Learning the Latent "Look": Unsupervised Discovery of a Style-Coherent Embedding from Fashion Images

Wei-Lin Hsiao
UT-Austin
kimhsiao@cs.utexas.edu

Kristen Grauman
UT-Austin
grauman@cs.utexas.edu

Abstract

What defines a visual style? Fashion styles emerge organically from how people assemble outfits of clothing, making them difficult to pin down with a computational model. Low-level visual similarity can be too specific to detect stylistically similar images, while manually crafted style categories can be too abstract to capture subtle style differences. We propose an unsupervised approach to learn a style-coherent representation. Our method leverages probabilistic polylingual topic models based on visual attributes to discover a set of latent style factors. Given a collection of unlabeled fashion images, our approach mines for the latent styles, then summarizes outfits by how they mix those styles. Our approach can organize galleries of outfits by style without requiring any style labels. Experiments on over 100K images demonstrate its promise for retrieving, mixing, and summarizing fashion images by their style.

1. Introduction

Computer vision methods that can understand fashion could transform how individual consumers shop for their clothing as well as how the fashion industry can analyze its own trends at scale. The scope for impact is high: fashion is already a $1.2 trillion USD global industry, popular social commerce websites like Chictopia and Polyvore draw millions of users, and online subscription services like StitchFix blend algorithms and stylists to personalize shopping selections. In sync with these growing possibilities, recent research explores new vision methods for fashion, with exciting advances for parsing clothing [39, 38], recognizing clothing, attributes [5, 40], and styles [20, 33], matching clothing seen on the street to catalogs [19, 25, 18, 37, 26], and recommending clothing [24, 28, 36, 14, 16].

Despite substantial progress on all these fronts, capturing the *style-coherent similarity* between outfits remains an important challenge. In particular, a visual representation with style coherency would capture the *relationship between clothing outfits that share a "look", even though they may differ in their specific composition of garments.* A style-coherent representation would be valuable for i) browsing, where a consumer wants to peruse diverse outfits similar in style, ii) recommendation, where a system should suggest new items that add novelty to a consumer's closet without straying from his/her personal style, and iii) style trend tracking, where analysts would like to understand the popularity of items over time.

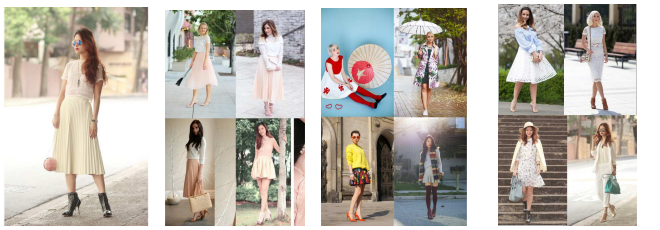

Figure 1: The leftmost query image relates to instances like those in (a),(b),(c) in distinct ways. In contrast to matching near-duplicate outfits (a) or classifying broad styles (b), we propose to discover the latent "looks"—compositions of clothing elements that are stylistically similar (c).

Style coherency differs from traditional notions of visual similarity. The problem of style coherency sits between the two extremes currently studied in the literature: on one end of the spectrum are methods that seek robust *instance matching*, e.g., to allow a photo of a garment seen on the street to be matched to a catalog [19, 25, 18, 37, 26] (see Figure 1(a)); on the other end of the spectrum are methods that seek *coarse style classification*, e.g., to label an outfit as one of a small number predefined categories like Hipster, Preppy, or Goth [20, 33] (see Figure 1(b)). In contrast to these two extremes, *style coherency* refers to consistent fine-grained trends exhibited by distinct assemblies of garments. In other words, coherent styles reflect some latent "look". See Figure 1(c).

We propose an unsupervised approach to learn a style-coherent representation. Given a large repository of unlabeled fashion images, the goal is to discover the latent factors that naturally guide how people dress—that is, the underlying compositions of visual attributes that define styles. To this end, we explore probabilistic topic models. Topic models in natural language processing [3, 31] represent text documents as distributions over concepts, each of which is a distribution over words. In our case, an outfit is a "document", a predicted visual attribute (e.g., *polka dotted*, *flowing*, *wool*) is a "word", and each style is a discovered "topic". Furthermore, we consider *polylingual* topic models [29] in order to model style consistency across multiple regions of the body, enforcing that the discovered style factors for each region of the body should interact compatibly. Our idea makes it possible to organize galleries of outfits by style without any style labels. Building on the proposed model, we develop methods to *mix* styles or *summarize* an image gallery by its styles.

The proposed approach is well-suited for the problem at hand, for several reasons. First, being unsupervised, our algorithm *discovers* the underlying elements of style, as opposed to us manually defining them. It is often difficult to manually craft labels, especially in the domain of fashion. Clothing styles emerge organically from instances of what people choose to wear—not from some top-down preordained bins of outfit types—and furthermore they evolve over time, meaning today's hand-crafted lexicon will eventually fade in relevance. Secondly, our approach accounts for the fact that style is about the Gestalt: individual items do not dictate a style; rather, it is their *composition* that creates a look [2]. Finally, our topic model approach also naturally accounts for the soft boundaries of style, describing outfits as mixtures of overlapping styles.

Our experiments on two challenging fashion datasets [20, 26] demonstrate the advantages of our unsupervised representation compared to style-based CNNs and more basic attribute descriptions. We validate that our styles align well with human perceived styles. We further show their value for retrieval, mixing, and summarization. Finally, we introduce a new dataset of 19K images labeled for fine-grained, body-localized attributes relevant for fashion analysis.

