Fashion Parsing With Weak Color-Category Labels

Si Liu, *Member, IEEE*, Jiashi Feng, *Student Member, IEEE*, Csaba Domokos, Hui Xu, Junshi Huang, Zhenzhen Hu, and Shuicheng Yan, *Senior Member, IEEE*

***Abstract*—In this paper we address the problem of automatically parsing the fashion images with weak supervision from the user-generated color-category tags such as "red jeans" and "white T-shirt". This problem is very challenging due to the large diversity of fashion items and the absence of pixel-level tags, which make the traditional fully supervised algorithms inapplicable. To solve the problem, we propose to combine the human pose estimation module, the MRF-based color and category inference module and the (super)pixel-level category classifier learning module to generate multiple well-performing category classifiers, which can be directly applied to parse the fashion items in the images. Besides, all the training images are parsed with color-category labels and the human poses of the images are estimated during the model learning phase in this work. We also construct a new fashion image dataset called Colorful-Fashion, in which all 2,682 images are labeled with pixel-level color-category labels. Extensive experiments on this dataset clearly show the effectiveness of the proposed method for the weakly supervised fashion parsing task.**

***Index Terms*—Fashion parsing, Markov random fields, weakly-supervised learning.**

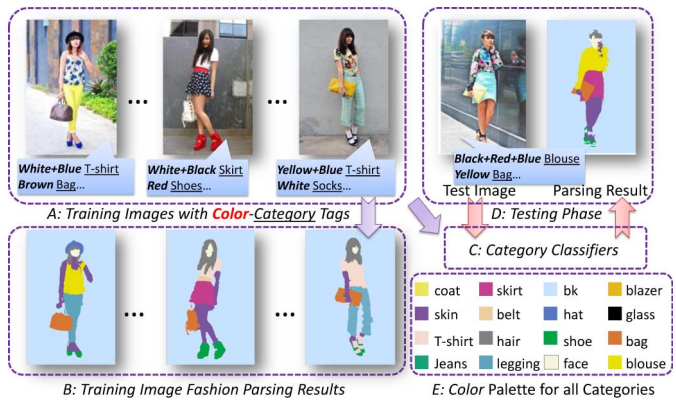


Fig. 1. An overview of weakly supervised fashion parsing problem. Given a set of training images with image-level color-category tags (A), we propose a fashion paring model to parse the fashion items in the training images (B) and output multiple well-performing category classifiers (C). The category classifiers (C) can be directly applied to parse any image with color-category tags (D). We assign a specific color for each fashion item (E). For better viewing, please see original color pdf.

I. Introduction

FASHION parsing is a relatively new research topic in computer vision, and is receiving growing attention of researchers due to the huge fashion market and the great potential of related applications, such as fashion attribute mining [2], clothes retrieval [21], clothes recommendation [20], clothes classification [3], [34] and clothes modeling [6]. The first work on fashion parsing was done by Hasan *et al.* [14]. They considered 4 categories, namely, shirt, jacket, tie, and face and skin, which is a rather limited number compared with the great diversity of fashion items. Later, Yamaguchi *et al.* [33] proposed an elegant framework for fashion parsing. However, their method requires pixel-level labels for model training, which costs enormous time and manual labor. According to [17], it may take 15–60 minutes to obtain pixelwise annotation for just one image. In order to perform fashion parsing with less time and labor cost, we propose a weakly supervised iterative approach utilizing color-category tags.

We observe that structured color-category tags are very common on fashion sharing websites. For example, we crawled 97,490 images with 292,541 tags from Chictopia.com, a popular fashion sharing website, and found **94**% of the tags have a color-category structure. Take the tag "brown bag" shown in Fig. 1(a) as a specific instance. The first term "brown" describes the color, and the second term "bag" indicates the category of the item in the image. Based on the correspondence between color and category tag in the structured tag "brown bag", roughly localizing the color tag "brown" can greatly assist finding the pixels belonging to category tag "bag". In this paper, given an image with (or without) color-category tags, fashion parsing means assigning both color and category tags to every pixel.

To parse the fashion images with color-category labels, we should pay attention to the following three aspects. Firstly, unlike the traditional fashion parsing studies [33] in a *supervised* setting, where pixel-level labels are provided in the training phase, our setting is *weakly supervised*, where only image-level tags are available in the training phase. For example, we only know the existence of "brown bag" in an image, but are unaware of which pixels belong to the "brown bag". Although compared with pixel-level labels, image-level labels are more common and natural in real scenarios, which makes our fashion parsing system more applicable, this setting also brings great challenge. Secondly, the intra-category variations are extremely large for fashion items. For example, the "bag" category have various

Manuscript received February 03, 2013; revised May 24, 2013 and August 20, 2013; accepted August 21, 2013. Date of publication October 11, 2013; date of current version December 12, 2013. This work was supported by the Singapore National Research Foundation under its International Research Centre @Singapore Funding Initiative and administered by the IDM Programme Office. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sheng-Wei (Kuan-Ta) Chen.

S. Liu, J. Feng, C. Domokos, J. Huang, and S. Yan are with the Department of Electrical and Computer Engineering, National University of Singapore (e-mail: dcslius@nus.edu.sg; jiashi@nus.edu.sg; eledc@nus.edu.sg; junshi.huang@nus.edu.sg; eleyans@nus.edu.sg).

