Relative Attributes

Devi Parikh
Toyota Technological Institute Chicago (TTIC)
dparikh@ttic.edu

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

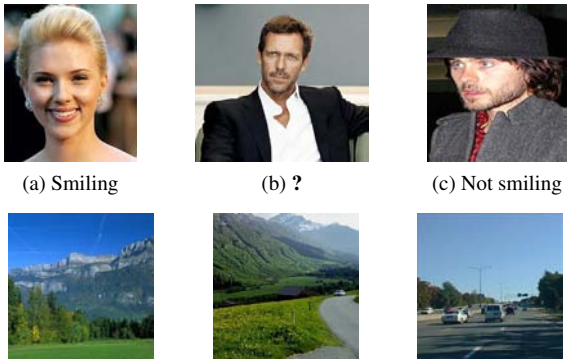
Abstract

Human-nameable visual "attributes" can benefit various recognition tasks. However, existing techniques restrict these properties to categorical labels (for example, a person is 'smiling' or not, a scene is 'dry' or not), and thus fail to capture more general semantic relationships. We propose to model relative *attributes. Given training data stating how object/scene categories relate according to different attributes, we learn a ranking function per attribute. The learned ranking functions predict the relative strength of each property in novel images. We then build a generative model over the joint space of attribute ranking outputs, and propose a novel form of zero-shot learning in which the supervisor* relates *the unseen object category to previously seen objects via attributes (for example, 'bears are furrier than giraffes'). We further show how the proposed relative attributes enable richer textual descriptions for new images, which in practice are more precise for human interpretation. We demonstrate the approach on datasets of faces and natural scenes, and show its clear advantages over traditional binary attribute prediction for these new tasks.*

1. Introduction

While traditional visual recognition approaches map low-level image features directly to object category labels, recent work proposes models using *visual attributes* [1–8]. Attributes are properties observable in images that have human-designated names (e.g., 'striped', 'four-legged'), and they are valuable as a new semantic cue in various problems. For example, researchers have shown their impact for strengthening facial verification [5], object recognition [6, 8, 16], generating descriptions of unfamiliar objects [1], and to facilitate "zero-shot" transfer learning [2], where one trains a classifier for an unseen object simply by specifying which attributes it has.

Problem: Most existing work focuses wholly on attributes as binary predicates indicating the presence (or absence) of a certain property in an image [1–8, 16]. This may suffice for part-based attributes (e.g., 'has a head') and some binary properties (e.g., 'spotted'). However, for a large variety of attributes, not only is this binary setting restrictive, but it is also unnatural. For instance, it is not clear if in Figure 1(b) Hugh Laurie is smiling or not; different people are likely to respond inconsistently in providing the presence or absence of the 'smiling' attribute for this image, or of the 'natural' attribute for Figure 1(e).



(a) Smiling (b) **?** (c) Not smiling

(d) Natural (e) **?** (f) Manmade

Figure 1. Binary attributes are an artificially restrictive way to describe images. While it is clear that (a) is smiling, and (c) is not, the more informative and intuitive description for (b) is via *relative* attributes: he is smiling more than (a) but less than (c). Similarly, scene (e) is less natural than (d), but more so than (f). Our main idea is to model relative attributes via learned ranking functions, and then demonstrate their impact on novel forms of zero-shot learning and generating image descriptions.

Indeed, we observe that *relative* visual properties are a semantically rich way by which humans describe and compare objects in the world. They are necessary, for instance, to refine an identifying description ("the 'rounder' pillow"; "the same except 'bluer'"), or to situate with respect to reference objects ("'brighter' than a candle; 'dimmer' than a flashlight"). Furthermore, they have potential to enhance active and interactive learning—for instance, offering a better guide for a visual search ("find me similar shoes, but 'shinier'." or "refine the retrieved images of downtown Chicago to those taken on 'sunnier' days").

Proposal: In this work, we propose to model *relative attributes*. As opposed to predicting the presence of an attribute, a relative attribute indicates the strength of an attribute in an image with respect to other images. For exam-

ple, in Figure 1, while it is difficult to assign a meaningful value to the binary attribute 'smiling', we could all agree on the relative attribute, *i.e*. Hugh Laurie is smiling less than Scarlett Johansson, but more than Jared Leto. In addition to being more natural, relative attributes would offer a richer mode of communication, thus allowing access to more detailed human supervision (and so potentially higher recognition accuracy), as well as the ability to generate more informative descriptions of novel images.

How can we learn relative properties? Whereas traditional supervised classification is appropriate to learn attributes that are intrinsically binary, it falls short when we want to represent visual properties that are nameable but not categorical. Our goal is instead to estimate the degree of that attribute's presence—which, importantly, differs from the probability of a binary classifier's prediction. To this end, we devise an approach that learns a *ranking function* for each attribute, given relative similarity constraints on pairs of examples (or more generally a partial ordering on some examples). The learned ranking function can estimate a *real-valued* rank[1] for images indicating the relative strength of the attribute presence in them. Then, we introduce novel forms of zero-shot learning and description that exploit the relative attribute predictions.

The proposed ranking approach accounts for a subtle but important difference between relative attributes and conceivable alternatives based on regression or multi-way classification. While such alternatives could also allow for a richer vocabulary, during training they could suffer from similar inconsistencies as binary attributes. For example, it is more difficult to define and perhaps more importantly, agree on, "With what strength is he smiling?" than "Is he smiling more than she is?". Thus, we expect the relative mode of supervision our approach permits to be more natural and consistent for human labelers.