2. Related Work

Attributes for fashion Describable attributes, such as *floral*, *denim*, *long-sleeved*, are often of interest in analyzing fashion images. Prior work explores a variety of recognition schemes [9, 40, 5, 4, 26, 7], including ways to jointly recognize attributes [40, 5] or simultaneously detect clothing articles and attributes [4]. In our work, attributes serve as a starting point for discovering styles (i.e., the "words" in our topic model), and improvements in attribute prediction from work like the above would also benefit our model.

Retrieval and matching for clothing Given an image of a garment, clothing retrieval methods identify exact or close matches, which supports shopping needs. Clothing deformations and complex body poses make retrieval challenging, so researchers develop ways to learn the similarity of outfits or garments [12, 25, 18, 37, 23]. Matching clothing seen "on the street" to instances stored in catalogs "at the shop" requires new ideas in domain adaptation [19, 15, 18, 25]. Existing methods are largely supervised, i.e., provided with street-shop pairs or pairs of outfits judged as similar by human annotators. In contrast, we develop an unsupervised approach that can leverage ample unlabeled data. More importantly, whereas retrieval work aims to match the same (or similar) garment(s), we aim to identify style-coherent complete outfits.

Models of style, fashionability, compatibility Previous work offers a few perspectives on the meaning of visual *style*. The Hipster Wars project defines five style categories (*Hipster*, *Goth*, *Preppy*, *Pinup*, *Bohemian*) and recognizes them based on patches on body part keypoints [20]. Another approach pre-trains a neural network for style using weak meta-data labels [33]. Compared to the retrieval work above, both (like us) aim to capture a broader notion of style. However, unlike [20, 33] we treat styles as discoverable latent factors rather than manually defined categories, which has the advantages discussed in the Introduction.

Whereas style refers to a characterization of whatever it is people wear, *compatibility* refers to how well-coordinated individual garments are [36, 16, 24], and *fashionability* refers to the popularity of clothing items, e.g., as judged by the number of "like" votes on an image posted online [32]. Recent work explores forecasting the popularity of styles [1].

Topic models Topic models originate in text processing. The well-known topic model Latent Dirichlet Allocation (LDA) [3] and its polylingual extension [29] use multinomial distributions to represent the generation of documents comprised of words. Polylingual topic models are applied to Web fashion data in [35] to discover links between textual design element meta-data and textual style meta-data, with no computer vision. Early uses of topic models in vision relied on "visual words" (quantized image patches) to discover representations for scene recognition [10] or perform object category discovery in unlabeled images [34]. More recently, topic models are used to recommend a color-coordinating garment in [16]. To our knowledge, we are the first to propose discovering visual styles for outfits using topic models. Deriving topic models on top of semantic visual attributes is also new, and has the added benefit of yielding interpretable latent topics.

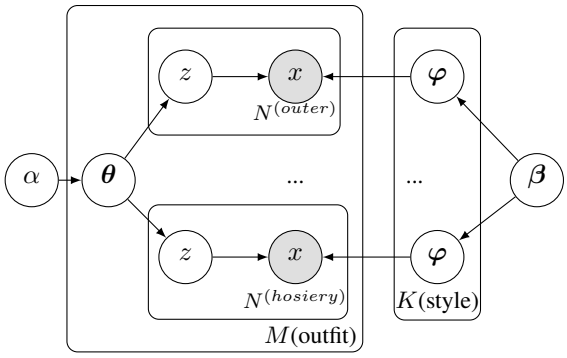

Figure 2: Graphical model of the polylingual visual style LDA.

3. Approach

After providing background on topic models in Section 3.1, we introduce our visual fashion topic model in Section 3.2. Next in Section 3.3 we overview the fine-grained localized attributes used in our model. Finally, we leverage the learned style-coherent embedding for retrieval, mixing, and summarization tasks in Section 3.4.

3.1. Background: Topic models

We explore unsupervised topic models originating from text analysis to discover visual styles. In particular, we employ Latent Dirichlet Allocation (LDA) [3]. LDA is a Bayesian multinomial mixture model that supposes a small number of K latent *topics* account for the distribution of observed words in any given document. It uses the following generative process for a corpus D consisting of M documents each of length N_i:

1. Choose $\boldsymbol{\theta_i} \sim \mathrm{Dir}(\alpha)$, where $i \in \{1, \ldots, M\}$ and $\mathrm{Dir}(\alpha)$ is the Dirichlet distribution for parameter α
2. Choose $\boldsymbol{\varphi_k} \sim \mathrm{Dir}(\beta)$, where $k \in \{1, \ldots, K\}$
3. For each word indexed by i, j, where $j \in \{1, \ldots, N_i\}$, and $i \in \{1, \ldots, M\}$
   (a) Choose a topic $z_{i,j} \sim \mathrm{Multinomial}(\boldsymbol{\theta_i})$
   (b) Choose a word $x_{i,j} \sim \mathrm{Multinomial}(\boldsymbol{\varphi_{z_{i,j}}})$

Only the word occurrences are observed.

Polylingual LDA [29] extends LDA (we will call it MonoLDA) to process an aligned corpus of documents expressed in multiple languages. The idea is to recover topics that preserve the ties between translated text. In particular, translated documents form a tuple, and all documents in a tuple have the same distribution over topics. Each topic is produced from a set of distributions over words, one distribution per language.

