attribute space (Sec. 3.2.1) with a local prior computed in the original image feature space (Sec. 3.2.2). See Figure 3.

#### 3.1. Relative Attribute Ranks

In all notation that follows, it is assumed that a single attribute is learned at a time (e.g., *seriousness*). For each attribute to be learned, we take as input two sets of annotated training image pairs. The first set consists of ordered pairs,  $\mathcal{P}_o = \{(i, j)\}$ , for which humans perceive image  $i$  to have the attribute more than image  $j$ . That is, each pair in  $\mathcal{P}_o$  has a “noticeable difference”. The second set consists of unordered, or “equal” pairs,  $\mathcal{P}_e = \{(p, q)\}$ , for which humans cannot perceive a difference in attribute strength.

We enforce stringent requirements to ensure the precision of these pair annotations, such that the training data reflects the common perception across multiple human ob-

servers (see Sec. 4 for details). This is critical, since a JND model demands that we correctly preserve the distinction between a “just barely orderable” pair and an equal pair.

Let  $\mathbf{x}_i \in X \subset \mathbb{R}^d$  be a  $d$ -dimensional image descriptor for image  $i$ . First we learn a ranking function  $R : X \rightarrow \mathbb{R}$  that maps an input image to (an initial estimate of) its attribute strength. Following [19], we use a large-margin approach based on the SVM-Rank framework [11]. The method optimizes the rank function parameters to preserve the orderings in  $\mathcal{P}_o$ , maintaining a margin between them in the 1D output space, while also minimizing the separation between the unordered pairs in  $\mathcal{P}_e$ . For the linear case, the parameters are simply a weight vector  $\mathbf{w}$ :

$$R(\mathbf{x}) = \mathbf{w}^T \mathbf{x}, \quad (1)$$

though non-linear ranking functions are also possible. The learning objective is as follows:

$$\begin{aligned} \text{minimize} \quad & \left( \frac{1}{2} \|\mathbf{w}\|_2^2 + C \left( \sum \xi_{ij}^2 + \sum \gamma_{p,q}^2 \right) \right) \quad (2) \\ \text{s.t.} \quad & \mathbf{w}^T (\mathbf{x}_i - \mathbf{x}_j) \geq 1 - \xi_{ij}; \forall (i, j) \in \mathcal{P}_o \\ & |\mathbf{w}^T (\mathbf{x}_p - \mathbf{x}_q)| \leq \gamma_{pq}; \forall (p, q) \in \mathcal{P}_e \\ & \xi_{ij} \geq 0; \gamma_{pq} \geq 0, \end{aligned}$$

where the constant  $C$  balances the margin regularizer and pair constraints. Step 1 in Figure 3 depicts a linear ranking function learned from the training pairs.

Given a novel image pair  $(\mathbf{x}_m, \mathbf{x}_n)$ , one can apply the rank function to predict their order. If  $R(\mathbf{x}_m) > R(\mathbf{x}_n)$ , then image  $m$  exhibits the attribute more than image  $n$ , and vice versa. As discussed above, despite the occasional use of unordered pairs for training<sup>1</sup>, it is assumed in prior work that all test images will be orderable. However, the real-valued output of the ranking function will virtually never be equal for two distinct inputs. Therefore, even though existing methods may learn to produce similar rank scores for equal pairs, it is non-trivial to determine when a novel pair is “close enough” to be considered un-orderable.

#### 3.2. A Local Bayesian Model of Distinguishability

The most straightforward approach to infer whether a novel image pair is distinguishable would be to impose a threshold on their rank differences, i.e., to predict “indistinguishable” if  $|R(\mathbf{x}_m) - R(\mathbf{x}_n)| \leq \epsilon$ . The problem is that unless the rank space is uniform, a global threshold  $\epsilon$  is inadequate. In other words, the rank margin for indistinguishable pairs need not be constant across the entire feature space. By testing multiple variants of this basic idea, our empirical results confirm this is indeed an issue, as we will see in Sec. 4.

<sup>1</sup>Empirically, we found the inclusion of unordered pairs during training in [19] to have negligible impact at test time.

