Trip Outfits Advisor: Location-Oriented Clothing Recommendation

Xishan Zhang, Jia Jia, Ke Gao, Yongdong Zhang, Dongming Zhang, Jintao Li and Qi Tian

Abstract—When packing for a journey, have you ever asked 'what clothes should I take with me?' Wearing appropriate and aesthetically pleasing clothing when traveling is a concern for many of us. Our data observation of photos from several popular travel websites reveals that people's choice of clothing items and their color combinations have strong correlations with the weather, the season, and also the main type of attraction at the destination. This leads to an interesting and novel problem: can the correlation between clothing and locations be automatically learned from social photos and leveraged for location-oriented clothing recommendations? In this paper, we systematically study this problem and propose a hybrid multi-label convolutional neural network combined with the support vector machine (mCNN-SVM) approach to capture the intrinsic and complex correlations between clothing attributes and location attributes. Specifically, we adapt the CNN architecture to multi-label learning and fine-tune it using each fine-grained clothing item. Then, the recognized items are fed to the SVM to learn the correlations. Experiments on three fashion datasets and a benchmark Journey Outfit Dataset show that our proposed approach outperforms several baselines by over 10.52-16.38% in terms of the mAP for clothing item recognition and outperforms several alternative methods by over 9.59-29.41% in terms of the mAP when ranking clothing by appropriateness for travel destinations. Finally, an interesting case study demonstrates the effectiveness of our method by answering what items to wear, how to match them, and how to dress in an aesthetically pleasing manner for a journey.

Index Terms—Fashion analysis, Clothing recommendation, Location-based multimedia system.

1 Introduction

When packing for a journey, choosing appropriate and aesthetically pleasing clothing is a concern for many of us. How do we dress appropriately? We usually consider the weather, the season and also the main attraction type (*i.e.*, cultural heritage, human landscape, or natural scenery) of the destinations. For example, wearing a scarf is recommended in the Sahara Desert because it has strong winds during the peak travel season. Moreover, visitors are not allowed to enter Angkor Wat wearing skirts or shorts because bare legs are disrespectful, and it is a religious temple. How do we dress in an aesthetically pleasing manner? In addition to sizes and shapes, which are considered when buying the clothes, we could further consider the visual match between the foreground clothing and the background scenery. For example, it is more attractive to wear bright color clothing in front of dark scenery such as Angkor Wat, as shown in Figure 1. Currently, an increasing number of travel websites (*e.g.*, tripadvisor.com, mafengwo.com, chanyouji.com, *etc.*) allow people to share their trip photos and are gaining increasing popularity. The worldwide travel site TripAdvisor found that more than 76 percent of travelers will share travel experiences via photos on social networks[1]. The large number of online photos actually provides ample and excellent references that travelers can use to plan their trips; people can vividly imagine which of their clothes will be appropriate and beautiful by exploring travel photos of their destinations. This leads to a novel problem: can we automatically learn the correlation between clothing and location from photos shared on travel websites? Studying the problem may benefit many applications such as location-oriented clothing recommendations.

Fulfilling this task is extremely challenging. **1) Location attribute definition**. There is no existing detailed investigation into or clear definition of the location attributes that are closely related to appropriate and aesthetic dress. Some pioneering works [1], [2] have demonstrated the strong correlation between location and fashionability. However, they often ignore the discrepant (*e.g.*, temperature, weather, *etc.*) and quantized descriptions of those correlations. As a result, only a rough fashion score at the whole country level was given. **2) Intrinsic correlation mining**. It is difficult to capture and model the intrinsic relationship between people's dress and specific locations. Most existing fashion analysis works focus on the correlation between fashion items only. For example, Jagadeesh *et al.* [2], [3] learned

Manuscript received July 12, 2016. This work is supported by National Nature Science Foundation of China under Grant 61525206,61271428,61672495,61429201,61370023,61521002, the National Key Research and Development Plan of China under Grant 2016YFB0801203,2016YFB1001200,2016YFB0801200,2016IM010200, Beijing Advanced Innovation Center for Imaging Technology under Grant BAICIT-2016009, ARO grant W911NF-15-1-0290 and Faculty Research Gift Awards by NEC Laboratories of America and Blippar

Xishan Zhang and Yongdong Zhang are with the Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, and also with the University of the Chinese Academy of Sciences, Beijing, China,(e-mail: zhangxishan@ict.ac.cn;zhyd@ict.ac.cn).
Jia Jia is with the Department of Computer Science and Technology, Tsinghua University, Beijing, China (e-mail: jjia@mail.tsinghua.edu.cn).
Ke Gao, Jintao Li are with the Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China. (e-mail: kegao@ict.ac.cn; jtli@ict.ac.cn).
Dongming Zhang is with National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100029, China.
Qi Tian is with Department of Computer Science, University of Texas at San Antonio,San Antonio, USA. (e-mail: qitian@cs.utsa.edu).

1. http://www.tripadvisor.com/PressCenter-c7-Survey_Insights.html

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2696825, IEEE Transactions on Multimedia

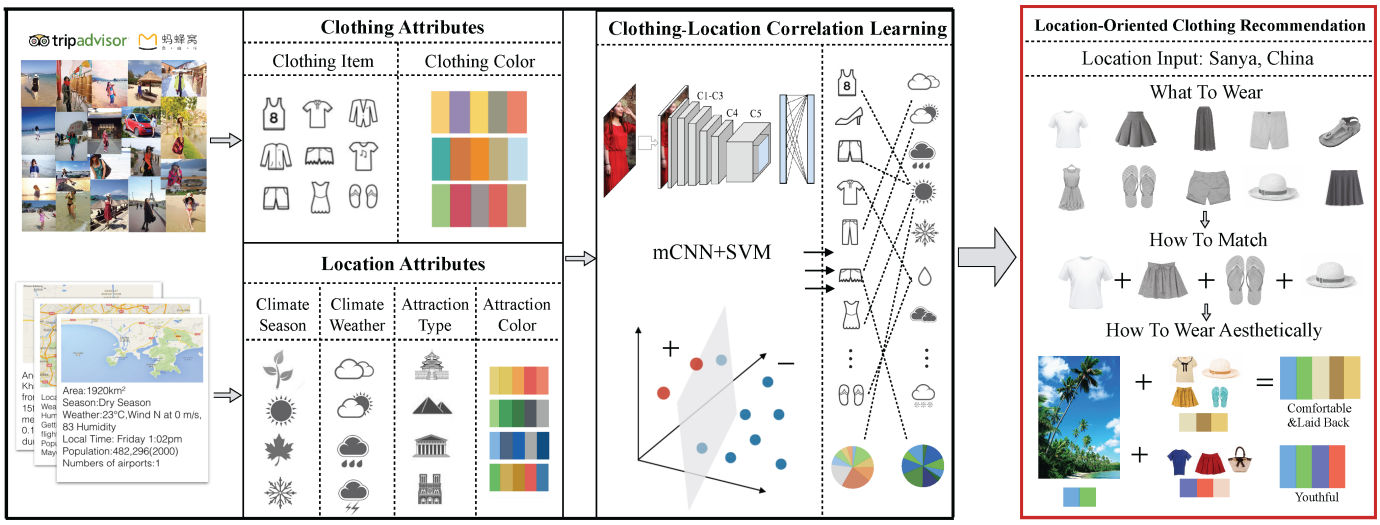

Fig. 2: Illustration of the proposed mCNN-SVM approach (in black box) and its application (in red box).

Fig. 1: Different dressing styles at different destinations.

the matching rules between different fashion items, and Liu *et al.* [4] considered the correlation between upper body and lower body. Meanwhile, learning correlations heavily depends on clothing recognition accuracy [4], [5], [6]. Despite recent advances in generic object recognition, there are relatively few studies focusing on fine-grained fashion item recognition (*i.e.*, sweaters, cardigans). **3) Practical application**. How to incorporate location information into an actual clothing recommendation system is still worth studying. Previous works [2], [4] generally used data-driven methods to recommend interdependent clothing items. However, as dress is quite personal, there is no universally correct criterion for an aesthetic match. Thus, providing different styles and color options for different people according to their preferences and location characteristics continues to be a problem.

In this paper, we first conduct a thorough data observation on photos from several international travel websites (*i.e.*, Mafengwo.com and Chanyouji.com) to determine the location attributes that influence dress. Based on the observations, four types of location attributes, '*Climate:Season*', '*Climate:Weather*', '*Attraction:Type*', and '*Attraction:Color*', are defined. Then we propose a hybrid multi-label convolutional neural network combined with the support vector machine (mCNN-SVM) approach to formulate the complex correlation between clothing and location attributes. The flowchart is shown in Figure 2. Specifically, we advance the CNN into a multi-label strategy and use it to recognize 53 fine-grained clothing items [5] by exploring the label correlation and considering the uneven distribution of clothing items. The structured SVM is adopted to discover the correlation between the location and the clothing attributes by learning the pairwise co-occurrence. Experiments on three fashion datasets (*i.e.*, Fashionista [7], CCP [8], Colorful-Fashion [9]) and a benchmark Journey Outfit Dataset show that our proposed approach outperforms several baselines by between 10.52-16.38% in terms of mean Average Precision (mAP) on clothing item recognition and outperforms several alternative methods by between 9.59-29.41% in terms of mAP on the appropriateness of clothing for travel destinations. This further demonstrates that our proposed approach is both classic and effective. Finally, an interesting case study on location-oriented clothing recommendations is presented to demonstrate the effectiveness of our method for determining what items to wear, how to match items, and how to dress in an aesthetically pleasing manner.

