Compare and Contrast: Learning Prominent Differences in Relative Attributes

by

Steven Ziqiu Chen

stevenchen@utexas.edu

Supervised by:
Dr. Kristen Grauman

Department of Computer Science

The University of Texas at Austin

Abstract

Relative attribute models can compare images in terms of all detected properties or attributes, exhaustively predicting which image is fancier, more colorful, more natural, and so on without any regard to ordering. However, when humans are asked to compare images, certain attribute differences will naturally stick out and come to mind first. These most noticeable differences, or *prominent differences* in relative attributes, are likely to be described first. In addition, certain differences in attributes, although present and true, may not be mentioned at all.

In this work, we introduce and model prominent differences, a rich new functionality for comparing images. Using instance-level human annotations of most noticeable differences, we build a model trained using relative attribute features that predicts prominent attributes for new image pairs. We test our model on the challenging UT-Zap50K shoes and LFW-10 faces datasets, and outperform an array of baseline methods. We then demonstrate how our prominence model improves two vision tasks, image search and textual description generation, enabling more natural communication between people and vision systems.

Acknowledgments

I am extremely fortunate and grateful to have Kristen Grauman as my advisor. Her generous support made this work possible, and I would like to thank her for her dedication and mentorship.

I would like to thank Aron Yu for patiently helping me through several technical challenges. I would also like to thank Calvin Lin for encouraging me and preparing me to conduct research as an undergraduate.

Finally, I would like to thank my family and friends for all their love and support. Thank you all so much for being there for me.

Contents

List of Figures

1 Introduction

Suppose you are asked to compare and contrast cats and dogs. You would likely quickly say that cats meow and dogs bark, then perhaps state that dogs are usually larger than cats. As soon as you are given the topic, cats and dogs, these differences stick out and are most noticeable and apparent. However, consider that cats and dogs have a huge number of actual differences. For instance, cats have retractable claws, whereas dogs' claws do not retract, and most cats will lick themselves clean, whereas dogs usually need a bath. Although these differences are certainly present and true, they are much less noticeable to us, and we would likely mention them later in our answer, or not mention them at all.

In general, when we perform any compare and contrast task on a pair of objects, certain differences in properties stick out as being most noticeable out of the space of all discernible differences. These most noticeable differences, or *prominent differences*, intuitively stick out to us as most noticeable and would be described first, while most other differences are not as noticeable and do not stand out. We can liken this to filling in a Venn Diagram, something many of us learned in grade school (see Figure 1). Given a topic, prominent differences that stick out to us are written first, and likely end up at the top of the circles, whereas other differences are not as noticeable, and generally appear later down the list or do not appear.

In this work, our main goal is to learn and model prominent differences in visual content. As a motivating example, many forms of visual media make use of prominent differences in their expression. For example, comparative advertising (Figure 2a and 2b) makes use of prominent differences to emphasize certain differences between competing products. In Apple's "Get a Mac" advertising campaign [35], a PC and a Mac are personified as two different individuals (Figure 2a). Looking at the advertisement, what sticks out is how *formal*, *old-fashioned*, and *businesslike* the PC personification looks compared to the more *casual* and *relaxed* Mac. Figures are also often used to

make statements between two different offerings: Verizon's 3G coverage ad [41] makes use of coverage maps of the United States to emphasize their advantage over AT&T (Figure 2b). Although the maps are certainly very different in their *color*, what sticks out as prominent and most noticeable is the wider *coverage area* of Verizon's map compared to AT&T's map. Understanding prominence reveals deep information on human perception of visual differences, and has significant potential to improve machine understanding of visual media and enhance communication between humans and vision systems.

In particular, in this work we focus on learning prominent differences in relative visual attributes. To provide some background on attributes in vision, when we compare images, we generally describe differences by their *attributes*. Attributes are human-nameable visual properties of images [1–3, 5–8, 11, 14, 17, 20, 21, 24, 26, 27, 29, 30, 32, 38, 39, 42–44, 48, 50, 51], and are used to describe anything from material properties (*smooth*, *furry*), parts (*4-legged*, *has glasses*), shapes (*round*, *boxy*), to styles (*sporty*, *formal*) and expressions (*smiling*, *angry*). Attributes are machine-learnable and can also be shared across different categories of images, making them valuable as semantic cues in many human-centric applications. For instance, attributes have made a significant impact on image search [21, 22, 44, 48], face verification [26, 30], description generation [8, 24, 34, 38, 42, 48], and human supervision for visual recognition [3, 14, 27, 28, 38, 48].

At their introduction, attributes work placed a focus on presence/absence predicates (e.g., *is standing, is not standing*) [8, 25, 28]. As such, these predicate properties are referred to as *binary attributes*. Binary attributes work well for describing clear-cut properties such as parts, where each image distinctly contains or does not contain the property (e.g., a zebra has four legs and a tail, but not wings), and have seen success at predicting importance of objects [1], predicting aesthetic quality of images [7], and ranking image search results using binary attribute-based queries [44]. However, binary attributes are limited in their expression, and do not work well with many attributes that are not binary present or absent (e.g., people's expressions, where many people are not simply always *smiling* or *not smiling*, or shoe styles, where many shoes are not simply always *formal* or *not formal*).

More recently, *relative attributes*, or attributes that indicate an image's attribute strength with respect to other images, were first introduced by Parikh and Grauman [38] as a more intuitive and meaningful representation of comparative image proper-

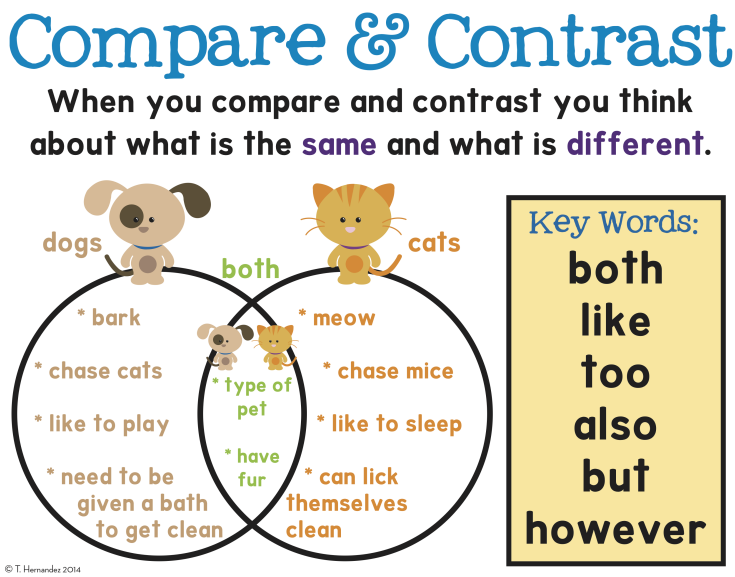
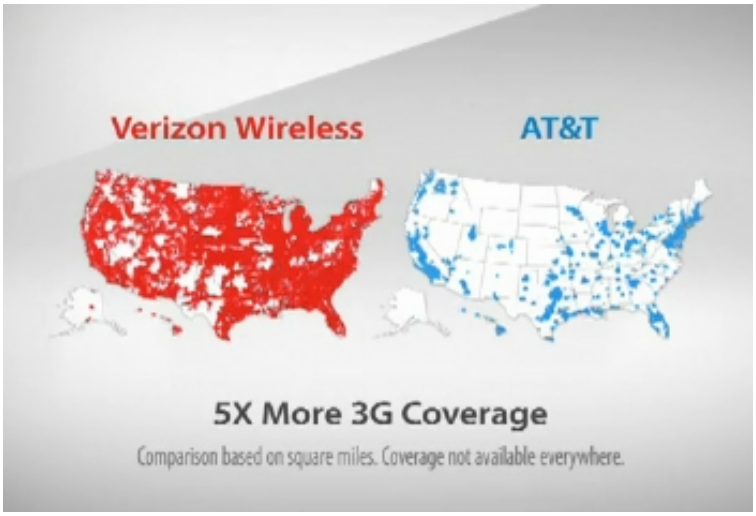


Figure 1: **Compare and Contrast Poster** - Compare and contrast is one of our most natural forms of thinking, and a skill we obtain at a young age and use for learning and language throughout our lives. This poster from a first-grade classroom [10] shows a Venn Diagram, and implicitly illustrates which differences are prominent from the order in which they are listed from top to bottom; out of all the points given, the prominent difference is that dogs bark while cats meow.

(a) **PC Versus Mac** - In Apple's "Get a Mac" campaign, which ran from 2006 to 2009 [35], a PC (left) and a Mac (right) are personified as two individuals. In comparing the two personifications, the PC and Mac prominently differ in how *formal* and *businesslike* they are: the PC person is very formal and businesslike, as seen in his rigid clothing, hairstyle, and gestures, whereas the Mac person is much more casual and relaxed.



(b) **Verizon Versus AT&T** - Verizon's "There's a Map For That" advertisements from 2009 [41] uses human perception of prominent differences to illustrate differences between AT&T and Verizon's network. Looking at the coverage maps, *coverage area* stands out as the prominent difference between the maps, as opposed to the difference in *color*, which is present simply for brand awareness.

Figure 2: **Comparative Advertising** - Two examples of comparative advertising, using prominent differences in visual attributes to convey differences in products to consumers.

ties, and have since been widely used for different vision tasks [6, 9, 21, 32, 42, 45, 46, 49–51]. Relative attributes express more/less comparisons of attribute strength between images (e.g., person X is *smiling more* than person Y, but *smiling less* than person Z), and are modeled as real-valued scores for each attribute indicating its strength in an image [38, 45, 46, 49]. Relative attributes allow for fine-grained comparisons between images of the same category (e.g., this high-heeled shoe is *more fancy*, *more shiny*, and *less tall* than the other high heel), and also allow one to measure the magnitude of differences in each attribute between two images. With these capabilities, relative attributes have been applied to describe images more naturally [42], discern fine-grained differences [50, 51], and predict an image's virality [6].

In visual compare and contrast tasks, *prominent differences in relative attributes* will stick out to us and be most noticeable for different image pairs. For instance, in Figure 3a, human judges point out *dark hair* as the prominent relative attribute difference between the two people: the person on the left has noticeably darker hair than the person on the right. In Figure 3d, *color* sticks out as the prominent difference between the two shoes: the left shoe is noticeably less colorful than the right shoe.

The proposed prominent differences have many applications in computer vision. Humans interact with vision systems as both users and supervisors, and communicate prominent differences through what they say. During an interactive search task, where humans provide comparative feedback (e.g, I would like to see images like this shoe, but *more formal* [21]), the relative attributes that people elect to comment on are likely prominent differences. Prominence, by definition, influences which attributes humans provide first when comparing and contrasting images, and consequently affects how humans understand computer-generated descriptions of image pairs. When humans act as supervisors, teaching a vision system about a new and unseen image category through comparisons with seen categories (e.g., donkeys are like horses, but with *longer ears*, *smaller hooves*, and *flatter backs*), prominent differences are likely to be presented first to the machine.

Thinking about why we perceive prominent differences, it is important to note that a large difference in relative attribute strength does not necessarily, or even regularly, correspond to a prominent difference. For instance, the individuals in Figure 3b differ significantly in the *dark hair* attribute, just like the individuals in Figure 3a: however, most human judges in our experiments observe *smiling* as the prominent difference for that image pair. Although the shoes in Figure 3f differ significantly in how colorful

(a) **dark hair (>)** (b) **smiling (<)** (c) **visible forehead (>)**

(d) **colorful (<)** (e) **fancy (>)** (f) **formal (<)**

Figure 3: **Prominent Differences in Image Pairs** - Prominent differences in relative attributes differ for different pairs of images. Even though 3a and 3b both differ in how dark their hair is, *dark hair* sticks out as the most prominent difference in 3a but not in 3b. The large difference in color palette makes *colorful* the prominent difference in 3d, while in 3f, a combination of visual attribute differences between the sneaker and dress shoe result in *formal* as the prominent difference.

they are, similar to the shoes in Figure 3d, their observed prominent difference is that the left shoe is noticeably *less formal* than the right shoe. A set of image pairs may have the same wide difference in a particular relative attribute, but that attribute may only be a prominent difference for some pairs but not for others. This suggests a trivial solution of declaring prominence as the most widely separated attribute in a pair is insufficient (as we will confirm later).

Indeed, there is a large variety of reasons why an attribute may stand out as a prominent difference in a pair of images: in this work, we seek to capture these causes in our approach. Large differences in relative attribute strength certainly play a role, as observed in Figure 3d. Absence of other significant attribute differences can also influence prominence: the people in Figure 3a have very similar expressions and complexions, so their hair color stands out, even though both have different degrees of greying hair. Unusual and uncommon occurrences also impact how we perceive prominence: for instance, the man in Figure 3c has an unusually high hairline and large forehead, making *visible forehead* the prominent difference. Interactions between visual properties and attributes also act as a complex influence on prominent differences: Figure 3e's left shoe has an elegant shape and combination of different materials, colors, and textures, compared to the relatively simple and bland right

shoe. In this instance, the left shoe is prominently *fancier* than the right shoe. In this work, we will train a model to learn these pairwise interactions between relative attributes in order to predict prominent differences.

Existing relative attribute rankers [38, 45, 46, 49] can predict the strength of relative attributes in individual images. However, no existing vision systems represent or predict prominent differences. Although relative attribute models can show which attributes have the largest difference in strength for image pairs, as noted previously, this is only one reason out of many that could contribute to prominence. Binary attribute dominance [48] represents how much a binary attribute stands out over others for categories of images. For example, binary attribute dominance can express that *is furry* stands out more for cats than *is big* or *swims*. However, binary attribute dominance cannot take into account pairwise relationships or relative comparisons in attributes (e.g., Persian cats are *more furry* and *less nimble* than Cornish Rex cats), and loses important fine-grained information by abstracting to categories (see Section 2.3 for more details).

