Discovering the Spatial Extent of Relative Attributes

Fanyi Xiao and Yong Jae Lee
University of California, Davis
{fanyix,yjlee}@cs.ucdavis.edu

Abstract

We present a weakly-supervised approach that discovers the spatial extent of relative attributes, given only pairs of ordered images. In contrast to traditional approaches that use global appearance features or rely on keypoint detectors, our goal is to automatically discover the image regions that are relevant to the attribute, even when the attribute's appearance changes drastically across its attribute spectrum. To accomplish this, we first develop a novel formulation that combines a detector with local smoothness to discover a set of coherent visual chains *across the image collection. We then introduce an efficient way to generate additional chains anchored on the initial discovered ones. Finally, we automatically identify the most relevant visual chains, and create an ensemble image representation to model the attribute. Through extensive experiments, we demonstrate our method's promise relative to several baselines in modeling relative attributes.*

1. Introduction

Visual attributes are human-nameable object properties that serve as an intermediate representation between low-level image features and high-level objects or scenes [23, 10, 21, 9, 30, 32, 33, 17]. They yield various useful applications including describing an unfamiliar object, retrieving images based on mid-level properties, "zero-shot" learning [29, 23, 30], and human-computer interaction [4, 5]. Researchers have developed systems that model binary attributes [23, 10, 21]—a property's presence/absence (e.g., "is furry/not furry")—and relative attributes [30, 35, 34]—a property's relative strength (e.g., "furrier than").

While most existing work use global image representations to model attributes (e.g., [23, 30]), recent work demonstrates the effectiveness of using localized part-based representations [3, 34, 43]. They show that attributes—be it global ("is male") or local ("smiling")—can be more accurately learned by first bringing the underlying object-parts into correspondence, and then modeling the attributes conditioned on those object-parts. For example, the attribute "wears glasses" can be more easily learned when people's faces are in correspondence. To compute such correspondences, pre-trained part detectors are used (e.g., faces [34] and people [3, 43]). However, because the part detectors are trained independently of the attribute, the learned parts may not necessarily be useful for modeling the desired attribute. Furthermore, some objects do not naturally have well-defined parts, which means modeling the part-based detector itself becomes a challenge.

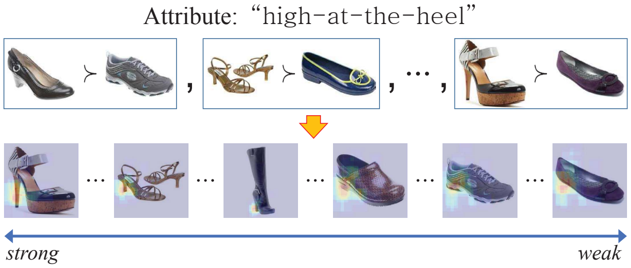

Figure 1. (**top**) Given pairs of images, each ordered according to relative attribute strength (e.g., "higher/lower-at-the-heel"), (**bottom**) our approach automatically discovers the attribute's spatial extent in each image, and learns a ranking function that orders the image collection according to predicted attribute strength.

The method in [7] addresses these issues by *discovering* useful, localized attributes. A drawback is that the system requires a human-in-the-loop to verify whether each discovered attribute is meaningful, limiting its scalability. More importantly, the system is restricted to modeling binary attributes; however, relative attributes often describe object properties better than binary ones [30], especially if the property exhibits large appearance variations (see Fig. 1).

So, how can we develop robust visual representations for *relative attributes*, without expensive and potentially uninformative pre-trained part detectors or humans-in-the-loop? To do so, we will need to automatically identify the visual patterns in each image whose appearance correlates with (i.e., changes as a function of) attribute strength. This is a challenging problem: as the strength of an attribute changes, the object's appearance can change drastically. For example, if the attribute describes how "high-heeled" a shoe is, then pumps and flats would be on opposite ends of the spectrum, and their heels would look completely different (see Fig. 1). Thus, identifying the visual patterns that characterize the attribute is very difficult without a priori knowledge of what a heel is. Moreover, it is even more difficult to do so given only samples of pairwise relative comparisons, which is the typical mode of relative attribute annotation.

In this work, we propose a method that automatically discovers the spatial extent of relative attributes in images across varying attribute strengths. The main idea is to leverage the fact that the visual concept underlying the attribute undergos a *gradual change* in appearance across the attribute spectrum. In this way, we propose to discover a set of local, transitive connections ("visual chains") that establish correspondences between the same object-part, even when its appearance changes drastically over long ranges. Given the candidate set of visual chains, we then automatically select those that together best model the changing appearance of the attribute across the attribute spectrum. Importantly, by combining a subset of the most-informative discovered visual chains, our approach aims to discover the full spatial extent of the attribute, whether it be concentrated on a particular object-part or spread across a larger spatial area.

Contributions. To our knowledge, no prior work discovers the spatial extent of attributes, given weakly-supervised pairwise relative attribute annotations. Towards this goal, important novel components include: (1) a new formulation for discovery that uses both a detector term and a smoothness term to discover a set of coherent visual chains, (2) a simple but effective way of quickly generating new visual chains anchored on the existing discovered ones, and (3) a method to rank and combine a subset of the visual chains that together best capture the attribute. We apply our approach to three datasets of faces and shoes, and outperform state-of-the-art methods that use global image features or require stronger supervision. Furthermore, we demonstrate an application of our approach, in which we can edit an object's appearance conditioned on the discovered spatial extent of the attribute.

2. Related Work

Visual attributes. Most existing work use global image representations to model attributes (e.g., [23, 30]). Others have demonstrated the effectiveness of *localized* representations. For example, the attribute "mouth open" can be more easily learned when people's mouths are localized. Early work showed how to localize simple color and shape attributes like "red" and "round" [12]. Recent approaches rely on pre-trained face/body landmark or "poselet" detectors [20, 21, 3, 16, 43], crowd-sourcing [7], or assume that the images are well-aligned and object/scene-centric [2, 41], which either restricts their usage to specific domains or limits their scalability. Unlike these methods that try to localize *binary* attributes, we instead aim to discover the spatial extent of *relative* attributes, while forgoing any pre-trained detector, crowd-sourcing, or object-centric assumptions.