The main contributions of this paper are three-fold:

- To the best of our knowledge, this is the first time that the correlation between clothing and location has automatically been learned from online travel photos. Unlike existing fashion analysis works [1], [2], [4], which are only based on clothing analysis, this work is the first to identify and leverage location-related attributes in the field of fashion analysis.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2696825, IEEE Transactions on Multimedia

TABLE 1: Clothing attributes

Clothing Item [5]									**Clothing Color**
accessories	bag	belt	blazer	blouse	bodysuit	boots	bikini top	bracelet	color theme
cape	cardigan	clogs	coat	dress	earrings	flats	glasses	gloves	
hat	heels	intimate	jacket	jeans	jumper	leggings	loafers	necklace	
pants	pumps	purse	ring	romper	sandals	scarf	shirt	shoes	
shorts	skirt	sneakers	socks	stockings	suit	sunglasses	sweater	sweatshirt	
t-shirt	tie	tights	top	vest	wallet	watch	wedges		

TABLE 2: Location attributes

Climate:Season			**Climate:Weather**			**Attraction:Type**		**Attraction:Color**
spring	summer	autumn	wind	rain	scorching	cultural heritage	human landscape	color theme
winter	dry	wet	cold	hot		natural scenery		

- We propose a hybrid mCNN-SVM approach to boost attribute extraction and explicitly formulate the correlation between clothing and location. The proposed mCNN fully explores the uneven distribution of clothing attributes, while the structured SVM explores the intrinsic clothing-location correlation based on well-defined location attributes.
- We construct a benchmark dataset for location-oriented clothing recommendations. This Journey Outfit Dataset contains 3,392 traveler images with annotations for 14 travel destinations. Meanwhile, we have released the data together with the code [2].

The rest of paper is organized as follows. Section 2 discusses related work. Section 3 presents the data observations. Section 4 formulates the problem. Section 5 proposes the clothing-location correlation learning. Section 6 reports the experimental results and the case study, followed by the conclusion in Section 7.

2 Related work

In this section, we review the related work in three areas: fashion analysis, clothing recognition (which is the basis for fashion analysis), and location-aware multimedia systems.

Fashion analysis: There has been increasing attention given to fashion analysis from the multimedia and computer vision communities. The notion of style and fashion is usually studied considering visual features such as clothing parsing [5], [6], clothing retrieval [10], [11], and clothing recommendations [2], [11], [12]. Little research has been devoted to introducing scenario-based attributes into fashion analysis. The pioneering work [4] considers occasions in dressing. However, there are few studies introducing the location attributes of travel destinations into fashion analysis. Meanwhile, the availability of enormous numbers of social images has immense potential to shift the research focus from low-level tasks to novel research topics relating to higher-level semantic analysis, such as personal style analysis [3], [13] and fashionability prediction [1], [14], [15], [16]. The closest work to ours is scenario-oriented clothing recommendations [4]. However, in these works, the aesthetic rules for dressing are constrained to clothing items only. In our scenario, we consider clothing when traveling to a specific scenic spot, which enables us to offer a global aesthetic matching rule between clothing and the distinct visual scenery.

Clothing recognition: The general object recognition algorithms can be roughly categorized into two types. The first is based on a deformable part model. For clothing recognition, typical methods [5], [8] recognize 53 fine-grained clothing items and demand expensive annotations for human poses and clothing locations, which hinder their training on weakly labeled large fashion datasets. The second type is convolutional neural network- (CNN-) based methods, which naturally fit large training datasets. Most works assume a single object in the image [11], [17] and reduce the clothing item categories to fewer (*i.e.*, 11, 18) super-categories [11], [18]. The most closely related work that recognizes multi-objects using CNN is Oquab's work [19], which focuses on general objects. However, the recognition of a clothing category is more challenging because of the large number of fine-grained clothing categories. In our scenario, fine-grained categories are needed for location-oriented clothing recommendations. Though it is hard to distinguish 'sandal' from 'heel', we demand accurate recognition of these shoes because they suit different destinations.

Location-aware multimedia systems: The vast amount of geo-tagged photos shared on social networks enable novel multimedia systems such as location recommendations [20], route-based travel recommendations [21], [22], [23] and landmark recognition [24], [25]. A recent survey of location-based recommendations can be found in [26]. Among these works, various types of geo-based data are investigated: GPS trajectories [21], check-ins [21], [23], and geo-tagged photos [20], [27]. As the GPS trajectory is relatively difficult to obtain [20], recent works identify points of interest (POI) from user-generated photos [28] or discover geo-informative attributes for location recognition [25]. There are few works jointly considering humans' clothing in location analysis [1], [2], [29]. Some pioneering works [1], [2] rate the fashionability score for different locations. For example, [1] focuses on a general country-level fashion analysis. Others [29] jointly considered the location and clothes to capture additional facial attribute features. Our work aims to extend fashion analysis to the fine-grained city level using the tagged photos from travel websites.

2. http://pan.baidu.com/s/1i47rFJB

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2696825, IEEE Transactions on Multimedia

3 DATA OBSERVATION

Before learning the correlation between clothing attributes and location attributes, we first conduct a series of data observations to reveal its existence and to further verify the feasibility of our task. We adopt two datasets: 1) the **Paper Doll Dataset** [5], which contains 339,797 fashion photos with annotations of clothing attributes listed in Table 1. (*c.f.* Section 6 for more details of the dataset) and 2) the **Journey Scene Dataset**, which contains 7,674 photos from two worldwide travel websites (*i.e.*, mafengwo.com and chanyouji.com) with annotations of location attributes for 14 destinations listed in Table 2. The 14 destinations are listed in Table 3. Note that unlike the **Journey Outfit Dataset** in Section 6, we do not impose restrictions on human presence in the **Journey Scene Dataset**. Based on these two datasets, we conduct four observations.

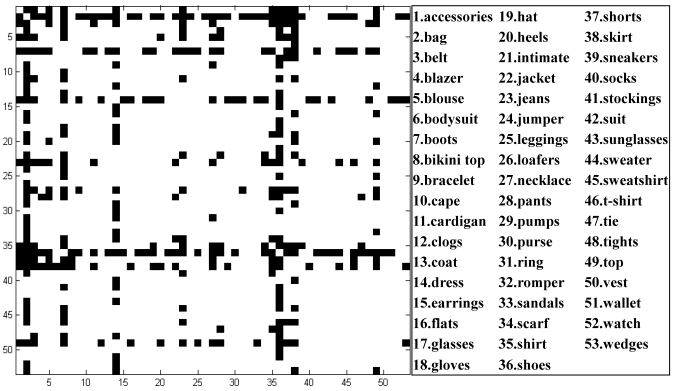

Fig. 3: The frequently matched clothing item pairs (black blocks) learned from the Paper Doll Dataset.

Observation 1. The co-occurrence of clothing items. We show the statistics on the co-occurrence of clothing items in the Paper Doll Dataset. The normalized co-occurrence [30] is defined as $\frac{\mathcal{N}(c_i,c_j)}{\mathcal{N}(c_i)\times\mathcal{N}(c_j)}$, in which $\mathcal{N}(c_i)$ counts the occurrence of the clothing item c_i, and $\mathcal{N}(c_i, c_j)$ counts the co-occurrence of c_i and c_j. The black color in Figure 3 shows the top 200 highly correlated clothing item pairs among all of the 53×53 pairs. For example, the top pairs are bag and dress, shorts and t-shirt, dress and sandals. The co-occurrence of clothing items is in accordance with our daily dressing habits, which verifies the rationality of the definition of clothing attributes.

Observation 2. The clustering of destinations. We use the four types of location attributes defined in Table 2 as the clustering features to show the embedding of the locations in attribute space. In particular, PCA and k-means are applied to the concatenation of the location attributes, and we visualize the first two principle components after PCA in Figure 4. For example, the red points represent a cluster for Southeast Asia. The blue points represent a cluster for higher latitude locations with cold weather and four clear seasons. The orange points represent some cities that enjoy famous human landscapes. Therefore, the definitions of location attributes in Table 2 perfectly reveal the similarity among different destinations and reflect discrimination between destinations, which verifies the rationality of the definition of location attributes.

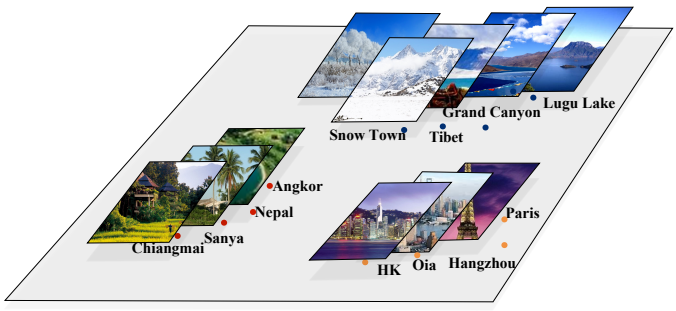

Fig. 4: The clusters reflect defined location attributes that provide a reasonable similarity measurement for 14 destinations.