Contributions: Our main contribution is the idea to learn relative visual attributes, which to our knowledge has not been explored in any prior work. Our other contribution is to devise and demonstrate two new tasks well-served by relative attributes: (1) zero-shot learning from relative comparisons, and (2) image description in reference to example images or categories. We demonstrate the approach for both tasks using the Outdoor Scenes dataset [11] and a subset of the Public Figure Face Database [12]. We find that relative attributes yield significantly better zero-shot learning accuracy when compared to their binary counterparts. In addition, we conduct human subject studies to evaluate the informativeness of the automatically generated image descriptions, and find that relative attributes are clearly more powerful than existing binary attributes in uniquely identifying an image.

[1]Throughout this paper we refer to rank as a real-valued score.

2. Related Work

We review related work on visual attributes, other uses of relative cues, and methods for learning comparisons.

Binary attributes: Learning attribute categories allows prediction of color or texture types [13], and can also provide a mid-level cue for object or face recognition [2, 5, 8]. Beyond object recognition, the semantics intrinsic to attributes enable zero-shot transfer [2, 6, 14], or description and part localization [1, 7, 15]. Rather than manually define attribute vocabularies, some work aims to discover attribute-related concepts on the Web [3, 4], extract them from existing knowledge sources [6, 16] or discover them interactively [9]. In contrast to our approach, all such methods restrict the attributes to be categorical (and in fact, binary).

Relative information: Relative information has been explored in vision in a variety of ways. Recent work on large-scale recognition exploits WordNet-based information to specify a semantic-distance sensitive classifier [18], or to make do with few labels by sharing training images among semantically similar classes [17]. Stemming from a related motivation of limited labeled data, Wang *et al*. [19] make use of explicit similarity-based supervision such as "A serval is like a leopard" or "A zebra is similar to the crosswalk in texture" to share training instances for categories with limited or no training instances. Unlike our approach, that method learns a model for each object category, and does not model attributes. In contrast, our attribute models are category-independent and transferrable, enabling relative descriptions between all classes. Moreover, whereas that technique captures similarity among object categories, ours models a general ordering of the images sorted by the strength of their attributes, as well as a joint space over multiple such relative attributes.

Kumar *et al*. [12] explore comparative facial attributes such as "Lips like Barack Obama" for face verification. These attributes, although comparative, are also modeled as binary classifiers and are similarity-based as opposed to an ordering. Gupta *et al*. [20] and Siddiquie *et al*. [21] use prepositions and adjectives to relate objects to each other for more effective contextual modeling and active learning, respectively. In contrast, our work involves relative modeling of attribute strengths for a richer vocabulary that enhances supervision and description of images.

Learning to rank: Learning to rank has received extensive attention in the machine learning literature [22–24], for information retrieval in general and image retrieval in particular [25, 26]. Given a query image, user preferences (often captured via click-data) are incorporated to learn a ranking function with the goal of retrieving more relevant images in the top search results. Learned distance metrics (e.g., [27, 28]) can induce a ranking on images; however,

this ranking is also specific to a query image, and typically intended for nearest-neighbor-based classifiers. Our work learns a ranking function on images based on constraints specifying the relative strength of attributes, and the resulting function is not relative to any other image in the dataset. Thus, unlike query-centric retrieval tasks, we can characterize individual images by the strength of the attributes present, which we show is valuable for new recognition and description applications.

3. Approach

We first present our approach for learning relative attributes (Section 3.1), and then explain how we can use relative attributes for enhanced zero-shot learning (Section 3.2) and image description generation (Section 3.3).

3.1. Learning Relative Attributes

We are given a set of training images $I = \{i\}$ represented in $\mathbb{R}^n$ by feature-vectors $\{\boldsymbol{x_i}\}$ and a set of M attributes $A = \{a_m\}$. In addition, for each attribute a_m, we are given a set of ordered pairs of images $O_m = \{(i,j)\}$ and a set of un-ordered pairs $S_m = \{(i,j)\}$ such that $(i,j) \in O_m \implies i \succ j$, *i.e.* image i has a stronger presence of attribute a_m than j, and $(i,j) \in S_m \implies i \sim j$, *i.e.* i and j have similar relative strengths of a_m. We note that O_m and S_m can be deduced from any partial ordering of the images I in the training data with respect to strength of a_m. Either O_m or S_m, but not both, can be empty.

Our goal is to learn M ranking functions

$$r_m(\boldsymbol{x_i}) = \boldsymbol{w_m^T x_i}, \tag{1}$$

for $m = 1, \ldots, M$, such that the maximum number of the following constraints is satisfied:

$$\forall (i,j) \in O_m : \boldsymbol{w_m^T x_i} > \boldsymbol{w_m^T x_j} \tag{2}$$

$$\forall (i,j) \in S_m : \boldsymbol{w_m^T x_i} = \boldsymbol{w_m^T x_j}. \tag{3}$$

While this is an NP hard problem [22], it is possible to approximate the solution with the introduction of non-negative slack variables, similar to SVM classification. We directly adapt the formulation proposed in [22], which was originally applied to web page ranking, except we use a quadratic loss function together with similarity constraints, leading to the following optimization problem:

$$\text{minimize} \quad \left(\frac{1}{2}||\boldsymbol{w_m^T}||_2^2 + C\left(\sum \xi_{ij}^2 + \sum \gamma_{ij}^2\right)\right) \tag{4}$$

$$s.t. \quad \boldsymbol{w_m^T x_i} \geq \boldsymbol{w_m^T x_j} + 1 - \xi_{ij}; \forall (i,j) \in O_m \tag{5}$$

$$|\boldsymbol{w_m^T x_i} - \boldsymbol{w_m^T x_j}| \leq \gamma_{ij}; \forall (i,j) \in S_m \tag{6}$$

$$\xi_{ij} \geq 0; \gamma_{ij} \geq 0. \tag{7}$$

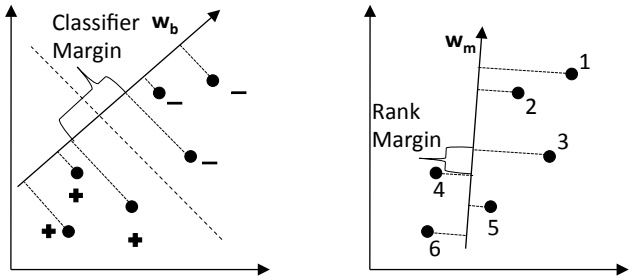


Figure 2. Distinction between learning a wide-margin ranking function (right) that enforces the desired ordering on training points (1-6) , and a wide-margin binary classifier (left) that only separates the two classes ($+$ and $-$), and does not necessarily preserve a desired ordering on the points.