In this work, we introduce and model prominent differences in attributes. We propose a model that, given a pair of novel images, predicts which attribute stands out as the most prominent difference for that image pair. To train our model, we create feature vectors for training image pairs using the outputs of state-of-the-art relative attribute ranking models, capturing the complex pairwise interactions between relative attributes that contribute to prominent differences. We collect prominent difference annotations at a large scale and transform these annotations into prominence ground truth, the first annotated dataset collected for this purpose.

We evaluate our prominence model on two unique and challenging domains: the UT-Zap50K shoes dataset [50] and the LFW10 faces dataset [43]. These datasets are designed for fine-grained recognition and comparisons between related images, making them suitable for prominent difference modeling. We assemble a new relative attribute vocabulary for UT-Zap50K and use the existing vocabulary from LFW10, collecting prominence data for image pairs from both datasets. We show that our model significantly outperforms an array of baselines for predicting prominent differences on novel image pairs, including the state-of-the-art binary dominance approach [48].

We then illustrate how our prominence model can be used to enhance two vision applications: interactive image search and description generation. For the first appli-

cation, we consider an interactive image search framework called WhittleSearch [21, 22], in which users provide feedback in the form of relative attribute comparisons to whittle away irrelevant images. Users of this system are prone to provide prominent differences as feedback, instead of arbitrary comparisons; we leverage this to return relevant images to users in fewer iterations. For our second task, we tackle generating textual descriptions of novel image pairs using their attributes. We demonstrate that describing prominent differences in relative attributes leads to more natural and expressive textual descriptions.

In this work, we introduce and learn prominent differences in relative visual attributes, and demonstrate how prominent differences can be used to enhance image search and description generation. In Chapter 2, we discuss related work, with a focus on attributes and the vision tasks to which we apply prominent differences. In Chapter 3, we present our novel approach for modeling and annotating prominent differences. In Chapter 4, we illustrate how we apply prominent differences to interactive image search and textual description generation. In Chapter 5, we present our experimental setup and results. We conclude in Chapter 6, where we summarize our work and suggest future work in the area.

2 Related Work

We now review related work on attributes, with more emphasis placed on relative attributes and models of visual importance. We then present work evaluated on the applications to which we apply prominent differences: image search and description generation.

2.1 Attributes

Human-nameable *semantic properties*, or *attributes*, have been used for a variety of applications [1–3, 5–8, 11, 14, 17, 20, 21, 24, 26, 27, 29, 30, 32, 38, 39, 42–44, 48, 50, 51]. Attributes are human-readable and machine-understandable properties of images (e.g., *smiling*, *shiny*) that are used by people to describe images. Attributes serve as expressive mid-level features for recognition in scene classification [39] and face verification [26, 30]. Attributes also express a vocabulary for human input: Branson *et al.* [3] conduct fine-grained recognition with interactive attribute guidance, while Lad and Parikh [27] conduct semi-supervised clustering using attribute descriptions. In addition, attributes have served as a bridge for learning unseen visual categories using human descriptions, known as zero-shot learning [14, 28, 38, 48].

In the past few years, deep neural networks have improved attribute prediction and performance in many areas, such as face retrieval [30] and clothing search [29]. Neural networks have also enabled new methods to produce attributes: Huang *et al.* [11] explore unsupervised learning and prediction of new attribute vocabularies with deep convolutional networks.

Broadly, attributes have been used in vision as a communication medium between humans and computers. Our approach seeks to improve this channel, by introducing and predicting prominent differences. In contrast to any of the above, instead of

detecting individual objects or properties in an image, we learn which properties among all stick out to people as prominent between image pairs.

2.2 Relative Attributes

More recently, relative attributes, first introduced in [38], have been widely used to indicate an image's attribute strength with respect to other images [6, 9, 21, 32, 42, 45, 46, 49–51], encoding more than just binary presence or absence. This richer representation helps to discern fine-grained differences [50, 51], describe images more naturally [42], or predict a person's age [9] or an image's virality [6]. Recent work has also explored using neural networks to predict relative attributes with success [45, 46, 49].

However, no prior work considers which relative attributes stand out over others, and what attributes humans use when comparing images. Our work introduces and models prominent differences, a novel functionality representing most noticeable differences in relative attributes. It is important to emphasize that prominent differences *do not* correspond to the relative attributes that have the largest differences in strength (See Section 5.3 for experimental results); relative attribute strength is just one factor out of many interactions between visual attributes that contributes to prominent differences.

2.3 Importance of Objects and Attributes

Different concepts of visual importance have used attributes as features and/or a vocabulary [1, 7, 37, 47, 48]. Object importance is defined as the likelihood that an object would be named first by a viewer out of the objects in an image [47]. Berg *et al.* [1] learn object importance from natural language descriptions of images, and use binary attribute scores as a feature for predicting object importance in novel images. In contrast, we predict which relative attributes are perceived as most noticeably different between image pairs, using pairwise relative attribute scores as input features. Dhar *et al.* [7] use binary attributes to predict the perceived aesthetic quality of new images, while Kong *et al.* [18] use a deep neural network incorporating joint learning of binary attributes and visual content to rank photo aesthetics. As opposed to

aesthetic quality, we consider the separate concept of prominent differences selected from a vocabulary of visual properties.

In their recent work, Turakhia and Parikh [48] introduce the related concept of binary attribute dominance. Attribute dominance is defined as a ranking score for each binary attribute indicating how much that attribute stands out over other attributes for each high-level object category. For example, given the category of high heeled shoes, the binary attribute *fancy* has a higher dominance score than *comfortable* or *rugged*. The authors collect attribute dominance ground truth for each category by presenting annotators with all possible pairs of binary attributes and observing the frequency each attribute was picked as standing out more over its corresponding pairs. Attribute dominance is modeled by projecting the categorical dominance values onto individual images from within the category and training a predictor using individual image binary classifier outputs as features.

Our work is distinct in several notable factors. (1) At a high level, the goals are different. We seek to learn which differences stick out as most noticeable in a given image comparison, while attribute dominance models which binary attributes are more noticeable than others per visual category. (2) We define prominent differences as natural "first impressions", the top k relative attributes out of the vocabulary that are most noticeable for an image pair, as opposed to a single strength score for each binary attribute intended to apply to all instances in a large object category. (3) We present annotators with individual image pairs and the entire relative attribute vocabulary and ask for a single prominent difference, a task similar to brainstorming for compare and contrast, whereas Turakhia and Parikh [48] annotate exhaustive binary attribute pairs for each category, which is much less scalable and not as intuitive. (4) We model prominent differences at the instance image level, capturing rich differences between individual images, whereas attribute dominance can only capture general category-level properties. For instance, given instance images of tennis shoes, the attribute dominance model may detect general trends that *sporty* and *comfortable* are more dominant than other binary attributes; however, our prominent differences model captures individual differences between specific tennis shoes, such as one shoe being prominently *more colorful* than other, in addition to learning categorical trends. (5) We evaluate our approach on challenging, single-domain datasets designed for fine-grained recognition and test on unseen instance pairs, whereas Turakhia and Parikh [48] evaluate on categorical datasets and test on unseen categories.

2.4 Image Saliency

Works modeling saliency (e.g., [13, 16, 37]) have used attributes and other visual features to predict which areas of an image attract human attention, or, in other words, where humans tend to look. Although saliency has an influence on prominent differences, for instance in how areas of fixation influence human perception of attributes, prominence is a higher-level, pairwise concept, and the result of a combination of visual attribute factors.

2.5 Image Search

Image search has benefited from approaches using attributes [21, 44, 48, 51]. Siddiquie *et al.* [44] introduce image search using binary attribute queries, while Turakhia and Parikh [48] improve attribute query-based search for image categories by using the order in which people name attributes. We also use attribute ordering, but apply it to the instance-level and interactive WhittleSearch framework [21, 22]. In WhittleSearch, human relative attribute feedback is used to whittle away images (i.e., I want images like this shoe, but *more formal*). Recent work adds equality selection to WhittleSearch using just noticeable differences [51]. In contrast, instead of introducing more mechanics to WhittleSearch, we use the existing order in which people select feedback comparisons as signals for prominent differences, leading to faster target retrieval without any extra feedback required.

2.6 Describing Images

As machine-understandable semantic properties, attributes are well-suited for visual description tasks [8, 24, 34, 38, 42, 48]. Recent work uses attributes to generate binary attribute descriptions of objects [8], full sentence descriptions [24], and includes improvements to these systems that list only more noticeable attributes [48].

Recent work also explores generating referring expressions, or phrases identifying a specific object in an image [33, 34]. Mitchell *et al.* [34] generate expressions using an attribute vocabulary, while Mao *et al.* [33] use deep learning to generate expressions from raw images. Instead, our work focuses on differences between images, which

could potentially be applied to enhance referring expressions by predicting prominent differences between objects in an image. Parikh and Grauman [38] describe image pairs by relative attribute comparisons, while Sadovnik *et al.* [42] explore whether to say a binary or relative statement for a particular attribute. However, both [38] and [42] use an arbitrary order to list all applicable comparison statements; we seek to improve these comparative descriptions by focusing on differences that are prominent and natural for human communication (e.g., describing the most prominent differences to create more intuitive and expressive descriptions of image pairs).

3 Approach

First, we present an overview of relative attribute models (Section 3.1). We discuss relative attributes and how they differ from binary attributes. We then present the general paradigm for learning a model for predicting relative visual attribute strength using labeled image pairs.

Next, we explain our approach for modeling prominent differences (Section 3.2). We present the main problem we seek to solve, then describe how we construct our model for learning prominent differences. We demonstrate how we use relative attribute outputs as inputs to our model, how we train our prominence model using prominence labeled image pairs, and how we predict prominent differences for new image pairs.

Finally, we detail how we annotate and collect prominent difference data to train and evaluate our approach (Section 3.3). We illustrate the interface and intuition behind our instance-level prominence data collection from human annotators, and how we transform these human annotations into ground truth prominent differences.

3.1 Relative Attribute Models

Relative attributes are visual semantic properties that represent the strength of an attribute in an image relative to other images [38]. While binary attributes only represent the presence or absence of a property in an image, i.e., *is smiling* or *is not smiling*, relative attributes encode comparisons between images, i.e., the person on the left is *smiling more* than the person on the right. This allows for comparisons between images (e.g., shoe X is *more fancy*, *less rugged*, and *more pointy* than shoe Y), and also reveals richer information for certain images and properties that cannot necessarily be represented by binary presence/absence.

Figure 4: **Relative Attributes** - Relative attributes [6, 9, 21, 32, 38, 42, 45, 50, 51] allow us to rank images across a range of relative strengths for a particular attribute. Using relative attributes, we can compare whether one shoe image is *more sporty* than another, or whether one face image is *more smiling* than another. In addition, we can also use relative attributes to compare how large the difference is between two images in terms of a particular attribute. However, as we will show, widest difference alone is not enough to capture prominent differences between image pairs.

Now, we describe a general framework for relative attribute predictors. Suppose we have a set of images $I = \{x_i\}$, along with a vocabulary of M relative attributes $A = \{a_m\}_{m=1}^{M}$. Let $\mathcal{D}(x_i) \in \mathbb{R}^D$ represent the image's D-dimensional visual descriptor. This descriptor could be comprised of GIST [36], color, part-based representations, convolutional neural network (CNN) features, or just the raw pixels of the image. Given a target attribute a_m from the vocabulary, along with a pair of images $y_{ij} = (x_i, x_j)$, the goal of the relative attribute ranker is to determine if one image contains more of attribute a_m than the other, or if both images have similar relative strengths of attribute a_m.

Relative attribute models currently use ordered and unordered pairs of images for supervised training [6, 9, 21, 32, 38, 42, 45, 50, 51]. A learning algorithm is given a set of ordered image pairs $O_m = \{(i, j)\}$ and a set of unordered image pairs $S_m = \{(i, j)\}$ such that $(i, j) \in O_m \implies i > j$, i.e., image i contains more of attribute a_m that image j, and $(i, j) \in S_m \implies i \sim j$, i.e., image i and image j have similar strengths of attribute a_m.

The idea of a relative attribute model is to learn a ranking function $\mathcal{R}_m(\mathcal{D}(x_i))$ for each attribute a_m such that the following constraints are satisfied as well as possible:

$$\forall (i,j) \in O_m : \mathcal{R}_m(\mathcal{D}(x_i)) > \mathcal{R}_m(\mathcal{D}(x_j)) \tag{3.1}$$

$$\forall (i,j) \in S_m : \mathcal{R}_m(\mathcal{D}(x_i)) = \mathcal{R}_m(\mathcal{D}(x_j)). \tag{3.2}$$

The definition of what satisfies the constraints best depends on the specific relative attribute model being used, such as a RankNet objective [4, 45, 46, 49] for a deep convolutional neural network based ranker, or wide margin paired classification objective [38] for a ranking SVM [15] based ranker. We employ both such models in our implementation, and review them briefly next.