3.2. Discovering a style-coherent embedding

We propose to learn a style-coherent embedding using a topic model. In our setting, the latent topics will be discovered from unlabeled full-body fashion images, meaning images of people wearing an entire outfit (as opposed to catalog images of individual garments). In this way, we aim to discover the compositions of lower-level visual cues that characterize the main visual themes—styles—emerging in how people choose to assemble their outfits.

The basic mapping from document topic models to our visual style topic models is as follows: an observed outfit is a "document", a predicted visual attribute (e.g., *polka dotted*, *flowing*, *wool*) is a "word", and each style is a discovered "topic".

A potential limitation of this mapping, however, is that it treats an outfit as a bag of attributes. Thus it loses valuable information about the attributes' associated articles of clothing. For example, learned topics could interchange the appearance of *wool* pants with a *wool* jacket, when the two may in reality signify distinct latent styles. A partial solution is to specify localized attributes. For example, we could expand *wool* to *wool pants* and *wool jacket*. However, this expansion may suffer from allowing LDA to decouple topics across different regions of the body. For example, in Figure 3 (left), MonoLDA dedicates topic 1 to shirt and topic 2 to skirt.

Thus, we consider a polylingual LDA (PolyLDA) model [29]. In this case, each region of the body is a "language", and an outfit is a "document tuple" in multiple languages. As above, latent styles are topics and inferred clothing attributes are words. The body regions (denoted as R) we consider are: *outer layer* (i.e., where a jacket or blazer goes), *upper body* (shirt/blouse/sweater), *lower body* (pants/skirt/shorts), and *hosiery* (tights/leggings). The polylingual topic model adds a structural constraint that forces body regions to *share* styles, such that we can learn styles consistent across body regions. The generative process of PolyLDA is as follows:

1. For each topic $k \in \{1, \ldots, K\}$
   (a) For each body part $r \in R$
      i. Choose attribute distribution $\boldsymbol{\varphi_k^{(r)}} \sim \mathrm{Dir}(\beta)$
2. For each outfit $i \in \{1, \ldots, M\}$
   (a) Choose style distribution $\boldsymbol{\theta_i} \sim \mathrm{Dir}(\alpha)$
   (b) For each body part $r \in R$
      i. For each attribute belonging to that body part, indexed by i, j, where $j \in \{1, 2, \ldots, N_i^{(r)}\}$
         – Draw a style $z_{ij}^{(r)} \sim \mathrm{Mult}(\boldsymbol{\theta_i})$
         – Draw an attribute $x_{ij}^{(r)} \sim \mathrm{Mult}(\boldsymbol{\varphi_{z_{ij}^{(r)}}^{(r)}})$

Figure 2 shows the associated graphical model. In contrast to MonoLDA, PolyLDA captures the interaction of garment regions, such that each style specifies a full-body trend (Figure 3, right).

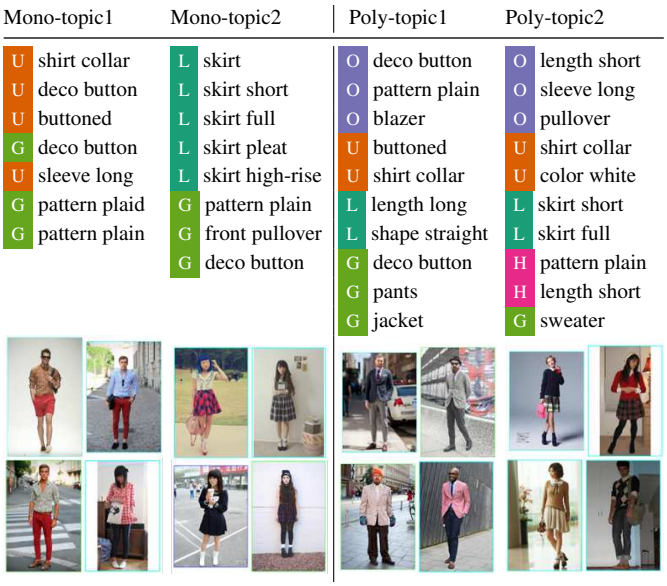

Figure 3: Mono vs. Polylingual LDA: U for *upper body*, O for *outer layer*, L for *lower body*, H for *hosiery* and G for *global*. MonoLDA (left) learns a topic either for U or L , while PolyLDA's styles (right) span the whole body.

By applying PolyLDA to a database of unlabeled outfit images, we obtain a set of discovered styles (see Figure 7 in our experiment section for examples) with which to encode novel images. Each topic k has its attribute probability $\boldsymbol{\varphi}_k^{(r)}$ depending on body region r. Given an outfit d, we represent it in a style-coherent embedding by its topic proportions:

$$\boldsymbol{\theta}_d = [\theta_{d1}, \ldots, \theta_{dK}], \quad (1)$$

where $\theta_{dk} \geq 0$, $\Sigma_k \theta_{dk} = 1$. The resulting embedding accounts for the fact that a *composition* of style elements defines a look [2].

We stress that our style-coherent embedding is fully unsupervised. Our method discovers styles from unlabeled images, as opposed to learning a style embedding with supervision. For example, one could gather pairs of fashion images and ask human annotators to label them as similar or dissimilar in style (or use noisy tags as labels [33]), then learn an embedding that keeps similar pairs close. Or, in the spirit of [19], one could train classifiers to target a predefined set of style categories. While the attribute models are trained on a disjoint pool of attribute labeled images, our style model runs on *predicted* attributes; annotators do not touch the images on which we perform discovery. Our unsupervised strategy saves manual effort. More importantly, it also addresses challenges specific to visual styles—namely, their ever-evolving nature, the difficulty in enumerating them with words, and their soft boundaries.