Rearranging the constraints reveals that the above formulation, without the similarity constraints in Eqn. 6, is quite similar to the SVM classification problem, but on pairwise difference vectors:

$$\text{minimize} \quad \left(\frac{1}{2}||\boldsymbol{w_m^T}||_2^2 + C\left(\sum \xi_{ij}^2 + \sum \gamma_{ij}^2\right)\right) \tag{8}$$

$$s.t. \quad \boldsymbol{w_m^T}(\boldsymbol{x_i} - \boldsymbol{x_j}) \geq 1 - \xi_{ij}; \forall (i,j) \in O_m \tag{9}$$

$$|\boldsymbol{w_m^T}(\boldsymbol{x_i} - \boldsymbol{x_j})| \leq \gamma_{ij}; \forall (i,j) \in S_m \tag{10}$$

$$\xi_{ij} \geq 0; \gamma_{ij} \geq 0, \tag{11}$$

where C is the trade-off constant between maximizing the margin and satisfying the pairwise relative constraints. We solve the above primal problem using Newton's method [29]. While we use a linear ranking function in our experiments, the above formulation can be easily extended to kernels.

We note that this learning-to-rank formulation learns a function that explicitly enforces a desired ordering on the training images; the margin is the distance between the closest two projections within all desired (training) rankings. In contrast, if one were to train a binary classifier, only the margin between the nearest binary-labeled examples is enforced; ordering among examples beyond those defining the margin is arbitrary. See Figure 2. Our experiments confirm this distinction does indeed matter in practice, as our learnt ranking function is more effective at capturing the relative strengths of the attributes than the score of a binary classifier (*i.e.*, the magnitude of the SVM decision function).

In addition, training with *comparisons* (image i is similar to j in terms of attribute a_m, or i exhibits a_m less than j) is well-suited to the task at hand. Attribute strengths are arguably more natural to express in relative terms, as opposed to requiring absolute judgments in isolation (*i.e.*, i represents a_m with degree 10).

We apply our learnt relative attributes in two new settings: (1) zero-shot learning with relative relationships and (2) generating image descriptions. We now introduce our approach to incorporate relative attributes for each of these applications in turn.

3.2. Zero-Shot Learning From Relationships

Consider N categories of interest. For example, each category may be an object class, or a type of scene. During training, S of these categories are 'seen' categories for which training images are provided, while the remaining $U = N - S$ categories are 'unseen', for which no training images are provided.

The S categories are described via relative attributes with respect to each other, be it pairwise relationships or partial orders. For example, "bears are furrier than giraffes but less furry than rabbits", "lions are larger than dogs, as large as tigers, but less large than elephants", etc. We note that all pairs of categories need not be related in the supervision, and different subsets of categories can be related for the different attributes.

The U unseen categories, on the other hand, are described relative to one or two seen categories for a subset of the attributes, *i.e.*, unseen class $c_j^{(u)}$ can be described as $c_i^{(s)} \succ c_j^{(u)} \succ c_k^{(s)}$ for attribute a_m, or $c_i^{(s)} \succ c_j^{(u)}$, or $c_j^{(u)} \succ c_k^{(s)}$, where $c_i^{(s)}$ and $c_k^{(s)}$ are seen categories. We note the simple and flexible supervision required for the categories, especially the unseen ones: for any attribute (not necessarily all), the user can select any seen category depicting a stronger and/or weaker presence of the attribute to which to relate the unseen category. While alternative list-based learning to rank techniques are available [23], we choose the pairwise learning technique as described in Section 3.1 to ensure this ease of supervision.

During testing, a novel image is to be classified into any of the N categories. Our zero-shot learning setting is more general than the model proposed by Lampert *et al*. [2], in that the supervisor may not only associate attributes with categories, but also express how the categories *relate* along any number of the attributes. We expect this richer representation to allow better divisions between both the unseen and seen categories, as we demonstrate in the experiments.

We propagate the category relationships provided during training to the corresponding images, *i.e.*, for seen classes $c_i^{(s)}$ and $c_j^{(s)}$, $c_i^{(s)} \succ c_j^{(s)} \implies i \succ j; \forall i \in c_i^{(s)}, \forall j \in c_j^{(s)}$ for attribute a_m.[2] We then learn all M relative attributes as described in Section 3.1. Predicting the real-valued rank of all images in the training dataset I allows us to transform $\boldsymbol{x_i} \in \mathbb{R}^n \rightarrow \tilde{\boldsymbol{x}}_{\boldsymbol{i}} \in \mathbb{R}^M$, such that each image i is now represented as an M-dimensional vector $\tilde{\boldsymbol{x}}_{\boldsymbol{i}}$ indicating its rank score for all M attributes.

We now build a generative model for each of the S seen categories in $\mathbb{R}^M$. We use a Gaussian distribution, and estimate the mean $\boldsymbol{\mu}_i^{(s)} \in \mathbb{R}^M$ and $M \times M$ covariance matrix $\Sigma_i^{(s)}$ from the ranking-scores of the training images from class $c_i^{(s)}$, so we have $c_i^{(s)} \sim \mathcal{N}(\boldsymbol{\mu}_i^{(s)}, \Sigma_i^{(s)})$, for $i = 1, \ldots, S$.