H. Xu is with the Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing, China (e-mail: xuhui@cigit.ac.cn).

Z. Hu is with the Hefei University of Technology, Hefei, China (e-mail: huzhen.ice@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2013.2285526

TABLE I
DIFFERENCES BETWEEN THE METHOD PROPOSED BY [33] AND OURS. THEIR MODEL WORKS IN A FULLY SUPERVISED SETTING WHILE OUR MODEL WORKS IN A WEAKLY-SUPERVISED SETTING. THE INPUTS AND OUTPUTS OF THE TWO MODELS ARE DIFFERENT, WHICH MAKES THEM NOT COMPARABLE

	Input			Output	
	Image-level Color	Image-level Category	Pixel-level Category	Pixel-level Color	Pixel-level Category
Yamaguchi et al. [33]	✗	✗	✓	✗	✓
Our	✓	✓	✗	✓	✓

kinds of colors, patterns and materials. This increases the difficulty of constructing reliable category classifiers, especially in the absence of pixel-level labels. Thirdly, the great variations of human poses make the fashion parsing difficult. In fashion images people usually pose quite freely. For example, their hands may be put around the waist, chin or forehead.

In this work, we propose an iterative framework to parse fashion images. The whole framework is shown in Fig. 1. In the training phase, a pose estimation module is performed to determine the locations of the human key points, such as head, neck, left/right shoulders, etc. Accurate pose estimation provides important spatial cues for fashion items. For example, identifying the location of head makes it easier to localize "hair". Then based on the training images with image-level color-category tags (Fig. 1(a)) and the estimated human poses, we infer the color and category labels for each pixel via Markov Random Fields (MRF). The joint inferring of colors and categories is critical for accurate fashion parsing. Supposing an image contains a "brown bag" tag, the "brown" and "bag" labels should be turned on/off simultaneously for each pixel. When we cannot correctly localize "bag" by category classifiers, we can easily find the label by identifying "brown". After the fashion parsing module is performed, the pixels as well as the inferred color-category tags are collected to update the category classifiers, which is linear Support Vector Machine (SVM) in our work. The human pose estimation module, MRF-based fashion parsing module and category classifier training module are iterated for several rounds. After the iterations converge, all training images are parsed (Fig. 1(b)) and the final category classifiers (Fig. 1(c)) are obtained. These classifiers can then be used for parsing new images (Fig. 1(d)).

The major contributions of this work can be summarized as follows:

- To the best of our knowledge, this work is the first attempt to explore the challenging weakly supervised fashion parsing problem. Our attempt is feasible and practical, since the fashion images with color-category tags are very common on the Internet. Fashion parsing is a fundamental problem, and accurate parsing can help other fashion related applications, such as fashion search and fashion editing.
- To investigate this problem, we construct a large Colorful-Fashion dataset. It consists of 2,682 images in total, and all the pixels in the images are annotated with both color labels (including 13 types) and category labels (including 23 types). It will serve as a good benchmark for fashion data analysis and can be downloaded from https://sites.google.com/site/fashionparsing/home.
- We propose a general fashion parsing framework which can: 1) select the best pose from the top-3 pose candidates for all training images; 2) simultaneously and thoroughly parse all training images with pixel-level color and category labels; and 3) obtain several well-performing category classifiers which can be applied directly to new images.

The paper is organized as follows. In Section II we briefly review the related work. In Section III, we introduce the Colorful-Category database constructed in our work. We describe the proposed fashion parsing model in Section IV. In Section V, we discuss the optimization process and implementation issues of the model. In Section VI, we conduct extensive experiments and demonstrate the effectiveness of our model. Finally, we conclude our paper in Section VII.

II. RELATED WORK

A. Fashion Parsing

Not much research work has been done for the fashion parsing problem. Hasan *et al.* [14] first worked on the fashion parsing problem, but they only considered 4 categories, including shirt, jacket, tie, and face and skin. Instead, we target at complete fashion parsing, involving 13 colors and 23 categories. The work of Yamaguchi *et al.* [33] is most related to ours. They proposed an elegant framework to iteratively refine the human detection results and fashion parsing alternatively. Our work differs from their work in that we do not use any pixel-level labels in the training phase. Based on this significant difference, our problem can be defined as a *weakly supervised* parsing problem, while what other works deal with as a *supervised* parsing problem. The differences between the two methods in terms of *input* and *output* are as listed in Table I. For the input, the method proposed by Yamaguchi *et al.* [33] requires pixel-level category labels for training, which cost the labelers great amount of time and labor, and is not applicable for large-scale data. To the contrary, our method only needs the image-level color-category labels, which can be easily obtained by crawling any fashion website. For the output, Yamaguchi *et al.* [33] only assign a category label for each pixel. However, our method can assign both a color and a category label for each pixel, which is more complete.

Besides fashion parsing, other fashion related problems, such as fashion attribute mining, clothes retrieval, clothes recommendation, clothes classification and clothes modeling, have received growing attention in the past few years. Berg *et al.* [2] dealt with the fashion attribute problem, but focused on automatically discovering common attribute terms and characterizing attributes according to their visual representation. Different from [2], we aim to assign the attribute and category labels to the corresponding pixels, which provides a comprehensive understanding of the fashion images. In our early work

[21], we proposed to enhance clothes retrieval via human parts alignment and auxiliary set. Later, we developed a clothes recommendation system called magic closet [20]. The key idea is to use latent SVM to model the relationships between different clothes items. The clothes classification task was first explored by Yang *et al.* [34] and later by Chen *et al.* [6] who refined the classification results by the contextual informations via a Conditional Random Fields model. The latest work of clothes classification is by Bossard *et al.* [3], who proposed a random forest based system and obtained very competitive results. Finally, clothes modeling was explored by Chen *et al.* [7] via And/Or Graphs. Actually, fashion parsing can be viewed as a pre-step for these fashion related applications, and boost their performances.