While the "relative parts" approach of [34] shares our goal of localizing relative attributes, it uses strongly-supervised pre-trained facial landmark detectors, and is thus limited to modeling only facial attributes. Importantly, because the detectors are trained independently of the attribute, the detected landmarks may not necessarily be optimal for modeling the desired attribute. In contrast, our approach aims to directly localize the attribute without relying on pre-trained detectors, and thus can be used to model attributes for any object.

Visual discovery. Existing approaches discover object categories [37, 8, 13, 31, 28], low-level foreground features [26], or mid-level visual elements [36, 6].

Recent work shows how to discover visual elements whose appearance is correlated with time or space, given images that are time-/geo-stamped [25]. Algorithmically, this is the closest work to ours. However, our work is different in three important ways. First, the goal is different: we aim to discover visual chains whose appearance is correlated with *attribute strength*. Second, the form of supervision is different: we are given pairs of images that are ordered according to their relative attribute strength, so unlike [25], we must infer a global ordering of the images. Finally, we introduce a novel formulation and efficient inference procedure that exploits the local smoothness of the varying appearance of the attribute, which we show in Sec. 4.3 leads to more coherent discoveries.

3. Approach

Given an image collection $S{=}\{I_1, \dots, I_N\}$ with pairwise ordered and unordered image-level relative comparisons of an attribute (i.e., in the form of $\Omega(I_i){>}\Omega(I_j)$ and $\Omega(I_i){\approx}\Omega(I_j)$, where $i,j{\in}\{1,\dots,N\}$ and $\Omega(I_i)$ is I_i's attribute strength), our goal is to discover the spatial extent of the attribute in each image and learn a ranking function that predicts the attribute strength for any new image.

This is a challenging problem for two main reasons: (1) we are not provided with any localized examples of the attribute so we must automatically *discover* the relevant regions in each image that correspond to it, and (2) the appearance of the attribute can change drastically over the attribute spectrum. To address these challenges, we exploit the fact that for many attributes, the appearance will change gradually across the attribute spectrum. To this end, we first discover a diverse set of candidate *visual chains*, each linking the patches (one from each image) whose appearance changes smoothly across the attribute spectrum. We then select among them the most relevant ones that agree with the provided relative attribute annotations.

There are three main steps to our approach: (1) initializing a candidate set of visual chains; (2) iteratively growing each visual chain along the attribute spectrum; and (3) ranking the chains according to their relevance to the target attribute to create an ensemble image representation. In the following, we describe each of these steps in turn.

3.1. Initializing candidate visual chains

A visual attribute can potentially exhibit large appearance variations across the attribute spectrum. Take the *high-at-the-heel* attribute as an example: high-heeled shoes have strong vertical gradients while flat-heeled shoes have strong horizontal gradients. However, the attribute's appearance will be quite similar in any local region of the attribute spectrum. Therefore, to capture the attribute across its entire spectrum, we sort the image collection based on (predicted) attribute strength and generate candidate *visual chains* via iterative refinement; i.e., we start with short but visually homogeneous chains of image regions in a local region of the attribute spectrum, and smoothly grow them out to cover the entire spectrum. We generate multiple chains because (1) appearance similarity does not guarantee relevance to the attribute (e.g., a chain of blank white patches satisfies this property perfectly but provides no information about the attribute), and (2) some attributes are better described with multiple image regions (e.g., the attribute "eyes open" may better be described with two patches, one on each eye). We will describe how to select the relevant chains in Sec. 3.3.

We start by first sorting the images in S in descending order of predicted attribute strength—with $\tilde{I}_1$ as the strongest image and $\tilde{I}_N$ as the weakest—using a linear SVM-ranker [15] trained with global image features, as in [30]. To initialize a single chain, we take the top N_{init} images and select a set of patches (one from each image) whose appearance varies smoothly with its neighbors in the chain, by minimizing the following objective function:

$$\min_P C(P) = \sum_{i=2}^{N_{init}} ||\phi(P_i) - \phi(P_{i-1})||_2, \qquad (1)$$

where $\phi(P_i)$ is the appearance feature of patch P_i in $\tilde{I}_i$, and $P = \{P_1, \ldots, P_{N_{init}}\}$ is the set of patches in a chain. Candidate patches for each image are densely sampled at multiple scales. This objective enforces *local smoothness*: the appearances of the patches in the images with neighboring indices should vary smoothly within a chain. Given the objective's chain structure, we can efficiently find its global optimum using Dynamic Programming (DP).

In the backtracking stage of DP, we obtain a large number of K-best solutions. We then perform a chain-level non-maximum-suppression (NMS) to remove redundant chains to retain a set of K_{init} diverse candidate chains. For NMS, we measure the distance between two chains as the sum of intersection-over-union scores for every pair of patches from the same image. This ensures that different initial chains not only contain different patches from any particular image, but also together spatially cover as much of each image as possible (see Fig. 2).

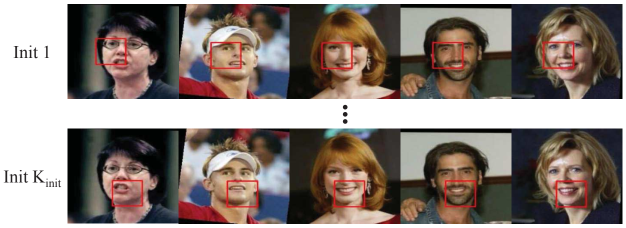

Figure 2. Our initialization consists of a set of diverse visual chains, each varying smoothly in appearance.