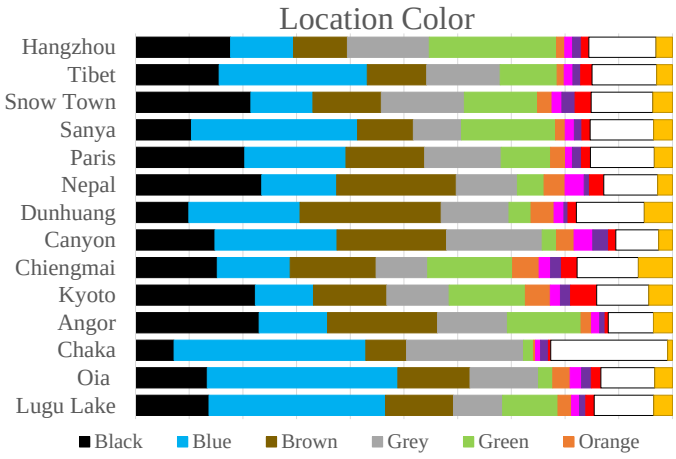

Fig. 5: The different color distribution of different destinations.

Observation 3. The correlation between clothing items and location attributes ('*Climate:Season*', '*Climate:Weather*', and '*Attraction:Type*'). The strong correlation between clothing items and '*Climate:Season*', '*Climate:Weather*', '*Attraction:Type*' is in accordance with common sense. In Figure 1, it is apparent that people wear warmer clothing (*e.g.*, coat, scarf) at high altitudes and latitudes such as Tibet. Using some well-defined location attributes instead of city names alone, we will impose high-level intuitions in the formulation of real world problems.

Observation 4. The correlation between clothing colors and location colors. Different locations have unique scenery. We use the attraction color extracted from the Journey Scene Dataset to investigate the color distribution in 14 different destinations. Figure 5 shows a color histogram based on 11 color names [31]. The color distribution is quite different for the 14 destinations. For example, in Angkor, the black color appears frequently. In order to take fabulous photos during a journey, the selected outfits should also be compatible with the location's visual appearance.

Summary. Based on the observations above, we have the following intuitions:

- Clothing attributes and location attributes do show a strong correlation, which supports the motivation of this work.
- The '*Climate:Season*', '*Climate:Weather*', and '*Attrac-*

tion: Type' will influence the traveler's appropriate choice of clothing items.
- The '*Attraction:Color*' will influence the traveler's aesthetic choice of clothing color.

4 Problem Formulation

Our goal is to derive the correlation between the clothing and the location. Each image sample is represented by $(\mathbf{x}, \mathbf{c}, \mathbf{l}, y)$, where $\mathbf{x}$ is the image visual content; $\mathbf{c} = (\mathbf{c}_v, \mathbf{c}_a)$ and $\mathbf{l} = (\mathbf{l}_v, \mathbf{l}_a)$ denote the clothing and location attributes, respectively; and y denotes the different destinations.

Definition 1. Clothing attributes $\mathbf{c} = (\mathbf{c}_v, \mathbf{c}_a)$. The details of clothing attributes are summarized in Table 1. $\mathbf{c}_v$ is the clothing color. $\mathbf{c}_a = (c_{a_1}, c_{a_2}, ..., c_{a_{53}})$ with ($|\mathbf{c}_a| = 53$) is the clothing item vector, whose each dimension shows the presence/absence of a specific clothing item in the photo.

Definition 2. Location attributes $\mathbf{l} = (\mathbf{l}_v, \mathbf{l}_a)$. The details of location attributes are summarized in Table 2. $\mathbf{l}_v$ is the '*Attraction: Color*'. $\mathbf{l}_a$ with $|\mathbf{l}_a| = 14$ being the vector comprising '*Climate: Season*', '*Climate: Weather*' and '*Attraction: Type*'. For example, the travel season for Angkor is the dry season, and it is hot and always raining. People enjoy the cultural heritage and natural scenery there. Therefore, for this specific location, $\mathbf{l}_a = (0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 1, 0, 1)$.

Learning task. We aim to learn the correlation between location and clothing attributes. In particular, a score function $g_{\mathbf{w}} : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ models the correlations, and the model parameter $\mathbf{w}$ in the function $g_{\mathbf{w}}(\mathbf{x}, \mathbf{c}, \mathbf{l}, y) = \mathbf{w}^T \Phi(\mathbf{x}, \mathbf{c}, \mathbf{l}, y)$ tends to assign the highest score to the most suitable clothing items $\mathbf{c}_a$ appearing in the given image $\mathbf{x}$ for location y. It also simultaneously considers the location attributes assignment $\mathbf{l}_a$. In application, $g_{\mathbf{w}}$ is used to recommend the most suitable clothing items and clothing colors given location y. As a linear model, this is similar to feature ranking or feature selection [32], [33] in Linear Support Vector Machines.

5 Clothing-Location Correlation Learning

An overview of our system is shown in Figure 2. Note that the images from the location-based social network have no annotations about clothing attributes, and the location attribute assignment for a specific place does not reflect the true situation for all images of the same place. To flexibly make use of the travel photos, we assume neither clothing annotation nor complete location attribute labels for the training images. The entire model is formed as a hierarchical structure. We need to recognize clothing attributes from the images and then predict the location attributes. Last, we use the combined information to determine the dress for a specific location. We define the probability $p(y, \mathbf{c}, \mathbf{l}|\mathbf{x})$ of matching clothing and location assignments to image $\mathbf{x}$, and it is given by Eq. 1,

$$p(y, \mathbf{c}, \mathbf{l}|\mathbf{x}, \mathbf{w}) \propto exp(\mathbf{w}^T \Phi(\mathbf{x}, \mathbf{c}, \mathbf{l}, y)) \quad (1)$$

We make use of the insights from the data observation to analyze the problem components in designing the potential function. The potential $\Phi(\mathbf{x}, \mathbf{c}, \mathbf{l}, y)$ consists of four potential parts as given in Eq. 2,

$$\begin{aligned} \mathbf{w}^T \Phi(\mathbf{x}, \mathbf{c}, \mathbf{l}, y) = \mathbf{w}_1 \varphi_1(\mathbf{x}, \mathbf{c}_a) + \mathbf{w}_2 \varphi_2(\mathbf{c}_a, \mathbf{c}_a) \\ + \mathbf{w}_3 \varphi_3(\mathbf{x}, \mathbf{c}_a, \mathbf{l}_a) + \mathbf{w}_4 \varphi_4(\mathbf{x}, \mathbf{c}_v, \mathbf{l}_v) \end{aligned} \quad (2)$$

We now explain each potential in detail:

Clothing item potential. $\mathbf{w}_1 \varphi_1(\mathbf{x}, \mathbf{c}_a)$: This potential function $\varphi_1(\mathbf{x}, \mathbf{c}_a) \in \mathbb{R}^{|\mathbf{c}_a|}$ determines the occurrence of each clothing item c_{a_i} from image $\mathbf{x}$. A pre-trained image-clothing classifier, denoted as mCNN, that leverages deep learning is used to construct these unary factors, $\varphi_1^{(i)}(\mathbf{x}, c_{a_i}) = f_i^{c_{a_i}}(\mathbf{x})$ (*c.f.* Section 5.1 for details). Note that we learn different sets of weights for each location class.

Pairwise clothing item potential. $\mathbf{w}_2 \varphi_2(\mathbf{c}_a, \mathbf{c}_a)$: This potential reflects the co-occurrence of different clothing items in location y. We only consider the top m frequently matched pairs as shown in black color in Figure 3. $\mathbf{w}_2 \in \mathbb{R}^{|m|}$ is the weight for joint clothing pair assignment to an image $\mathbf{x}$ in location y. In this way, we generate the location-specific frequently matched pairs. For example, 'bikini top' and 'shorts' frequently appear at the sunny beach area of 'Sanya'.

Location potential. $\mathbf{w}_3 \varphi_3(\mathbf{x}, \mathbf{c}_a, \mathbf{l}_a)$: This potential function $\varphi_3(\mathbf{x}, \mathbf{c}_a, \mathbf{l}_a)$ determines each l_{a_i} based on a person wearing $\mathbf{c}_a$ in image $\mathbf{x}$. Therefore, this potential function implies the correlation between $\mathbf{l}_a$ and $\mathbf{c}_a$. The value of the location potential is the confidence score produced by a location attribute detector $f_i^{l_{a_i}}(\mathbf{x}, \mathbf{c}_a)$ (*c.f.* Section 5.2 for details).

Global appearance potential. $\mathbf{w}_4 \varphi_4(\mathbf{x}, \mathbf{c}_v, \mathbf{l}_v)$: This potential is the concatenation of the visual features, including generic feature $\mathbf{x}$, attraction color $\mathbf{l}_v$ and clothing color $\mathbf{c}_v$. The parameters $\mathbf{w}_4$ model the aesthetic match between attraction color $\mathbf{l}_v$ and clothing color $\mathbf{c}_v$.