B. Weakly Supervised Image Parsing

In recent years, with the popularity of the images with weak labels on the Internet, researchers have started to pay more attention to the problem of weakly supervised image parsing [29], [31]. Vezhnevets *et al.* [29] improved the parsing performance by importing information from a geometrical task within a multi-task learning framework. Later, they built a multi-image model [31] by exploiting the manifold structure of patches. Based on the specific characteristics of the fashion domain, our method considers color and category labels simultaneously, which significantly distinguishes our work from the large body of related works considering category parsing only.

Some other works [13], [22], [23], [36] adopt sparse coding related methods (with different structure priors) to parse the testing image without training process. More specifically, Liu *et al.* [22] used bi-layer sparse coding for better reconstruction. Later, they proposed to refine the image parsing by Internet searching results [23]. Han *et al.* [13] and Yang *et al.* [36] considered spatial information between patches to achieve robust label transfer. All these methods need a possibly large-scale training data set in the testing phase, while our model does not.

Wang *et al.* [32] presented a method to learn color attributes and object classes together under a weakly-supervised setting. Their method is based on multiple instance learning while our method is based on MRF inference. In addition, we focus on the specific domain of fashion parsing. Thus, we can use many kinds of domain specific techniques, such as human pose estimation, to achieve better parsing.

III. Dataset Construction

Actually, our database is much larger than most of the existing semantic labeling database. The MRSC dataset [27] contains 591 images, while CamVid dataset [4] contains 711 images. LabelMe subset (also known as SIFT Flow dataset) [19] includes 2688 images. Recently, Yamaguchi *et al.* constructed a fashion dataset [33]. However, this dataset is relatively small-scale, containing only 685 images, and does not provide any color labels. To learn and evaluate the color-category fashion parsing model, we collect our own fashion dataset in this work, called Colorful-Fashion dataset, which is about 3 times larger than the dataset from Yamaguchi *et al.* [33]. The dataset can be downloaded from https://sites.google.com/site/fashionparsing/home.

Fig. 2. An example as well as its associated tags crawled from Chictopia.com. Color tags are highlighted in their representative colors, while category tags are highlighted by underline.

A. Image Collection

All the images in the Colorful-Fashion dataset are collected from Chictopia.com, a fashion sharing website. We crawl 97,490 photos in total along with their associated tags. One example of the dataset is shown in Fig. 2. All the 97,490 photos have good visibility of the full body and we randomly select a subset containing 2,682 photos to form the final Colorful-Fashion dataset. Then, we randomly select half of the 2,628 images as the training set, and take the other half as the testing set.

B. Color-Category Tags

Fig. 2 shows an example photo and its associated tags from Chictopia.com. We can see that each fashion image is described by multiple rows of tags. Generally, each row describes a specific fashion item, such as "Camel Asos Bag". The three terms indicate color ("Camel"), brand ("Asos") and category ("Bag"), respectively. The user-generated tags on the website are usually incomplete and may contain some errors, so we manually annotate all the 2,682 images with complete color-category tags.

In the dataset construction process, we only keep 13 colors and 23 categories of tags. To balance the labeling efficiency and accuracy, we first over-segment each image into about 400 patches by using a state-of-the-art image over-segmentation method [1]. Then, we ask the labelers to judge the main color and main category for each patch.

Table II shows the numbers of different colors in both training and testing subsets. For each color, the number of patches is shown without bracket and the number of corresponding images is shown in bracket. We select 13 colors by referring to *color naming* research [28] and previous fashion studies [20], [21]. Note that our dataset contains both solid color clothes and multiple-color clothes. For multiple-color clothes, for example, a T-shirt with black and white stripes, we label its black stripes as "black T-shirt" and label its white stripes as "white T-shirt". For the clothes with fully mixed or even messy colors, we cannot accurately label the colors of the clothes since the current labeling is conducted at the patch level. That is the limitation of our current system.

Table III shows the numbers of different categories in both training and testing subsets. For each category, the number of patches is shown without bracket and the number of corresponding images is shown in bracket. We summarize all the categories into five groups: head, upper body, lower body,

TABLE II
FOR EACH COLOR, THE NUMBERS OF PATCHES IN TRAINING AND TESTING SUBSET ARE SHOWN IN THE FIRST AND SECOND ROW RESPECTIVELY. THE NUMBERS OF IMAGES CONTAINING THE COLOR ARE SHOWN IN BRACKET

	beige	*black*	*blue*	*brown*	*gray*	*green*	*orange*
train	31651 (1304)	19673 (1102)	11490 (520)	10139 (833)	3528 (224)	3989 (178)	2014 (111)
test	32085 (1307)	20157 (1118)	10928 (541)	9564 (830)	3296(202)	3960 (178)	1742 (108)
	pink	*purple*	*red*	*white*	*yellow*	*bk*	
train	6666 (328)	2158 (106)	6249 (310)	13592 (793)	5817 (361)	452003 (1341)	
test	6864 (368)	2236 (100)	5879 (302)	13958 (803)	6096 (401)	452338 (1341)	