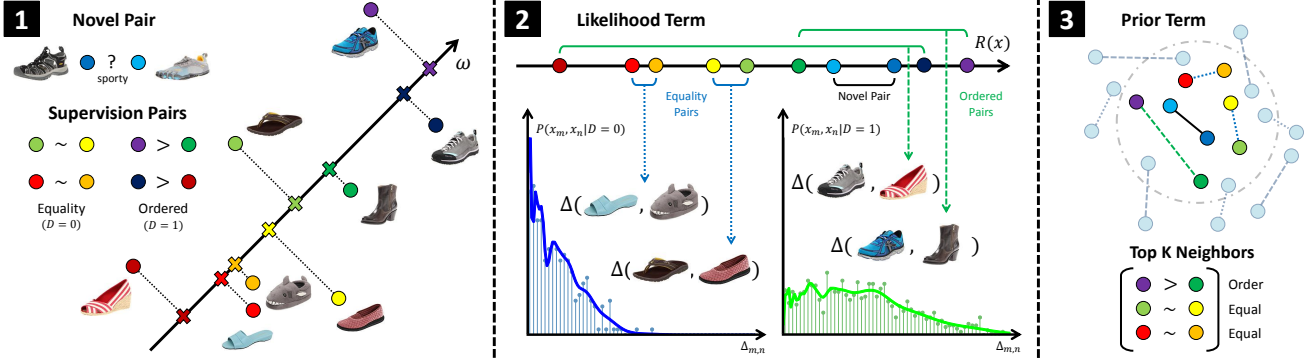


Figure 3: Overview of our approach. (1) Learn a ranking function  $R$  using all annotated training pairs. (2) Estimate the likelihood densities of the equal and ordered pairs, respectively, using the pairwise distances in relative attribute space. (3) Determine the local prior by counting the labels of the analogous pairs in the image descriptor space. (4) Combine the results to predict whether the novel pair is distinguishable (not depicted). Best viewed in color.

Our key insight is to formulate distinguishability prediction in a probabilistic, local learning manner. Mindful of the non-uniformity of relative attribute space, our approach uses distributions tailored to the data in the proximity of a novel test pair. Furthermore, we treat the relative attribute ranks as an imperfect mid-level representation on top of which we can learn to target the actual (sparse) human judgments about distinguishability.

Let  $D \in \{0, 1\}$  be a binary random variable representing the distinguishability of an image pair. For a distinguishable pair,  $D = 1$ . Given a novel test pair  $(\mathbf{x}_m, \mathbf{x}_n)$ , we are interested in the posterior:

$$P(D|\mathbf{x}_m, \mathbf{x}_n) \propto P(\mathbf{x}_m, \mathbf{x}_n|D)P(D), \quad (3)$$

to estimate of how likely two images are distinguishable. To make a hard decision we take the maximum a posteriori estimate over the two classes, i.e.,  $d^* = \operatorname{argmax}_d P(D = d|\mathbf{x}_m, \mathbf{x}_n)$ .

At test time, our method can further be used in a two-stage cascade. If the test pair appears distinguishable, we return the response “more” or “less” according to whether  $R(\mathbf{x}_m) < R(\mathbf{x}_n)$ . Otherwise, we say the test pair is indistinguishable. In this way we unify relative attributes with JND, generating partially ordered predictions in spite of the ranker’s inherent totally ordered outputs.

Next, we derive models for the likelihood and prior in Eq. 3, accounting for the challenges described above.

### 3.2.1 Likelihood model

We use a kernel density estimator (KDE) to represent the distinguishability likelihood over image pairs. The likelihood captures the link between the observed rank differences and the human-judged just noticeable differences.