Ranking SVM relative attribute models optimize $\mathcal{R}_m^{(svm)}(\mathcal{D}(x_i)) = w_m^T \mathcal{D}(x_i)$ to preserve the ordering of constraints while maximizing the distance between the closest data points $(\mathcal{D}(x_i), \mathcal{D}(x_j))$ when projected onto w. $w \in \mathbb{R}^D$ is the weight vector to be learned, and in this case, is linear; however, nonlinear models are also possible using the kernel method. Ranking SVM relative attributes have seen wide use [6, 9, 21, 32, 38, 42, 50, 51], are flexible in their choice of input image descriptors, and generally require fewer training observations to achieve reasonable performance on relative attribute prediction.

Deep CNN based relative attribute rankers have recently emerged in the literature as an alternative to train strong predictors of relative attribute strength [45, 46, 49]. These models generally combine a CNN optimized for paired ranking loss [4], and use raw image pixels and image crops as input. Deep CNN rankers have seen higher prediction accuracy for relative attributes over ranking SVM models; however, they require significantly more time to train the network and generally need and benefit more from larger amounts of training data.

Given a novel pair of images $y_{uv} = (x_u, x_v)$, we can compare their relative attribute scores $r_m^u = \mathcal{R}_m(\mathcal{D}(x_u))$ and $r_m^v = \mathcal{R}_m(\mathcal{D}(x_v))$ to determine whether x_u contains more of attribute a_m, x_v contains more of attribute a_m, or x_u and x_v are similar in terms of attribute a_m. In addition, we can also compute the "relative difference" in relative attribute scores by computing $|r_m^u - r_m^v|$, and use this as a measure of how different two images are in terms of attribute a_m. By computing this difference for all attributes

and taking the maximum

$$\mathcal{W}^{uv} = \arg\max_m |r_m^u - r_m^v| \tag{3.3}$$

we can obtain the attribute $\mathcal{W}^{uv}$ with the widest difference in relative attribute score for the image pair y_{uv}. In later sections we will refer to this quantity as the *widest relative attribute difference*.

Although the widest relative attribute difference $\mathcal{W}^{uv}$ indicates the attribute with the largest strength difference between an image, we hypothesize that this attribute is not necessarily the same attribute that humans perceive as the most prominent difference for image pairs. As discussed in the introduction (Section 1), a set of image pairs may have some attribute as the widest score difference, but that attribute would be prominent for some pairs but not for others. There are many different reasons that contribute to an attribute standing out as prominent, including unusual occurrences or absences of attributes, interactions between attributes in one image, and pairwise interactions between the attributes of the image pair. Our results in Section 5.3 will support this, demonstrating that just selecting widest relative attribute difference is inadequate for predicting prominence.

3.2 Modeling Prominent Differences

We now introduce our model for representing and predicting prominent differences. Our approach uses pairwise relative attribute scores from the entire attribute vocabulary as input features, exposing the complex interplay between the attributes of a specific image pair that results in prominent differences. We illustrate how we train our model using prominent difference annotations of instance image pairs, and demonstrate how our model is used to predict the most prominent difference for new image pairs.

Suppose we have a set of images $I = \{x_i\}$, along with a vocabulary of M relative attributes $A = \{a_m\}_{m=1}^M$ as defined before. These relative attributes could be specific to the image set in consideration, or could be generic relative visual attributes for any type of image. For instance, a domain-specific vocabulary for shoes could contain relative attributes such as *ruggedness, sportiness, tallness*, etc., while a generic vocabulary could contain relative attributes such as *colorfulness, blurriness, interest-*

ingness.

In addition, for each attribute a_m in the vocabulary, we are given a set of unordered prominence image pairs $U_m = \{(x_i, x_j)\}$ such that the most prominent difference when comparing image x_i and image x_j is the relative attribute a_m. Note that U_m is distinct from O_m and S_m, the sets of ordered and unordered relative attribute pairs used to train a relative attribute ranker. While O_m and S_m represent relative attribute strength comparisons for one attribute, U_m represents most prominent differences in terms of relative attributes for image pairs across the whole vocabulary.

Our goal is to construct a model that, given a novel pair of images $y_{uv} = (x_u, x_v)$, predicts which single attribute $\mathcal{A}^{uv}$ is the most prominent difference for that image pair.

To do this, we build M predictors

$$\mathcal{P}_m(y_{uv}) \tag{3.4}$$

for $m = 1, \ldots, M$ such that $\mathcal{P}_m(y_{uv})$ is the predicted confidence score that the prominent difference between image pair y_{uv} is the attribute a_m.

In order to represent an image pair y_{ij} as an unordered pair for training and testing, we need a pairwise invariant transformation $\phi(y_{ij})$ that transforms the features of the two images into a single, joint representation. This representation needs to be symmetric (i.e., $\phi(y_{ij}) = \phi(y_{ji})$) so that the model always predicts the same most prominent difference for a specific pair of images.

To create our pairwise representation, we first obtain the relative attribute ranking score

$$r_m^i = \mathcal{R}_m(\mathcal{D}(x_i)) \tag{3.5}$$

for each image x_i in the pair using a relative attribute ranker $\mathcal{R}_m$ as described before, for all attributes in the vocabulary, resulting in M relative attribute scores $r_1^i, \ldots, r_M^i$. We use relative attribute scores to represent each image so that our model captures the pairwise interactions between all relative attributes that result in prominent differences.

Then, we convert the M relative attribute scores for each image into a pairwise invariant representation. We experiment with different pairwise invariant transforma-

tions, including element-wise product, absolute difference, and average. For its strong performance on prominence prediction, we select the average of the pair's scores for each attribute and concatenate the absolute difference between the pair's attribute scores, creating a feature vector of length $2M$:

$$\phi(y_{ij}) = (\frac{r_1^i + r_1^j}{2}, \ldots, \frac{r_M^i + r_M^j}{2}, \left|r_1^i - r_1^j\right|, \ldots, \left|r_M^i - r_M^j\right|) \tag{3.6}$$

This feature representation captures the individual relative attribute scores while maintaining symmetry: pair scores for each attribute can be reconstructed simply by $average \pm \frac{absdifference}{2}$. Additionally, we standardize all pairwise feature vectors before input into our model.

We experiment with two different relative attribute rankers $\mathcal{R}(\mathcal{D}(x_i))$ for generating relative attribute scores r^i; one ranking SVM model and one deep CNN model, as defined above (Section 3.1). The first model we use is the ranking SVM relative attribute model with similarity first introduced by Parikh and Grauman [38], while the second is the the deep convolutional neural network with a spatial transformer network introduced by Singh and Lee [45]. We experiment with both rankers as the relative attribute score predictor for our prominence model, and report results from both.

Given our pairwise relative attribute features $\phi(y_{uv})$, we now predict the confidence score for each attribute a_m using

$$\mathcal{P}_m(y_{uv}) = \mathcal{S}_m(w_m^T \phi(y_{uv})) \tag{3.7}$$

where w_m^T are weights learned by a binary linear classifier, and $\mathcal{S}_m$ is a function mapping linear classifier scores to confidence values.

To learn the linear classifier predictor weights w_m^T for each attribute a_m, we first mark all training pairs from U_m as positive examples, and training pairs from other prominence sets $\{y_{ij} | y_{ij} \notin U_m\}$ as negative examples. We use a single binary classifier for each attribute using its positive and negative training pairs. Specifically, we use a linear SVM classifier for its strong performance in practice, though certainly other classifiers would be applicable.

To address the class imbalance problem, where the number of negative examples, i.e., all image pairs that are not most prominent in a particular attribute, outweighs

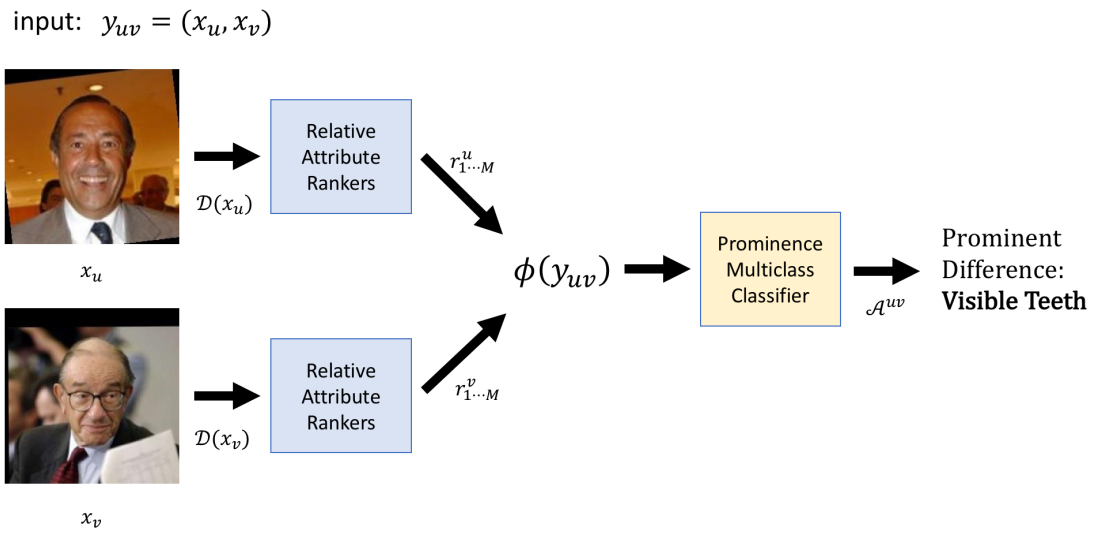


Figure 5: **Pipeline for Prominent Difference Prediction** - To predict the most prominent difference given a new image pair, our approach first computes the relative attribute scores over the entire vocabulary for each individual image, then combines these scores into a pairwise representation, which is used as feature input into the multiclass prominence classifier. The classifier returns the predicted most prominent difference for the image pair.

the number of positive examples, we adjust the SVM misclassification cost to give a higher penalty to misclassifying positive examples.

Raw linear SVM classifier outputs generally provide poor estimates of probability, producing distorted probability distributions. So, to transform the classifier output $w_m^T \phi(y_{ij})$ to a confidence score $\mathcal{P}_m(y_{uv})$, we use Platt's method [40], indicated by $\mathcal{S}_m$, which transforms each SVM classifier output into the posterior probability that a_m is prominent given input $\phi(y_{ij})$.

Platt scaling produces probability estimates by fitting a logistic transformation of classifier scores:

$$\mathcal{S}_m(w_m^T \phi(y_{ij})) = \frac{1}{1 + \exp(A_m w_m^T \phi(y_{ij}) + B_m)}. \tag{3.8}$$

The parameters A_m and B_m are learned using maximum likelihood estimation, optimized on the same training data as the original SVM classifier.

Now, given our set of M confidence predictors $\mathcal{P}_m$, we extract the most prominent

attribute $\mathcal{A}^{uv}$ for a novel image pair y_{uv} by choosing the attribute model with the highest outputted confidence level:

$$\mathcal{A}^{uv} = \arg\max_{m}(\mathcal{P}_m(y_{uv}))). \tag{3.9}$$

In addition, we can also return the top k prominent differences for an image pair using our model by sorting attribute confidence values and selecting the k attributes with the highest confidence scores. This can be used, for instance, to generate a textual description for a pair of images comparing their k most prominent differences.

Our model follows the structure of a one versus all multiclass classifier, with our relative attribute features $\phi(y_{uv})$ as input features and the most prominent attribute as input class labels. Other models could certainly be considered for predicting prominence, such as a one versus one multiclass classifier, which trains $\binom{M}{2}$ binary classifiers, one between each unique pair of classes, or a ranker model, which would return a full ranking of attributes by how prominent they are.

We experiment with both multiclass classification models, and chose the one versus all model for several reasons: its strong performance for prominence prediction, its easy interpretability for individual attributes as compared to a one versus one multiclass model, and its efficiency (only requiring one classifier per attribute in the vocabulary). We choose a classifier approach versus a ranker for its ease of collecting intuitive and natural human perceptions of "first impression" prominence (see the next section, Section 3.3), as opposed to exhaustive comparisons between all different pairs of attributes in the vocabulary, which is more cumbersome for humans and can lead to more noisy results.

3.3 Annotating Prominent Differences

In order to build a set of ground truth values for prominence training and evaluation, we collect human annotations of prominent differences for image pairs at the image pair instance level. These annotations are used as the target prominence labels during training, as well as the ground truth prominence labels to evaluate our approach. We collect annotations at a large scale using Amazon Mechanical Turk, a crowd-sourcing platform for workers to complete jobs, known as Human Intelligence Tasks or HITs.

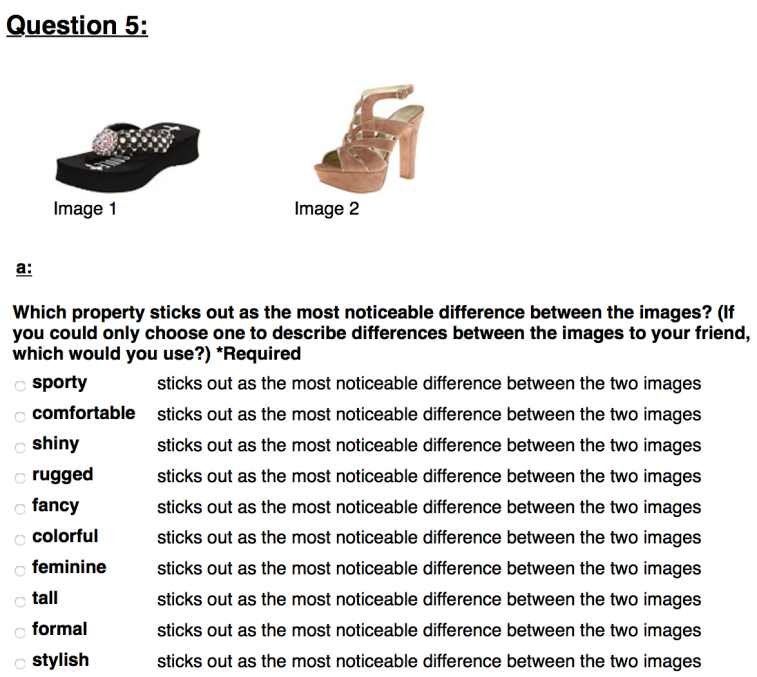


Figure 6: **Annotation Interface for Prominence** - We provide annotators with a pair of instance images, and ask for the single most noticeable difference between the two images. We provide the entire relative attribute vocabulary as choices.