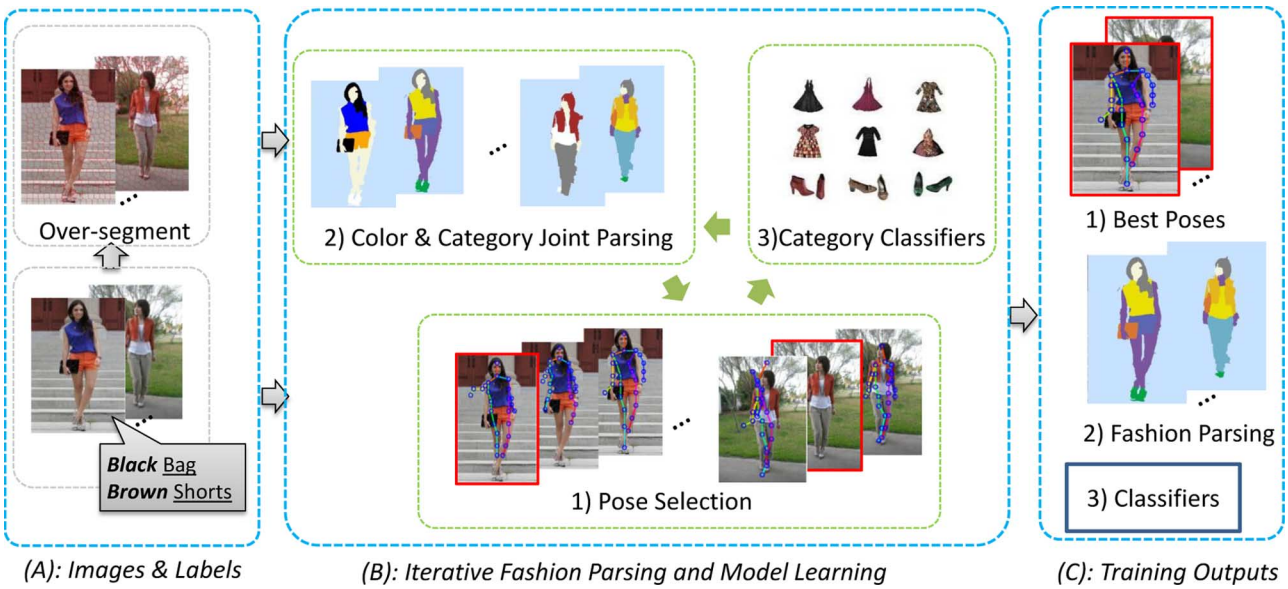


Fig. 3. The framework of our weakly-supervised fashion parsing system (A) Images are over-segmented [1] to decrease computational cost (B) Three components, namely human pose estimation, training images fashion parsing and category classifier training, are iteratively refined. (C) After the iteration converges, the optimal human poses and the fashion parsing results of the whole training image set are obtained. Other outputs of the training process are the well-performing category classifiers which can be used to parse new images.

TABLE III
FOR EACH CATEGORY, THE NUMBERS OF PATCHES IN TRAINING AND TESTING SUBSET ARE SHOWN IN THE FIRST AND SECOND ROW RESPECTIVELY. THE NUMBERS OF IMAGES CONTAINING THE CATEGORY ARE SHOWN IN BRACKET

	face	*sunglass*	*hat*	*scarf*	*hair*
HEAD	3629 (1337)	292 (220)	782 (190)	965 (116)	10806 (1330)
	3675 (1341)	272 (205)	620 (160)	890 (91)	10732 (1332)
	blazer	*T-shirt*	*blouse*	coat	*sweater*
UPPER	3811 (178)	6172 (460)	8669 (474)	3999 (170)	3030 (133)
	3917 (176)	6068 (457)	8198 (461)	4201 (186)	2795 (115)
	jeans	*legging*	*pants*	*shorts*	*skirt*
LOWER	2860 (113)	3124 (123)	3418 (103)	2451 (226)	10214 (438)
	2388 (87)	2699 (110)	2825 (86)	3021 (276)	10353 (445)
	shoe	*socks*	*stocking*		
FOOT	5184 (1300)	242 (83)	1831 (131)		
	5287 (1301)	284 (87)	2316 (157)		
	skin	*belt*	*bag*	*dress*	*bk*
OTHER	26830 (1324)	1056 (390)	6301 (723)	11300 (342)	452003 (1341)
	26690 (1330)	1174 (432)	6692 (739)	11680 (336)	452338 (1341)

foot-area and others. Following Hasan *et al.* [14], we divide the skin area into "face" and "skin". In the training subset, the number of faces is smaller than the image number, because faces are occluded by hat or black oversize sunglasses sometimes.

IV. FASHION PARSING MODEL

A. Overview of the Proposed Approach

Our framework is shown in Fig. 3. For the training set, we first adopt SLIC patch segmentation package [1] and obtain about 400 patches for each image, as shown in Fig. 3(a). Fig. 3(b) indicates the key part of our framework, i.e., the iteration among three modules: 1) pose selection, 2) color and category joint parsing and 3) category classifiers learning. One module in the three enhances the performance of the subsequent module. More specifically, accurate human pose selection will greatly facilitate fashion parsing, which will provide high-quality training samples for category classifiers training, and then the responses of the updated classifiers serve as criteria for human pose selection in the next iteration.

Since in our database, all the people in the images take relatively simple poses, i.e., standing almost in the frontal view, the state-of-the-art full body pose estimator [35] produces acceptable results. For each image, we keep the top-N poses estimated by [35], and select the optimal one, which is highlighted by a red bounding box in Fig. 3(b). The newly estimated human pose serves as the input to generate the location-aware features which are then fed into the fashion parsing module. The MRF based parsing module can embed image-level constraints and output both the color and category labels for each pixel in the image. The newly obtained parsing results are used to re-train the category classifiers, shown in Fig. 3(b).

After several rounds of iterations, the system should converge. As shown in Fig. 3(c), we can get 1) the best poses of all the training images, 2) fashion parsing results and 3) category

classifiers. The well-performing category classifiers can be directly applied for general fashion parsing in the testing phase.