Let  $\Delta_{m,n}$  denote the difference in attribute ranks for images  $m$  and  $n$ :

$$\Delta_{m,n} = |R(\mathbf{x}_m) - R(\mathbf{x}_n)|. \quad (4)$$

We compute the rank differences for all training pairs in  $\mathcal{P}_o$  and  $\mathcal{P}_e$ , and fit a non-parametric Parzen density:

$$P(\mathbf{x}_m, \mathbf{x}_n|D) = \frac{1}{|\mathcal{P}|} \sum_{i,j \in \mathcal{P}} K_h(\Delta_{i,j} - \Delta_{m,n}), \quad (5)$$

for each set in turn. Here  $\mathcal{P}$  refers to the ordered pairs  $\mathcal{P}_o$  when representing distinguishability ( $D = 1$ ), and the equal pairs  $\mathcal{P}_e$  when representing indistinguishability ( $D = 0$ ). The Parzen density estimator [20] superimposes a kernel function  $K_h$  at each data pair. It integrates local estimates of the distribution and resists overfitting. The KDE has a smoothing parameter  $h$  that controls the model complexity. To ensure that all density is contained within the positive absolute margins, we apply a positive support to the estimator. Namely, we transform  $\Delta_{i,j}$  using a log function, estimate the density of the transformed values, and then transform back to the original scale. See step 2 in Figure 3.

The likelihood reflects how well the equal and ordered pairs are separated in the attribute space. However, critically,  $P(\mathbf{x}_m, \mathbf{x}_n|D = 1)$  need not decrease monotonically as a function of rank differences. In other words, the model permits returning a higher likelihood for certain pairs separated by smaller margins. This is a direct consequence of our choice of the non-parametric KDE, which preserves local models of the original training data. This is valuable for our problem setting because in principle it means our method can correct imperfections in the original learned ranks and account for the non-uniformity of the space.

### 3.2.2 Prior model

Finally, we need to represent the prior over distinguishability. The prior could simply count the training pairs, i.e., let  $P(D = 1)$  be the fraction of all training pairs that were distinguishable. However, we again aim to account for the non-uniformity of the visual feature space. Thus, we estimate the prior based only on a subset of data near the input images. Intuitively, this achieves a simple prior for the label

distribution in multiple pockets of the feature space:

$$P(D = 1) = \frac{1}{K} |\mathcal{P}'_o|, \quad (6)$$

where  $\mathcal{P}'_o \subset \mathcal{P}_o$  denotes the set of  $K$  neighboring ordered training pairs.  $P(D = 0)$  is defined similarly for the indistinguishable pairs  $\mathcal{P}_e$ . Note that while the likelihood is computed over the pair’s rank difference, the locality of the prior is with respect to the image descriptor space. See step 3 in Figure 3.

To localize the relevant pocket of the image space, we adopt the metric learning strategy developed in prior work for comparing fine-grained attributes [27]. Briefly, it works as follows. First, a Mahalanobis distance metric  $f : X \times X \rightarrow \mathfrak{R}$  is trained to return small distances for images perceptually similar according to the attribute, and large distances for images that are dissimilar. Using that metric, pairs analogous to  $(x_m, x_n)$  are retrieved based on a product of their individual Mahalanobis distances, so as to find pairs whose members both align. See [27] for details.

### 3.3. Discussion

An alternative approach to represent partial orders is ordinal regression, where training data would consist of ordered equivalence classes of data. However, ordinal regression has severe shortcomings for our problem setting. First, it requires a consistent ordering of all training data (via the equivalence classes). This is less convenient for human annotators and more challenging to scale than the distributed approach offered by learning-to-rank, which pools any available paired comparisons. For similar reasons, learning-to-rank is much better suited to crowdsourcing annotations and learning universal (as opposed to person-specific [3, 1]) predictors. Finally, ordinal regression requires committing to a fixed number of buckets. This makes incremental supervision updates problematic. Furthermore, to represent very subtle differences, the number of buckets would need to be quite large.

Our approach offers a way to learn a computational model for just noticeable differences. While we borrow the term JND from psychophysics to motivate our task, of course the analogy is not 100% faithful. In particular, psychophysical experiments to elicit JND often permit systematically varying a perceptual signal until a human detects a change, e.g., a color light source, a sound wave amplitude, or a compression factor. In contrast, the space of all visual attribute instantiations does not permit such a simple generative sampling. Instead, our method extrapolates from relatively few human-provided comparisons (fewer than 1,000 per attribute in our experiments) to obtain a statistical model for distinguishability, which generalizes to novel pairs based on their visual properties.