5.1 Clothing item recognition

In this section, we focus on predicting accurate image-level labels indicating the presence/absence of clothing item $f_i^{c_{a_i}}(\mathbf{x})$. Inspired by the huge success of convolutional neural networks (CNNs) [17], we develop an end-to-end learning method from image-level labels. In our scenario, multiple clothing items appear in one image, and we have no clothing-location information. To adapt CNN architecture to multi-label clothing learning, we focus on two issues. First, we expect that the information for one clothing item should be helpful for the learning of other items, so exploiting clothing item correlation is important for multi-label learning. Second, positive training examples for some rare items (*e.g.*, gloves, loafers) are quite limited compared to other commonly occurring items (*e.g.*, dresses) because of the highly uneven distribution of clothing items. To solve these problems, training our clothing item recognition network, denoted as mCNN, requires two stages: multi-label pre-training and single category fine-tuning, as shown in Figure 6.

In multi-label pre-training, we consider AlexNet [17] as our basic architecture. A straightforward solution to multi-label learning is to decompose the problem into a series of binary classifications. However, such a solution neglects the fact that recognition of one label may be helpful for learning other labels, for example 'shorts' and 'bikini top' frequently co-occur. Inspired by the idea of the local label correlations, we minimize the Euclidean loss between the multi-label ground truth and the prediction as in [34],

$$\ell(f^{\mathbf{c}_a}(\mathbf{x}), \mathbf{c}_a) = \|f^{\mathbf{c}_a}(\mathbf{x}) - \mathbf{c}_a\|_2^2 = \|\mathbf{W}^T \mathbf{z} - \mathbf{c}_a\|_2^2 \quad (3)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2696825, IEEE Transactions on Multimedia

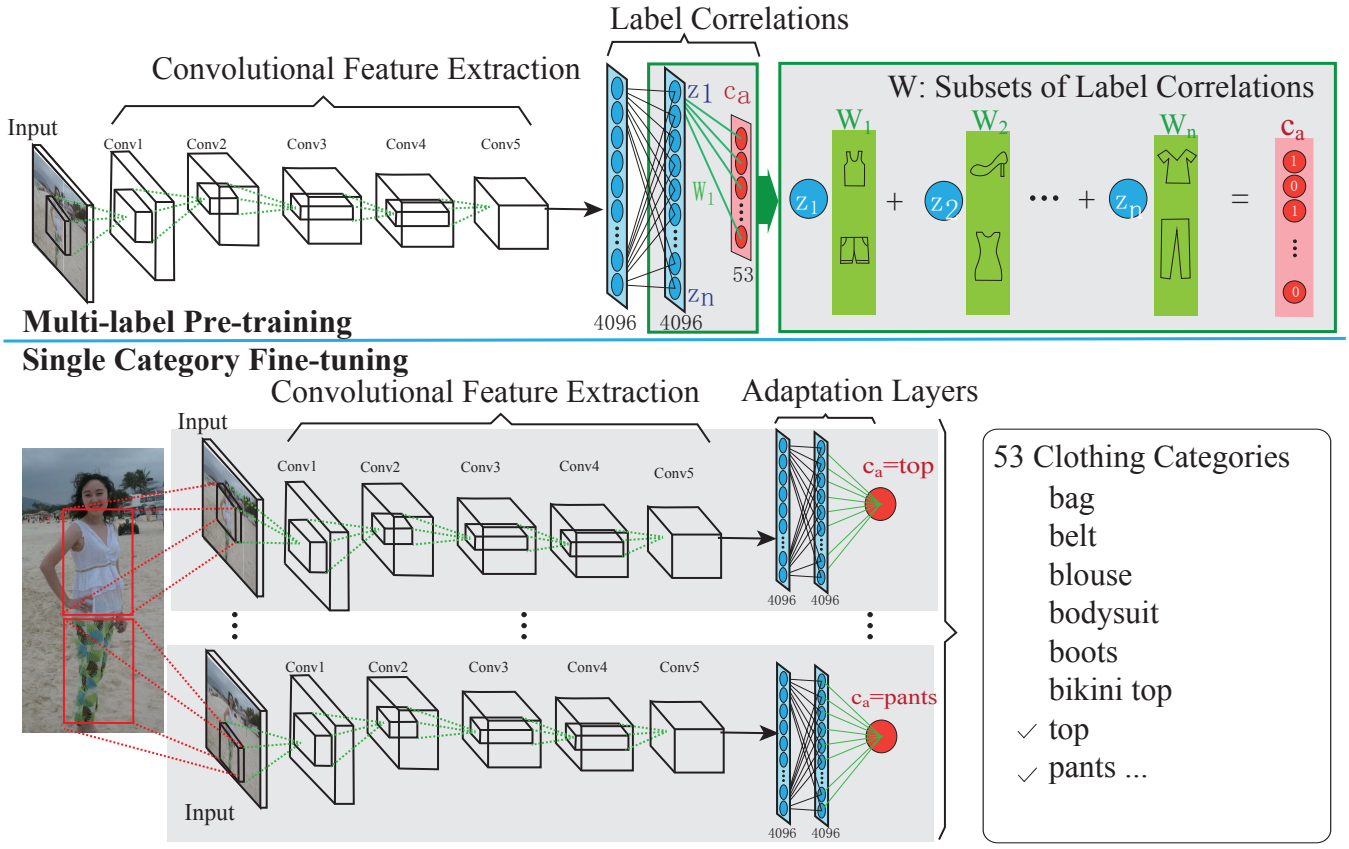

Fig. 6: The multiple clothing item recognition convolutional neural networks (mCNN).

where $f^{\mathbf{c}_a}(\mathbf{x})$ is the output of the network for the input image; $\mathbf{x}$ and $\mathbf{c}_a$ are the clothing labels annotated on image $c_{a_i} \in \{0, 1\}$; $\mathbf{z} = (z_1, z_2, ..., z_{4096})^T$ represents the feature extracted on the penultimate layer; and $\mathbf{W} = (w_{ij})_{4096\times 53}$ represents the weights between the last two layers. The instances can be separated into different groups, and each group shares a subset of label correlations. For example, as shown in Figure 6, each row $\mathbf{W}_i = (w_{i1}, w_{i2}, ...w_{i53})$ learns a subset of frequently co-occurring labels, and z_i encodes the local influence of label correlations on the input instance. The implementation of the last fully connected layer in the mCNN network is formulated as:

$$\mathbf{W}^T\mathbf{z} = \sum_{i=1}^{4096} z_i \mathbf{W}_i^T \tag{4}$$

As all of the categories are considered jointly when minimizing the loss, we can learn the subsets of label correlations $\mathbf{W}_i$ and search for a linear combination of $\mathbf{W}_i$ for each instance. The optimal coefficients in the linear combination are the transformed feature value $\mathbf{z}$. As instances with similar annotations are expected to have similar linear combinations, the learned visual representations $\mathbf{z}$ should be similar. By tuning the overlap of the shared hidden state distribution, the network captures the intrinsic structure among training samples in both feature space and label space, thus modeling feature and label correlations.

However, for some small items (*e.g.*, gloves, loafers) with fewer positive instances, when using the above architecture, the recognition decision is taken primarily based on the label correlation and not on the true appearance and position of the clothing items [19]. Therefore, we fine-tune the network, which records the label correlations, for each clothing category to learn the category-specific features $\mathbf{z}$ and classifiers $\mathbf{W}$. Binary logistic regression loss is used to fine-tune the networks,

$$\ell(f^{c_{ai}}(\mathbf{x}), c_{ai}) = log(1 + e^{-c_{ai}f^{c_{ai}}(\mathbf{x})}) \tag{5}$$

As a result, the convolutional networks are able to focus on accurately positioning each item (*c.f.* Section 6 and Figure 7 for more examples). Although there are improvements for each category, the small items benefit the most. Due to the uneven distribution of clothing labels, we fine-tune for different iterations. For some categories with positive instances over 10,000, more iterations (10,000 iterations) are needed to learn the large intra-class variances. For categories with positive instances below 10,000, we train for 500 iterations, in case of over-fitting. The trained model for each class has been released [3].