To collect human perception of prominent differences, we show Mechanical Turk annotators a pair of randomly selected instance images, along with a list of all M attributes $\{a_m\}, m \in \{1, \ldots, M\}$. We ask the annotators which single attribute out of the list sticks out as being the most noticeable difference for that image pair.

To help annotators better understand the task at hand, we present the following situation as intuition: "Imagine that your friend cannot see the pair of images. You would like to tell your friend the most noticeable difference between the images." To moderate more thoughtful answers, we also ask annotators to justify in a short sentence why they chose their answer for a subset of questions. See Figure 6 for our annotation interface.

It is important to highlight that we ask each annotator to select *just one* attribute as prominent. This enables Mechanical Turk annotators to provide their first impression of the most noticeable difference when comparing and contrasting a new pair of images. Additionally, we provide the entire vocabulary of M attributes to choose from for every sample image pair, which aids in ensuring that at least a subset of

attribute choices are noticeably different for almost all image pairs.

Our annotation system has several strengths when compared to the format used by Turakhia and Parikh [48] for annotating binary attribute dominance. In our annotation task, we ask annotators to select one attribute out of all as prominent, which is more natural and intuitive than the pairwise attribute task used in their dominance work, where two arbitrary attributes are given and annotators must choose one of the two, even if neither are dominant. Finally, our annotation system is much more scalable: it requires only one annotation question per image pair, regardless of the number of attributes in the vocabulary, versus $\binom{M}{2}$ combinations of attribute pair questions required to annotate one instance of dominance. This allows us to collect and train prominent differences at the instance level, capturing fine-grained information on which specific images and features lead to prominence, whereas Turakhia and Parikh collect dominance at the category-level, projecting the same dominance strengths for all instance images in a category.

To obtain our prominent difference ground truth labels for image pairs, we first collect annotations from seven Mechanical Turk annotators for each pair in our sample. This gives us a set of prominent attribute choices. For each image pair, we order the attributes by their frequency chosen as prominent, creating a ranking of attributes for each image pair. We use the most frequently chosen attribute, i.e., highest ranked attribute r_{i1}, as the ground truth label for each input image pair to our model.

We experiment with two different datasets, the UT Zappos50K (UT-Zap50K) Shoes Dataset [50] and LFW10 Faces Dataset [43], and collect prominence annotations for both datasets. These datasets were designed for evaluating fine-grained recognition and comparison tasks, where two images of a similar type (e.g., two pairs of shoes, or two people's faces) are compared in terms of their relative attributes. We create a new relative attribute vocabulary of 10 attributes for UT-Zap50K using human annotation responses from [52], with the attributes (1) *sporty*, (2) *comfortable*, (3) *shiny*, (4) *rugged*, (5) *fancy*, (6) *colorful*, (7) *feminine*, (8) *tall*, (9) *formal*, (10) *stylish*, and use the existing relative attribute vocabulary of 10 attributes for LFW, (1) *bald head*, (2) *dark hair*, (3) *eyes open*, (4) *good looking*, (5) *masculine*, (6) *mouth open*, (7) *smiling*, (8) *visible teeth*, (9) *visible forehead*, (10) *young*. We select these relative attribute vocabularies for their size (10 attributes), so that there are a large variety of properties for making comparisons, and aim for as little overlap as possible between attributes, so that prominent differences can be more clear-cut. For more

detail on the datasets used, as well as training and testing splits, see Section 5.1.

We collect prominence data for $4,990$ sample image pairs for each of UT-Zap50K and LFW10, with seven annotators per pair, and transform these pairs into ground truth. In terms of annotator agreement, 77 percent of image pairs had three or more annotators agree on the top prominent difference, and 51 percent of pairs had four or more annotators agree on the top prominent difference for UT-Zap50K, with 87 percent and 53 percent for LFW-10, respectively. On average, 3.8 unique attributes were chosen as most noticeable for each image pair for UT-Zap50K, with 3.3 unique attributes chosen per pair for LFW10. This high level of agreement shows that prominent differences are in fact consistent for most image pair comparisons, and that people tend to agree on prominent differences when shown an image pair.

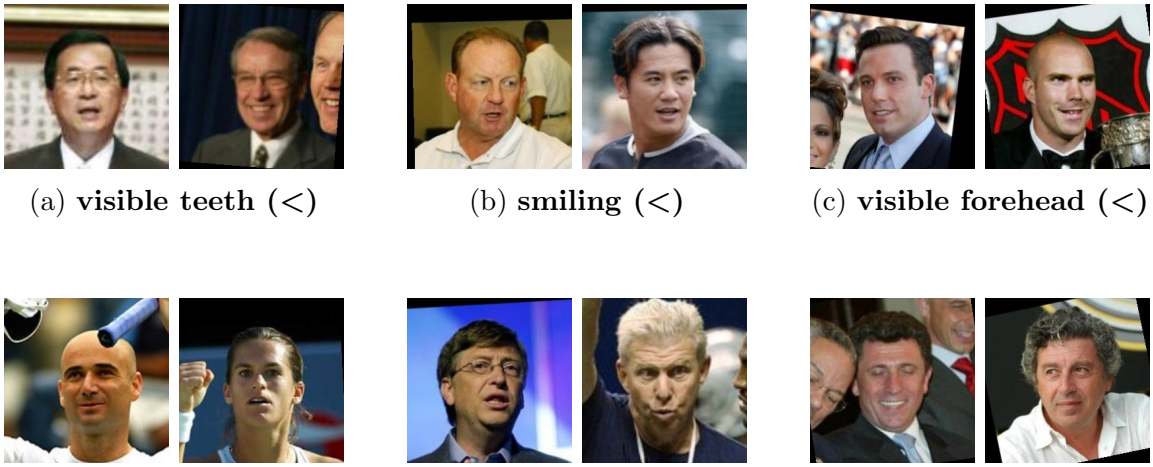
We illustrate examples of ground truth most prominent differences in Figure 7 and 8. From these examples, we observe different annotator-provided rationales for prominent differences, provided as common explanations in our data collection studies. For instance, Figure 8e's prominent difference is *dark hair* and Figure 7b's prominent difference is *tall* because annotators state the image pairs are most different in that attribute. In Figure 8f and Figure 7f, *eyes open* and *stylish* stand out to human judges, respectively, because of the unusualness of having eyes closed in a picture and for the unusual style of the high heel. Finally, *visible teeth* and *shiny* stand out for Figure 8a and Figure 7e even though those differences are not stark, because annotators stated that other attributes had less noticeable differences. These rationales show that our feature design for learning prominent differences is well motivated: interactions between all relative attributes present in a pair of images cause human perceptions of prominence.

We have now illustrated how we model and learn prominent differences using relative attribute features, as well as how we annotate prominent differences and transform human annotations into ground truth. Next, we will introduce our approaches for applying prominent differences on two vision tasks: image search and description generation.

(a) **sporty (>)** (b) **tall (<)** (c) **rugged (>)**

(d) **sporty (>)** (e) **shiny (<)** (f) **stylish (>)**

Figure 7: **Ground Truth Prominent Differences for UT-Zap50K** - Sample image pairs from the UT-Zap50K shoes dataset along with their ground truth most prominent attribute differences.



(a) **visible teeth (<)** (b) **smiling (<)** (c) **visible forehead (<)**

(d) **bald (>)** (e) **dark hair (>)** (f) **eyes open (<)**

Figure 8: **Ground Truth Prominent Differences for LFW10** - Sample image pairs are shown along with the most prominent attribute difference for each, determined as the annotators' most chosen attribute out of the vocabulary for that image pair.

4 Applications

We now illustrate our approaches for applying our prominent difference predictor on two human-centric applications. In Section 4.1, we present an application of prominence on interactive image search. In Section 4.2, we show use of prominent differences to generate textual comparisons of image pairs. For experimental details and results, see Section 5.4 and 5.5.

4.1 Image Search

We consider applying prominent differences to WhittleSearch [21, 22], an interactive image search framework that allows users to provide relative attribute feedback in the form of comparisons (e.g., I would like images that are *more formal* than reference image X) to refine search results.

WhittleSearch considers the scenario in which a user has a target image in mind (e.g., a specific shoe or specific image of a person), and would like to find that image, or ones similar to it, in a database of images. At each iteration of WhittleSearch, the user is shown a page of K top-ranked image results. The user selects some subset of reference images from the page, and specifies feedback using these reference images in the form "What I am looking for is *more/less* a_m than image x_{ref}," where a_m is an attribute name and x_{ref} is a reference image. The user is able to freely choose the reference images and attributes on which they wish to specify feedback.

User feedback from the current search iteration, along with feedback from all previous iterations, is then converted into C relative attribute constraints, where, for all feedback statements of the form "What I am looking for is *more* a_m than image

x_{ref}," a relevant image x_i should satisfy the constraint

$$r_m^i > r_m^{ref}, \tag{4.1}$$

and for all statements of the form "What I am looking for is *less* a_m than image x_{ref}," relevant image x_i should satisfy the constraint

$$r_m^i < r_m^{ref}, \tag{4.2}$$

where r_m^i and r_m^{ref} are relative attribute scores for attribute a_m predicted by the ranker $\mathcal{R}_m$ for the database image and reference image, respectively.

At the end of each iteration, WhittleSearch orders the images in the database according to how many of the C relative attribute constraints are satisfied by each image: the group of images that satisfy all C constraints appear first, followed by images that satisfy $C-1$, etc. It is important to note that in the method of [21], images within groups that satisfy the same number of constraints are ordered randomly. The top ranked K results from the new ordering are shown, corresponding to the group(s) satisfying the most constraints, marking the start of a new iteration. The user provides additional feedback to add to the set of constraints, "whittling away" images not meeting the user's requirements, until the target image is found. A probabilistic extension is given in [19].

When users provide feedback in the form of "What I am looking for is *more/less* a_m than image x_{ref}", it is likely they will provide the most prominent attribute differences between the chosen reference image x_{ref} and their mental target (Figure 9). Thus, images are more likely to be relevant if they are prominently different in attribute a_m with image x_{ref}, for all C user-specified feedback constraints. We model this in our approach by introducing a target probability term $p(x_i)$ for each database image x_i using prominence values:

$$p(x_i) \propto \prod_{c=1}^{C} \mathcal{P}_{m_c}((x_i, x_{ref_c})), \tag{4.3}$$

where $\mathcal{P}_{m_c}$ is our prominence predictor for attribute m_c, and attribute m_c and reference image ref_c are the constraint parameters from constraint c, for all constraints $c = 1, \ldots, C$.

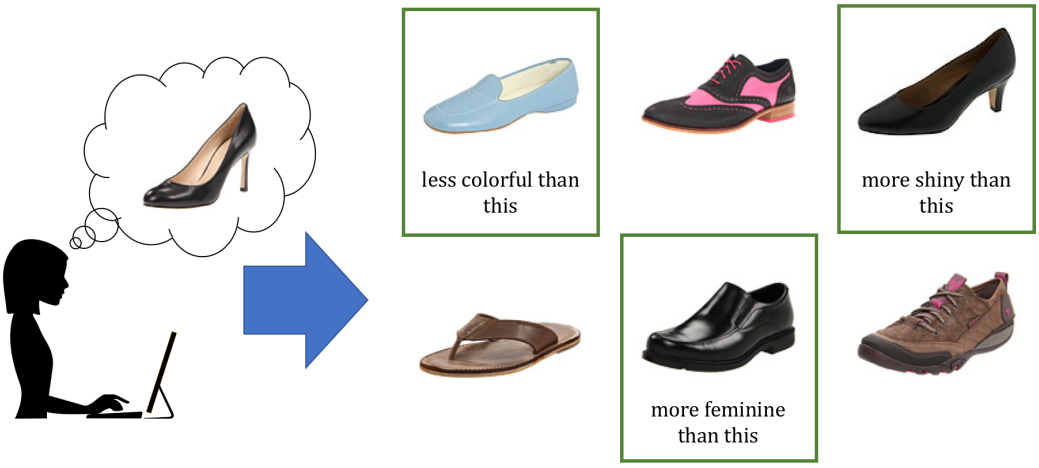


Figure 9: **WhittleSearch Relative Attribute Feedback** - In WhittleSearch [21, 22], a user has in their mind a target image or concept that they are searching for. The user is shown a page of results, and chooses reference images and feedback constraints from the page (shown in the boxes). Our hypothesis is that users will not randomly select different relative attributes as comparisons between reference images and their target; instead, they will likely provide *prominent differences* between the reference and their target as feedback. We leverage this in our approach.

We use $p(x_i)$ to rank the images *within each group of images satisfying N constraints*, by listing them in descending order of $p(x_i)$, with images satisfying prominence relationships best listed first in each group. We use this approach to maintain the overall constraint group ordering of WhittleSearch, while significantly improving the ordering of images within constraint brackets.