The parameters of the generative model corresponding to each of the U unseen categories are selected under the guidance of the input relative descriptions. In particular, given an unseen category $c^{(u)}$, we employ the following:

1. If $c_j^{(u)}$ is described as $c_i^{(s)} \succ c_j^{(u)} \succ c_k^{(s)}$, where $c_i^{(s)}$ and $c_k^{(s)}$ are seen categories, then we set the m-th component of the mean $\boldsymbol{\mu}_{jm}^{(u)}$ to $\frac{1}{2}(\boldsymbol{\mu}_{im}^{(s)} + \boldsymbol{\mu}_{km}^{(s)})$.
2. If $c_j^{(u)}$ is described as $c_i^{(s)} \succ c_j^{(u)}$, we set $\boldsymbol{\mu}_{jm}^{(u)}$ to $\boldsymbol{\mu}_{im}^{(s)} - d_m$, where d_m is the average distance between the sorted mean ranking-scores $\boldsymbol{\mu}_{im}^{(s)}$'s of seen classes for attribute a_m. It is reasonable to expect the unseen class to be as far from the specified seen class as other seen classes tend to be from each other.
3. Similarly, if $c_j^{(u)}$ is described as $c_j^{(u)} \succ c_k^{(s)}$, we set $\boldsymbol{\mu}_{jm}^{(u)}$ to $\boldsymbol{\mu}_{im}^{(s)} + d_m$.
4. If a_m is not used to describe $c_j^{(u)}$, we set $\boldsymbol{\mu}_{jm}^{(u)}$ to be the mean across all training image ranks for a_m and the m-th diagonal entry of $\Sigma_j^{(u)}$ to be the variance of the same.

In the first three cases, we simply set $\Sigma_j^{(u)} = \frac{1}{S}\sum_{i=1}^{S} \Sigma_i^{(s)}$.

Given a test image i, we compute $\tilde{\boldsymbol{x}}_i \in \mathbb{R}^M$ indicating the relative attribute ranking-scores for the image. It is then assigned to the seen *or* unseen category that assigns it the highest likelihood:

$$c^* = \underset{j \in \{1,\ldots,N\}}{\operatorname{argmax}} \; P\left(\tilde{\boldsymbol{x}}_i \mid \boldsymbol{\mu}_j, \Sigma_j\right). \tag{12}$$

From a Bayesian perspective, our approach to setting the parameters of the unseen categories' generative models can be considered to be priors transferred from the knowledge of the models for the seen categories. Under reasonable priors, the choice of mean and covariances correspond to the minimum mean-squared error and maximum likelihood estimates. Related formulations of transfer through parameter sharing have been studied by Fei-Fei *et al*. [30] and Stark *et al*. [31] for learning shape-based object models with few training images, though no prior models consider transferring knowledge based on relative comparisons, as we do here.

We note that if one or more images from the unseen categories were subsequently to become available, our estimated parameters could easily be updated in light of the additional evidence. Furthermore, our general approach could potentially support more specific supervision about the relative relationships, should it be available (e.g., bears (unseen) are significantly more furry than cows (seen)).

[2] This generalizes naturally to allow stronger supervision per image instance, when available.

3.3. Describing Images in Relative Terms

The second application of relative attributes that we propose is that of describing novel images. The goal is to be able to relate any new example to other images according to different properties—whether its class happens to be familiar or not. This basic functionality would allow, for instance, the meaningful search example applications given in the introduction. (See recent work in [32] for other forms of image description based on object-action-scene tags.)

During training, we are given a set of training images $I = \{i\}$, each represented by a feature-vector $\boldsymbol{x_i} \in \mathbb{R}^n$, a list $A = \{a_m\}$ of M attributes along with $O_m = \{(i,j)\}$ $s.t.$ $i \succ j$ and $S_m = \{(i,j)\}$ $s.t.$ $i \sim j$ in relative strength of a_m.[3]. We learn M ranking-functions as described in Section 3.1, and evaluate them on all training images in I.

Given a novel image j to be described, we evaluate all learnt ranking functions $r_m(\boldsymbol{x_j})$. For each attribute a_m, we identify two reference images i and k from I that will be used to describe j via relative attributes. In principle, with a good ranking function any reference images could be informative. In our implementation, we adhere to the following guidelines. To avoid generating an overly precise description, we wish to select i and k such that they are not very similar to j in terms of attribute strength. However, to avoid trivial descriptions, they must not be too far from j, either.

Hence, we pick i and k such that $i \succ j$ and $j \succ k$ in strength of a_m, and $\frac{1}{8}^{th}$ of the images in I lie between i and j, as well as between j and k. In the case of boundary conditions where no such i or k exist, i is chosen to be the image in I with the least strength of a_m, and k is set to the image in I with the highest strength of a_m. The image j can then be described in terms of all or a subset of the M attributes, relative to any identified pairs (i,k). Figure 7 shows an example description generated by our approach, as well as an illustration of selected pairs (i,k).

While more elaborate analysis of the dataset distribution—and even psychophysics knowledge of the sensitivity of humans to change in different attributes—could make the selection of reference images more effective, we employ this straightforward technique as a proof-of-concept and leave such analysis for future work.

4. Experiments

We evaluate our approach on two datasets: (1) **Outdoor Scene Recognition (OSR) Dataset** [11] containing 2688 images from 8 categories. We use the 512-dimensional gist [11] descriptor as our image features. (2) A subset of the **Public Figure Face Database (PubFig)** [12] containing 800 images from 8 random identities (100 images each). We use a concatenation of the gist descriptor and a 45-dimensional Lab color histogram as our image features.