5.2 Location attribute learning

In this section, we focus on location attribute detector $f_i^{l_{a_i}}(\mathbf{x}, \mathbf{c}_a)$, which is denoted as LocationC. The recognition of each l_{a_i} is based on the visual feature $\mathbf{x}$ extracted from the human area and the clothing items $\mathbf{c}_a$ as learned in Section 5.1. Therefore, we guarantee that the location attributes are inferred only from what humans are wearing and not from

3. http://pan.baidu.com/s/1i47rFJB

Algorithm 1 Clothing-location correlation learning

Input: $(\mathbf{x}^1, y^1), ..., (\mathbf{x}^N, y^N)$, N is the number of training samples
Output: Learned weight vector $\mathbf{w}$

for $i = 1, 2, ..., N$ **do**
  learning clothing items $\varphi_1(\mathbf{x}^i, \mathbf{c}_a^i) = f^{\mathbf{c}_a^i}(\mathbf{x}^i)$
  learning location attributes $\varphi_2(\mathbf{x}^i, \mathbf{c}_a^i, \mathbf{l}_a^i) = f_a^l(\mathbf{x}^i, \mathbf{c}_a^i)$
end for
repeat
  $\mathbf{w}_t = \mathbf{w}_{t-1}$
  for $i = 1, 2, ..., N$ **do**
    $\hat{y}^i = \max_{\hat{y}^i} \mathbf{w}_t^T \Phi(\mathbf{x}^i, \mathbf{c}^i, \mathbf{l}^i, \hat{y}^i)$
  end for
  update $\mathbf{w}_{t+1}$ using Eq. (6)
until $\mathbf{w}_{t+1} - \mathbf{w}_t < \varepsilon$

the background scenes. In this work, we choose SVM as our location attribute detector model, and LIBLINEAR [4] is used to implement it. In SVM training, the location attribute labels should be $l_{a_i} \in \{-1, 1\}$, so in testing the outputs, l_{a_i} are also in the range of $[-1, 1]$, and we normalize them into $[0, 1]$ as the final location potential.

5.3 Large margin learning

The whole procedure for training the clothing-location correlation learning model is outlined in Algorithm 1. First, we train independent per-node classifiers for clothing and location attributes $f^{\mathbf{c}_a}, f^{\mathbf{l}_a}$ (*c.f.* Section 5.1 for details of $f^{\mathbf{c}_a}$ and Section 5.2 for details of $f^{\mathbf{l}_a}$). Then, the outputs of these classifiers are used as feature maps $\varphi_1(\mathbf{x}^i, \mathbf{c}_a^i), \varphi_2(\mathbf{x}^i, \mathbf{c}_a^i, \mathbf{l}_a^i)$. The potential function $\Phi(\mathbf{x}, \mathbf{c}, \mathbf{l}, y)$ is constructed based on Eq. 2. We use a standard structural SVM to solve the margin-based parameter learning problem [35],

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}} \{\|\mathbf{w}_t\|_2^2 + C \sum_{i=1}^{N} \{\max_{\hat{y}^i} [\Delta(y^i, \hat{y}^i) + \mathbf{w}_t^T \Phi(\mathbf{x}^i, \mathbf{c}^i, \mathbf{l}^i, \hat{y}^i)] - \mathbf{w}_t^T \Phi(\mathbf{x}^i, \mathbf{c}^i, \mathbf{l}^i, y^i)\}\}, \tag{6}$$

where $\hat{y}^i = \max_{\hat{y}^i} \mathbf{w}_t^T \Phi(\mathbf{x}^i, \mathbf{c}^i, \mathbf{l}^i, \hat{y}^i)$ with a zero-one loss,

$$\Delta(y^i, \hat{y}^i) = \begin{cases} 0, & if \quad y^i = \hat{y}^i \\ 1, & otherwise \end{cases}$$

This optimization problem can be solved efficiently using the cutting-plane method [36] [5]. In optimizing Equation 6, model parameters $\mathbf{w}$ tend to assign the highest score to the most suitable clothing for a location y, so the parameters reflect the clothing-location correlation naturally. In the analysis of the learned $\mathbf{w}$, we can clearly know the most suitable clothing item and clothing pairs for a specific location y. After learning, based on image content $\mathbf{x}$ and learned parameter $\mathbf{w}$, we can obtain the location probability $p(y, \mathbf{c}, \mathbf{l}|\mathbf{x}, \mathbf{w})$ and the location-based clothing image ranking. The higher this output probability is, the more appropriate these outfits are for this location y.

4. https://www.csie.ntu.edu.tw/~cjlin/liblinear/
5. https://vision.princeton.edu/pvt/OnlineStructuralSVM/

TABLE 3: Journey Outfit (JO) dataset

Location	Country	Num of images
Tibet	China	346
Lugu Lake	China	325
Dunhuang	China	277
Snow Town	China	204
Hangzhou	China	247
Sanya	China	276
Chaka(salt lake)	China	219
Angor	Cambodia	214
Oia	Greece	239
Chiengmai	Thailand	188
Nepal	Nepal	200
Grand Canyon	U.S.A.	180
Paris	France	233
Kyoto	Japan	244

6 Experiments

6.1 Dataset and Features

6.1.1 Dataset

Clothing Recognition Datasets. This is a combination of four published clothing datasets and provides annotation for clothing items. The first is the **Paper Doll Dataset** [5], which contains 339,797 photos with annotations for 53 clothing items listed in Table 1. The second is the **Fashionista Dataset** [7]. It provides 685 images from the same source as the Paper Doll Dataset, and it has fully parsed images with 53 labels. The third is the **Colorful-Fashion Dataset** [9]. This dataset provides 2,682 images, and it annotates 20 labels (a subset of 53 labels). The fourth is the **CCP Dataset** [8]. This dataset provides 2,098 images and provides annotation of all 53 labels.

Journey Outfit Dataset. We establish a benchmark dataset, the Journey Outfit (JO) Dataset, which consists of travel images from online travel review websites (*e.g.*, mafengwo.com and chanyouji.com). This dataset contains 3,392 photos with various people wearing clothing in different scenery with ground truth annotation for 14 locations. We only select photos that are favored by many users, which guarantees the appropriacy of dress in the photos. To study the visual features in the foreground dress and background scenery separately, we run a state-of-the-art human detector [5] to automatically select photos with only one person. The details for the number of images for each location are shown in Table 3. We randomly split the images for each location into three subsets: 40% for learning location attributes (JO Train-1), 40% for learning clothing-location correlation (JO Train-2) and 20% for testing (JO Test).

6.1.2 Features

Visual Features. The visual features include generic features $\mathbf{x}$, attraction color $\mathbf{l}_v$ and clothing color $\mathbf{c}_v$. For generic visual features, we adopt a 4,096 dimensional Decaf CNN features [39] extracted from the human area of the photos. This is fed as the input of the proposed clothing-location correlation learning to compute the location potential and the global appearance potential. For color features, we extract five dominant colors [40] from both the human area ($\mathbf{c}_v$) and the background scenery area ($\mathbf{l}_v$) and use color names [41] to quantify the color space into 50 base colors, denoted as the color theme. We use the color names [41] because they

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2696825, IEEE Transactions on Multimedia

TABLE 4: Clothing item recognition performance on three datasets.

dataset		MIMLWEL [37]	MLLOC [38]	Paper Doll [5]	CNN [17]	mCNN
Fashionista [7]	macro-F	0.1616	0.1217	0.1700	0.1377	**0.3169**
	mAP	0.2016	0.1776	0.2468	0.2479	**0.4117**
	AUC	0.6015	0.5318	0.6345	0.6263	**0.8013**
Colorful-Fashion [9]	macro-F	0.4404	0.3995	0.5416	0.3763	**0.6073**
	mAP	0.4083	0.3532	0.5620	0.5149	**0.7037**
	AUC	0.6988	0.6058	0.7896	0.7666	**0.8969**
CCP [8]	macro-F	0.1931	0.1673	0.1330	0.1118	**0.2251**
	mAP	0.2013	0.1845	0.1762	0.2099	**0.3151**
	AUC	0.6158	0.5304	0.5890	0.5911	**0.7405**

are robust to illumination. The color feature will be used to compute the global appearance potential.

Text Feature. The '*Climate:Season*', '*Climate:Weather*' and '*Attraction:Type*' are crawled from web pages introducing the destinations on travel websites (*i.e.*, mafengwo.com).

6.2 Experimental Setup and Results

The whole clothing-location correlation learning approach is formed as a hierarchical structure, with the clothing item potential and location potential as its main components. Therefore, we evaluate at three stages: clothing item recognition, location attribute learning and the clothing-location correlation learning as reflected by the location-based clothing ranking.

6.2.1 Clothing item recognition

This experiment aims to evaluate the clothing item potential as illustrated in Section 5.1. We make use of the Paper Doll Dataset for clothing item recognition. The dataset is split into 9 folds for training and 1 fold for validation. In pre-training, we use all 53 clothing item labels to supervise the training of a unified network (learning rate 0.01, momentum 0.9). In fine-tuning, we train 53 category-independent networks. We then monitor using the validation set to determine when to stop training for each category. We compare four methods for the Clothing Recognition Dataset:

1) **MIMLWEL** [37]: the multi-label learning method assumes that the untagged labels are not necessarily negative;
2) **MLLOC** [38]: the multi-label learning method exploits label correlations;
3) **Paper Doll**[6] [5]: the clothing recognition method explores the K nearest neighbors in the large Paper Doll Dataset;
4) **CNN** [17]: the state-of-the-art convolutional neural network (AlexNet) uses the multi-label loss function (Euclidean loss) and is similar to the pre-training step in our mCNN.

We evaluate the performances using three common multi-label criteria [42]: Macro-F, mean Average Precision (mAP) and AUC. Detailed definitions of these can be found in [42].

Table 4 lists the clothing recognition results for the Clothing Recognition Dataset. Clearly, mCNN outperforms the other methods. On average, our method achieves a 10.52-16.38% improvement in terms of mAP. In particular, for the Fashionista Dataset, which has 53 fine-grained classes, we obtain a 86%-163% improvement in terms of macro-F, indicating that our method offers strong improvement over those fine-grained classes with low occurrence.