B. Human Pose Estimation

Fashion parsing and human pose estimation are two closely related problems. On the one hand, accurate fashion parsing generally highly relies on precise human pose estimation. A fashion item usually appears regularly w.r.t. human skeleton. For example, T-shirts are worn on the upper body while shoes are usually on the feet. Moreover, restricting the fashion items near the human body can effectively suppress the false alarms in the background area. On the other hand, fashion parsing can also help infer better human pose configurations. For example, identifying a region as "T-shirt" implies that the region is highly likely to be the upper-torso.

For human pose estimation, we use the state-of-the-art full body pose estimator [35] which can output the locations of 26 human key points, such as head, neck, left/right shoulder and left/right knees. Most of the images in our dataset contain one single person with relatively simple pose (mostly standing), against relatively clean background. The pose estimator is acceptable in our relatively restricted scenario. To further decrease the effects of imperfect pose estimation results, following the similar ideas in [25], we keep the top-N (here $N = 3$) pose proposals with the highest confidences for each image. And in each round of model learning, the best pose proposal under the current model will be selected automatically. Note that our method is different from that proposed by Yamaguchi *et al.* [33]. They need to *estimate* the best pose configuration in each round, which may be time-consuming, while we only need to *select* the optimal one among the top-N possible pose configurations, which can generally cover the true positive and be computed in advance.

C. Image Patches and Features

Due to the large number of pixels for an image, directly operating on pixels will be computationally expensive. For the trade-off between the computational efficiency and parsing accuracy, we first over-segment the image into patches by the popular SLIC segmentation method proposed in [1]. Following the usual practice in related works [29], [31], we assume that each patch only contains one color and belongs to one category.

We extract four types of features for each patch, including color, SIFT [24], HOG [8] and location features. The color feature is composed of 3-dim mean color value, 24-dim color histogram and 100-dim color Bag of Words (BoWs) in both RGB and LAB color space. For SIFT and HOG features, we apply dense sampling strategy with 4×4 step size and then generate the BoWs features. The dictionary sizes for SIFT and HOG are both set as 300 [5]. Since the numbers of pixels and key points in one patch are very small, enlarging the vocabulary size will make the BoWs feature very sparse, which may decrease the performance. So 300 is a suitable value. Moreover, we extract two kinds of location features: 1) a 2-dim absolute location feature, which consists of mean horizontal and vertical coordinate of the patch, normalized by the image width and height; 2) a relative location feature by the distances between the patch and each human key point, normalized by the estimated human shoulder width. We then calculate the square of each dimension of the relative location feature. Since we have 26 human key points, the dimension of the final relative location feature is 26×4, which is 104-dim. The four kinds of features are normalized independently and then concatenated to form the final feature of the super pixel, which is 960-dim.

D. Fashion Parsing Model

In this work, we build a joint MRF inference model for the fashion parsing. We first introduce the notations. Supposing we have N images, the color-category tags for the i^{th} image are denoted by $Y_i = \{C_i, A_i\}$, where C_i is the category tags, and A_i is the color tags. After over-segmentation, the i^{th} image is partitioned into $n_i \approx 400$ patches $\{x_{i,1}, x_{i,2}, \dots x_{i,n_i}\}$. When there is no ambiguity, we denote both the patch itself and its feature vector as x. And the tags for the patch $x_{i,j}$ are denoted $y_{i,j} = \{c_{i,j}, a_{i,j}\}$.

We treat each patch as a node in the MRF model and each node is associated with two types of labels, i.e., the color and category labels. The unary and pairwise term of the adopted MRF model are defined as follows.

Unary Term: Classifier Responses In the proposed MRF model, unary term contains two potentials. The first one is called category potential. Suppose μ_m is the classifier parameter for the m^{th} category, and the potential for the j^{th} patch $x_{i,j}$ w.r.t. the class m is defined as:

$$U_c(x_{i,j}, m; p_i) = -\log\big(f_c(\mu_m, x_{i,j})\big). \tag{1}$$

Here p_i denotes the estimated human pose and f_c is the multi-class category classifiers. Note that the feature $x_{i,j}$ (more precisely, the location feature in $x_{i,j}$) depends on the current pose. Thus this term is conditioned on the pose p_i.

The second potential is called color potential. Similarly, the potential for the j^{th} patch $x_{i,j}$ w.r.t. the color n is defined as:

$$U_a(x_{i,j}, n) = -\log\big(f_a(\eta_n, x_{i,j})\big), \tag{2}$$

where f_a is the multi-class color classifiers, and η_n is the classifier parameters.

Joint Unary Potential: Mutual Constraints Between Color and Category This term is used to incorporate the mutual constraints between color and category labels for each patch. The constraint is from the image level annotation. For example, in Fig. 4, the image contains "black skirt". From this cue, we can infer that "black" and "skirt" should be turned on/off simultaneously. The term is defined based on the image labels as follows,

$$\pi(Y_i, \{c_{i,j}, a_{i,j}\}) = \begin{cases} 0, & \{c_{i,j}, a_{i,j}\} \in Y_i \\ \infty, & \{c_{i,j}, a_{i,j}\} \notin Y_i \end{cases}. \tag{3}$$

Equation (3) can well handle the cases where all the color-category tags are correct. If the tags contain noise, we need to modify it. Instead of setting the potential to positive infinity, we set it as a parameter κ. Therefore, (3) is modified to:

$$\pi(Y_i, \{c_{i,j}, a_{i,j}\}) = \begin{cases} 0, & \{c_{i,j}, a_{i,j}\} \in Y_i \\ \kappa, & \{c_{i,j}, a_{i,j}\} \notin Y_i. \end{cases} \tag{4}$$