JND models for attributes appear most relevant for category-specific attributes. Within a category domain (e.g., faces, cars, handbags, etc.), attributes describe fine-grained properties, and it is valuable to represent any perceptible differences (or realize there are none). In contrast, comparative questions about very unrelated things or extra-domain attributes can be nonsensical. For example, do we need to model whether the shoes and the table are “equally ornate”? or whether the dog or the towel is “more fluffy”? Accordingly, we focus our experiments below on two domains with rich vocabularies of fine-grained attributes, faces and shoes.

## 4. Experiments

With two challenging datasets, we present results on the core JND detection task (Sec. 4.1) and demonstrate its impact on an existing image search application (Sec. 4.2).

**Datasets and establishing JND ground truth** Our task requires attribute datasets that (1) have instance-level relative supervision, meaning annotators were asked to judge attribute comparisons on individual pairs of images, not object categories as a whole and (2) have pairs labeled as “equal” and “more/less”. To our knowledge, UT-Zap50K [27] and LFW-10 [23] are the only existing datasets satisfying those conditions.

To train and evaluate just noticeable differences, we must have annotations of utmost precision. Therefore, we take extra care in establishing the (in)distinguishable ground truth for both datasets. We perform pre-processing steps to discard unreliable pairs, as we explain next. This decreases the total volume of available data, but it is essential to have trustworthy results.

The **UT-Zap50K** dataset [27] consists of 50,025 total catalog shoes images from Zappos.com.<sup>2</sup> It contains 4 relative attributes, *open*, *pointy*, *sporty*, and *comfort*, with 3,000 annotated pairs each. Each pair was labeled by 5 workers on Mechanical Turk (MTurk). The labeled image pairs are partitioned into two sets—coarser pairs and fine-grained pairs—as determined by a two-stage crowdsourcing procedure to discover subtle pairs. As ordered pairs  $\mathcal{P}_o$ , we use all coarse and fine-grained pairs for which all 5 workers agreed and had high confidence. Even though the fine-grained pairs might be visually similar, if all 5 workers could come to agreement with high confidence, then the images are most likely distinguishable. As equal pairs  $\mathcal{P}_e$ , we use all fine-grained pairs with 3 or 4 workers in agreement and only medium confidence. Since the fine-grained pairs have already been presented to the workers twice (see [27]), if the workers are still unable to come to an consensus with high confidence, then the images are most likely indistinguishable. The resulting dataset has 4,778 total annotated pairs, consisting of on average 800 ordered and 350 indistinguishable (equal) pairs per attribute.

<sup>2</sup>[vision.cs.utexas.edu/projects/finegrained/utzap50k](http://vision.cs.utexas.edu/projects/finegrained/utzap50k)

The **LFW-10** dataset [23] consists of 2,000 face images, taken from the Labeled Faces in the Wild [10] dataset.<sup>3</sup> It contains 10 relative attributes, like *smiling*, *big eyes*, etc., with 1,000 labeled pairs each. Each pair was labeled by 5 people. As ordered pairs  $\mathcal{P}_o$ , we use all pairs labeled “more” or “less” by at least 4 workers. As equal pairs  $\mathcal{P}_e$ , we use pairs where at least 4 workers said “equal”, as well as pairs with the same number of “more” and “less” votes. The latter reflects that a split in decision signals indistinguishability. Due to the smaller scale of LFW-10, we could not perform as strict of a pre-processing step as in UT-Zap50K; requiring full agreement on ordered pairs would eliminate most of the labeled data. The resulting dataset has 5,543 total annotated pairs, on average 230 ordered and 320 indistinguishable pairs per attribute.