[3]While this application does not *require* category labels, the relative supervision can be provided for categories which is propagated to images.

	Binary	Relative
OSR	T I S HC OMF	
natural	0 0 0 0 1 1 1 1	T≺I∼S≺H≺C∼O∼M∼F
open	0 0 0 1 1 1 1 0	T∼F≺I∼S≺M≺H∼C∼O
perspective	1 1 1 1 0 0 0 0	O≺C≺M∼F≺H≺I≺S≺T
large-objects	1 1 1 0 0 0 0 0	F≺O∼M≺I∼S≺H∼C≺T
diagonal-plane	1 1 1 1 0 0 0 0	F≺O∼M≺C≺I∼S≺H≺T
close-depth	1 1 1 1 0 0 0 1	C≺M≺O≺T∼I∼S∼H∼F
PubFig	ACHJ MS V Z	
Masculine-looking	1 1 1 1 0 0 1 1	S≺M≺Z≺V≺J≺A≺H≺C
White	0 1 1 1 1 1 1 1	A≺C≺H≺Z≺J≺S≺M≺V
Young	0 0 0 0 1 1 0 1	V≺H≺C≺J≺A≺S≺Z≺M
Smiling	1 1 1 0 1 1 0 1	J≺V≺H≺A∼C≺S∼Z≺M
Chubby	1 0 0 0 0 0 0 0	V≺J≺H≺C≺Z≺M≺S≺A
Visible-forehead	1 1 1 0 1 1 1 0	J≺Z≺M≺S≺A∼C∼H∼V
Bushy-eyebrows	0 1 0 1 0 0 0 0	M≺S≺Z≺V≺H≺A≺C≺J
Narrow-eyes	0 1 1 0 0 0 1 1	M≺J≺S≺A≺H≺C≺V≺Z
Pointy-nose	0 0 1 0 0 0 0 1	A≺C≺J∼M∼V≺S≺Z≺H
Big-lips	1 0 0 0 1 1 0 0	H≺J≺V≺Z≺C≺M≺A≺S
Round-face	1 0 0 0 1 1 0 0	H≺V≺J≺C≺Z≺A≺S≺M

Table 1. Binary and relative attribute assignments used in our experiments. Note that none of the relative orderings violate the binary memberships. The OSR dataset includes images from the following categories: coast (C), forest (F), highway (H), inside-city (I), mountain (M), open-country (O), street (S) and tall-building (T). The 8 attributes shown above are listed in [11] as the properties subjects used to organize the images.The PubFig dataset includes images of: Alex Rodriguez (A), Clive Owen (C), Hugh Laurie (H), Jared Leto (J), Miley Cyrus (M), Scarlett Johansson (S), Viggo Mortensen (V) and Zac Efron (Z). The 11 attributes shown above are a subset of the attributes provided with the dataset. They were chosen for their simplicity, sufficient variation among the 8 categories, and to avoid redundancy (*e.g.* using Young instead of Old, Middle Aged, Youth, Child).

Table 1 provides more details about the datasets, and shows the binary memberships and relative orderings of categories by attributes. These were collected using the judgements of a colleague unfamiliar with the details of this work. We see the limitation of binary attributes in distinguishing between some categories, while the same set of attributes used *relatively* tease them apart. Although we have a full ordering, in our experiments we sample random pairs of categories as supervision (as noted below). Recall that different pairs of categories can be related for different attributes. Note that we collect the binary supervision only to train baseline approaches; our approach uses only the relative supervision.

As a sanity check, we first demonstrate the superiority of our learnt ranks to capture relative orderings, as compared to an approach that treats the score of binary classifiers as a rank (Section 4.1). Then we evaluate the use of relative attributes for the two new tasks (Sections 4.2 and 4.3).

4.1. Learned Ranking vs. Classifier Scores

We train a binary linear SVM h_m by transferring the binary supervision listed in Table 1 to the training images for each attribute. For an image-pair (i,j) in a held-out test set (2648 images for OSR, 560 for PubFig), we evaluate the learnt classifier, and if $h_m(\boldsymbol{x_i}) > h_m(\boldsymbol{x_j})$ we predict $i \succ j$,

else $i \prec j$ for a_m. For comparison, we learn a linear ranking function r_m for each attribute using the relative constraints in Table 1, and compare $r_m(\boldsymbol{x_i})$ to $r_m(\boldsymbol{x_j})$ on the same test pairs. Both methods' predictions are then compared to the ground-truth relative ordering.

The learnt ranking function's accuracy is 89% and 82% on the OSR and PubFig datasets, respectively, as compared to 80% and 67% if using the binary classifier scores, confirming the advantage of a ranking function to effectively capture relative information.

4.2. Zero-Shot Learning Results

We compare our zero-shot approach to two baselines:

Baselines: Our first baseline is the Direct Attribute Prediction (DAP) model of Lampert *et al*. [2], which uses binary attribute descriptions for all categories. We train linear SVMs by transferring the binary supervision in Table 1 to training images from the seen categories. A test image $\boldsymbol{x}$ is assigned to a category using

$$c^* = \underset{c \in \{1,\ldots,N\}}{\operatorname{argmax}} \prod_{m=1}^{M} P(a_m = b_m^c \mid \boldsymbol{x}), \qquad (13)$$

where $P(a_m = b_m^c \mid \boldsymbol{x})$ is computed by transforming the binary classifier score via a sigmoid function, and b_m^c is the ground-truth binary bit taken by attribute a_m for class c as seen in Table 1. If a_m is not used to describe an unseen category, $P(a_m = b_m^c \mid \boldsymbol{x})$ is uniform (0.5).