The MIMLWEL and MLLOC, as traditional shallow learning algorithms, are trained using the training splits of each clothing dataset without using the larger Paper Doll Dataset, thus leading to relatively poor performance. Due to the large deformation of clothing, the intra-class variance is large, so it is generally beneficial to use a larger training dataset.

Our mCNN contains two steps: pre-training(CNN [17]) and fine-tuning. In Table 5, we compare the overall results from using only the multi-label pre-training method (CNN) without fine-tuning and using the method without pre-training and only fine-tuned on the ImageNet CNN model. Both the fine-tuning method and our method show great improvement for small items such as 'sandals' and 'sunglasses'. For example, we have a 161% relative improvement over the pre-training CNN method for recognizing 'gloves' and a 104% relative improvement over the Paper Doll method for recognizing 'stockings'. This is because after fine-tuning for each category, the lower layers of the networks can focus on the appearance and position of each clothing item more precisely. As shown in Figure 7, we use the method in [43] to visualize the feature maps for pre-training CNN and mCNN. It is clear that we locate small objects such as 'belt' and 'boots' more accurately. Moreover, our mCNN shows great improvement for infrequently occurring categories such as 'clogs' and 'tie' compared to the fine-tuning method alone. Data imbalance is a big problem in multi-label classification. For example, in the Paper Doll Dataset, there are 127,028 positive instances of dress, but only 2,364 positive instances of clogs and 2,070 positive instances of tie. The great improvements in identifying these infrequently occurring categories indicate that our pre-training method, which explores the multi-label correlations, is helpful for training, especially with fewer positive instances. Note that the original CNN method (our pre-training method) is good for recognizing large and general clothing items such as 'pants' and 'dress'.

6.2.2 Location attributes learning

This experiment aims to evaluate the location potential as illustrated in Section 5.2. We make use of the Journey Outfit

6. http://vision.is.tohoku.ac.jp/~kyamagu/research/paperdoll/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2696825, IEEE Transactions on Multimedia

TABLE 6: The mined correlation between location attributes and clothing attributes

winter	LocationC Paper Doll [5]	coat bikini top	boots sweatshirt	dress tie	sweater bodysuit	scarf socks
summer	LocationC Paper Doll [5]	sunglasses shoes	bodysuit jeans	jeans jacket	loafers top	glasses boots
scorching	LocationC Paper Doll [5]	bodysuit shoes	sunglasses dress	hat shirt	shorts skirt	sandals jeans
boots	LocationC Paper Doll [5]	cold natural scenery	wind cold	winter summer	rain hot	culture heritage winter
sandals	LocationC Paper Doll [5]	wet season wet season	natural scenery dry season	hot rain	dry season wind	scorching winter
jacket	LocationC Paper Doll [5]	rain summer	culture heritage cold	spring natural scenery	dry season autumn	wind winter

TABLE 5: Per-category recognition performance in term of Average Precision (%) on CCP dataset

	Paper Doll [5]	pre-training (CNN [17])	fine-tuning	mCNN
clogs	1.44	2.04	4.83	**11.68**
gloves	5.42	8.37	11.80	**30.81**
loafers	5.93	4.40	7.78	**26.54**
tie	7.57	9.35	11.64	**36.58**
sandals	19.59	15.69	38.75	**48.48**
sunglasses	36.33	34.46	69.62	**94.20**
stockings	18.58	13.16	29.41	**37.89**
top	12.23	**14.41**	12.24	11.96
pants	53.65	**82.91**	74.99	81.66
dress	47.36	**61.10**	51.08	60.75
accessories	20.27	**21.46**	17.19	19.27

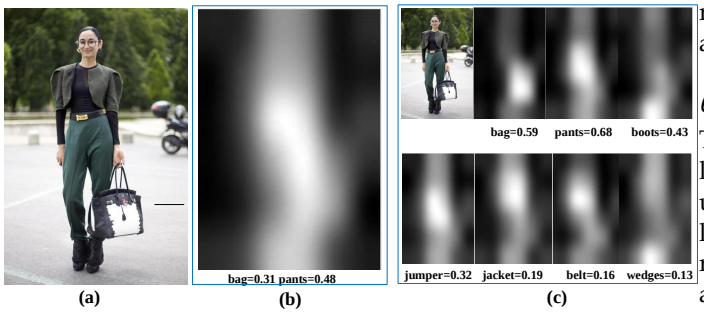

Fig. 7: (a) Original photos (b) Pre-training map and (c) mCNN feature maps. The white colored areas correspond to the predicted position of the clothing items. mCNN identifies and locates more small clothing items (*i.e.*, boots).

Dataset for location attribute learning. We compare our **LocationC** with the **Paper Doll** [5] method. Both **LocationC** and **Paper Doll** are trained on JO Train-1. In **LocationC**, $\mathbf{x}$ is the Decaf CNN features described in Section 6.1. $\mathbf{c}_a$ is the detected clothing item, using our proposed **mCNN**. For the **Paper Doll** method, $\mathbf{x}$ is the style features specifically in terms of clothing design, and $\mathbf{c}_a$ is the clothing item scores. For both methods, we only use features extracted from the foreground human area because we want to infer the location attributes from the attire and not the scene. We use mAP [42] as the evaluation metric.

Table 6 shows the top related clothing items for some of the location attributes and the top related location attributes for clothing items. On average, in recognizing the location attributes, the mAP of our method (*i.e.*, 0.72) outperforms that of Paper Doll (*i.e.*, 0.60). The correlation between location attributes and clothing items is generated by the linear parameter in $f_i^{l_{a_i}}(\mathbf{x}, \mathbf{c}_a)$. As shown in Table 6, compared with the results of Paper Doll, our relevant attributes are more consistent with common sense. For example, in summer and scorching weather, people usually wear sunglasses; we also wear boots during cold windy winter. However, bikini tops and bodysuits are not frequently seen in winter. The unsatisfactory results of Paper Doll are due to its poor performance in recognizing small clothing items. The advantage of our method demonstrates the mCNN-SVM's ability to describe the strong correlations between the location and clothing attributes. Meanwhile, the accurate recognition of clothing items contributes to the location attribute learning.

6.2.3 *Location-based clothing ranking*

This experiment aims to evaluate the overall clothing-location correlation as illustrated in Section 5.3. We make use of the Journey Outfit Dataset for clothing-location correlation learning (JO Train-2) and the location-based clothing ranking (JO Test) for all of the methods. Given a location, all of the clothing photos in the testing repository are ranked according to the appropriateness score generated by different approaches. We evaluate their performances using Average Precision (AP), which is commonly used to evaluate ranking systems [42]. We compared our proposed method (**mCNN-SVM**) with several baseline methods:

1) **Color Names [41] + SVM**: it uses discriminative color descriptors to rank the appropriateness for locations;
2) **Paper Doll [5] + SVM**: it uses clothing style descriptors based on pose detection together with clothing parsing results and has been widely used in fashion analysis [13], [44];
3) **Magic Closet [4] (CNN feature + latentSVM)**: It uses latent SVM [45], which takes clothing and location attributes as latent variables and infers them through the learning process.

We also analyze the contributions of different components as proposed in Section. 5. The components include global appearance potential, clothing potential (individual and

pairwise) and location potential. All of the methods using different components are listed in Table. 7.

TABLE 7: Different components in clothing-location correlation learning methods

	Appearance potential	Clothing potential	Location potential
Color Names [41] + SVM	✓color [41]	-	-
Paper Doll [5] + SVM	✓style [5]	✓Paper Doll [5]	-
Magic Closet [4]	✓visual features	✓latentSVM [4]	✓latentSVM [4]
mCNN-SVM-1	✓visual features	-	-
mCNN-SVM-2	✓visual features	✓mCNN	-
mCNN-SVM-3	✓visual features	-	✓LocationC
mCNN-SVM	✓visual features	✓mCNN	✓LocationC

The evaluation results for the clothing recommendations by location are shown in Table 8. On average, the mCNN-SVM outperforms the other methods by over 9.59-29.41% in terms of mAP when ranking the appropriateness of clothing for travel destinations. The poor performance of Color Names, which explores the visual scenery, verifies that the scenery is not the reason for the performance gain. It is clear that without considering appropriate dress, visual cues alone are inadequate to recommend satisfactory clothing. Though the Paper Doll method considers the clothing items, its features are not discriminative enough to describe the clothing-location correlations because it overlooks the location aspects. Magic Closet is capable of learning the correlations between location and clothing. However, there is no guarantee of the reasonable semantic meaning of the inferred latent attributes. Our method outperforms Magic Closet because we explicitly describe the correlations based on not only color features but also on a group of well-defined location attributes (climate, attraction) and clothing attributes (clothing item, color).