A significant strength of our approach is that it does not require any additional user input: we use simply use the attributes chosen in the existing feedback method. We hypothesize that this approach is especially impactful for reducing the rank of the target image in the first few iterations of WhittleSearch, when many images from the database satisfy all or most of the user's chosen feedback constraints.

4.2 Description Generation

For our second application, we consider applying prominent differences to generate descriptions of images with respect to each other. In particular, given a novel image pair, we would like to generate a textual description comparing the two images in terms of their relative attribute differences (e.g., Image X is *more sporty*, *less formal*, and *more colorful* than Image Y).

As explained in the introduction, when humans are asked to describe two images with respect to each other, such as in a compare and contrast task, they will likely state most prominent differences first. In addition, humans will not name all possible differences that are present between the images; instead, humans will focus on only a subset of the most noticeable differences in their expression.

Currently, relative attribute models can generate textual descriptions comparing two images in terms of all relative attributes in the vocabulary, in arbitrary order [38]. We argue that this is not natural or intuitive. With a large attribute vocabulary, listing out all differences generates descriptions that are too long and impractical for real-world use. To reduce the size of generated descriptions, [38] output a randomly chosen subset of k attributes to form the description. This approach can miss important differences and place less noticeable, or even irrelevant, differences first, producing less descriptive comparisons.

We propose generating descriptions using the k most prominent differences. Namely, given a novel image pair y_{uv}, we compute prominent difference confidence scores $\mathcal{P}_m(y_{uv})$ for all M attributes in the vocabulary. We then sort all attributes in de-

scending order of their prominence confidence scores. Using this ordering, we can generate a description with k comparison statements by using the top k attributes from the ranking. We can either state the prominent differences directly without relative attribute direction, or use predicted relative attribute scores or ground truth to state comparison directions. For example, given two shoe images, our model can generate the description "The left shoe is *more sporty*, *less stylish*, and *less shiny* than the right shoe," stating the three most prominent differences between the instance images.

We have thus illustrated our approaches for applying prominent differences to two applications, image search and description generation. We now present our experimental results, including an overview of the datasets used, experimental setup, and results of evaluating our approach on predicting prominent differences, followed by setup and results on image search and description generation.

5 Results

We first introduce the two image datasets we use, the annotations that we collect as ground truth, as well as the splits used during evaluation (Section 5.1). We then introduce the four baselines that we use to compare our approach: binary attribute dominance [48], widest relative attribute difference, single image prominence, and prior (Section 5.2). Finally, we show the evaluation of our approach on prominent difference prediction (Section 5.3), as well as our experimental setup and results on image search (Section 5.4) and description generation (Section 5.5).

5.1 Datasets

5.1.1 UT-Zap50K Shoes Dataset

The UT Zappos50K (UT-Zap50K) Dataset [50] is a dataset of 50,025 shoe catalog images and four relative attributes from Zappos.com. The shoe images are divided into four major types (Boots, Sandals, Shoes, Slippers), followed by subdivision into 19 functional categories (e.g., Ankle Boots, Mid-Calf Boots, Flat Sandals, Oxfords Shoes, Heels Shoes). Visually, the shoe images are centered, oriented, and placed on a white background. The UT-Zap50K dataset was created in the context of fine-grained recognition tasks such as online shopping and fine-grained comparisons, where users are comparing similar images, such as two pairs of high heels, in terms of relative attributes, making it well suited for prominence evaluation and our comparison-based applications.

The dataset comes with instance-level comparison labels for four relative attributes: however, to increase our vocabulary and variety for prominence prediction, we introduce a new vocabulary of ten relative attributes to conduct our experiments: (1) *sporty*, (2) *comfortable*, (3) *shiny*, (4) *rugged*, (5) *fancy*, (6) *colorful*, (7) *feminine*,

Figure 10: **Annotation Interface for Relative Attributes** - We collect relative attribute annotations for UT-Zap50K using our new attribute vocabulary to train and validate our relative attribute rankers. For LFW10, we use the relative attribute annotations provided in the dataset [43].

(8) *tall*, (9) *formal*, (10) *stylish* (see Figure 11). These attributes were selected from Amazon Mechanical Turk data collected by Yu and Grauman [52], in which users were presented with pairs of UT-Zap50K images and asked to provide the first adjective difference that comes to mind, filling in the sentence "Image A is a little more *adjective* than Image B." We select our 10 attributes out of the most frequently stated words by users, ensuring that no attributes are synonyms of each other, and that all are understandable to the average Mechanical Turk user. We use the 19 functional categories as described above as categories for the binary attribute dominance baseline [48], for a balance between number of categories and number of instance images per category.

For our experiments, we randomly sample 2,000 images from the dataset and use this subset for all subsequent data collection and sampling. We sample 4,990 pairs

randomly from our pool of 2,000 images, and collect prominent difference data for each pair. Each image pair is labeled by seven annotators, using the annotation approach in Section 3.3. This prominence data is used for training and as the ground truth for prominence evaluation.

In order to train and validate the relative attribute rankers, we sample 1,600 image pairs and collect relative attribute annotations for each of the 10 attributes in our vocabulary for each pair, for a total of 16,000 relative attribute annotations. For the relative attribute annotation task, we present users with an image pair and ask whether Image 1 has more or less of each attribute than Image 2 (see Figure 10). Each attribute is labeled by five annotators. We transform the raw relative attribute annotations into ground truth by labeling image pairs with only 3 people in agreement for an attribute label as ground truth equal, and image pairs with 4 or more people in agreement for an attribute label as ground truth more/less.

For models that use image features, such as the wide margin relative attribute ranker [38] and the binary attribute classifiers, we generate state-of-the-art CNN features from the fc7 fully-connected layer of AlexNet [23], a deep convolutional neural network trained on the ImageNet dataset, a large-scale database for visual object recognition. These 4096-dimension CNN features outperform the 960-dimension GIST and 30-dimension Lab color features given in the UT-Zap50K dataset in our experiments, so we report results for the CNN features only. For the deep CNN+STN relative attribute ranker, we use raw images as input, scaling and cropping using the method presented by Singh and Lee [45].

For model evaluation on prominence, we use 10-fold cross validation, splitting on individual images. In particular, for each split, we use 1,800 images for training and 200 for testing. For binary (single image) models, we train using all instances from the training set. For pairwise models such as prominent difference and relative attribute models, we train using image pairs contained within the training set and test with image pairs contained within the testing set. All models receive on average the same proportion of training instances to testing instances.

5.1.2 LFW10 Faces Dataset

The LFW10 Dataset [43] is a collection of 2000 images randomly selected from the Labeled Faces in the Wild dataset (LFW) [12], along with 10,000 relative attribute

Figure 11: **UT-Zap50K Dataset Attributes** - Visual examples of the ten different UT-Zap50K attributes used in our experiments. For each attribute, we show a sample image with less of the attribute on the left, more of the attribute on the right, and a median example in the middle.

image pairs collected over ten different attributes, (1) *bald head*, (2) *dark hair*, (3) *eyes open*, (4) *good looking*, (5) *masculine*, (6) *mouth open*, (7) *smiling*, (8) *visible teeth*, (9) *visible forehead*, (10) *young* (see Figure 12). Each relative attribute pair is annotated by five different annotators, who are given the task of answering whether or not one image is more/less/the same for a particular attribute compared to another image. The face images from Labeled Faces in the Wild are collected from the web and detected by the Viola-Jones face detector and aligned. The dataset images contain a variety of poses, backgrounds, lighting, and other difficult conditions.

For our experiments, we directly use the LFW10 relative attributes as our vocabulary. We create categories for the binary attribute dominance baseline [48] by matching the LFW10 images to their people labels, and use individual people as categories. We keep all individuals with three or more instance images in the dataset, resulting in a total set of 1,064 images belonging to 150 individual categories.

We sample 1,463 image pairs randomly from our image set and collect prominent difference data from each pair, with seven annotators per pair. We use the given relative attribute annotations in LFW10, consisting of 500 training and testing annotations per attribute. We discard this split, combining both training and testing pairs for LFW10 into one set, then selecting only the annotated pairs that are contained within our image set, resulting in a total of 2,675 image pairs over ten attributes. We transform the individual Mechanical Turk labels by choosing the majority chosen label of the five annotators as the ground truth for training and evaluation.

For image features for LFW10, we use the 8,300 dimension part-based features learned on dense SIFT [31] bag of words features, provided in [43]. These features isolate local regions of the face, and have been shown to significantly outperform global descriptor representations for relative attribute prediction. We reduce the dimensionality of these features to 200 using principal components analysis (PCA) to avoid overfitting. As with UT-Zap50K, for the deep CNN relative attribute ranker, we use scaled and cropped raw images as input. We evaluate prominence using 5-fold cross validation, splitting on individual images in the same method used for UT-Zap50K.

(a) ⟵ **bald head** ⟶ (b) ⟵ **dark hair** ⟶

(c) ⟵ **eyes open** ⟶ (d) ⟵ **good looking** ⟶

(e) ⟵ **masculine** ⟶ (f) ⟵ **mouth open** ⟶

(g) ⟵ **smiling** ⟶ (h) ⟵ **visible teeth** ⟶

(i) ⟵ **visible forehead** ⟶ (j) ⟵ **young** ⟶

Figure 12: **LFW10 Dataset Attributes** - Visual examples of the ten different LFW10 attributes used in our experiments. For each attribute, we show a sample image with less of each attribute on the left, more of the attribute on the right, and a median example in the middle.

5.2 Baselines

We now introduce the four baselines that we use to compare our approach: binary attribute dominance [48] (Section 5.2.1), widest relative attribute difference (Section 5.2.2), single image prominence (Section 5.2.3), and prior (Section 5.2.4).

5.2.1 Binary Attribute Dominance

Our first baseline is Turakhia and Parikh's binary attribute dominance model [48], as introduced in Related Work (Section 2.3).

To ensure a fair baseline, we follow the approach of Turakhia and Parikh [48] 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 for each dataset 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 [48]. For the UT-Zap50K shoes dataset, we use the functional categories as stated in Section 5.1.1, resulting in 19 categories (e.g., Ankle Boots, Flat Sandals, Sneakers and Athletic Shoes, Clogs and Mules) across a total of 2,000 images. For the LFW10 faces dataset, we use individual people as categories, with 150 categories across a total of 1,463 images. For each category, and for each possible combination pair of attributes, we ask annotators to choose which attribute pops out more (see Figure 13). Dominance ground truth, as defined by [48], is the number of annotators that selected the attribute when it appeared as one of the options for that category.

We show examples of ground truth attribute dominance from UT-Zap50K and LFW10 in Figure 14 and 15. For each category, we show the top 3 highest ranked dominant binary attributes. From these examples, we can see that attribute dom-

Figure 13: **Annotation Interface for Attribute Dominance [48]** - To gather data for training this baseline, we use the same interface and method as used by Turakhia and Parikh [48]. We show users a montage of images from a category, and a pair of binary attributes from the vocabulary. We then ask users which of the two presence/absence attributes stands out more.

(a) BoatShoes
stylish, comfortable, colorful

(b) OxfordsShoes
comfortable, formal, stylish

(c) KneeHighBoots
tall, colorful, feminine

(d) SlipperFlats
colorful, comfortable, fancy

(e) HeelsShoes
stylish, feminine, fancy

(f) LoafersShoes
comfortable, formal, stylish

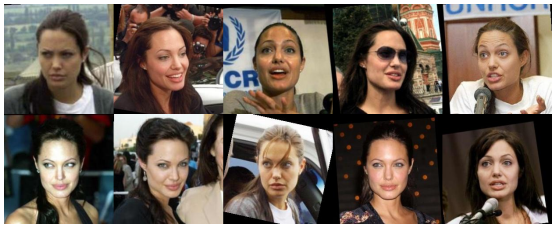
Figure 14: **Ground Truth Attribute Dominance [48] for UT-Zap50K** - Six categories out of the 19 total are shown with their Mechanical Turk annotation montages, along with the ranked top three dominant attributes for each category.

inance captures general attribute trends within a category, such as *formal* for the OxfordsShoes in 14b, *stylish* for the HeelsShoes in 14e, and *masculine* for the man in 15a. However, many attribute differences among instances in a category are lost by the generalization. Although the KneeHighBoots category in Figure 14c is labeled with *colorful* as dominant, only certain instance images within the category are colorful, and the others are quite dull. Although the woman in Figure 15d is labeled with *smiling* as the most dominant attribute, she is only smiling in some instances but not in others.

We follow the approach of Turakhia and Parikh [48] for training, projecting the category-level dominance ground truth to each training image in the split. We represent the images by their Platt scaled [40] binary attribute SVM classifier outputs. In

(a) **masculine, good looking, smiling**

(b) **not dark hair, smiling, not young**



(c) **good looking, dark hair, eyes open**

(d) **smiling, visible teeth, good looking**

(e) **eyes open, not bald head, dark hair**

(f) **smiling, dark hair, visible teeth**

Figure 15: **Ground Truth Attribute Dominance [48] for LFW10** - Six categories out of the 150 total are shown with their Mechanical Turk annotation montages, along with the ranked top three dominant attributes for each category.

our experiments, the dominance model is trained on all images in the training split, which usually results in it learning from all categories.

Note that the method of [48] 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 $y_{uv} = (x_u, x_v)$, we first compute binary attribute dominance values for each image in the pair, resulting in dominance values $d_u = (d_1^u, d_2^u, \ldots, d_M^u)$ for x_u and $d_v = (d_1^v, d_2^v, \ldots, d_M^v)$ for x_v. We select the attribute with the highest dominance value among both images $a_d^{uv} = \arg\max_m([d_u, d_v])$ as the predicted most prominent attribute difference for that pair. This method selects the attribute that sticks out as most dominant from either of the single images in the input pair.