We call our second baseline "score-based relative attributes" (SRA). It follows the same approach as in Section 3.2, except that it replaces rank values with the binary classifier output score. It is a stronger baseline than DAP, as it has the same benefits of the generative modeling of seen classes and relative descriptions of unseen classes as our approach. It is not limited by the binary description of the categories, which may be deprived as seen in Table 1.

Set up: We compare all methods in several different scenarios. Unless specified, we use 2 unseen and 6 seen categories. To train the ranking functions, we use 4 category-pairs among seen categories, and unseen categories are described relative to the two closest seen categories for each attribute (one stronger, one weaker). We use 30 training images per class, and the rest for testing, and report mean per-class accuracy over 10 random train/test and seen/unseen splits.

Proportion of unseen categories: We first study zero-shot learning accuracy as the proportion of unseen categories increases. Figure 3 shows the results.

First, we see even when all 8 categories are seen (0 unseen), our approach significantly outperforms both baselines. This validates the power of relative attributes for the classical recognition task. Also, SRA's gains over DAP with

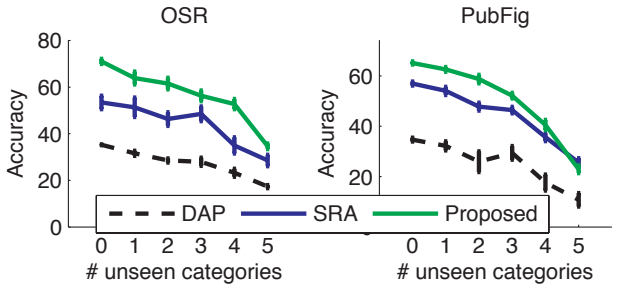


Figure 3. Zero-shot learning performance as the proportion of unseen categories increases. Total number of classes N remains constant at 8.

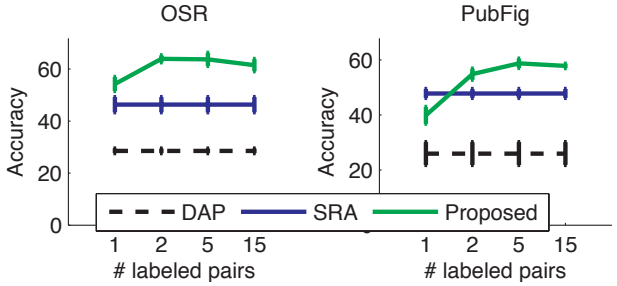


Figure 4. Zero-shot learning performance as more pairs of seen categories are related (*i.e*. labeled) during training.

0 unseen categories demonstrate the benefit of the generative modeling of the categories in SRA.

Further, as we would expect, accuracy for all three approaches decreases with more unseen categories. However, our method remains better than the baselines for most of the spectrum, until only 3 seen categories remain, at which point it performs similarly to SRA. This is expected, since beyond that with only 2 seen categories, the relative and binary supervision becomes equivalent. Both still compare favorably to DAP due to the benefit of relative description.

In general, we can expect that with even more total categories, the description power of relative attributes will also increase, as unseen categories would have more categories to be related to (even with a fixed number of attributes). A binary description, on the other hand, can only lose discriminative power as more categories are added.

Amount of supervision: We next study the impact of varying the amount of supervision.

Figure 4 shows the results as we increase the *number of pairs of seen categories* used to generate relative constraints, where for each attribute we randomly select the category pairs to be related from all $\binom{6}{2}$ possibilities. Our performance is quite robust to the number and choice of pairs; as few as two pairs suffice.[4] When using only one pair, our method receives significantly less supervision than the two baselines, for which all six categories are labeled (hence their flat curves). In spite of this, our approach performs favorably on OSR, though suffers compared to SRA on PubFig.

[4]While there are a total of 15 possible pairs to be labeled, as few as 5 of them could determine a unique ordering on all categories.

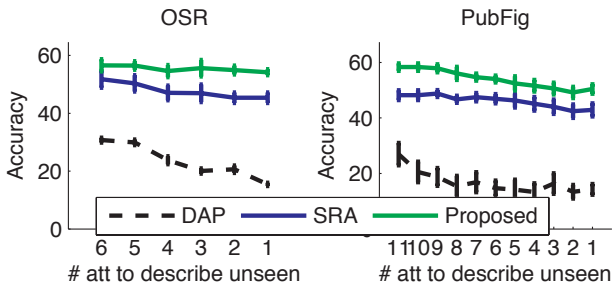


Figure 5. Zero-shot learning performance as fewer attributes are used to describe the unseen categories.

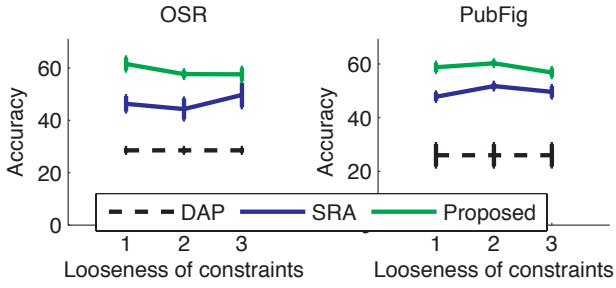


Figure 6. Zero-shot learning performance as the unseen categories are described via looser relationships.

Figure 5 shows the results as we decrease the *number of attributes* used to describe the unseen category during training. Note that the number of attributes used to describe the *seen* categories during training remains the same (see item 4 in Sec. 3.2). The accuracy of all methods degrades; however, the approaches using relative attributes (SRA and ours) decay gracefully, whereas DAP suffers more dramatically. This illustrates how each attribute conveys stronger distinguishing power when used relatively. This is a key result. This scenario exemplifies the high level of flexibility in supervision of unseen categories that our approach enables, which is crucial for practical applications.