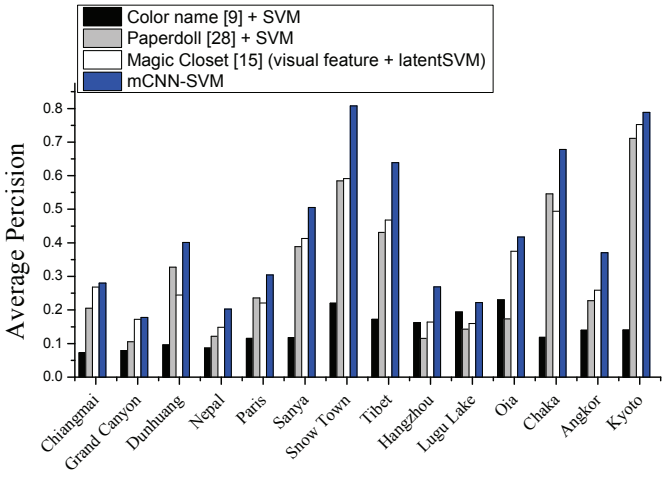

Fig. 8: Comparison of Average Precision for location-oriented clothing ranking on Journey Outfit Dataset.

We also present an empirical analysis to demonstrate the contributions of each component of our approach in Table 8. We find that by adding the location and clothing attributes, the performance is better than when using the visual features. This demonstrates that our selected mid-level attributes serve as good bridges. Finally, adding all of the attributes, the CNN-SVM generates the best results.

TABLE 8: Comparison of mAP for location-oriented clothing ranking on Journey Outfit Dataset.

Color Names [41] + SVM	0.1389
Paper Doll [5] + SVM	0.3083
Magic Closet [4] (visual features + latentSVM)	0.3371
mCNN-SVM-1 (visual features)	0.2999
mCNN-SVM-2 (with clothing attributes)	0.3687
mCNN-SVM-3 (with location attributes)	0.3987
mCNN-SVM (with clothing, location attributes)	0.4330

Figure 8 shows a comparison with the other methods for each of the 14 locations. For example, in Snow Town, the mCNN-SVM has a huge performance gain because it is always cold during the travel season and the frequently appearing scenery is white snow, so the location attributes ('*Climate:Season*', '*Climate:Weather*', '*Attraction:Type*') show little variation. However, for a location such as the 'Grand Canyon', which has four distinct seasons suitable for traveling, the intra-class variation for climate is greater. Therefore, our approach is more suitable for locations with distinct location attributes.

6.3 Case Study

In this section, we present an interesting case study on location-oriented clothing recommendations to demonstrate the effectiveness of our proposed approach. Now, let's assume that a young woman is preparing for a journey to Sanya in China, which is renowned for its tropical climate and sunny beaches. We will show how our proposed approach helps her to prepare her outfits. In particular, we provide 1) clothing item recommendations, 2) dressing pair recommendations, and 3) visual aesthetic recommendations.

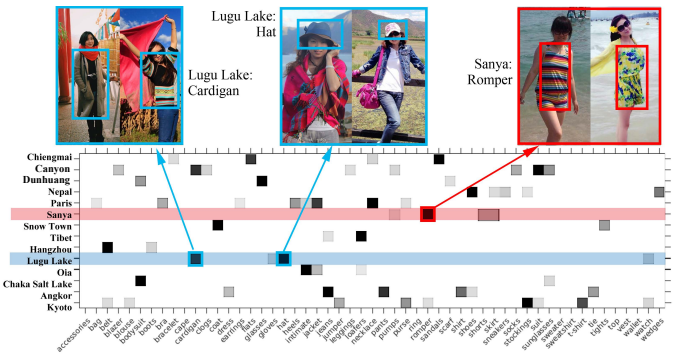

Fig. 9: Examples of the recommended clothing items for 14 destinations. In each row, the darker the nodes, the more appropriate the clothing items are for each destination.

Clothing item recommendations. By leveraging the proposed mCNN-SVM, we can provide the best correlated clothing items for a specific destination. Figure 9 shows the frequently occurring items for 14 destinations based on our mCNN-SVM, which reflects the exact correlations between locations and clothing attributes and verifies the rationality of this work. It is interesting that locations with similar climates may suggest different outfits. For example, though Angkor and Sanya are both hot places, in Angkor, it

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2696825, IEEE Transactions on Multimedia

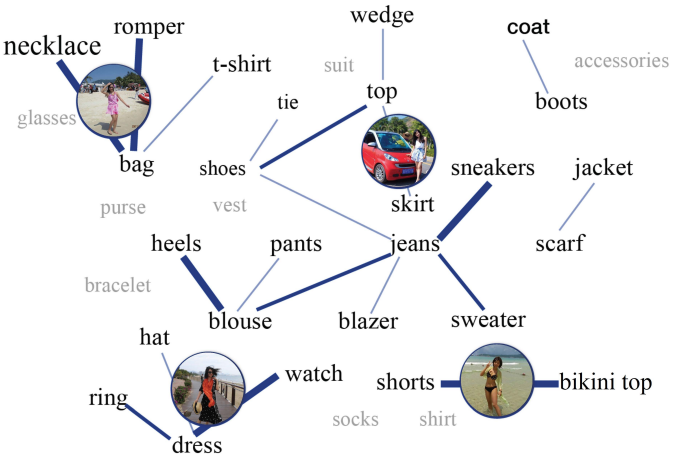

Fig. 10: The learned co-occurrence structure for 'Sanya'. The bolder the links are, the stronger the correlations. Different pairs reflect different recommended dressing styles. For example, 'bikini top' and 'shorts' is quite different from 'sneakers' and 'jeans'.

is recommended that visitors wear a dress or jeans covering the legs, while in Sanya, shorts and skirts are more typical. For different locations, the most discriminative attributes (climate or attraction) are different.

Since this woman is going to Sanya, it is appropriate to wear a 'romper', 'shorts' and a 'skirt' according to our approach. Why is a 'bikini top' less popular at such a sunny beach? This is because Sanya is located in China, and our approach has learned appropriate dress mostly from conservative Chinese travelers. If you do not want to stick out as a traveler, take our thoughtful advice and take a 'romper' to Sanya. Next, if the woman wants to go to Lugu Lake, the proper clothing is a 'cardigan' and a 'hat'.

Dressing pairs recommendation. We will further discuss how to match the selected clothing for your destination. Figure 10 shows the learned clothing co-occurrence structure for Sanya. Different combinations of clothing items reflect different dressing styles. As expected, our method successfully assigns greater weights to the location-related clothing pairs. For example, in Sanya, 'romper', 'bag' and 'necklace' are frequently seen. Moreover, users can choose a dressing style that fits their own needs. For example, the woman can choose the 'bikini top' and 'shorts' pair if she is confident about her body shape or choose the 'sneakers' and 'jeans' pair if she does not want too much exposure.

Visual aesthetic recommendation. We believe that the young woman has more than one pair of shorts; which one should she take? Our approach has already learned the main colors of Sanya and the majority of colors chosen for clothing by travelers from the photos of Sanya, with the top colors being shown in Figure 11. The woman can select altogether 5 colors from the background and clothing colors according to her preference. The system then automatically maps different color combinations into 16 aesthetic effect categories, using the methods in [46]. In Figure 11, the aesthetic effect is represented as a two-dimensional space (warm-cool and hard-soft), where each point in the space corresponds to a five color combination [47]. For example,

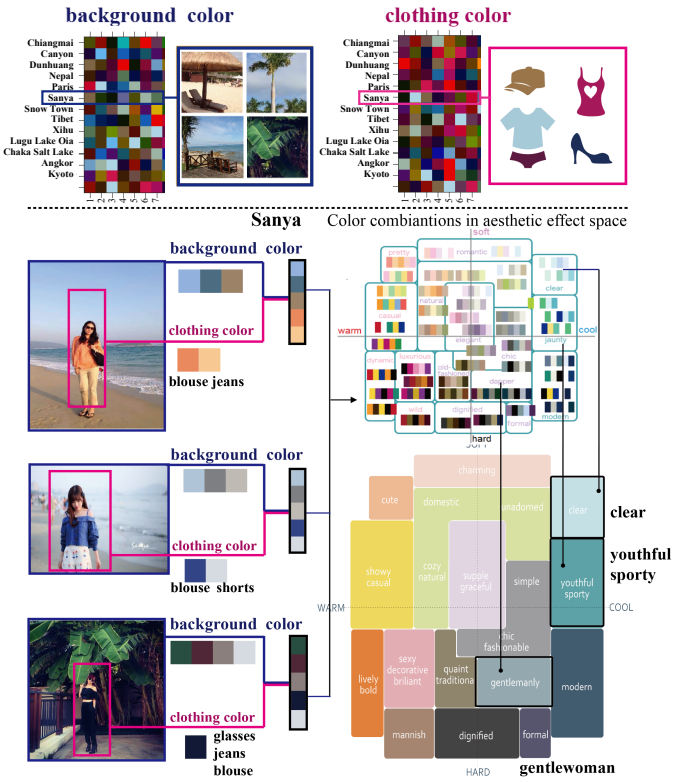

Fig. 11: Color recommendations. The most frequent background colors and clothing colors identified are listed for each location. Users can select clothing color combinations according to their personal tastes. For example, there are three combinations shown in example photos for Sanya representing 'youthful and sporty', 'clear', and 'gentlewoman'.

if the woman chooses the yellow blouse and the blue background, it gives a 'youthful and sporty' feeling. Therefore, our aesthetic guidance is not only based on a data-driven method but also obeys aesthetic rules in accordance with users' preferences.