3.2. Iteratively growing each visual chain

The initial set of K_{init} chains are visually homogeneous but cover only a tiny fraction of the attribute spectrum. We next iteratively grow each chain to cover the entire attribute spectrum by training a model that adapts to the attribute's smoothly changing appearance. This idea is related to *self-paced learning* in the machine learning literature [22, 1], which has been applied to various computer vision tasks such as object discovery and tracking [25, 27, 38].

Specifically, for each chain, we iteratively train a detector and in each iteration use it to grow the chain while simultaneously refining it. To grow the chain, we again minimize Eqn. 1 but now with an additional term:

$$\min_P C(P) = \sum_{i=2}^{t*N_{iter}} ||\phi(P_i) - \phi(P_{i-1})||_2 - \lambda \sum_{i=1}^{t*N_{iter}} \mathbf{w}_t^T \phi(P_i), \qquad (2)$$

where $\mathbf{w}_t$ is a linear SVM detector learned from the patches in the chain from the $(t-1)$-th iteration[1], $P = \{P_1, \ldots, P_{t*N_{iter}}\}$ is the set of patches in a chain, and N_{iter} is the number of images considered in each iteration (explained in detail below). As before, the first term enforces local smoothness. The second term is the *detection* term: since the ordering of the images in the chain is only a rough estimate and thus possibly noisy (recall we computed the ordering using an SVM-ranker trained with global image features), $\mathbf{w}_t$ prevents the inference from drifting in the cases where local smoothness does not strictly hold. λ is a constant that trades-off the two terms. We use the same DP inference procedure used to optimize Eqn. 1.

Once P is found, we train a new detector with all of its patches as positive instances. The negative instances consist of randomly sampled patches whose intersection-over-union scores are lower than 0.3 with any of the patches in P. We use this new detector $\mathbf{w}_t$ in the next growing iteration. We repeat the above procedure T times to cover the entire attribute spectrum. Fig. 3 (a) illustrates the process of iterative chain growing for the "high-at-the-heel" and "smile" attributes. By iteratively growing the chain, we are able to coherently connect the attribute despite large appearance variations across its spectrum. There are two important considerations to make when growing the chain:

[1] For $t = 1$, we use the initial patches found in Sec. 3.1.

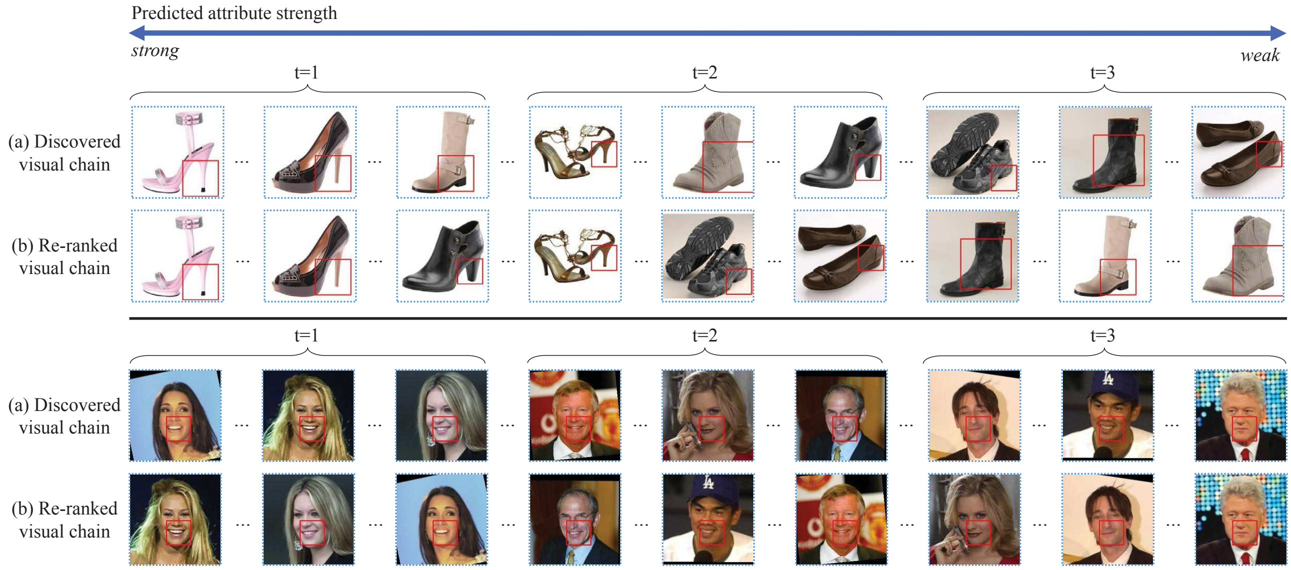
Predicted attribute strength
strong
weak
t=1
t=2
t=3
(a) Discovered visual chain
(b) Re-ranked visual chain
t=1
t=2
t=3
(a) Discovered visual chain
(b) Re-ranked visual chain

Figure 3. Top: "high-at-the-heel"; bottom: "smile". (**a**) We iteratively grow candidate visual chains along the direction of decreasing attribute strength, as predicted by the ranker trained with global image features [30]. (**b**) Once we obtain an accurate alignment of the attribute across the images, we can train a new ranker conditioned on the discovered patches to obtain a more accurate image ordering.

Multimodality of the image dataset. Not all images will exhibit the attribute due to pose/viewpoint changes or occlusion. We therefore need a mechanism to rule out such irrelevant images. For this, we use the detector $\mathbf{w}_t$. Specifically, we divide the image set S—now ordered in decreasing attribute strength as $\{\tilde{I}_1, \ldots, \tilde{I}_N\}$—into T *process sets*, each with size N/T. In the t-th iteration, we fire the detector $\mathbf{w}_t$ trained from the $(t-1)$-th iteration across each image in the t-th process set in a sliding window fashion. We then add the N_{iter} images with the highest maximum patch detection scores for chain growing in the next iteration.