3.3. Fine-grained localized fashion attributes

We next discuss our approach to infer attributes in full-body fashion images. We consider both global and localized attributes. Global attributes indicate the presence of a property somewhere on the body (e.g., *floral*), whereas a localized attribute links it specifically to a body region (e.g., *floral-shirt* and *floral-skirt* are distinct words). Figure 4 shows an example image and the attributes from either vocabulary that are present, as well as the body region association for the localized ones.

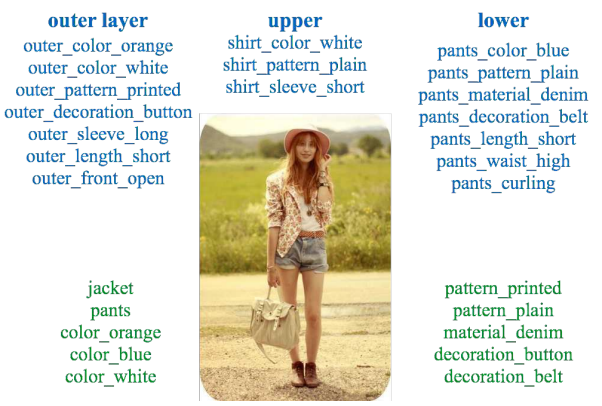

Figure 4: Attributes present for an outfit in the *localized* vocabulary (top) or *global* vocabulary (bottom).

Vocabulary and data collection As input to our style discovery model, we need a rich attribute vocabulary that is both localized and fine-grained. In existing fashion datasets, the attributes lack one or both of these aspects [4, 5, 9, 40] or are not publicly available [25]. Thus, we curate a new dataset for attribute training.

For the vocabulary, we build on the 53 attributes enumerated in [25]. First we remove those too subtle for most annotators to discern (*chiffon*; *jewel* collar). Then we add missing but frequently appearing attributes (e.g., *pink*; *polka-dot*). Finally, we expand the set so that *color*, *material*, and *pattern* are localized to each body region. This yields 195 total attributes (see Supp for details).

To gather images, we use keyword search with the attribute name on Google, then manually prune those where the attribute is absent. This yields 70 to 600 positive training images per attribute. We also gather 2000 random street images from chictopia.com (manually pruned for false-negatives) to serve as negative examples. In total, the new dataset has 18,878 images.[1]

Training with multilabel outfits The clothing outfit images are multilabel in terms of their attributes. To circumvent the expense of labeling all 19K images for all 195 attributes, and to deal appropriately with highly localized attributes, we develop a piecewise training procedure. First we group the attributes into six types: *pattern*, *material*, *shape*, *collar type*, *clothing article*, and *color*. The types

[1] vision.cs.utexas.edu/projects/StyleEmbedding/

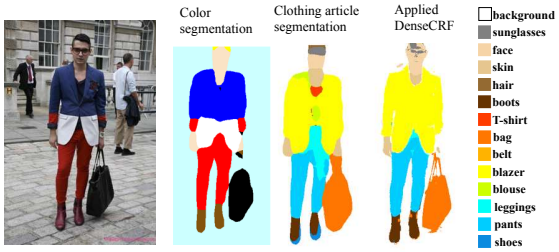

Figure 5: We intersect the article and color labels to generate the final *blazer-color-blue*, *pants-color-red* prediction.

have 105, 15, 20, 8, 27, 13 attributes, respectively.[2] Then we train separate convolutional neural networks (CNN) per type. This allows us to directly use the positive examples for each attribute, and all others from other keywords as negatives, while still yielding predictions at test time that are multilabel. We find it is also important during training to have negatives with none of the named attributes present, since such outfits are rather common.

Our attribute learning framework accounts for the challenge that many attributes occupy a small portion of a fashion image. First, we detect people using faster-RCNN [30], and extract a crop from the detected person bounding box according to whether the image is a training image for an upper or lower body attribute. We give the network both the whole-body person box and the crop as two instances with the same attribute label, allowing it to leverage any useful cues. We fine-tune our attribute prediction networks (one per type for the first four types) on ResNet-50 [13] pre-trained with ImageNet [8].

For both the *clothing article* and *color* types, we train a segmentation network. For both, we fine-tune DeepLab's [6] repurposed VGG-16 network and apply DenseCRF [22]. The networks target 27 pixel-wise clothing article labels and 13 pixel-wise color labels from the Fashionista data [38], respectively. At test time, we i) record the detected clothing article names, and ii) intersect the color and clothing semantic segmentations to produce article-specific color attributes, e.g., *shirt-color-blue* (Fig. 5).

The resulting attribute classifiers offer a fairly reliable basis for style discovery. For the validation split of our 19K image dataset, they attain 90% average precision. Figure 6 shows attribute predictions on novel test images.

3.4. Using the style-coherent embedding

Our method produces a style-coherent representation for fashion images. Our experiments consider three tasks leveraging this representation, defined next.

Retrieval of style-related images For retrieval, the system is given a query image and must return database images

[2] Attributes within *pattern* and *collar* types are mutually exclusive, thus their multilabeling can be done efficiently. We obtain complete 20-label and 15-label multilabeling for types *material* and *shapes*.