**Baselines** We are the first to address the attribute JND task. Prior relative attributes work evaluates only the “more/less” decision task [13, 17, 19, 22, 23, 27]. No prior methods infer indistinguishability at test time. Therefore, we develop multiple baselines to compare to our approach:

- **Rank Margin:** Use the magnitude of  $\Delta_{m,n}$  as a confidence measure that the pair  $m, n$  is distinguishable. This baseline assumes the learned rank function produces a uniform feature space, such that a *global threshold* on rank margins would be sufficient to identify indistinguishable pairs. To compute a hard decision for this method (for F1-scores), we threshold the Parzen window likelihood estimated from the training pairs by  $\epsilon$ , the mid-point of the likelihood means.
- **Logistic Classifier** [13]: Train a logistic regression classifier to distinguish training pairs in  $\mathcal{P}_o$  from those in  $\mathcal{P}_e$ , where the pairs are represented by their rank differences  $\Delta_{i,j}$ . To compute a hard decision, we threshold the posterior at 0.5. This is the method used in [13] to obtain a probabilistic measure of attribute equality. It is the closest attempt we can find in the literature to represent equality predictions, though the authors do not evaluate its accuracy. This baseline also maintains a global view of attribute space.
- **SVM Classifier:** Train a nonlinear SVM classifier with a RBF kernel to distinguish ordered and equal pairs. We encode pairs of images as single points by concatenating their image descriptors. To ensure symmetry, we include training instances with the two images in either order.<sup>4</sup>
- **Mean Shift:** Perform mean shift clustering on the predicted attribute scores  $R(x_i)$  for all training images.

<sup>3</sup>[cvit.iit.ac.in/projects/relativeParts](http://cvit.iit.ac.in/projects/relativeParts)

<sup>4</sup>We also implemented other encoding variants, such as taking the difference of the image descriptors or using the predicted attribute scores  $R(x_i)$  as features, and they performed similarly or worse.

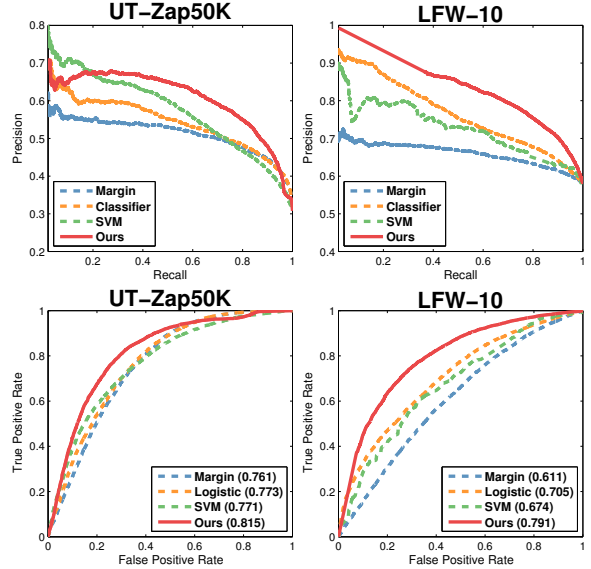


Figure 4: Just noticeable difference detection accuracy for all attributes. We show the precision-recall (top row) and ROC curves (bottom row) for the shoes (left) and faces (right) datasets. Legends show AUC values for ROC curves. Note that the Mean Shift baseline does not appear here, since it does not produce confidence values.

Images falling in the same cluster are deemed indistinguishable. Since mean shift clusters can vary in size, this baseline does *not* assume a uniform space. Though unlike our method, it fails to leverage distinguishability supervision as it processes the ranker outputs.

**Implementation details** We use the image descriptors kindly provided by the authors of each dataset. For UT-Zap50K, they are 960-dim GIST and 30-bin Lab color histograms. For LFW-10, they are 8,300-dim part-based features learned on top of dense SIFT bag of words features. We reduce their dimensionality to 100 with PCA to prevent overfitting. The part-based features [23] isolate localized regions of the face (e.g., exposing cues specific to the eyes vs. hair). We experimented with both linear and RBF kernels for  $R$ . Since initial results were similar, we use linear kernels for efficiency. We use Gaussian kernels for the Parzen windows. We set all hyperparameters ( $h$  for the KDE, bandwidth for Mean Shift,  $K$  for the prior) on held-out validation data. To maximize the use of training data, in all results below, we use leave-one-out evaluation and report results over 4 folds of random training-validation splits.