Overfitting of the detector. The detector can overfit to the existing chain during iterative growing, which means that mistakes in the chain may not be fixed. To combat this, we adopt the *cross-validation* scheme introduced in [36]. Specifically, we split our image collection S into S_1 and S_2, and in each iteration, we run the above procedure first on S_1, and then take the resulting detector and use it to mine the chain in S_2. This produces more coherent chains, and also cleans up any errors introduced in either previous iterations or during chain initialization.

3.3. Ranking and creating a chain ensemble

We now have a set of K_{init} chains, each pertaining to a unique visual concept and each covering the entire range of the attribute spectrum. However, some image regions that capture the attribute could have still been missed because they are not easily detectable on their own (e.g., forehead region for "visible forehead"). Thus, we next describe a simple and efficient way to further diversify the pool of chains to increase the chance that such regions are selected. We then describe how to select the most relevant chains to create an ensemble that together best models the attribute.

Generating new chains anchored on existing ones. Since the patches in a chain capture the same visual concept across the attribute spectrum, we can use them as *anchors* to generate new chains by perturbing the patches *locally* in each image with the same perturbation parameters $(\Delta_x, \Delta_y, \Delta_s)$. More specifically, perturbing a patch centered at (x, y) with size (w, h) using parameter $(\Delta_x, \Delta_y, \Delta_s)$ leads to a new patch at location $(x+\Delta_x w, y+\Delta_y h)$, with size $(w \times \Delta_s, h \times \Delta_s)$ (see Fig. 4). Note that we get the alignment for the patches in the newly generated chains for free, as they are *anchored* on an existing chain. We generate K_{pert} chains for each of the K_{init} chains with Δ_x and Δ_y each sampled from $[-\delta_{xy}, \delta_{xy}]$ and Δ_s sampled from a discrete set χ, which results in $K_{pert} \times K_{init}$ chains in total. To detect the visual concept corresponding to a perturbed chain on any new unseen image, we take the detector of the anchoring chain and perturb its detection using the corresponding perturbation parameters.

Creating a chain ensemble. Different chains characterize different visual concepts. Not all of them are relevant to the attribute of interest and some are noisy. To select the relevant chains, we rank all the chains according to their relatedness with the target attribute using the image-level relative attribute annotations. For this, we split the original training data into two subsets: one for training and the other for validation. For each of the $K_{pert} \times K_{init}$ candidate chains, we train a linear SVM detector and linear SVM ranker [15, 30]. We then fire the detector on each validation image in a sliding window fashion and apply the ranker on

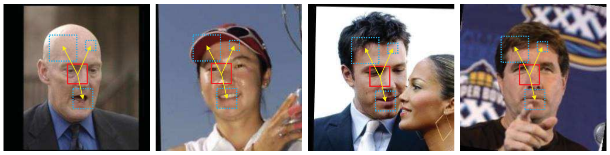

Figure 4. We generate new chains (blue dashed patches) anchored on existing ones (red solid patches). Each new chain is sampled at some location and scale relative to the chain anchoring it. This not only allows us to efficiently generate more chains, but also allows us to capture visual concepts that are hard to detect in isolation yet still important to model the attribute (e.g., 1st image: the patch at the top of the head is barely detectable due to its low gradient energy, even though it is very informative for "Bald head").

the patch with the maximum detection score to get an estimated attribute strength $\hat{\Omega}(I_i)$ for each image I_i. Finally, we count how many of the pairwise ground-truth attribute orderings agree with our predicted attribute orderings:

$$acc(R, \hat{\Omega}) = \frac{1}{|R|} \sum_{(i,j) \in R} \mathbb{1}[\hat{\Omega}(I_i) - \hat{\Omega}(I_j) \geq 0], \qquad (3)$$

where $|R|$ is the cardinality of the relative attribute annotation set on the validation data, and $\mathbb{1}[\cdot]$ is the indicator function. We rank each chain according to this validation set accuracy, and select the top K_{ens} chains. To form the final image-level representation for an image, we simply concatenate the feature vectors extracted from the detected patches, each weighted by its chain's validation accuracy. We then train a final linear SVM ranker using this ensemble image-level representation to model the attribute.

4. Results

We analyze our method's discovered spatial extent of relative attributes, pairwise ranking accuracy, and contribution of local smoothness and perturbed visual chains.

Implementation details. The feature ϕ we use for detection and local smoothness is HOG [11], with size 8×8 and 4 scales (patches ranging from 40×40 to 100×100 of the original image). For ranker learning, we use both the LLC encoding of dense-SIFT [40] stacked with a two-layer Spatial Pyramid (SP) grid [24], and *pool-5* activation features from the ImageNet pre-trained CNN (Alexnet architecture) implemented using Caffe [19, 14].[2] We set $\lambda = 0.05$, $N_{init} = 5$, $N_{iter} = 80$, $K_{init} = 20$, $K_{pert} = 20$, $K_{ens} = 60$, $\delta_{xy} = 0.6$, and $\chi = \{1/4, 1\}$. We find $T = 3$ iterations to be a good balance between chain quality and computation.

Baselines. Our main baseline is the method of [30] (Global), which learns a relative attribute ranker using global features computed over the whole image. We also compare to the approach of [34] (Keypoints), which learns a ranker with dense-SIFT features computed on facial keypoints detected using the supervised detector of [44], and to the local learning method of [42], which learns a ranker using only the training samples that are close to a given testing sample. For Global [30], we use the authors' code with the same features as our approach (dense-SIFT+LLC+SP and *pool-5* CNN features). For Keypoints [34] and [42], we compare to their reported numbers computed using dense-SIFT and GIST+color-histogram features, respectively.

Datasets. *LFW-10* [34] is a subset of the *Labeled faces in the wild* (LFW) dataset. It consists of 2000 images: 1000 for training and 1000 for testing. Annotations are available for 10 attributes, with 500 training and testing pairs per attribute. The attributes are listed in Table 1.