5.2.2 Widest Relative Attribute Difference

For our second baseline, we consider using the widest relative attribute difference $\mathcal{W}^{uv}$ to predict prominent differences (cf. Section 3.2). To reiterate, given a novel pair of images $y_{uv} = (x_u, x_v)$, we use their relative attribute strengths $r^u = (r_1^u, r_2^u, \ldots, r_M^u)$ and $r^v = (r_1^v, r_2^v, \ldots, r_M^v)$ as predicted by relative attribute rankers, and select the attribute with the widest pairwise difference in strength $\mathcal{W}^{uv} = \arg\max_m(|r_m^u - r_m^v|)$.

We use the same relative attribute models, tuning parameters, and output scores as used by our prominent difference model. We tune the relative attribute models for strong performance on ordered relative attribute prediction, with 86% accuracy on UT-Zap50K and 80% accuracy on LFW10 using the SVM ranker, and 89% accuracy and 86% accuracy using the CNN ranker. We experiment with manipulating relative attribute scores using per-attribute standardization to a standard normal and normalization to $[0, 1]$ before widest difference computation for better comparisons between different attribute outputs, and select normalization for both rankers and datasets for its strongest performance on prominence evaluation.

5.2.3 Single Image Prominence

Our third baseline is a "single image prominence" model, which uses as ground truth the projection of prominent difference annotations onto both images in each

labeled pair. This model observes the same prominence ground truth as our proposed approach, but learns from individual images, without any pairwise knowledge.

We train a linear SVM for each attribute, using relative attribute scores of individual images as input features, and label an image as positive if it is part of an image pair that is ground truth prominent in the target attribute. We convert the outputs of each SVM into posterior probabilities using Platt's method [40], resulting in predicted confidence scores $p_m(x_i)$ for each attribute a_m.

Given a novel image pair (x_i, x_j), we predict the most prominent attribute using this baseline model by computing the predicted prominence confidence levels of each image in the pair for all attributes, and selecting the attribute with the highest posterior probability:

$$\arg\max_m(p_m(x_i), p_m(x_j)). \tag{5.1}$$

The model follows the general structure of a one versus all multiclass SVM classifier, but has overlap between different classes for certain data points, because certain individual images may be part of multiple image pairs with different prominent attribute ground truths, and will be labeled as prominent in multiple attributes.

5.2.4 Prior Frequency

Our final baseline is a simple "prior frequency" model, which predicts prominent differences proportionally according to their frequency of occurrence in the ground truth. For instance, if 20% of image pairs were labeled as most prominently different in *sporty* in the ground truth, the prior model will predict *sporty* as most prominent 20% of the time. This baseline performs stronger on average than a purely random model, and provides a reference from which to see improvement by other baselines and our approach.

5.3 Prominent Differences Evaluation

We evaluate prominent difference prediction using both datasets (UT-Zap50K and LFW10), for our prominence model and all baselines. As our main evaluation

measure, we consider the prediction accuracy of each model. We have each model predict a single, most prominent attribute difference for each image pair, using the approaches previously described.

Recall that seven annotators supply ground truth prominence on each image pair. Because there is not always a unanimous prominent attribute difference, we evaluate accuracy over a range of k maximum attributes marked as ground truth correct, to help account for variance in human perception. Specifically, we take the prominence annotations for each image and sort attributes by the number of times each was marked, creating a partial ranking of c attributes (due to many attributes not being selected for each image pair). We take the $min(k, c)$ top ranked attributes as ground truth prominent, and mark a pair as correctly predicted if the prediction $\mathcal{A}^{uv}$ is present in the ground truth. At $k = 1$, only the ground truth most prominent attribute is considered correct.

We show our accuracy results in Figure 16. We divide results for each dataset into two plots for clarity, one corresponding to each ranker (ranking SVM and deep CNN+STN) used to produce relative attribute scores. Between the two plots for both datasets, the prior and binary attribute dominance baselines are shared; results from single image prominence, widest relative difference, and our approach are different across the two plots because of the two relative attribute rankers employed (i.e., CNN for right plots; ranking SVM for left plots).

Overall, our approach significantly outperforms all baselines for prominence prediction. We observe sizable gains of roughly 20-22% on Zap50K, and 6-15% on LFW10 for prediction accuracy over the strongest baselines, across all sizes of ground truth. This clearly demonstrates the advantage of our approach, which uses pairwise relative attribute features to learn the complex interactions between attributes that result in prominent differences.

Our results show that the baselines of widest relative difference, binary attribute dominance, and single image prominence are not enough to predict prominent differences. In the case of widest relative difference, its lower accuracy compared to our approach demonstrates that the widest difference in attribute strength is only one contributing factor for prominent differences: our approach is able to capture other interactions between the visual properties of the images, and therefore predict prominence more accurately.

We also outperform the binary attribute dominance model of Turakhia and Parikh

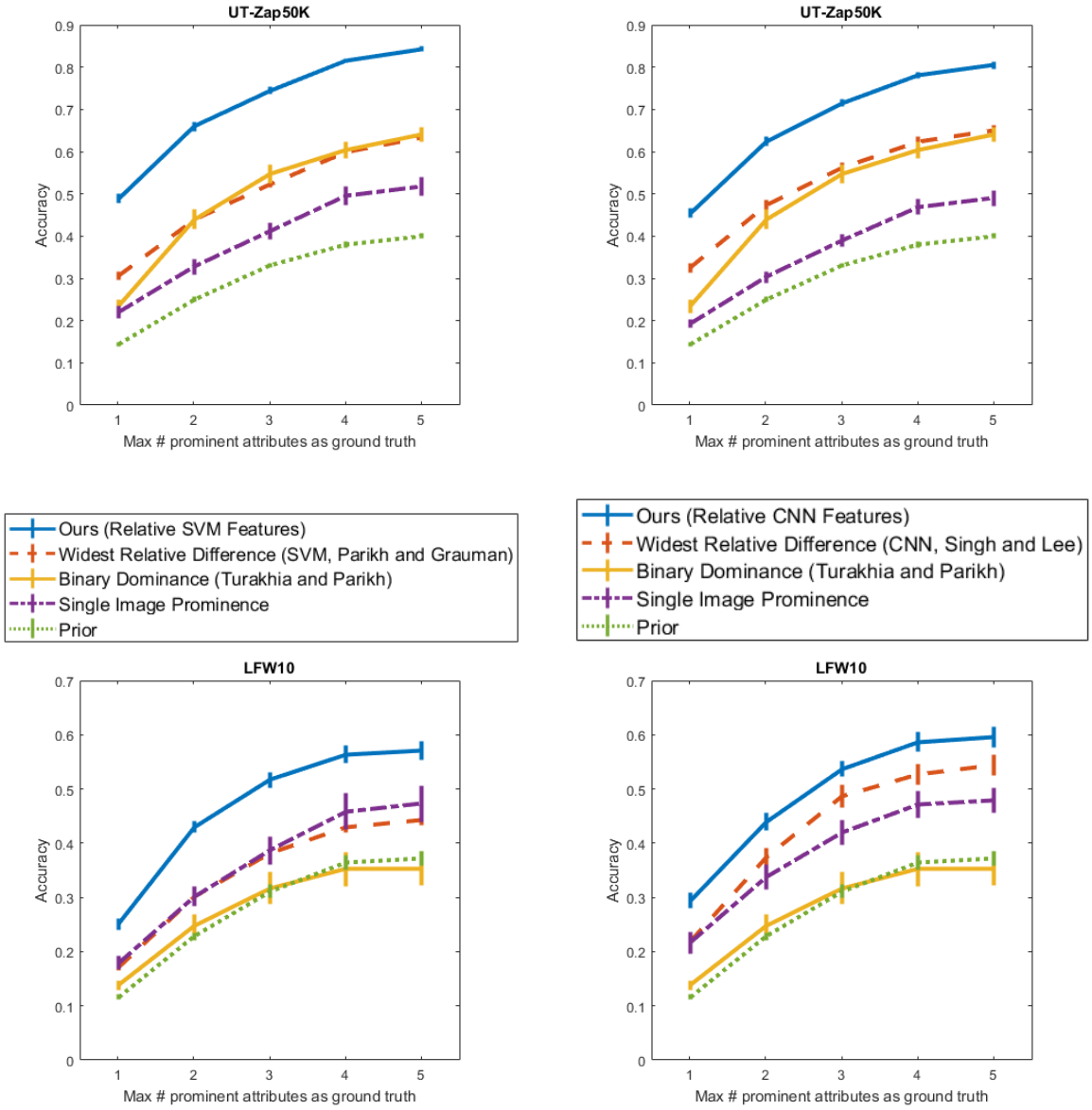


Figure 16: **Prominence Evaluation Accuracy** - Prominence prediction accuracy results for our model and baselines. UT-Zap50K shown on top, LFW10 shown on bottom, with ranking SVM relative attribute scores used on the left and CNN relative attribute scores used on the right. Our model significantly outperforms all baselines.

[48] by a significant margin: attribute dominance performs roughly in line with the widest relative difference for UT-Zap50K, but has very poor performance on LFW10, in line with the Prior baseline. We infer that the binary dominance model's better performance on UT-Zap50K is due to the more homogeneous shoe categories in terms of binary attributes; for instance, most high heels are *stylish* and *formal*, most boots are *rugged*, and most athletic shoes and sneakers are *sporty*. Thus, binary dominance data is more consistent for these categories, and, when projected onto individual images, has greater success. With LFW10, where the binary dominance model learns from categories of specific individuals, many of the dominant attributes in LFW10 differ within images in one group. For instance, although Arnold Schwarzenegger would be referred to as *masculine* in most of his images, it is not possible for a categorical dominance model to accurately learn attributes such as *smiling*, *eyes open*, *visible teeth*, *mouth open*, or *young* just from a collage of images with various expressions.

Based on the performance of our third baseline, the single image prominence model, we demonstrate that prominent differences are a pairwise phenomenon and must be modeled as the relationship between two images, instead of per each image in the pair. Comparisons between images require both images as context; modeling using single images is not sufficient.

Comparing the CNN relative attribute ranker scores to the ranking SVM scores, our approach achieves similar performance on UT-Zap50K but benefits from the CNN ranker scores for LFW10. The widest relative difference baseline performs slightly stronger with CNN ranker scores compared to ranking SVM scores, but still performs significantly below our approach. It is important to highlight that our idea and contributions are orthogonal to the choice of relative attribute ranking model: our approach can learn from relative attribute scores generated from any ranker model.

For a secondary evaluation of prominence, we show individual attribute average precision (AP) for our model compared to the three baselines, in Tables 5.1, 5.2, 5.3, and 5.4 (excluding the Prior baseline, because it does not output confidence scores per attribute). We generate precision-recall curves for each attribute by marking all image pairs with that attribute in the top 3 annotated prominent differences as ground truth positive, and all other image pairs as ground truth negative. Since prominent differences are not a per-attribute phenomenon, and are thus more suited for full-model accuracy evaluation as shown in Figure 16, we show these AP values as a supplement to the accuracy figures, to reveal more about the per-attribute parts of our

	Sporty	Comfort	Shiny	Rugged	Fancy	Color.	Femin.	Tall	Formal	Stylish	All Attributes
Single	55.72	43.01	23.29	18.02	27.00	43.19	27.63	36.83	18.28	22.67	31.56 ± 3.97
Dominance [48]	57.44	41.09	30.09	27.68	33.33	49.28	36.21	52.00	24.46	24.86	37.64 ± 3.74
Widest RA	60.44	37.55	29.48	25.55	32.29	57.37	49.41	71.23	26.26	21.57	41.11 ± 5.46
Ours	**76.29**	**59.87**	**49.60**	**45.01**	**40.18**	**73.34**	**54.05**	**80.59**	**39.28**	**33.74**	**55.19 ± 5.29**

Table 5.1: **Average Precision for UT-Zap50K, SVM Models**

	Sporty	Comfort	Shiny	Rugged	Fancy	Color.	Femin.	Tall	Formal	Stylish	All Attributes
Single	34.97	40.71	22.41	19.35	26.66	42.04	26.97	37.34	18.03	21.43	28.99 ± 2.86
Dominance [48]	57.44	41.09	30.09	27.68	**33.33**	49.28	36.21	52.00	24.46	**24.86**	37.64 ± 3.74
Widest RA	59.15	38.75	26.41	23.20	32.10	58.88	43.61	68.35	21.11	21.66	39.32 ± 5.54
Ours	**75.59**	**57.22**	**45.48**	**34.79**	31.03	**72.12**	**52.23**	**78.62**	**26.80**	24.58	**49.85 ± 6.51**

Table 5.2: **Average Precision for UT-Zap50K, CNN Models**

	Bald	DarkHair	EyeOpen	Looks	Mascul.	Mouth	Smile	Teeth	Foreh.	Young	All Attributes
Single	24.66	38.79	12.17	23.46	36.90	38.59	44.97	40.95	12.16	19.59	29.22 ± 5.48
Dominance [48]	24.35	36.55	12.93	25.66	27.58	39.25	52.53	38.28	10.68	18.90	28.68 ± 5.83
Widest RA	29.57	42.19	13.13	23.70	49.25	36.61	54.24	**50.31**	**15.07**	22.64	33.68 ± 6.72
Ours	**34.24**	**42.43**	**14.83**	**26.19**	**52.01**	**47.54**	**55.71**	48.16	12.82	**23.98**	**35.79 ± 6.99**

Table 5.3: **Average Precision for LFW10, SVM Models**

	Bald	DarkHair	EyeOpen	Looks	Mascul.	Mouth	Smile	Teeth	Foreh.	Young	All Attributes
Single	**43.48**	42.21	13.83	22.85	45.83	40.02	48.04	44.68	11.44	19.50	33.19 ± 6.48
Dominance [48]	24.35	36.55	12.94	**25.66**	27.58	39.25	52.53	38.29	10.69	18.90	28.68 ± 5.83
Widest RA	41.05	51.52	18.30	24.78	**60.31**	40.12	**58.73**	56.41	**18.05**	27.95	**39.72 ± 7.45**
Ours	43.66	**52.35**	**18.49**	23.33	55.69	**44.78**	57.64	**57.91**	14.12	**28.36**	39.48 ± 7.67

Table 5.4: **Average Precision for LFW10, CNN Models**

model and baselines. Results show that our model has higher average precision over all attributes for UT-Zap50K and both relative attribute rankers, as well as LFW10 with the ranking SVM relative attribute ranker. We obtain similar AP values to the widest relative attribute-difference baseline using the deep CNN relative attribute ranker.