	Upper	Lower	Hosiery	Global	
	blue purple white	short dr. loose dr. flat dr. white dr. blue dr. purple dr.	blue	floral translucent dress shoe stocking red	blue purple white beige cardigan
	floral blue green beige	floral dr. tight dr. flat dr. blue dr. green dr. beige dr.		floral dress blue green beige	

Figure 6: Example of predicted attributes on a HipsterWars [20] (top) image and a DeepFashion [26] (bottom) image (dr.=dress)

that illustrate similar style. Here we simply use our learned embedding to retrieve images close to a query image, i.e., nearest neighbors in the space of $\boldsymbol{\theta}_q$ for query image q.

Our embedding retrieves images that maintain style coherence with the query. While conventional embeddings (e.g., CNNs) can return the examples *closest in appearance*, our embedding can return those that are *close in style*. Whereas the former is preferable when doing street-to-shop visual matching [19, 15, 18, 25], the latter is preferable for many browsing scenarios, e.g., to view recommendations related to past purchases.

Mixing and "traversing between" styles A new task supported by our approach is to *mix* fashion styles. In this scenario, a user identifies T styles of interest $\mathbf{S} := \{S_1, \ldots, S_T\}$, $S_t \in \{1, \ldots, K\}$, and queries for images that exhibit a blend of those styles. For example, the user could manually select styles of interest by viewing images associated with each discovered style, or the styles for mixing could be automatically discovered based on the dominant styles in his photo album or shopping history. We measure the relevance of an image I_i as:

$$\text{MixRelevance}(\boldsymbol{\theta}_i, \mathbf{S}) = \min_{t \in \mathbf{S}} \theta_{it}, \tag{2}$$

where $\boldsymbol{\theta}_i$ is the style embedding for I_i. The min assures that an image is only as relevant to the requested style mix as it is close to its most distant style. Similarly, we can offer new browsing capabilities by depicting a gradual transition from one style to another (see Figure 9).

Summarizing styles A third application uses our style-coherent embedding to summarize the styles in an image collection. Given images $\{I_1, \ldots, I_N\}$, we calculate the relative influence of each style k as $\text{Influence}(S_k) = \sum_{i=1}^{N} \boldsymbol{\theta}_{ik}$. With these frequencies we can visualize the collection compactly by sampling images dominant for each influential style (Figure 11 in our experiment section).

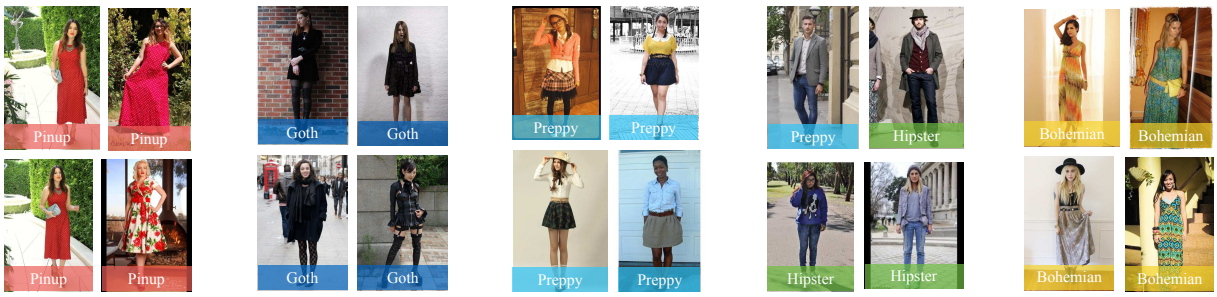

Figure 7: Top images for the five **discovered style topics** with PolyLDA. Labels indicate human-assigned styles from HipsterWars [20], which are *not* seen by our algorithm. Our approach successfully discovers the five human-perceived styles. Please see Supp for more examples and baseline cluster results.

	Avg AP	NMI
MonoLDA	0.48	0.30
PolyLDA	**0.53**	**0.31**

Table 1: Mono vs. poly LDA discovery accuracy judged against the manual style labels of HipsterWars, using GT attributes.

4. Experiments

We first show that our discovered topics align with human-perceived styles (Section 4.1). Then, we apply the embedding for retrieval (Section 4.2), mixing (Section 4.3), and summarizing styles (Section 4.4).

Datasets We use two datasets: (i) HipsterWars [20], which has 1,893 images, each labeled by one of 5 style labels: Hipster, Preppy, Goth, Pinup, Bohemian; (ii) DeepFashion [26], from which we take all 108,145 images that have at least one of the 230 style labels *and* a fully-visible person. Because the 230 style labels in DeepFashion are noisy labels, we collapse them into 42 higher level styles by affinity propagation [11] using cosine similarity to measure co-occurrence of styles in an outfit (see Supp). We use the attribute networks trained with our new 19K image dataset (Section 3.3) to predict the attributes in the HipsterWars and DeepFashion images.

Baselines We compare with four baseline: (i) StyleNet [33]: a state-of-the-art feature for clothing that fine-tunes a CNN using 123 metadata labels (e.g., red-sweater) on images from the Fashion 144K dataset [32], (ii) vanilla ResNet-50 [13]: the last layer of a state-of-the-art CNN pretrained for ImageNet, (iii) Attr-ResNet: ResNet-50 fine-tuned to classify the same 148 attributes[3] used by our method with the same training data, and (iv) Attributes: indicator vectors using the same attributes as our model.