UT-Zap50K [42] is a large collection of 50025 shoe images. We use the UT-Zap50K-1 annotation set, which provides on average 1388 training and 300 testing pairs of relative attribute annotations for each of 4 attributes: "Open", "Sporty", "Pointy", and "Comfort". (See supp. for UT-Zap50K-2 results.)

Shoes-with-Attributes [18] contains 14658 shoe images from *like.com* and 10 attributes, of which 3 are overlapping with UT-Zap50K: "Open", "Sporty", and "Pointy". Because each attribute has only about 140 pairs of relative attribute annotations, we use this dataset only to evaluate cross-dataset generalization performance in Sec. 4.2.

4.1. Visualization of discovered spatial extent

In this section, we show qualitative results of our approach's discovered spatial extent for each attribute in LFW-10 and UT-Zap50K. For each image, we use a heatmap to display the final discovered spatial extent, where red/blue indicates strong/weak attribute relevance. To create the heatmap, the spatial region for each visual chain is overlaid by its predicted attribute relevance (as described in Sec. 3.3), and then summed up. Fig. 5 shows the resulting heatmaps on a uniformly sampled set of unseen test images per attribute, sorted according to predicted attribute strength using our final ensemble representation model.

Clearly, our approach has understood where in the image to look to find the attribute. For almost all attributes, our approach correctly discovers the relevant spatial extent (e.g., for localizable attributes like "Mouth open", "Eyes open", "Dark hair", and "Open", it discovers the corresponding object-part). Since our approach is data-driven, it can sometimes go beyond common human perception to discover non-trivial relationships: for "Pointy", it discovers not only the toe of the shoe, but also the heel, because pointy shoes are often high-heeled (i.e., the signals are highly correlated). For "Comfort", it has discovered that the lack or presence of heels can be an indication of how comfortable a shoe is. Each attribute's precisely discovered spatial extent also leads to an accurate image ordering by our en-

[2] We find the *pool-5* activations, which preserve more spatial information, to be more useful in our tasks than the fully-connected layers.

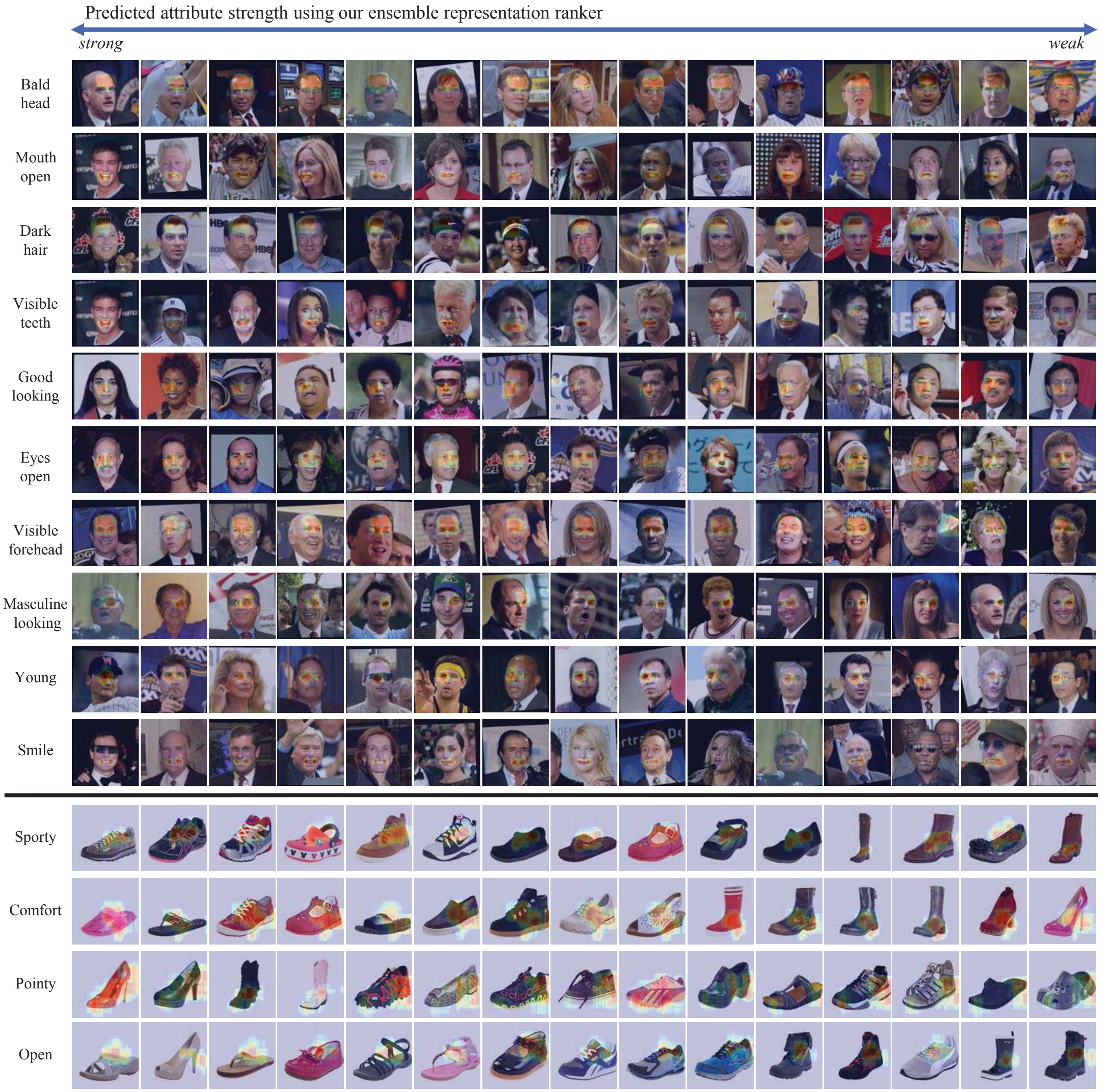

Figure 5. Qualitative results showing our discovered spatial extent and ranking of relative attributes on LFW-10 (top) and UT-Zap50K (bottom). We visualize our discoveries as heatmaps, where red/blue indicates strong/weak predicted attribute relevance. For most attributes, our method correctly discovers the relevant spatial extent (e.g., for "Mouth open", "Dark hair", and "Eyes open", it discovers the corresponding object-part), which leads to accurate attribute orderings. Our approach is sometimes able to discover what may not be immediately obvious to humans: for "Pointy", it discovers not only the toe of the shoe, but also the heel, because pointy shoes are often high-heeled (i.e., the signals are highly correlated). There are limitations as well, especially for atypical images: e.g., "Visible teeth" (12th image) and "Visible forehead" (13th image) are incorrect due to mis-detections resulting from extreme pose or clutter. **Best viewed on pdf.**