Quality of supervision: What happens if the relationships described for an unseen class are 'looser'? That is, what if the annotator relates it to seen classes whose attribute strengths are more distant, e.g., says "Miley is younger than Vitto" rather than "Miley is younger than Scarlett" (a person closer in age)? Ideally, the supervisor would have freedom to specify any reference categories; that is the most natural form of description, and does not require the supervisor to know the exhaustive list of seen categories. Thus, we next evaluate performance as we increase the number of relative ranks away (looseness) from the seen categories used to describe the unseen category.[5]

Figure 6 shows the results. We see our approach is very robust to the looseness of the constraints. We attribute this

[5]At any level of looseness, if there exists no seen category at a desired distance from the unseen category in either direction, we simply use a one-ended constraint. Hence, when the constraints are at a looseness of 3, since only 6 out of 8 categories are seen, some of which often have similar attribute strengths, a large percentage of the constraints are one-sided.

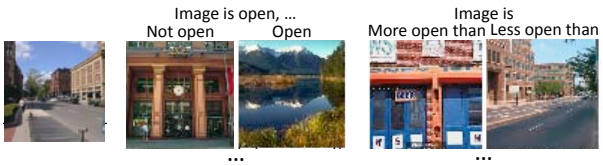


Figure 7. Part of example description generated for left image by binary attribute baseline (middle) and our method (right). See text for details.

to the power of relative attributes to *jointly* carve out regions in the space of attribute strengths corresponding to the unseen category. This makes the distance of the reference categories less relevant, as long as the relationships are correctly indicated.

4.3. Describing Images Results

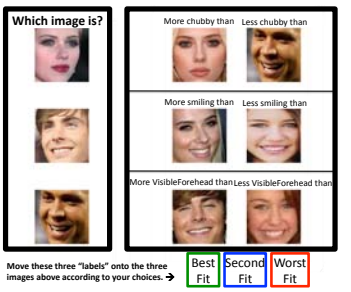
Next we demonstrate our approach to generate relative descriptions of novel images. To quantify their effectiveness, we perform a human subject study that pits the binary attribute baseline against our relative approach. Our method reports the properties predicted relative to reference images (see Sec. 3.3), while the baseline reports the predicted presence/absence of attributes only. The human subject must guess which image led to the auto-generated descriptions. To our knowledge, these are the first results to quantify how well algorithm-generated attribute descriptions can communicate to humans.

We recruited 18 subjects, only some familiar with vision. We randomly selected 20 PubFig and 10 OSR images. For each of the 30 test cases, we present the subject a description using three randomly selected attributes plus a multiple-choice set of three images, one of which is correct. The subject is asked to rank their guesses for which fits the description best. See Figure 8(a). To avoid bias, we divided the subjects into two groups; each group saw either the binary or the relative attributes, but not both. Further, we display reference images for either group's task, to help subjects understand the attribute meanings.

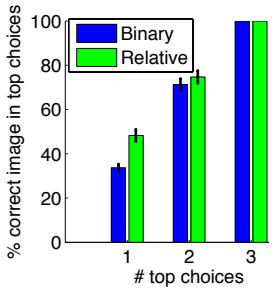
Figure 8(b) shows the results. Subjects are significantly more likely to identify the correct image using our method's description, i.e., 48% vs. 34% in the first choice. This reinforces our claim that relative attributes can better capture the "concept" of the image, and suggests their real promise for improved guided search or interactive learning.

We note that we augmented the baseline's binary descriptions with prototype images (showing stark contrast of attribute presence), even though, unlike our approach, they are not an intrinsic part of the generated description. We suspect that subjects would perform even worse with purely textual binary descriptions. Thus, the human study is, if anything, generous to the baseline.

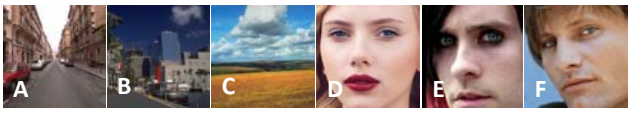
Our approach can be used to generate purely textual descriptions as well, where an image is described relative to other categories instead of images. Figure 8 (c) shows examples. Here our method selects the categories to compare



(a) Human Study Interface



(b) Results (both datasets)



(c) Images described relative to categories

Figure 8. Auto-generated descriptions of images in (c) **A: (bin)** not natural, not open, perspective **(rel)** more natural than tallbuilding; less natural than forest; more open than tallbuilding; less open than coast; more perspective than tallbuilding; **B: (bin)** not natural, not open, perspective **(rel)** more natural than insidecity; less natural than highway; more open than street; less open than coast; more perspective than highway; less perspective than insidecity; **C: (bin)** natural, open, perspective **(rel)** more natural than tallbuilding; less natural than mountain; more open than mountain; less perspective than opencountry; **D: (bin)** White, not Smiling, VisibleForehead **(rel)** more White than AlexRodriguez; more Smiling than JaredLeto; less Smiling than ZacEfron; more VisibleForehead than JaredLeto; less VisibleForehead than MileyCyrus; **E: (bin)** White, not Smiling, not VisibleForehead **(rel)** more White than AlexRodriguez; less White than MileyCyrus; less Smiling than HughLaurie; more VisibleForehead than ZacEfron; less VisibleForehead than MileyCyrus; **F: (bin)** not Young, BushyEyebrows, RoundFace **(rel)** more Young than CliveOwen; less Young than ScarlettJohansson; more BushyEyebrows than ZacEfron; less BushyEyebrows than AlexRodriguez; more RoundFace than CliveOwen; less RoundFace than ZacEfron.

to such that at least 50% of the images in the category have an attribute strength larger than (less than) that computed for the image to be described. Echoing our quantitative results, we can qualitatively see that the relative descriptions are more precise and informative than the binary ones. More results can be found on the authors' websites.