#### 4.1. Just Noticeable Difference Detection

We evaluate just noticeable difference detection accuracy for all methods on both datasets. Figure 4 shows the precision-recall curves and ROC curves, where we pool the results from all 4 and 10 attributes in UT-Zap50K and LFW-10, respectively. Tables 1 and 2 report the summary F1-scores and standard deviations for each individual attribute (see Supp for per-attribute curves). The F1-score is a useful

	Bald	DarkHair	BigEyes	GdLook	Masc.	Mouth	Smile	Teeth	Forehead	Young	All Attributes
Margin	71.10	55.81	74.16	61.36	82.38	62.89	60.56	65.26	67.49	<b>34.20</b>	63.52 ± 2.67
Logistic	75.77	53.26	86.71	64.27	87.29	63.41	59.66	64.83	75.00	NaN	63.02 ± 1.84
SVM	79.06	32.43	89.70	70.98	87.35	70.27	55.01	39.09	79.74	NaN	60.36 ± 9.81
M. Shift	66.37	56.69	54.50	51.29	69.73	68.38	61.34	65.73	73.99	23.19	59.12 ± 10.51
Ours	<b>81.75</b>	<b>69.03</b>	<b>89.59</b>	<b>75.79</b>	<b>89.86</b>	<b>72.69</b>	<b>73.30</b>	<b>74.80</b>	<b>80.49</b>	32.89	<b>74.02 ± 1.66</b>

Table 1: Just noticeable difference detection on LFW-10 (F1 scores). NaN occurs when recall=0 and precision=inf.

	Open	Pointy	Sporty	Comf.	All Attributes
Margin	48.95	67.48	66.93	57.09	60.11 ± 1.89
Logistic	10.49	62.95	63.04	45.76	45.56 ± 4.13
SVM	48.82	50.97	47.60	40.12	46.88 ± 5.73
M. Shift	54.14	58.23	60.76	<b>61.60</b>	58.68 ± 8.01
Ours	<b>62.02</b>	<b>69.45</b>	<b>68.89</b>	54.63	<b>63.75 ± 3.02</b>

Table 2: Just noticeable difference detection on UT-Zap50K (F1 scores).

summary statistic for our data due to the unbalanced nature of the test set: 25% of the shoe pairs and 80% of the face pairs are indistinguishable for some attribute.

Overall, our method outperforms all baselines. We obtain sizeable gains—roughly 4-18% on UT-Zap50K and 10-15% on LFW-10. This clearly demonstrates the advantages of our local learning approach, which accounts for the non-uniformity of attribute space. The “global approaches”, Rank Margin and Logistic Classifier, reveal that a uniform mapping of the relative attribute predictions is insufficient. In spite of the fact that they include equal pairs during training, simply assigning similar scores to indistinguishable pairs is inadequate. Their weakness is likely due both to noise in those mid-level predictions as well as the existence of JND regions that vary in scale. Furthermore, the results suggest that even for challenging, realistic image data, we can identify just noticeable differences at a high precision and recall, up to nearly 90% in some cases.

The SVM baseline is much weaker than our approach, indicating that discriminatively learning what indistinguishable image pairs look like is insufficient. This result underscores the difficulty of learning subtle differences in a high-dimensional image descriptor space, and supports our use of the compact rank space for our likelihood model.

Looking at the per-attribute results (Tables 1 and 2), we see that our method also outperforms the Mean Shift baseline. While Mean Shift captures dominant clusters in the spectrum of predicted attribute ranks for certain attributes, for others (like *pointy* or *masculine*) we find that the distribution of output predictions are more evenly spread. Despite the fact that the rankers are optimized to minimize margins for equal pairs, simple post-processing of their outputs is inadequate.

The tables also show that our method is nearly always best, except for two attributes: *comfort* in UT-Zap50K and *young* in LFW-10. Of the shoe attributes, *comfort* is perhaps the most subjective; we suspect that all methods may have suffered due to label noise for that attribute. While *young* would not appear to be subjective, it is clearly a more dif-