semble representation ranker (Fig. 5 rows are sorted by predicted attribute strength). There are limitations as well, especially for atypical images: e.g., "Visible teeth" (12th image) and "Visible forehead" (13th image) are incorrect due to mis-detections resulting from extreme pose/clutter. Finally, while the qualitative results are harder to interpret for

	BH	DH	EO	GL	ML	MO	S	VT	VF	Y	Mean
Keypoints [34]+DSIFT	**82.04**	80.56	83.52	68.98	**90.94**	82.04	**85.01**	82.63	83.52	71.36	81.06
Global [30]+DSIFT	68.98	73.89	59.40	62.23	87.93	57.05	65.82	58.77	71.48	66.74	67.23
Ours+DSIFT	78.47	**84.57**	**84.21**	**71.21**	90.80	**86.24**	83.90	**87.38**	**83.98**	**75.48**	**82.62**
Global [30]+CNN	78.10	83.09	71.43	68.73	**95.40**	65.77	63.84	66.46	81.25	72.07	74.61
Ours+CNN	**83.21**	**88.13**	**82.71**	**72.76**	93.68	**88.26**	**86.16**	**86.46**	**90.23**	**75.05**	**84.66**

Table 1. Attribute ranking accuracy (%) on LFW-10. Our approach outperforms the baselines for both dense-SIFT (first 3 rows) and CNN (last 2 rows) features. In particular, the largest performance gap between our approach and the Global [30] baseline occurs for attributes with localizable nature, e.g., "Mouth open". Using the same dense-SIFT features, our approach outperforms the Keypoints [34] baseline on 7 of 10 attributes but with less supervision; we do not use any facial landmark annotations for training. *BH–Bald head; DH–Dark hair; EO–Eyes open; GL–Good looking; ML–Masculine looking; MO–Mouth open; S–Smile; VT–Visible teeth; VF–Visible forehead; Y–Young.*

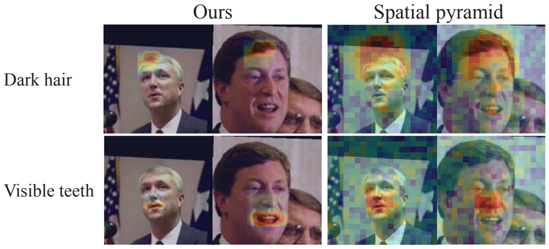

Figure 6. Spatial extent of attributes discovered by our approach vs. a spatial pyramid baseline. Red/blue indicates strong/weak attribute relevance. Spatial pyramid uses a fixed rigid grid (here 20x20), and so cannot deal with translation and scale changes of the attribute across images. Our approach is translation and scale invariant, and so its discoveries are much more precise.

the more global attributes like "Good looking" and "Masculine looking", quantitative analysis shows that they occupy a larger spatial extent than the more localizable attributes like "Mouth open" and "Smile" (see supp. for details).

In Fig. 6, we compare against the Global baseline. We purposely use a higher spatial resolution (20x20) grid for the baseline to make the visualization comparison fair. Since the baseline uses a fixed spatial pyramid rigid grid, it cannot deal with changes in translation or scale of the attribute across different images; it discovers the background clutter to be relevant to "Dark hair" (1st row, 3rd column) and the nose region to be relevant to "Visible teeth" (2nd row, 4th column). Our approach is translation and scale invariant, and hence its discoveries are much more precise.

4.2. Relative attribute ranking accuracy

We next evaluate relative attribute ranking accuracy, as measured by the percentage of test image pairs whose pairwise orderings are correctly predicted (see Eqn. 3).

We first report results on LFW-10 (Table 1). We use the same train/test split as in [34], and compare to the Global [30] and Keypoints [34] baselines. Our approach consistently outperforms the baselines for both feature types. Notably, even with the weaker dense-SIFT features, our method outperforms Global [30] that uses the more powerful CNN features for all attributes except "Masculine-looking", which may be better described with a global feature. This result demonstrates the importance of accurately discovering the spatial extent for relative attribute modeling. Compared to Keypoints [34], which also argues for the value of localization, our approach performs better but with less supervision; we do not use any facial landmark annotations during training. This is likely due to our approach being able to discover regions beyond pre-defined facial landmarks, which may not be sufficient in modeling the attributes.

	Open	Pointy	Sporty	Comfort	Mean
Yu and Grauman [42]	90.67	90.83	92.67	92.37	91.64
Global [30]+DSIFT	93.07	92.37	94.50	94.53	93.62
Ours+DSIFT	92.53	**93.97**	95.17	94.23	93.97
Ours+Global+DSIFT	**93.57**	93.83	**95.53**	**94.87**	**94.45**
Global [30]+CNN	94.37	93.97	95.40	95.03	94.69
Ours+CNN	93.80	94.00	96.37	95.17	94.83
Ours+Global+CNN	**95.03**	**94.80**	**96.47**	**95.60**	**95.47**

Table 2. Attribute ranking accuracy (%) on UT-Zap50K.