Through these three steps, we can help this young woman be the most fashionable woman on her trip.

7 Conclusion

Recommending location-oriented clothing is an interesting and novel task. This paper is the first to study the correlation between clothing and location automatically using online travel photos. We propose a hybrid mCNN-SVM approach, which is classic and effective at improved attribute extraction and explicitly formulates correlations. A benchmark dataset for location-oriented clothing recommendations is proposed. We further provide a location-oriented clothing recommendation in three steps, giving the answers to what items to wear, how to match items and how to dress in an aesthetically pleasing manner. Our method has great extensibility and can support clothing recommendations for new locations that are not in the training dataset because the well-defined location attributes provide an efficient method for measuring the similarity between locations. In the future, we will enlarge the Journey Outfit Dataset in order to investigate sophisticated deep learning methods for learning clothing-location correlations.

REFERENCES

[1] E. Simo-Serra, S. Fidler, F. Moreno-Noguer, and R. Urtasun, "Neuroaesthetics in fashion: Modeling the perception of fashionability," pp. 869–877, 2015.

[2] V. Jagadeesh, R. Piramuthu, A. Bhardwaj, W. Di, and N. Sundaresan, "Large scale visual recommendations from street fashion images," in *SIGKDD*. ACM, 2014, pp. 1925–1934.

[3] 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, pp. 4642–4650.

[4] S. Liu, J. Feng, Z. Song, T. Zhang, H. Lu, C. Xu, and S. Yan, "Hi, magic closet, tell me what to wear!" in *MM*. ACM, 2012, pp. 619–628.

[5] K. Yamaguchi, M. H. Kiapour, and T. Berg, "Paper doll parsing: Retrieving similar styles to parse clothing items," in *ICCV*. IEEE, 2013, pp. 3519–3526.

[6] S. Liu, X. Liang, L. Liu, K. Lu, L. Lin, and S. Yan, "Fashion parsing with video context," in *MM*. ACM, 2014, pp. 467–476.

[7] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg, "Parsing clothing in fashion photographs," in *CVPR*. IEEE, 2012, pp. 3570–3577.

[8] W. Yang, P. Luo, and L. Lin, "Clothing co-parsing by joint image segmentation and labeling," in *CVPR*, 2014, pp. 3182–3189.

[9] S. Liu, J. Feng, C. Domokos, H. Xu, J. Huang, Z. Hu, and S. Yan, "Fashion parsing with weak color-category labels," *TMM*, vol. 16, no. 1, pp. 253–265, 2014.

[10] 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*. IEEE, 2012, pp. 3330–3337.

[11] M. H. Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg, "Where to buy it: Matching street clothing photos in online shops," in *ICCV*. IEEE, 2015, pp. 3343–3351.

[12] Y. Hu, X. Yi, and L. S. Davis, "Collaborative fashion recommendation: A functional tensor factorization approach," in *MM*. ACM, 2015, pp. 129–138.

[13] M. H. Kiapour, K. Yamaguchi, A. C. Berg, and T. L. Berg, "Hipster wars: Discovering elements of fashion styles," in *ECCV*. Springer, 2014, pp. 472–488.

[14] S. C. Hidayati, K.-L. Hua, W.-H. Cheng, and S.-W. Sun, "What are the fashion trends in new york?" in *MM*. ACM, 2014, pp. 197–200.

[15] S. Vittayakorn, K. Yamaguchi, A. C. Berg, and T. L. Berg, "Runway to realway: Visual analysis of fashion," in *WACV*. IEEE, 2015, pp. 951–958.

[16] K. Chen, K. Chen, P. Cong, W. H. Hsu, and J. Luo, "Who are the devils wearing prada in new york city?" in *MM*. ACM, 2015, pp. 177–180.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[18] S. Liu, X. Liang, L. Liu, X. Shen, J. Yang, C. Xu, L. Lin, X. Cao, and S. Yan, "Matching-cnn meets knn: quasi-parametric human parsing," in *CVPR*, 2015, pp. 1419–1427.

[19] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Is object localization for free?-weakly-supervised learning with convolutional neural networks," in *CVPR*, 2015, pp. 685–694.

[20] S. Jiang, X. Qian, J. Shen, Y. Fu, and T. Mei, "Author topic model-based collaborative filtering for personalized poi recommendations," *IEEE Transactions on Multimedia*, vol. 17, no. 6, pp. 907–918, 2015.

[21] L.-Y. Wei, Y. Zheng, and W.-C. Peng, "Constructing popular routes from uncertain trajectories," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 195–203.

[22] Z. Yu, H. Xu, Z. Yang, and B. Guo, "Personalized travel package with multi-point-of-interest recommendation based on crowd sourced user footprints," *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 1, pp. 151–158, 2016.

[23] J. Sang, T. Mei, J.-T. Sun, C. Xu, and S. Li, "Probabilistic sequential pois recommendation via check-in data," in *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*. ACM, 2012, pp. 402–405.

[24] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3128–3137.

[25] Q. Fang, J. Sang, and C. Xu, "Giant: Geo-informative attributes for location recognition and exploration," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 13–22.

[26] J. Bao, Y. Zheng, D. Wilkie, and M. Mokbel, "Recommendations in location-based social networks: a survey," *Geoinformatica*, vol. 19, no. 3, pp. 525–565, 2015.

[27] N. M. Kou, Y. Yang, Z. Gong *et al.*, "Travel topic analysis: a mutually reinforcing method for geo-tagged photos," *GeoInformatica*, vol. 19, no. 4, pp. 693–721, 2015.

[28] Q. Fang, J. Sang, C. Xu, and K. Lu, "Paint the city colorfully: Location visualization from multiple themes," in *International Conference on Multimedia Modeling*. Springer, 2013, pp. 92–105.

[29] Y. C. J. Wang, R. S. Feris, "Walk and learn: Facial attribute representation learning from egocentric video and contextual data," in *CVPR*, 2016.

[30] H. Izadinia and M. Shah, "Recognizing complex events using large margin joint low-level event model," in *ECCV*. Springer, 2012, pp. 430–444.

[31] J. van de Weijer, C. Schmid, J. Verbeek, and D. Larlus, "Learning color names for real-world applications," *Image Processing, IEEE Transactions on*, vol. 18, no. 7, pp. 1512–1523, July 2009.

[32] X. Wen, L. Shao, Y. Xue, and W. Fang, "A rapid learning algorithm for vehicle classification," *Information Sciences*, vol. 295, pp. 395–406, 2015.

[33] B. Gu, V. S. Sheng, K. Y. Tay, W. Romano, and S. Li, "Incremental support vector learning for ordinal regression," *IEEE Transactions on Neural networks and learning systems*, vol. 26, no. 7, pp. 1403–1416, 2015.

[34] L. Jing, L. Yang, J. Yu, and M. K. Ng, "Semi-supervised low-rank mapping learning for multi-label classification," in *CVPR*, 2015, pp. 1483–1491.

[35] S. Nowozin and C. H. Lampert, "Structured learning and prediction in computer vision," *Foundations and Trends® in Computer Graphics and Vision*, vol. 6, no. 3–4, pp. 185–365, 2011.

[36] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun, "Support vector machine learning for interdependent and structured output spaces," in *ICML*. ACM, 2004, p. 104.

[37] S.-J. Yang, Y. Jiang, and Z.-H. Zhou, "Multi-instance multi-label learning with weak label," in *AAAI*, 2013, pp. 1862–1868.

[38] S.-J. Huang, Z.-H. Zhou, and Z. Zhou, "Multi-label learning by exploiting label correlations locally." in *AAAI*, 2012.

[39] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *MM*. ACM, 2014, pp. 675–678.

[40] X. Wang, J. Jia, J. Yin, and L. Cai, "Interpretable aesthetic features for affective image classification," in *ICIP*. IEEE, 2013, pp. 3230–3234.

[41] R. Khan, J. Weijer, F. Khan, D. Muselet, C. Ducottet, and C. Barat, "Discriminative color descriptors," in *CVPR*, 2013, pp. 2866–2873.

[42] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *TKDE*, vol. 26, no. 8, pp. 1819–1837, 2014.

[43] P. O. Pinheiro and R. Collobert, "From image-level to pixel-level labeling with convolutional networks," in *CVPR*, 2015, pp. 1713–1721.

[44] K. Yamaguchi, T. L. Berg, and L. E. Ortiz, "Chic or social: Visual popularity analysis in online fashion networks," in *MM*. ACM, 2014, pp. 773–776.

[45] C.-N. J. Yu and T. Joachims, "Learning structural svms with latent variables," in *ICML*. ACM, 2009, pp. 1169–1176.

[46] J. Jia, H. Jie, S. Guangyao, H. Tao, L. Zhiyuan, L. Huanbo, and Y. Chao, "Learning to appreciate the aesthetic effects of clothing," in *AAAI*, 2016.

[47] S. Kobayashi, "Art of color combinations," *Kosdansha International*, 1995.