[3]The attributes in types: *pattern*, *material*, *shape*, *collar type*; we did not include *clothing article* and *color* because they are predicted differently, from a segmentation network.

4.1. Consistency with human labeled styles

First we analyze how well our discovered styles align with human perception, as captured by the datasets' style labels (never seen by our approach). We use two metrics: (i) Normalized Mutual Information (NMI), which captures the overall alignment of topics with ground truth (GT) styles; and (ii) averaged maximal average precision (AP) per style, which uses each topic's probability as a relevance score for a style to sort all images, then records the AP per topic per style. The best (max) AP a style has in all topics is that style's final score. We average the max AP scores of all styles (5 in HipsterWars and 200+ in DeepFashion) to get "avg maximal AP". To extract clusters for the baseline representations we use K-means clustering. We also tried GMM and AP-clustering and found the clustering algorithm itself has negligible impact on their results. For all methods, we set the number of clusters/topics to be the number of style labels in the respective dataset.

First we examine the impact of our polylingual model. Table 1 shows the results for both LDA variants on HipsterWars. Here we use ground truth attributes, in order to evaluate the LDA models independent of attribute prediction quality. We see that the polylingual model has an advantage, and thus adopt it as our model for all experiments.

Next we quantify discovery accuracy for our approach against the baselines. Table 2 shows the results on both datasets. For the attribute-indicator baseline and our approach, we show results with predicted and ground truth attributes in order to separate the success of discovery from the success of attribute prediction.[4] Overall, PolyLDA is the strongest. Both PolyLDA and Attributes perform better with perfect attributes, reinforcing that attribute precision is an important ongoing research challenge [40, 5, 4]. However, even with predicted attributes, we outperform all baselines on both datasets for both metrics. Despite having been pretrained to capture noisy fashion labels, the StyleNet CNN [33] does not discover the human-perceived styles as

[4]DeepFashion has GT for only global attributes.

well, though it does soundly outperform the vanilla ResNet baseline. Attr-ResNet falls in between StyleNet and vanilla ResNet, as expected. The absolute accuracy on DeepFashion is much lower for all methods, a function of its larger size and more varied and noisy style labels. Whereas the Hipster style labels are manually curated through a rigorous crowdsourcing procedure [20], the DeepFashion style labels are gleaned from text meta-data [26].

Figure 7 shows the most central images for our discovered styles on HipsterWars (see Supp for DeepFashion). The qualitative examples reinforce the quantitative result above. Our model discovers the human-perceived styles better than the CNN and attributes clusters (see Supp). Our style-coherent embedding better tolerates superficial visual differences in intra-style images.

4.2. Style-coherent retrieval

Having shown that the discovered styles are meaningful, next we evaluate the style embedding for retrieval. In this task, a user queries by example for images related by style, e.g., for recommendation or catalog browsing relative to some currently viewed item (query). Recall, this is distinct from instance retrieval for near-duplicates. Thus, we evaluate performance simultaneously by *style coherence* and *diversity* or *novelty*. Diversity refers to the retrieved images' mutual visual dissimilarity with each other, and novelty refers to their collective dissimilarity to the query. The goal is to obtain retrieval results that maintain style coherence while avoiding redundancy. For style coherence, we evaluate NDCG [27] against ground truth style labels. For diversity/novelty, we learn a metric to mimic human-perceived dissimilarity, following [21]. In particular, we collect 350 triplets labeled by 5 human annotators and learn a ranking function [17] on top of the attribute and CNN descriptors that respects human-given judgments.

Figure 8 shows the results for both datasets. For HipsterWars (top), we treat each image as a query in turn, and for DeepFashion (bottom) we sample 2,000 of the 108,145 images as queries. Our model offers a good combination of coherency and diversity/novelty. On HipsterWars, it maintains diversity/novelty while maintaining a similar or better level of coherence as the baselines. As before, predicted attributes diminish coherence, yet the topic model coherence appears to degrade more gracefully.

4.3. Mixing styles

Next we consider mixing styles. Since evaluation of mixing requires images labeled for multiple human-perceived styles as well as instances exhibiting exclusively one style, we collect a ground truthed test set of 177 Web images using the HipsterWars style names (see Supp). While our mixing approach (Sec 3.4) can blend arbitrary selected styles, for sake of evaluation we focus on blending pairs of GT-labeled styles, then score the AP against the ground truth, i.e., images exhibiting both the initial selected styles. For the baselines, we use their clusters analogously to our topics, creating K-dim embeddings that record the distance of the image to each cluster's centroid. We use $K = 25$ topics/clusters; $K \in (15, 30)$ gives similar results.

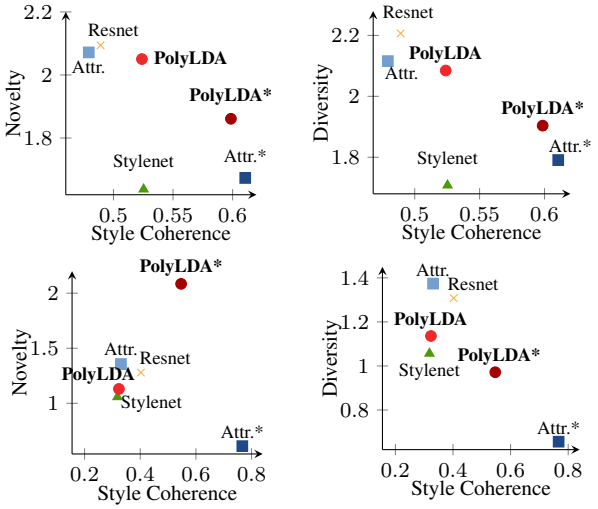

Figure 8: Style retrieval on HipsterWars (top) and DeepFashion (bottom). * denotes use of GT attributes. The ideal method would sit in the top right corner of the plots. Our embedding offers a good trade-off in style coherency and diversity/novelty.