We also report ranking accuracy on UT-Zap50K (Table 2). We use the same train/test splits as in [42], and compare again to Global [30], as well as to the local learning method of [42]. Note that Keypoints [34] cannot be easily applied to this dataset since it makes use of pre-trained landmark detectors, which are not available (and much more difficult to define) for shoes. While our approach produces the highest mean accuracy, the performance gain over the baselines is not as significant compared to LFW-10. This is mainly because all of the shoe images in this dataset have similar scale, are centered on a clear white background, and face the same direction. Since the objects are so well-aligned, a spatial pyramid is enough to capture detailed spatial alignment. Indeed, concatenating the global spatial pyramid feature with our discovered features produces even better results (Ours+Global+DSIFT/CNN).[3]

Finally, we conduct a cross-dataset generalization experiment to demonstrate that our method is more robust to dataset bias [39] compared to Global [30]. We take the detectors and rankers trained on UT-Zap50K, and use them to make predictions on Shoes-with-Attributes. Table 3 shows the results. The performance for both methods is much lower because this dataset exhibits shoes with very different styles and much wider variation in scale and orientation.

[3] Produces worse results on LFW-10; images are not as well-aligned.

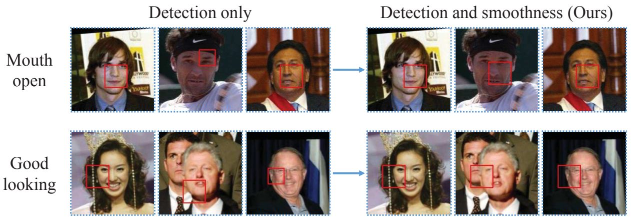

Figure 7. Three consecutive patches in two different visual chains, for "Mouth open" and "Good looking". (**left**) The middle patches are mis-localized due to the confusing patterns at the incorrect locations. (**right**) These errors are corrected by propagating information from neighbors when local smoothness is considered.

	Open	Pointy	Sporty	Mean
Global [30]+DSIFT	55.73	50.00	47.71	51.15
Ours+DSIFT	**63.36**	**62.50**	**55.96**	**60.61**
Global [30]+CNN	77.10	72.50	71.56	73.72
Ours+CNN	**80.15**	**82.50**	**88.07**	**83.58**

Table 3. Cross-dataset ranking accuracy (%), training on UT-Zap50K and testing on Shoes-with-Attributes.

Still, our method generalizes much better than Global [30] due to its translation and scale invariance.

4.3. Ablation Studies

Contribution of each term in Eqn. 2. We conduct an ablation study comparing our chains with those mined by two baselines that use either only the detection term or only the local smoothness term in Eqn. 2. For each attribute in LFW-10 and UT-Zap50K, we select the single top-ranked visual chain. We then take the same N_{init} initial patches for each chain, and re-do the iterative chain growing, but *without* the detection or smoothness term. Algorithmically, the detection-only baseline is similar to the style-aware mid-level visual element mining approach of [25].

We then ask a human annotator to mark the outlier detections that do not visually agree with the majority detections, for both our chains and the baselines'. On a total of 14 visual chains across the two datasets, on average, our approach produces 3.64 outliers per chain while the detection-only and smoothness-only baselines produce 5 and 76.3 outliers, respectively. The smoothness-only baseline often drifts during chain growing to develop multiple modes. Fig. 7 contrasts the detection-only baseline with ours.

Visual chain perturbations. As argued in Sec. 3.3, generating additional chains by perturbing the originally mined visual chains is not only an efficient way of increasing the size of the candidate chain pool, but also allows the discovery of non-distinctive regions that are hard to localize but potentially informative to the attribute. Indeed, we find that for each attribute, on average only 4.25 and 2.3 selected in the final 60-chain ensemble are the original mined chains, for UT-Zap50K and LFW-10, respectively. For example, in Fig. 5 "Open", the high response on the shoe opening (which may even be lacking as with the shoes in the 1st and 3rd columns) is due to the perturbed chains being anchored on more consistent shoe parts such as the tongue and heel.

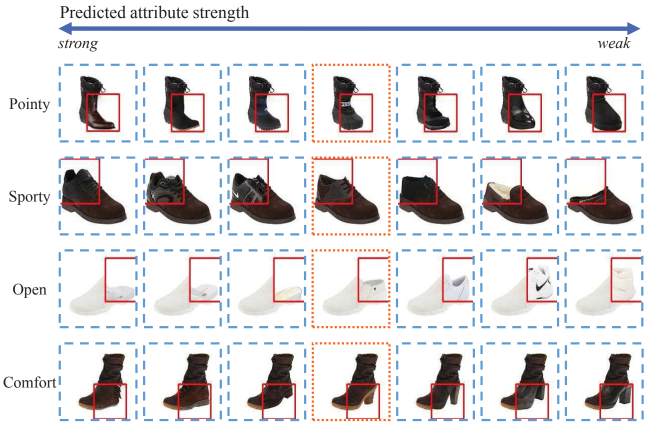

Figure 8. The middle column shows the query image whose attribute (automatically localized in red box) we want to edit. We synthesize new shoes of varying predicted attribute strengths by replacing the red box, which is predicted to be highly-relevant to the attribute, while keeping the rest of the query image fixed.

4.4. Application: Attribute Editor

One application of our approach is the *Attribute Editor*, which could be used by designers. The idea is to synthesize a new image, say of a shoe, by editing an attribute to have stronger/weaker strength. This allows the user to visualize the same shoe but e.g., with a pointier toe or sportier look. Fig. 8 shows four examples in which a user has edited the query image (shown in the middle column) to synthesize new images that have varying attribute strengths. To do this, we take the highest-ranked visual chain for the attribute, and replace the corresponding patch in the query image with a patch from a different image that has a stronger/weaker predicted attribute strength. For color compatibility, we retrieve only those patches that have similar color along its boundary as that of the query patch. We then blend in the retrieved patch using poisson blending.

Our application is similar to the 3D model editor of [5], which changes only the object-parts that are related to the attribute and keeps the remaining parts fixed. However, the relevant parts in [5] are determined manually, whereas our algorithm discovers them automatically.