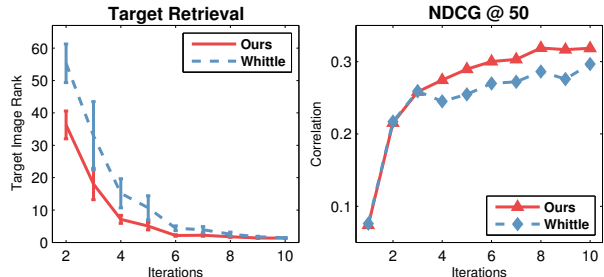


Figure 5: Image search results. We enhance an existing relative attribute search technique called WhittleSearch [14] with our JND detection model. The resulting system finds target images more quickly (left) and produces a better overall ranking of the database images (right).

ficult attribute to learn. This makes sense, as youth would be a function of multiple subtle visual cues like face shape, skin texture, hair color, etc., whereas something like *baldness* or *smiling* has a better visual focus captured well by the part features of [23]. Indeed, upon inspection we find that the likelihoods insufficiently separate the equal and distinguishable pairs. For similar reasons, the Logistic Classifier baseline [13] fails dramatically on both *open* and *young*.

Figure 6 shows qualitative prediction examples. Here we see the subtleties of JND. Whereas past methods would be artificially forced to make a comparison for the left panel of image pairs, our method declares them indistinguishable. Pairs may look very different overall (e.g., different hair, race, headgear) yet still be indistinguishable *in the context of a specific attribute*. Meanwhile, those that are distinguishable (right panel) may have only subtle differences.

Figure 7 illustrates examples of just noticeable difference “trajectories” computed by our method. We see how our method can correctly predict that various instances are indistinguishable, even though the raw images can be quite diverse (e.g., a strappy sandal and a flat dress shoe are equally *sporty*). Similarly, it can detect a difference even when the image pair is fairly similar (e.g., a lace-up sneaker and smooth-front sneaker are distinguishable for *openness* even though the shapes are close).

## 4.2. Image Search Application

Finally, we demonstrate how JND detection can enhance an image search application. Specifically, we incorporate our model into the existing WhittleSearch framework [14]. WhittleSearch is an interactive method that allows a user to provide relative attribute feedback, e.g., by telling the system that he wants images “more *sporty*” than some reference image. The method works by intersecting the relative

	Indistinguishable				Distinguishable			
Pointy								
Sporty								
Big Eyes								
Smiling								
Error Cases								

Figure 6: Example predictions. The top four rows are pairs our method correctly classifies as indistinguishable (left panel) and distinguishable (right panel), whereas the Rank Margin baseline fails. Each row shows pairs for a particular attribute. The bottom row shows failure cases by our method; i.e., the bottom left pair is indistinguishable for pointiness, but we predict distinguishable.



Figure 7: Example just noticeable differences. In each row, we take leftmost image as a starting point, then walk through nearest neighbors in relative attribute space until we hit an image that is distinguishable, as predicted by our method. For example, in row 2, our method finds the left block of images to be indistinguishable for *sportiness*; it flags the transition from the flat dress shoe to the pink “loafer-like sneaker” as being a noticeable difference.

attribute constraints, scoring database images by how many constraints they satisfy, then displaying the top scoring images for the user to review. See [14] for details.

We augment that pipeline such that the user can express not only “more/less” preferences, but also “equal” preferences. For example, the user can now say, “I want images that are equally *sporty* as image  $x$ .” Intuitively, enriching the feedback in this manner should help the user more quickly zero in on relevant images that match his envisioned target. To test this idea, we mimic the method and experimental setup of [14] as closely as possible, including their feedback generation simulator. See Supp for all details.

We evaluate a proof-of-concept experiment on UT-Zap50K, which is large enough to allow us to sequester disjoint data splits for training our method and performing the searches (LFW-10 is too small). We select 200 images at random to serve as the mental targets a user wants to find in the database, and reserve 5,000 images for the database. The user is shown 16 reference images and expresses 8 feedback constraints per iteration.

Figure 5 shows the results. Following [14], we measure the relevance rank of the target as a function of feedback iterations (left, lower is better), as well as the similarity of all

top-ranked results compared to the target (right, higher is better). We see that JNDs substantially bolster the search task. In short, the user gets to the target in fewer iterations because he has a more complete way to express his preferences—and the system understands what “equally” means in terms of attribute perception.