	HipsterWars		DeepFashion	
	Avg AP	NMI	Avg AP	NMI
StyleNet [33]	0.39	0.20	0.0501	0.0011
ResNet [13]	0.30	0.16	0.0524	0.0004
Attr-ResNet	0.35	0.18	0.0615	0.0002
Attributes	0.28 / 0.32	0.19 / 0.28	0.0560 / 0.1294	0.0017 / 0.0082
PolyLDA	0.50 / **0.53**	0.21 / **0.31**	0.0647 / **0.1762**	0.0116 / **0.0227**

Table 2: Discovery accuracy for both datasets. Attributes and PolyLDA show result if using either predicted attributes (first) or ground truth attributes (second).

Table 3 shows the results. On the whole, our approach does better than the baselines, and in most cases this is true even using predicted attributes. This result highlights the power of the topic model over the raw attributes, which are too low-level for adequate mixing.

Figure 10 shows example images predicted as strong exemplars for two style blends. Figure 9 shows an example gradually mixing from a source style to a target style.

4.4. Style summaries

Finally, we demonstrate the power of our model to organize galleries of outfits. As proof of concept, we select two users from chictopia.com, and download 200 photos from each of their albums. Figure 11 shows the results. We show snapshots from their albums along with summary piecharts

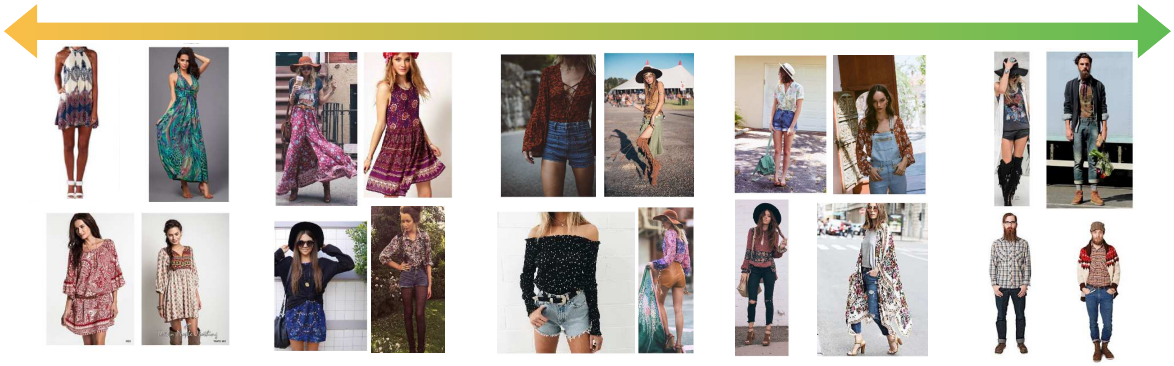

Figure 9: Visualization generated by mixing our style topics, gradually traversing from one style (Bohemian) to another (Hipster).

(a) Hipster×Bohemian

(b) Hipster×Goth

Figure 10: Top retrievals for two mixes. Incorrect images are labeled with red actual labels: ***H**ipster*, ***G**oth*, ***P**reppy*.

	Preppy× Goth	Hipster× Goth	Preppy× Hipster	Goth× Bohemian	Bohemian× Hipster
StyleNet [33]	0.133	0.187	0.128	0.141	0.113
ResNet [13]	0.267	0.172	0.162	0.197	0.097
Attributes	0.175 / 0.136	0.172 / 0.115	0.050 / 0.096	0.185 / 0.132	0.090 / 0.198
PolyLDA	0.178 / **0.303**	**0.424** / 0.180	0.191 / **0.266**	0.130 / **0.281**	0.139 / **0.394**

Table 3: Accuracy (AP) of retrieving a mixture of styles.

Figure 11: Style summarization for two users. Left is the user's album, right is the visual style summary breaking down the main trends discovered in the album.

computed by our approach to highlight the dominant styles. Gray pie slices indicate insignificant styles for a user. Our summaries convey the user's tastes in a glance. In contrast, the status quo would entail manually paging through all 200 photos in the album in an arbitrary order.

5. Conclusion

This work explores unsupervised discovery of complex styles in fashion. We develop an approach based on polylingual topic models to model the composition of outfits from visual attributes. The resulting styles offer a fine-grained representation valuable for organizing unlabeled fashion photos beyond their superficial visual ties (e.g., same literal garments or attributes). While by necessity our results rely on external style labels for evaluation, we stress that the generality of the discovered styles is an asset, and they offer representational power beyond what traditional (supervised) classification schemes can do. Our example results highlighting blended styles, trajectories between styles, and style summaries suggest a few such applications of interest.

Acknowledgements: We thank Suyog Jain and Chao-Yuan Wu for helpful discussions. This research is supported in part by NSF IIS-1065390 and a gift from Amazon.