Conclusion. We presented an approach that discovers the spatial extent of relative attributes. It uses a novel formulation that combines a detector with local smoothness to discover chains of visually coherent patches, efficiently generates additional candidate chains, and ranks each chain according to its relevance to the attribute. We demonstrated our method's effectiveness on several datasets, and showed that it better models relative attributes than baselines that either use global appearance features or stronger supervision.

Acknowledgements. This work was supported in part by an Amazon Web Services Education Research Grant and GPUs donated by NVIDIA.

References

[1] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *ICML*, 2009. 3

[2] T. L. Berg, A. C. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. In *ECCV*, 2010. 2

[3] L. Bourdev, S. Maji, and J. Malik. Describing People: Poselet-Based Approach to Attribute Classification. In *ICCV*, 2011. 1, 2

[4] S. Branson, C. Wah, F. Schroff, B. Babenko, P. Welinder, P. Perona, and S. Belongie. Visual recognition with humans in the loop. In *ECCV*, 2010. 1

[5] S. Chaudhuri, E. Kalogerakis, S. Giguere, and T. Funkhouser. Attribit: content creation with semantic attributes. In *UIST*, 2013. 1, 8

[6] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. A. Efros. What makes paris look like paris? *SIGGRAPH*, 2012. 2

[7] K. Duan, D. Parikh, D. Crandall, and K. Grauman. Discovering localized attributes for fine-grained recognition. In *CVPR*, 2012. 1, 2

[8] A. Faktor and M. Irani. "Clustering by Composition"–Unsupervised Discovery of Image Categories. In *ECCV*, 2012. 2

[9] A. Farhadi, I. Endres, and D. Hoiem. Attribute-centric recognition for cross-category generalization. In *CVPR*, 2010. 1

[10] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009. 1

[11] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *TPAMI*, 32(9):1627–1645, 2010. 5

[12] V. Ferrari and A. Zisserman. Learning visual attributes. In *NIPS*, 2008. 2

[13] K. Grauman and T. Darrell. Unsupervised learning of categories from sets of partially matching image features. In *CVPR*, 2006. 2

[14] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. B. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv:1408.5093*, 2014. 5

[15] T. Joachims. Optimizing Search Engines using Clickthrough Data. In *KDD*, 2002. 3, 4

[16] M. Kiapour, K. Yamaguchi, A. C. Berg, and T. L. Berg. Hipster wars: Discovering elements of fashion styles. In *ECCV*, 2014. 2

[17] A. Kovashka and K. Grauman. Attribute adaptation for personalized image search. In *ICCV*, 2013. 1

[18] A. Kovashka, D. Parikh, and K. Grauman. Whittlesearch: Image Search with Relative Attribute Feedback. In *CVPR*, 2012. 5

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 5

[20] N. Kumar, P. Belhumeur, and S. Nayar. FaceTracer: A Search Engine for Large Collections of Images with Faces. In *ECCV*, 2008. 2

[21] N. Kumar, A. Berg, P.Belhumeur, and S.Nayar. Attribute and simile classifiers for face verification. In *ICCV*, 2009. 1, 2

[22] P. Kumar, B. Packer, and D. Koller. Self-paced learning for latent variable models. In *NIPS*, 2010. 3

[23] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009. 1, 2

[24] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006. 5

[25] Y. J. Lee, A. A. Efros, and M. Hebert. Style-Aware Mid-level Representation for Discovering Visual Connections in Space and Time. In *ICCV*, 2013. 2, 3, 8

[26] Y. J. Lee and K. Grauman. Foreground focus: Finding meaningful features in unlabeled images. *IJCV*, 2008. 2

[27] Y. J. Lee and K. Grauman. Learning the easy things first: Self-paced visual category discovery. In *CVPR*, 2011. 3

[28] Q. Li, J. Wu, and Z. Tu. Harvesting mid-level visual concepts from large-scale internet images. In *CVPR*, 2013. 2

[29] M. Palatucci, D. Pomerleau, G. Hinton, and T. Mitchell. Zero-Shot Learning with Semantic Output Codes. In *NIPS*, 2009. 1

[30] D. Parikh and K. Grauman. Relative Attributes. In *ICCV*, 2011. 1, 2, 3, 4, 5, 7, 8

[31] N. Payet and S. Todorovic. From a set of shapes to object discovery. In *ECCV*, 2010. 2

[32] M. Rastegari, A. Farhadi, and D. Forsyth. Attribute discovery via predictable discriminative binary codes. In *ECCV*, 2012. 1

[33] B. Saleh, A. Farhadi, and A. Elgammal. Object-Centric Anomaly Detection by Attribute-Based Reasoning. In *CVPR*, 2013. 1

[34] R. N. Sandeep, Y. Verma, and C. V. Jawahar. Relative parts: Distinctive parts for learning relative attributes. In *CVPR*, 2014. 1, 2, 5, 7

[35] A. Shrivastava, S. Singh, and A. Gupta. Constrained Semi-supervised Learning using Attributes and Comparative Attributes. In *ECCV*, 2012. 1

[36] S. Singh, A. Gupta, and A. A. Efros. Unsupervised discovery of mid-level discriminative patches. In *ECCV*, 2012. 2, 4

[37] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering object categories in image collections. In *CVPR*, 2005. 2

[38] J. Supancic and D. Ramanan. Self-paced learning for long-term tracking. In *CVPR*, 2013. 3

[39] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *CVPR*, 2011. 7

[40] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*, 2010. 5

[41] S. Wang, J. Joo, Y. Wang, and S. C. Zhu. Weakly supervised learning for attribute localization in outdoor scenes. In *CVPR*, 2013. 2

[42] A. Yu and K. Grauman. Fine-Grained Visual Comparisons with Local Learning. In *CVPR*, 2014. 5, 7

[43] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev. PANDA: Pose Aligned Networks for Deep Attribute Modeling. In *CVPR*, 2014. 1, 2

[44] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, 2012. 5