## 5. Conclusion

This work explores the challenging task of deciding whether a difference in attributes is perceptible. We present a simple, easily reproducible approach. Our method leverages local statistics in order to respect the perceptual non-uniformity of relative attribute space. Empirical results on two distinct domains with fine-grained visual properties demonstrate its advantages over multiple alternative strategies. In future work, we will investigate ways to blend our findings about JND with personalization, so as to account for heterogenous observer sensitivities that may exist for certain subjective attributes.

**Acknowledgements** We thank Naga Sandeep for providing the part-based features for LFW-10. This research is supported in part by ONR YIP Award N00014-12-1-0754.



## References

- [1] H. Altwaijry and S. Belongie. Relative ranking of facial attractiveness. In *WACV*, 2012.
- [2] A. Biswas and D. Parikh. Simultaneous active learning of classifiers and attributes via relative feedback. In *CVPR*, 2013.
- [3] C. Cao, I. Kwak, S. Belongie, D. Kriegman, and H. Ai. Adaptive ranking of facial attractiveness. In *ICME*, 2014.
- [4] K. Chen, S. Gong, T. Xiang, and C. Loy. Cumulative attribute space for age and crowd density estimation. In *CVPR*, 2013.
- [5] A. Datta, R. Feris, and D. Vaquero. Hierarchical ranking of facial attributes. In *FG*, 2011.
- [6] Q. Fan, P. Gabbur, and S. Pankanti. Relative attributes for large-scale abandoned object detection. In *ICCV*, 2013.
- [7] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing Objects by their Attributes. In *CVPR*, 2009.
- [8] D. Forsyth and J. Ponce. *Computer Vision: A Modern Approach*. Prentice Hall, 2002.
- [9] A. Gupta and L. Davis. Beyond Nouns: Exploiting Prepositions and Comparative Adjectives for Learning Visual Classifiers. In *ECCV*, 2008.
- [10] 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, 2007.
- [11] T. Joachims. Optimizing search engines using clickthrough data. In *SIGKDD*, 2002.
- [12] A. Kovashka and K. Grauman. Attribute adaptation for personalized image search. In *ICCV*, 2013.
- [13] A. Kovashka and K. Grauman. Attribute pivots for guiding relevance feedback in image search. In *ICCV*, 2013.
- [14] A. Kovashka, D. Parikh, and K. Grauman. WhittleSearch: Image search with relative attribute feedback. In *CVPR*, 2012.
- [15] N. Kumar, P. Belhumeur, and S. Nayar. Facetracer: A search engine for large collections of images with faces. In *ECCV*, 2008.
- [16] C. Lampert, H. Nickisch, and S. Harmeling. Learning to Detect Unseen Object Classes by Between-Class Attribute Transfer. In *CVPR*, 2009.
- [17] S. Li, S. Shan, and X. Chen. Relative forest for attribute prediction. In *ACCV*, 2012.
- [18] T. Matthews, M. Nixon, and M. Niranjan. Enriching texture analysis with semantic data. In *CVPR*, 2013.
- [19] D. Parikh and K. Grauman. Relative attributes. In *ICCV*, 2011.
- [20] E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.
- [21] D. Reid and M. Nixon. Using comparative human descriptions for soft biometrics. In *IJCB*, 2011.
- [22] A. Sadovnik, A. Gallagher, D. Parikh, and T. Chen. Spoken attributes: Mixing binary and relative attributes to say the right thing. In *ICCV*, 2013.
- [23] R. Sandeep, Y. Verma, and C. Jawahar. Relative parts: Distinctive parts for learning relative attributes. In *CVPR*, 2014.
- [24] W. Scheirer, N. Kumar, P. Belhumeur, and T. Boult. Multi-attribute spaces: Calibration for attribute fusion and similarity search. In *CVPR*, 2012.
- [25] A. Shrivastava, S. Singh, and A. Gupta. Constrained semi-supervised learning using attributes and comparative attributes. In *ECCV*, 2012.
- [26] B. Siddiquie, R. S. Feris, and L. S. Davis. Image ranking and retrieval based on multi-attribute queries. In *CVPR*, 2011.
- [27] A. Yu and K. Grauman. Fine-grained visual comparisons with local learning. In *CVPR*, 2014.