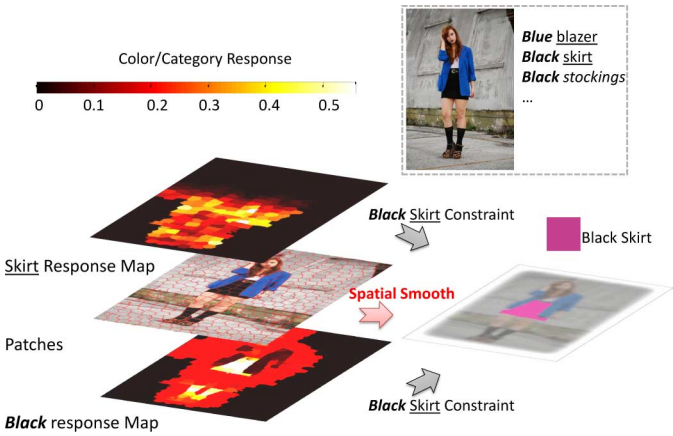


Fig. 4. An illustration of the proposed joint MRF model for fashion parsing and an exemplar result. Here the heat of the category ("skirt") response map and color ("black") response map indicates the confidence score. In the "skirt" response map, the strongest response does not correspond to the true "skirt" patches. However, with the help of the "black" response map, the MRF model can fuse the color information and spatial smooth prior to correctly infer the "black skirt" patches.

The value of κ is determined via 5-fold cross validation on the training set.

Pairwise Term: Spatial Smoothness A useful prior in the fashion parsing task is that the spatially neighboring patches in the image usually share the same label. Such a spatial smooth regularization term is widely adopted in other MRF models [5], [9], [30], [31], [33].

The corresponding pairwise term is defined as:

$$\sigma(y_{i,j}, y_{i,k}, x_{i,j}, x_{i,k}) = \begin{cases} -\ln \rho(B_{x_{i,j},x_{i,k}}), & y_{i,j} \neq y_{i,k} \\ 0, & y_{i,j} = y_{i,k} \end{cases} \tag{5}$$

where $\rho(B_{x_{i,j},x_{i,k}})$ is the boundary strength and estimated by the method introduced in [9], [15]. Typically, a penalty is enforced for assigning different labels to adjacent patches. The penalty is lowered for label disagreement when there is a strong boundary strength $\rho(B_{x_{i,j},x_{i,k}})$ between adjacent patches. If the labels are the same, there is no penalty.

Total Energy Function The total energy function of the proposed MRF model is defined as the weighted summation of the above potentials over all the training samples as follows:

$$\begin{aligned} \mathcal{E} = & \sum_{i,j} U_c(x_{i,j}, m; p_i) + \alpha_1 \sum_{i,j} U_a(x_{i,j}, n) \\ & + \alpha_2 \sum_{i,j} \pi(Y_i, \{c_{i,j}.a_{i,j}\}) \\ & + \alpha_3 \sum_{i,j,k} \sigma(y_{i,j}, y_{i,k}, x_{i,j}, x_{i,k}). \end{aligned} \tag{6}$$

The weight parameters α_1, α_2 and α_3 are determined by 5-fold cross validation on the training set. The optimization and implementation details are introduced in the following section.

V. Optimization and Implementation

A. Training Phase

As aforementioned, the model training process is the iterations among three modules, namely 1) pose selection; 2) patch labels inference and 3) category model update. Supposing the model has been initialized appropriately (the details will be introduced in Section V-C), the above three training modules are conducted iteratively as follows.

Pose Selection: By using the off-the-shelf human pose detector [35], for each training image, we can obtain a set of human pose candidates along with their confidence scores. Then we rank the poses according to their confidence scores. Due to the large diversity of human poses and the limitations of all current pose estimation algorithms, the top pose may be inaccurate. However in this paper, since the images in the database contain not so complicated backgrounds and near frontal view people, we assume that the accurate pose is very likely to be among the top-N poses. Therefore, we propose to select the best pose from the top-N poses, which best fits the category classifiers. Mathematically, the human pose p_i^* for each training image X_i is determined by:

$$p_i^* = \arg \min_{p_i \in \mathcal{P}_i} \sum_j U_c(x_{i,j}, m; p_i). \tag{7}$$

Here $x_{i,j}$ incorporates the location features defined by the current pose estimation p_i. And $\mathcal{P}_i$ is the set of top-N candidate human poses.

Training Images Parsing: After the optimal human pose p_i^* is determined and all the features are fixed, we apply MRF on the training images to infer the color and category labels for each patch. The patch labels $\{c_{i,j}^*, a_{i,j}^*\}$ for the training image X_i with image-level label Y_i are obtained by minimizing the image specific energy function defined as follows,

$$\begin{aligned} \{c_{i,j}^*, a_{i,j}^*\} &= \arg \min_{\{c_{i,j}, a_{i,j}\}} \mathcal{E}(\{c_{i,j}, a_{i,j}\} | X_i, Y_i) \\ &= \arg \min_{\{c_{i,j}, a_{i,j}\}} \Big\{ \sum_j U_c(x_{i,j}, m; p_i^*) \\ &\quad + \alpha_1 \sum_j U_a(x_{i,j}, n) \\ &\quad + \alpha_2 \sum_j \pi(Y_i, \{c_{i,j}.a_{i,j}\}) \\ &\quad + \alpha_3 \sum_{j,k} \sigma(y_{i,j}, y_{i,k}, x_{i,j}, x_{i,k}) \Big\}. \end{aligned} \tag{8}$$

The terms U_c, U_a, π and σ in (8) are introduced in Section IV-D. Note that in the model updating process, we have implicitly embedded the mutual constraints between the color and category labels in the joint unary term π.

Category Model Update: The fashion parsing results provide training samples with patch-level annotations for updating category models. Since the initial color model η_n is quite robust, we only update category model μ_m in this step. We re-train the linear SVM classifiers with the newly obtained patch data.