We hypothesize that widest attribute difference works well in predicting prominence for certain individual attributes, such as *masculine*, *good looking*, or *visible forehead*, because these attributes tend to be prominent when they show a very wide difference in strength. In addition, we hypothesize that our model's weaker performance on certain attributes are due to these attributes rarely being marked as prominent in the ground truth: thus, our predictor's individual binary classifiers $\mathcal{P}_m$ for these rarely prominent attributes see very few positive examples and likely output a noisier, low confidence value. However, in the context of our full model and goal, *to predict the most prominent difference in an image pair*, low confidence scores from attributes that are rarely prominent are not a large issue, because these attributes

should naturally have lower confidence scores than prominent attributes for a large majority of pairs.

Finally, we show qualitative examples of strong prominent differences as predicted by our approach, in Figure 17. We show the model's predicted most prominent attribute, along with its relative attribute direction for reference, and also report the two "runner-up" attributes with the next highest confidence values. From these results, we can observe the capability of our model to accurately predict prominence in images with large numbers of complex differences, as well as similar images with few differences. For example, the shoes in Figure 17a are very different in many attributes; despite these difficulties, our model accurately predicts *colorful* as the most prominent difference. Although the images in Figure 17o are of the same person with a very similar expression, our model is able to accurately predict *visible teeth* as the most prominent difference.

In Figure 18, we show weak predictions and mistakes made by our model; our model's prediction is shown in bold, and the ground truth top prominent attributes are shown in parentheses. In Figure 18b, our model mistakenly identifies *tall* as most prominent, whereas more human annotators perceived *comfortable* as the most prominent difference over tall. In Figure 18f, our model identifies *visible teeth* as the prominent difference; although the two individuals certainly differ in that attribute, what sticks out to humans is their difference in *mouth open*.

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

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

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

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

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

(f) **feminine (>)**,
comfortable, shiny

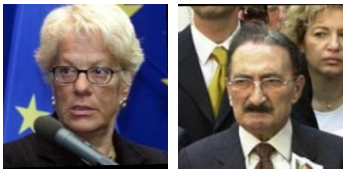
(g) **colorful (>)**,
sporty, comfortable

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

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

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

(k) **bald head (<)**,
dark hair, visible teeth



(l) **dark hair (<)**,
mouth open, smiling

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

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

(o) **visible teeth (>)**,
mouth open, smiling

Figure 17: **Sample Prominent Difference Prediction, Strong Results** - Sample image pairs from UT-Zap50K and LFW10 showing strong prediction results that agree with human annotated ground truth. The predicted most prominent attribute and its relative direction is shown in bold; the next two strongest attributes are also shown for each pair.

(a) **sporty (>)**
(feminine, shiny, comfortable)

(b) **tall (<)**
(comfortable, colorful, tall)

(c) **sporty (>)**
(comfortable, stylish)

(d) **tall (>)**
(feminine, comfortable, rugged)

(e) **masculine (<)**
(dark hair, good looking, young)

(f) **visible teeth (>)**
(mouth open)

(g) **dark hair (>)**
(bald, mouth open, visible forehead)

(h) **smiling (>)**
(visible teeth, young)

Figure 18: **Sample Prominent Difference Prediction, Failure Cases** - Sample image pairs from UT-Zap50K and LFW10 showing weak prediction results made by our method. The predicted most prominent attribute is shown in bold. Ranked ground truth prominent differences are shown in parentheses.

5.4 Image Search

For our image search application on WhittleSearch [21, 22], we evaluate a proof-of-concept experiment using the UT-Zap50K shoes dataset. We use the UT-Zap50K shoes dataset for its large size, which is suitable as a simulated database. We sample 5,000 new images outside of our experimental sample from Zap50K, and use this as our image database for the search experiment.

Due to the size and cost of obtaining human feedback for each possible pair of images, we generate user search feedback automatically using a modified version of the automatic feedback approach described in [21]. A random subset of images from the top results page are chosen as reference images. For the user's feedback between the target x_t and each reference image x_{ref}, the user selects the most prominent difference $\mathcal{A}_{t,ref}$ to provide feedback upon. To simulate variance in human perception, we add noise by randomly selecting 75% of user-specified feedback using the prominent difference method, and 25% as random differences from the attribute vocabulary.

For our experiment, we select 200 random images as the user's mental targets. At each iteration, the user is shown the top 16 results as ranked by the search algorithm, selects 8 images from the results as reference images, and provides 8 feedback constraints using these references.

In Figure 19, we show the average target image ranking (lower is better) for each iteration of WhittleSearch, between the baseline WhittleSearch ranking implementation [21] and our proposed prominence ranking approach. From these results, we observe that our proposed approach substantially improves the rank of the target image in the first few iterations of WhittleSearch. In these first iterations, many database images satisfy most of the user-specified constraints, leading to random orderings within constraint groups for the baseline; our prominence ranking method is able to intelligently order these groups and find the user's target in fewer total iterations, using the same feedback method as the baseline.

We also display examples of top search results produced by our model and the baseline in Figure 20, after two feedback iterations with four feedback constraints each iteration. From these qualitative examples, we can observe that our model produces more relevant top results. In Figure 20a, where the user's target image is a dark, formal flat-style shoe, our approach returns shoes that are a similar style and color to the user's target as top results, whereas the baseline returns an array of less

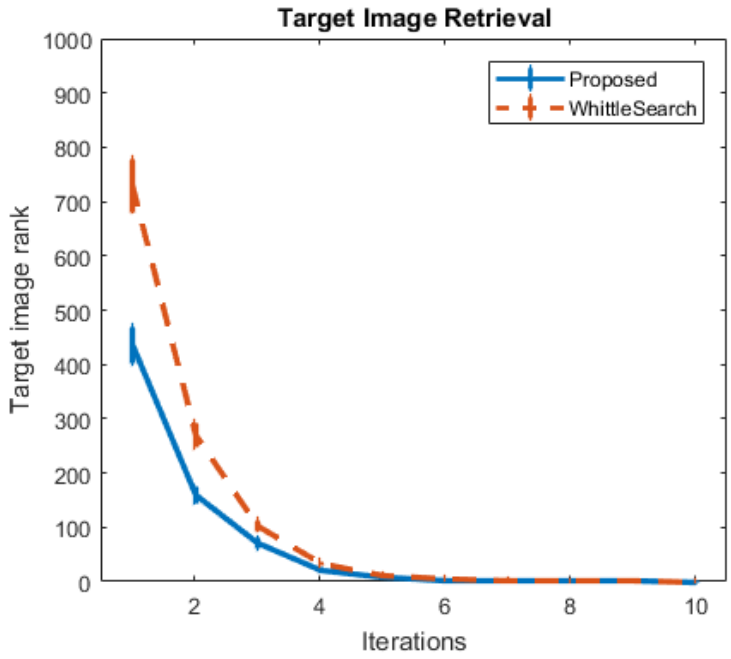


Figure 19: **Image Search Results** - We show the target image rank over multiple iterations of WhittleSearch [21, 22], for both our proposed approach and the standard WhittleSearch baseline. Our approach significantly lowers the ranking of the target image in the first iterations of search, and finds the target image in fewer total iterations.

relevant shoes of different styles. In Figure 20b, our approach finds colorful and casual sneakers similar in concept to the user target, whereas the baseline returns various boots, clogs, and formal shoes alongside some sneakers. By predicting which images are most prominently perceived as different when compared to the user's selected reference images and attributes, the search is able to deliver more relevant images without requiring additional user feedback.

(a)

(b)

Figure 20: **Qualitative WhittleSearch Rankings** - We show the target image along with the top eight ranked images produced by the baseline WhittleSearch [21] and our prominence approach, with both methods receiving two feedback iterations with four identical feedback constraints each. Our approach brings more relevant images to the top results by using prominent differences, without requiring any additional user input.

5.5 Description Generation

We evaluate the comparison descriptions generated by our model in one offline experiment and one online experiment. For our offline experiment, we have our model and baselines output the top k most prominent attributes that would be present in a description, and check what percentage of the k ground truth prominent attributes are described by our generated descriptions. We compare our approach to three of the baselines we use for prominence evaluation: widest relative difference, binary dominance [48], and single image prominence. We report our results in Figure 21. Our model outperforms all baselines, demonstrating that our predictor is able to generate descriptions with more stated prominent differences.

For our online experiment, we ask annotators on Mechanical Turk to judge our generated descriptions. Specifically, we present two descriptions to the annotator: our generated description with predicted most prominent attribute differences, and a baseline description with randomly chosen attribute differences selected among all true attributes. We ask annotators to select the description that is most natural and appropriate for the image pair. We sample 200 image pairs from UT-Zap50K and 100 image pairs from LFW10, all unseen by our model, generate descriptions with three stated differences each, and have seven Mechanical Turk annotators provide their feedback for each image pair. We take the majority vote of the seven annotators for each image pair.

For UT-Zap50K, 69% of people preferred our description, compared to 31% for the baseline random description, with a p-value < 0.0001, while for LFW10, 61% of people preferred our description, compared to 39% for the baseline random description, with a p-value of 0.01. We also ran the same experiment, using annotator ground truth prominence rankings instead of our prominence predictor: people preferred the ground truth description 69% of the time for UT-Zap50K and 70% of the time for LFW10. See Table 5.5 for a summary of these study results. We hypothesize the descriptions generated by our approach for UT-Zap50K are more preferred and closer to the ground truth due to the higher prediction accuracy of the UT-Zap50K prominence model compared to the LFW10 prominence model.

In Figure 22, we show qualitative examples of descriptions generated by our approach and the random baseline, with the first three rows as success cases (i.e., a majority of human judges prefer our description), and the last row as failure cases

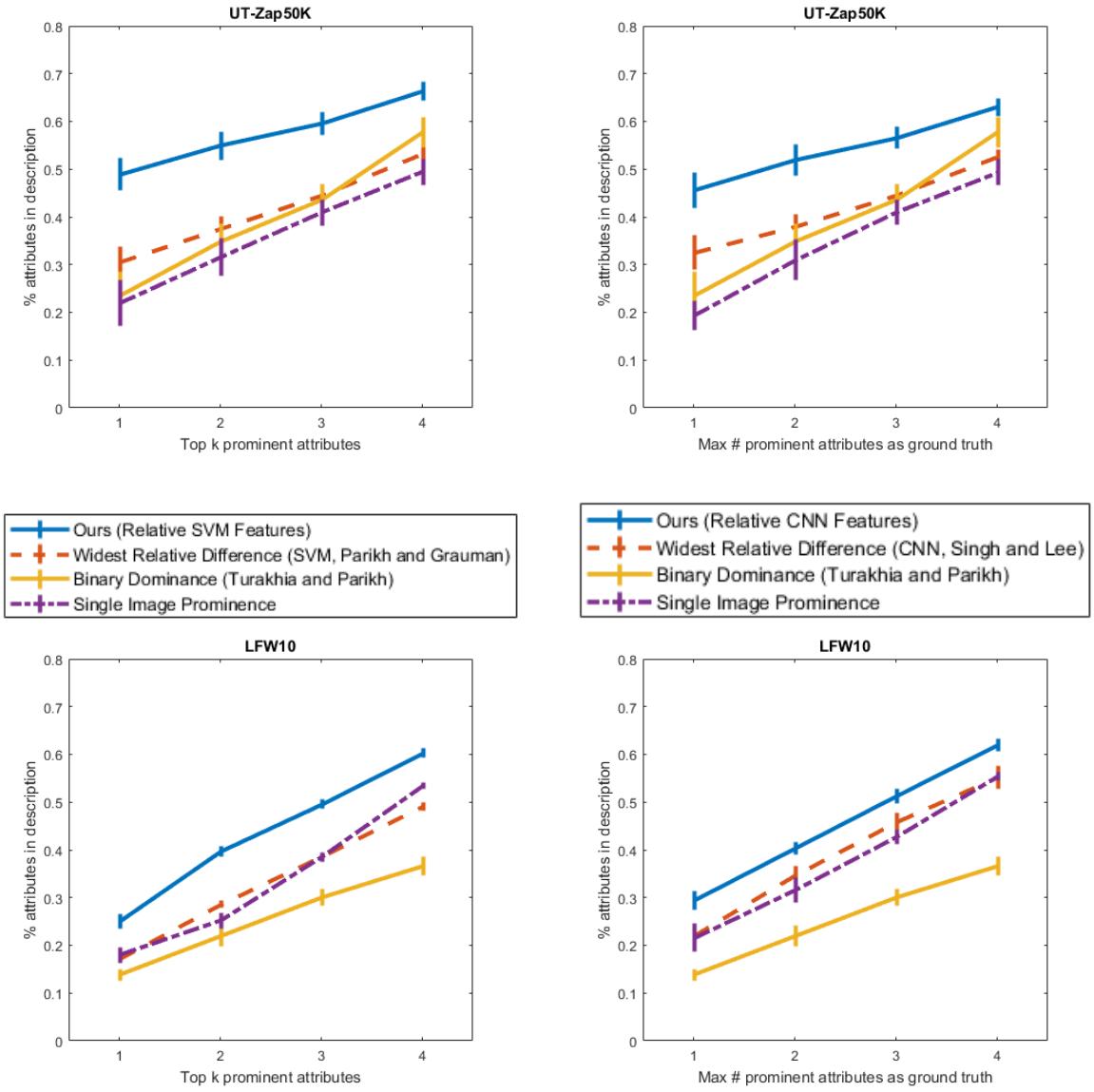


Figure 21: **Description Generation Accuracy** - We show the offline accuracy of our approach on describing images with the k most prominent attributes, using prominence annotations as ground truth. UT-Zap50K is shown on top, LFW10 is shown on bottom, with ranking SVM relative attribute scores used on the left and CNN relative attribute scores used on the right. Our approach outperforms all baselines.