5. Conclusion

We introduced relative attributes, which allow for a richer language of supervision and description than the commonly used categorical (binary) attributes. We presented two novel applications: zero-shot learning based on relationships and describing images relative to other images or categories. Through extensive experiments as well as a human subject study, we clearly demonstrated the advantages of our idea. Future work includes exploring more novel applications of relative attributes, such as guided search or interactive learning, and automatic discovery of relative attributes.

Acknowledgements: We thank the subjects of our human studies for their time. This research is supported in part by NSF IIS-1065390, ONR ATL N00014-11-1-0105, and the Luce Foundation.

References

[1] A. Farhadi, I. Endres, D. Hoiem, and D. A. Forsyth. Describing Objects by their Attributes. *CVPR*, 2009.

[2] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning To Detect Unseen Object Classes by Between-Class Attribute Transfer. *CVPR*, 2009.

[3] M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych and B. Schiele. What Helps Where And Why? Semantic Relatedness for Knowledge Transfer. *CVPR*, 2010.

[4] T. L. Berg, A. C. Berg and J. Shih. Automatic Attribute Discovery and Characterization from Noisy Web Data. *ECCV*, 2010.

[5] N. Kumar, A. Berg, P. Belhumeur and S. K. Nayar. Attribute and Simile Classifiers for Face Verification. *ICCV*, 2009.

[6] J. Wang, K. Markert and M. Everingham. Learning Models for Object Recognition from Natural Language Descriptions. *BMVC*, 2009.

[7] G. Wang and D. Forsyth. Joint Learning of Visual Attributes, Object Classes and Visual Saliency. *ICCV*, 2009.

[8] Y. Wang and G. Mori. A Discriminative Latent Model of Object Classes and Attributes. *ECCV*, 2010.

[9] D. Parikh and K. Grauman. Interactively Building a Discriminative Vocabulary of Nameable Attributes. *CVPR*, 2011.

[10] M. Palatucci, D. Pomerleau, G. Hinton and T. Mitchell. Zero-Shot Learning with Semantic Output Codes. *NIPS*, 2009.

[11] A. Oliva and A. Torralba. Modeling the Shape of the Scene: a Holistic Representation of the Spatial Envelope. *IJCV*, 2001

[12] N. Kumar, A. C. Berg, P. N. Belhumeur and S. K. Nayar. Attribute and Simile Classifiers for Face Verification. *ICCV*, 2009.

[13] V. Ferrari and A. Zisserman. Learning Visual Attributes. *NIPS*, 2007.

[14] O. Russakovsky and L. Fei-Fei. Attribute Learning in Large-scale Datasets. *Workshop on Parts and Attributes, ECCV*, 2010.

[15] A. Farhadi, I. Endres and D. Hoiem. Attribute-centric Recognition for Cross-category Generalization. *CVPR*, 2010.

[16] S. Branson, C. Wah, B. Babenko, F. Schroff, P. Welinder, P. Perona P and S. Belongie. Visual Recognition with Humans in the Loop. *ECCV*, 2010.

[17] R. Fergus, H. Bernal, Y. Weiss and A. Torralba. Semantic Label Sharing for Learning with Many Categories. *ECCV*, 2010.

[18] J. Deng, A. C. Berg, K. Li and L. Fei-Fei. What Does Classifying More Than 10,000 Image Categories Tell Us? *ECCV*, 2010.

[19] G. Wang, D. Forsyth and D. Hoiem. Comparative Object Similarity for Improved Recognition with Few or No Examples. *CVPR*, 2010.

[20] A. Gupta and L. S. Davis. Beyond Nouns: Exploiting Prepositions and Comparative Adjectives for Learning Visual Classifiers. *ECCV*, 2008.

[21] B. Siddiquie and A. Gupta. Beyond Active Noun Tagging: Modeling Contextual Interactions for Multi-Class Active Learning. *CVPR*, 2010.

[22] T. Joachims. Optimizing Search Engines using Clickthrough Data. *KDD*, 2002.

[23] Z. Cao, T. Qin, T. Liu, M. Tsai and H. Li. Learning to Rank: From Pairwise Approach to Listwise Approach. *ICML*, 2007.

[24] T. Liu. Learning to Rank for Information Retrieval. *Foundations and Trends in Information Retrieval*, 2009.

[25] V. Jain and M. Varma. Learning to Re-Rank: Query-Dependent Image Re-Ranking Using Click Data. *WWW*, 2011.

[26] Y. Hu, M. Li and N. Yu. Multiple-Instance Ranking: Learning to Rank Images for Image Retrieval. *CVPR*, 2008.

[27] E. P. Xing, A. Y. Ng, M. I. Jordan and S. Russell. Distance Metric Learning, with Application to Clustering with Side-Information. *NIPS*, 2002.

[28] A. Frome, Y. Singer, F. Sha, and J. Malik. Learning Globally-Consistent Local Distance Functions for Shape-based Image retrieval and Classification. *ICCV*, 2007.

[29] O. Chapelle. Training a Support Vector Machine in the Primal. *Neural Computation*, 2007.

[30] L. Fei-Fei, R. Fergus and P. Perona. A Bayesian Approach to Unsupervised One-Shot Learning of Object Categories *ICCV*, 2003.

[31] M. Stark, M. Goesele and B. Schiele. A Shape-Based Object Class Model for Knowledge Transfer. *ICCV*, 2009.

[32] A. Farhadi, M. Hejrati, M. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every Picture Tells a Story: Generating Sentences from Images. *ECCV*, 2010.

[33] C. Dwork, R. Kumar, M. Naor and D. Sivakumar. Rank Aggregation Methods for the Web. *WWW*, 2001.