Category Classifiers: The training output is the category classifiers μ_m for $\forall m$, which can be applied to any new image for fashion parsing.

B. Testing Phase

Given a test image X_t along with the color-category tags Y_t, we first over-segment it into small patches $x_{t,j}$, where j de-

notes the j^{th} patch of X_t. Then the category and color labels $\{c^*_{t,j}, a^*_{t,j}\}$ for each patch $x_{t,j}$ along with the human pose p^*_t are obtained by minimizing the image specific energy function as (6) in the training phase. Note that the category classifiers μ_m used here are obtained in the training phase. Thus, no iteration is needed in the testing phase, which makes our system efficient in the practical point of view. Based on the classifiers, we first enumerate each pose p_t among all possible pose set $\mathcal{P}_t$ for the testing image X_t, and we pick the optimal pose p^*_t which produces the minimum value via (7).

After determining the optimal pose p^*_t, we further infer all the patches' color labels $a^*_{t,j}$ and category labels $c^*_{t,j}$ via (8).

Our model can also be extended to handle images without any tags. In this case, the mutual constraint in the joint unary potential (3) is set as zero.

C. Initialization and Implementation Details

1) Initialization: The initialization process contains two steps, namely, first initializing color classifiers and then category classifiers.

First, we initialize the color classifiers utilizing the popular and well-performing search engines. We query Google with each color name for all the 12 foreground colors and crawl the top 200 returned images. Then we train some color classifiers using the color features introduced in Section IV-C. We apply the color classifiers on the Colorful-Fashion dataset to predict the color labels of the patches and average precision of all colors is 0.753.

Second, we get several category labels of patches based on the inferred color labels of patches and the color-category labels of images. For example, given image labels {"red skirt", "black hair", "black shoes" }, where "black" corresponds to both "hair" and "shoes" while "red" corresponds to "skirt" only, all patches predicted as "red" should also be labeled as "skirt" due to the one-one correspondence between color and category labels. Due to the large scale of Colorful-Fashion containing the one-one correspondence between color and category tags, sufficient patches can be collected. In our experiments, for each foreground category, averagely we can get 4,747 patches which are used to train the initial category classifier (linear SVM). We use off-the-shelf LibLinear package [11] to train the L2-regularized L2-loss linear multi-class SVM due to its fast training speed. The package adopts the one-vs-the-rest strategy by Crammer and Singer [16]. The average precision $(\#correct\ samples\ /\ \#samples)$ for foreground categories is 68.3%, which guarantees the performance of subsequent iterations. To further refine the category classifier, we include spatial constraints to filter out noises. For example, "shoe" only appears in the lower part of the image. The final average precision for foreground categories is increased to 77.0%. Note that only the foreground categories collect training samples in the aforementioned way due to the consistent appearance of each foreground category across the dataset. For example, "face" always has similar colors. However, the backgrounds are diverse in different images. For instance, "bk" can be city street, countryside, walls, etc. We collect background data separately for each image. The details are introduced in the following.

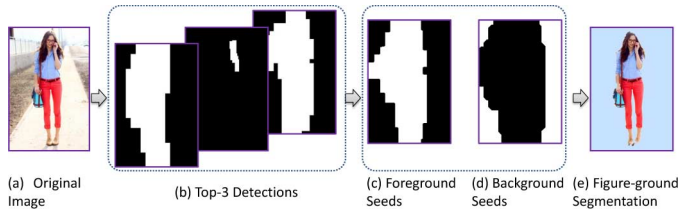


Fig. 5. Figure-ground segmentation. The original image (a) is fed into the pose estimation module. The bounding boxes of top-3 poses are shown in (b), where white regions correspond to figure (foreground) regions while black seeds are ground (background) regions. Based on the pose estimation results in (b), we can get the foreground (c) and background regions (d) which are the input of the figure-ground segmentation module. The results are shown in (e), where the background regions are marked in light blue color. Better view in color.

2) Implementation Details: Here, we introduce several implementation details. First, as aforementioned, we consider the patches generated by over-segmentation [1] as processing units. After a patch's label is inferred, all the pixels inside the patch inherit the label.

Second, figure-ground segmentation is conducted for each image. Here, figure-ground segmentation [18] means that the image is split into foreground regions and background regions. The flowchart of the figure-ground segmentation is show in Fig. 5. For the image shown in Fig. 5(a), its top-3 poses are estimated [35] and shown in Fig. 5(b). The union of the top-3 poses is eroded and treated as the foreground seeds (see Fig. 5(c)). Meanwhile, the rest regions are eroded and considered as background seeds (see Fig. 5(d)). The foreground and background regions are fed into the GrabCut algorihtm [26] to generate the final figure-ground segmentation (see Fig. 5(e)). Then patches inside the final foreground regions are further classified into the 23 categories (including "bk") by our category classifiers while all patches inside the final background region are fixed as "bk". It is reasonable, since all fashion items should be near human body which is always covered by the top-3 human pose estimation results.

The main computational cost of our fashion parsing model comes from the MRF inference step. We use alpha-expansion implemented by GCMex [12] to solve the Graph Cut problem, and the computational complexity is O($|\mathrm{P}|\,|\alpha|$), where $|\alpha|$ is the number of labels. In our experiment, it is number of colors multiplied by the number of categories. So in summary, the computational cost is linear in terms of the number of patches and also linear in terms of the number of images.

VI. Experiments

In this section, we conduct extensive experiments on the collected Colorful-Fashion Dataset to evaluate the effectiveness of our proposed model.