UT-Zap50K	Ours:	69%	Baseline:	31%
	Ground Truth:	69%	Baseline:	31%
LFW10	Ours:	61%	Baseline:	39%
	Ground Truth:	70%	Baseline:	30%

Table 5.5: **Description Generation Study Results** - Results of the online experiment, where we show our generated description and the baseline description to human judges, and ask which is more natural and appropriate. We also conduct an experiment with the ground truth prominent differences compared to the baseline. Human judges significantly prefer our generated descriptions over the baseline.

(i.e., a majority of human judges prefer the baseline description). Our generated descriptions are generally more natural than the random baseline, because the descriptions focus on the prominent differences that would be natural and appropriate in human descriptions.

These sets of results, both offline and online, show that describing images using prominent differences results in significantly more natural descriptions. With more accurate predictors of prominent differences and a larger attribute vocabulary, even stronger description results should be obtained using our description generation approach.

(a) **Left is more tall, less sporty, and less rugged than the right.**
(less colorful, more shiny, more feminine)

(b) **Left is less shiny, less formal, and more colorful than the right.**
(more feminine, more rugged, more tall)

(c) **Left is more colorful, more sporty, and less rugged than the right.**
(more fancy, less rugged, more stylish)

(d) **Left is less feminine, more rugged, and less shiny than the right.**
(less stylish, more comfortable, more rugged)

(e) **Left has less dark hair, more bald head, and more mouth open than the right.**
(more good looking, more mouth open, less dark hair)

(f) **Left is more masculine, less smiling, and less visible teeth than the right.**
(more bald head, less good looking, less young)

(g) **Left is less colorful, less comfortable, and more sporty than the right.**
(more shiny, more fancy, more formal)

(h) **Left is more masculine, less smiling, and more visible teeth than the right.**
(less smiling, less young, more visible forehead)

Figure 22: **Sample Textual Descriptions** - Sample descriptions generated by our approach in bold, with baseline result shown in parentheses. First three rows display success cases, where annotators chose our description as more natural, with the last row displaying failure cases, where the baseline was chosen over our approach.

6 Conclusion and Future Work

In this work, we introduce and model prominent differences in relative attributes, a novel high-level functionality for comparing images. When humans describe images with respect to each other, certain prominent differences in attributes naturally stick out and are likely to be described first, while other differences, although present, may not be mentioned. We present a novel approach for modeling prominent differences at the image pair level, using relative attribute features to capture the interactions between visual properties that result in prominent differences. Experimental results on the UT-Zap50K shoes and LFW10 faces datasets show that our proposed approach significantly outperforms an array of baseline methods for predicting prominence. In addition, we demonstrate how prominent differences as predicted by our model can be used to improve communication between humans and vision systems in two applications: interactive image search and textual description generation.

There is strong potential for future work using prominence. Prominent differences are naturally expressed by humans when describing different visual concepts; this can be used to improve other human-centric vision tasks. For instance, in zero-shot learning using relative relationships [38], where a human supervisor teaches a machine about an unseen image category using relative attributes, humans will likely provide prominent differences to the machine. This information, if modeled in a zero-shot learning framework, could result in improved classification without requiring any additional human supervision effort. In addition, prominent differences could be used to improve referring expressions [33, 34]. Referring expressions are phrases identifying specific objects in an image, e.g., "The man standing on the right, with the white shirt and long hair." Prominent differences could be used to better identify most noticeable visual differences to help identify one object over others. Finally, prominent similarities in relative attributes can be explored, as the similarities between images that stand out as most noticeable. Prominent similarities may be beneficial for fine-

grained image clustering according to natural human perception of visual properties, and, along with prominent difference and visual importance models, could help paint a fuller picture on what high-level properties and objects humans perceive in visual content.

7 Bibliography

[1] A. Berg, T. Berg, H. Daume, J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, A. Sood, K. Stratos, and K. Yamaguchi. "Understanding and Predicting Importance in Images". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2012.

[2] T. Berg, A. Berg, and J. Shih. "Automatic Attribute Discovery and Characterization from Noisy Web Data". In: *European Conference on Computer Vision (ECCV)*. 2010.

[3] S. Branson, C. Wah, F. Schroff, B. Babenko, P. Welinder, P. Perona, and S. Belongie. "Visual Recognition with Humans in the Loop". In: *European Conference on Computer Vision (ECCV)*. 2010.

[4] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. "Learning to Rank using Gradient Descent". In: *ICML*. 2005.

[5] H. Chen, A. Gallagher, and B. Girod. "Describing Clothing by Semantic Attributes". In: *European Conference on Computer Vision (ECCV)*. 2012.

[6] A. Deza and D. Parikh. "Understanding Image Virality". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015.

[7] S. Dhar, V. Ordonez, and T. L. Berg. "High Level Describable Attributes for Predicting Aesthetics and Interestingness". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2011.

[8] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. "Describing Objects by their Attributes". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2009.

[9] Y. Fu, T. M. Hospedales, T. Xiang, J. Xiong, S. Gong, Y. Wang, and Y. Yao. "Robust Subjective Visual Property Prediction from Crowdsourced Pairwise Labels". In: *CoRR* abs/1501.06202 (2015).

[10] T. Hernandez. *Compare and Contrast Poster*. Accessed: 2017-04-10. 2014. URL: `http://www.teachersnotebook.com/product/theresasmith24/compare-amp-contrast-poster`.

[11] C. Huang, C. C. Loy, and X. Tang. "Unsupervised Learning of Discriminative Attributes and Visual Representations". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[12] 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*. Tech. rep. 07-49. University of Massachusetts, Amherst, Oct. 2007.

[13] L. Itti, C. Koch, and E. Niebur. "A Model of Saliency-based Visual Attention for Rapid Scene Analysis". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20.11 (Nov. 1998), pp. 1254–1259.

[14] D. Jayaraman and K. Grauman. "Zero-Shot Recognition with Unreliable Attributes". In: *Proceedings of Advances in Neural Processing Systems (NIPS)*. 2014.

[15] T. Joachims. "Optimizing Search Engines Using Clickthrough Data". In: *SIGKDD*. 2002.

[16] T. Judd, K. Ehinger, F. Durand, and A. Torralba. "Learning to Predict Where Humans Look". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2009.

[17] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. "ReferIt Game: Referring to Objects in Photographs of Natural Scenes". In: *EMNLP*. 2014.

[18] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes. "Deep Understanding of Image Aesthetics". In: *ECCV*. 2016.

[19] A. Kovashka and K. Grauman. "Attribute Pivots for Guiding Relevance Feedback in Image Search". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Dec. 2013.

[20] A. Kovashka and K. Grauman. “Discovering Attribute Shades of Meaning with the Crowd”. In: *International Journal of Computer Vision (IJCV)* 114.1 (Aug. 2015), pp. 56–73.

[21] A. Kovashka, D. Parikh, and K. Grauman. “WhittleSearch: Image Search with Relative Attribute Feedback”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2012.

[22] A. Kovashka, D. Parikh, and K. Grauman. “WhittleSearch: Interactive Image Search with Relative Attribute Feedback”. In: *International Journal of Computer Vision (IJCV)* 115.2 (Nov. 2015), pp. 185–210.

[23] A. Krizhevsky, I. Sutskever, and G. Hinton. “ImageNet Classification with Deep Convolutional Neural Networks”. In: *NIPS*. 2012, pp. 1097–1105.

[24] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. “Baby Talk: Understanding and Generating Image Descriptions”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2011.

[25] N. Kumar, P. Belhumeur, and S. Nayar. “FaceTracer: A Search Engine for Large Collections of Images with Faces”. In: *European Conference on Computer Vision (ECCV)*. Oct. 2008, pp. 340–353.

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

[27] S. Lad and D. Parikh. “Interactively Guiding Semi-Supervised Clustering via Attribute-based Explanations”. In: *European Conference on Computer Vision (ECCV)*. 2014.

[28] C. Lampert, H. Nickisch, and S. Harmeling. “Learning To Detect Unseen Object Classes by Between-Class Attribute Transfer”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2009.

[29] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. “DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[30] Z. Liu, P. Luo, X. Wang, and X. Tang. "Deep Learning Face Attributes in the Wild". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2015.

[31] D. Lowe. "Distinctive Image Features from Scale-Invariant Keypoints". In: *International Journal of Computer Vision (IJCV)* 60.2 (Nov. 2004), pp. 91–110.

[32] S. Maji and G. Shakhnarovich. "Part and Attribute Discovery from Relative Annotations". In: *International Journal of Computer Vision (IJCV)* 108.1 (May 2014), pp. 82–96.

[33] J. Mao, J. Huang, A. Toshev, O. Camburu, A. Yuille, and K. Murphy. "Generation and Comprehension of Unambiguous Object Descriptions". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[34] M. Mitchell, K. Deemter, and E. Reiter. "Generating Expressions that Refer to Visible Objects". In: *North American Chapter of the Asssociation for Computational Linguistics (NAACL)*. 2013.

[35] T. Nudd. *Apple's 'Get a Mac,' the Complete Campaign*. Ed. by Adweek. Accessed: 2017-04-13. 2011. URL: `http://www.adweek.com/creativity/apples-get-mac-complete-campaign-130552`.

[36] A. Oliva and A. Torralba. "Modeling the shape of the scene: A holistic representation of the spatial envelope". In: *International Journal of Computer Vision (IJCV)* 42.3 (2001), pp. 145–175.

[37] Junting Pan, Elisa Sayrol, Xavier Giro-i-Nieto, Kevin McGuinness, and Noel E. O'Connor. "Shallow and Deep Convolutional Networks for Saliency Prediction". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2016.

[38] D. Parikh and K. Grauman. "Relative Attributes". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2011, pp. 502–510.

[39] G. Patterson and J. Hays. "SUN Attribute Database: Discovering, Annotating, and Recognizing Scene Attributes". In: *International Journal of Computer Vision (IJCV)* 108.1 (May 2014), pp. 59–81.

[40] John C. Platt. "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods". In: *Advances in Large Margin Classifiers*. MIT Press, 1999, pp. 61–74.

[41] J. R. Raphael. *AT&T's Verizon Ad Battle: Who's Being Hurt Worse?* Ed. by PCWorld. Accessed: 2017-04-13. 2009. URL: `http://www.pcworld.com/article/182185/ATTs_Verizon_Ad_Battle_Whos_Being_Hurt_Worse.html`.

[42] A. Sadovnik, A. Gallagher, D. Parikh, and T. Chen. "Spoken Attributes: Mixing Binary and Relative Attributes to Say the Right Thing". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Dec. 2013.

[43] R. Sandeep, Y. Verma, and C. Jawahar. "Relative Parts: Distinctive Parts for Learning Relative Attributes". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014.

[44] B. Siddiquie, R. S. Feris, and L. S. Davis. "Image Ranking and Retrieval based on Multi-Attribute Queries". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2011.

[45] K. Singh and Y. J. Lee. "End-to-End Localization and Ranking for Relative Attributes". In: *European Conference on Computer Vision (ECCV)*. 2016.

[46] Y. Souri, E. Noury, and E. Adeli. "Deep Relative Attributes". In: *ACCV*. 2016.

[47] M. Spain and P. Perona. "Measuring and Predicting Object Importance". In: *International Journal of Computer Vision (IJCV)* 91.1 (Aug. 2011), pp. 59–76.

[48] N. Turakhia and D. Parikh. "Attribute Dominance: What Pops Out?" In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Dec. 2013.

[49] X. Yang, T. Zhang, C. Xu, S. Yan, M. S. Hossain, and A. Ghoneim. "Deep Relative Attributes". In: *IEEE Transactions on Multimedia* 18.9 (Sept. 2016), pp. 1832–1842.

[50] 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)*. June 2014.

[51] A. Yu and K. Grauman. "Just Noticeable Differences in Visual Attributes". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Dec. 2015.

[52] Aron Yu and Kristen Grauman. "Semantic Jitter: Dense Supervision for Visual Comparisons via Synthetic Images". In: *CoRR* abs/1612.06341 (2016). URL: http://arxiv.org/abs/1612.06341.