References

[1] Z. Al-Halah, R. Stiefelhagen, and K. Grauman. Fashion forward: Forecasting visual style in fashion. In *ICCV*, 2017. 2

[2] R. Barthes. *The language of fashion*. A&C Black, 2013. 2, 4

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *JMLR*, 2003. 2, 3

[4] L. Bossard, M. Dantone, C. Leistner, C. Wengert, T. Quack, and L. Van Gool. Apparel classification with style. In *ACCV*, 2012. 2, 4, 6

[5] H. Chen, A. Gallagher, and B. Girod. Describing clothing by semantic attributes. In *ECCV*, 2012. 1, 2, 4, 6

[6] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *ICLR*, 2015. 5

[7] Q. Chen, J. Huang, R. Feris, L. M. Brown, J. Dong, and S. Yan. Deep domain adaptation for describing people based on fine-grained clothing attributes. In *CVPR*, 2015. 2

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 5

[9] W. Di, C. Wah, A. Bhardwaj, R. Piramuthu, and N. Sundaresan. Style finder: Fine-grained clothing style detection and retrieval. In *CVPR*, 2013. 2, 4

[10] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, 2005. 2

[11] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 2007. 6

[12] J. Fu, J. Wang, Z. Li, M. Xu, and H. Lu. Efficient clothing retrieval with semantic-preserving visual phrases. In *ACCV*, 2012. 2

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5, 6, 7, 8

[14] Y. Hu, X. Yi, and L. Davis. Collaborative fashion recommendation: A functional tensor factorization approach. In *ACM MM*, 2015. 1

[15] J. Huang, R. Feris, Q. Chen, and S. Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In *ICCV*, 2015. 2, 5

[16] T. Iwata, S. Watanabe, and H. Sawada. Fashion coordinates recommender system using photographs from fashion magazines. In *IJCAI*, 2011. 1, 2

[17] T. Joachims. Training linear svms in linear time. In *ACM SIGKDD*, 2006. 7

[18] Y. Kalantidis, L. Kennedy, and L.-J. Li. Getting the look: Clothing recognition and segmentation for automatic product suggestions in everyday photos. In *ICMR*, 2013. 1, 2, 5

[19] M. H. Kiapour, X. Han, and S. Lazebnik. Where to buy it: Matching street clothing photos in online shops. In *ICCV*, 2015. 1, 2, 5

[20] M. H. Kiapour, K. Yamaguchi, A. Berg, and T. Berg. Hipster wars: Discovering elements of fashion styles. In *ECCV*, 2014. 1, 2, 5, 6, 7

[21] A. Kovashka, D. Parikh, and K. Grauman. Whittlesearch: Image search with relative attribute feedback. In *CVPR*, 2012. 7

[22] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, 2011. 5

[23] K. Lin, H.-F. Yang, K.-H. Liu, J.-H. Hsiao, and C.-S. Chen. Rapid clothing retrieval via deep learning of binary codes and hierarchical search. In *ACM ICMR*, pages 499–502, 2015. 2

[24] S. Liu, J. Feng, Z. Song, T. Zheng, H. Lu, C. Xu, and S. Yan. Hi, magic closet, tell me what to wear! In *ACM MM*, 2012. 1, 2

[25] S. Liu, Z. Song, G. Liu, C. Xu, H. Lu, and S. Yan. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *CVPR*, 2012. 1, 2, 4, 5

[26] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016. 1, 2, 5, 6, 7

[27] C. D. Manning, P. Raghavan, H. Schütze, et al. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008. 7

[28] J. McAuley, C. Targett, Q. Shi, and A. van den Hengel. Image-based recommendations on styles and substitutes. In *SIGIR*, 2015. 1

[29] D. Mimno, H. M. Wallach, J. Naradowsky, D. A. Smith, and A. McCallum. Polylingual topic models. In *EMNLP*, 2009. 2, 3

[30] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*. 2015. 5

[31] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *UAI*, 2004. 2

[32] E. Simo-Serra, S. Fidler, F. Moreno-Noguer, and R. Urtasun. Neuroaesthetics in Fashion: Modeling the Perception of Fashionability. In *CVPR*, 2015. 2, 6

[33] E. Simo-Serra and H. Ishikawa. Fashion Style in 128 Floats: Joint Ranking and Classification using Weak Data for Feature Extraction. In *CVPR*, 2016. 1, 2, 6, 7, 8

[34] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman. Discovering Objects and their location in Images. In *ICCV*, 2005. 2

[35] K. Vaccaro, S. Shivakumar, Z. Ding, K. Karahalios, and R. Kumar. The elements of fashion style. In *ACM UIST*, 2016. 2

[36] A. Veit, B. Kovacs, S. Bell, J. McAuley, K. Bala, and S. Belongie. Learning visual clothing style with heterogeneous dyadic co-occurrences. In *ICCV*, 2015. 1, 2

[37] S. Vittayakorn, K. Yamaguchi, A. C. Berg, and T. L. Berg. Runway to realway: Visual analysis of fashion. In *WACV*, 2015. 1, 2

[38] K. Yamaguchi, M. Hadi Kiapour, and T. L. Berg. Paper doll parsing: Retrieving similar styles to parse clothing items. In *ICCV*, 2013. 1, 5

[39] K. Yamaguchi, H. Kiapour, L. Ortiz, and T. Berg. Parsing clothing in fashion photographs. In *CVPR*, 2012. 1

[40] K. Yamaguchi, T. Okatani, K. Sudo, K. Murasaki, and Y. Taniguchi. Mix and match: Joint model for clothing and attribute recognition. In *BMVC*, 2015. 1, 2, 4, 6