A. Experimental Setting

The whole Colorful-Fashion dataset is divided into a training set and a testing set. In the model learning process, we adopt

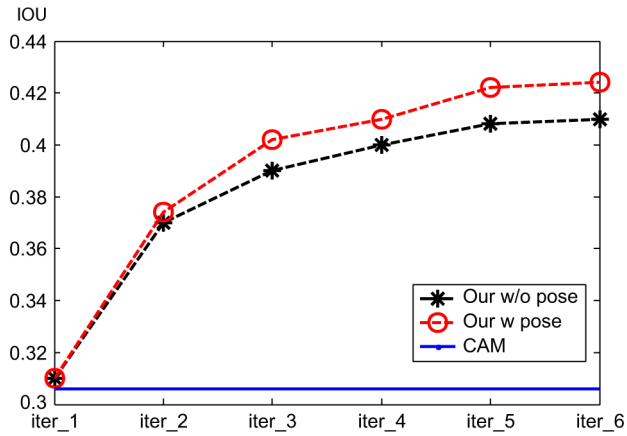


Fig. 6. Parsing performance of three methods against the iteration number, namely our model with pose selection (red dashed line with circle masks), our model without pose selection (black dashed line with asterisk masks) and baseline CAM (blue solid line). Our two models are improved along with the iterations.

5-fold cross-validation to tune the parameters. Experimental results are reported for both training set and testing set.

Following existing semantic parsing studies [10], [31], we use Intersection Over Union (IOU) to evaluate different algorithms. It measures the overlap between our parsing results and the ground truth. It is widely used in evaluating segmentation and detection performance. The IOU is calculated as:

$$\mathrm{IOU} = \frac{\#\text{true positive}}{\#\text{false positive} + \#\text{false negative} + \#\text{true positive}}.$$

B. Quantitative Analysis of Fashion Parsing W.R.T Iteration

In this subsection, we show the evaluation results based on the IOU metric to demonstrate the parsing model evolution along with the iterations. The performance of the proposed model on the training set w.r.t. the number of iterations is plotted in Fig. 6. We can see that the parsing performance of the model is improved steadily. This demonstrates that our proposed model can well utilize the tight correlation between color and category information. Generally, the optimization for learning parsing model converges quite fast. As shown in the figure, it converges in about 5 iterations. Note that the iteration is only necessary in the training phase to gradually 1) estimate human poses, 2) conduct fashion parsing and 3) update category classifiers. However, in the testing phase, as shown in Section V-B, no iteration is needed.

C. Qualitative Analysis of Fashion Parsing W.R.T Iteration

To prove that color can assist category classification in the parsing process, we freeze the pose selection module and always adopt the top pose. In this way, the effect of pose estimation can be removed. Then we evaluate the performance of the proposed model w.r.t. the iterations to investigate the effectiveness of color and category classification. Some exemplar results are presented in Fig. 7, from which we can observe that the results from the initial model are not satisfactory. Some patches of the background are mis-classified as fashion items. However, along with the iterations, the color information can help identify

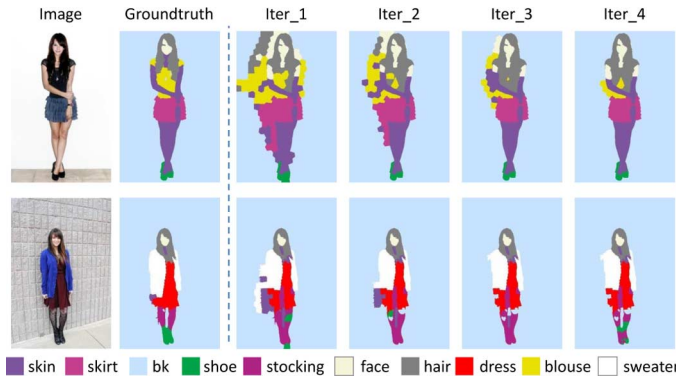


Fig. 7. Parsing results on the training set in the first 4 iterations. Both the ground truth and parsing results are shown along with the original images. Different categories are indicated by different colors, as defined at the bottom.

Fig. 8. Two examples of pose selection results. Upper row is an image from the training set while lower row is an image from the testing set. From left to right are the original image and its top-3 pose estimation results. Human key points are represented by blue circles while the skeleton is represented by the lines linking the circles. The selected poses are highlighted by red bounding boxes, which are more precise than the top poses.

fashion items more accurately, which will facilitate the category classifiers learning. As shown in the figure, the parsing results are improved significantly. More patches of the background are correctly classified, and the fashion items are parsed more accurately.

D. Pose Selection Evaluation

In this subsection, we first compare the performances with and without pose selection in the iteration step. From Fig. 6, it can be observed that the results with pose selection are consistently better than those without pose selection.

Some exemplar results are shown in Fig. 8, from which we can see that for some images with complex background (as shown in the top row), the top pose is not correct (the person in the background is detected). Our proposed model can select the second pose, which correctly locates the target person. This correct pose estimation benefits from the updated category classifiers, and can improve fashion parsing.

E. Quantitative Fashion Parings Results

We consider two baselines. One baseline is an intuitive but strong method called Color-Assisted-Model (CAM). CAM gets several initial color classifiers and initial category classifiers

TABLE IV
COMPARISON AMONG CAM, YAMAGUCHI [33] AND OUR METHOD IN TRAINING SET

Category	hat	glass	face	hair	scarf	T-shirt	blazer	blouse	sweater	coat	bag	belt
CAM	0.143	0.052	0.180	0.223	0.275	0.318	0.429	0.295	0.300	0.285	0.176	0.152
Our w/o	0.223	0.105	0.208	0.339	0.280	0.414	0.463	0.433	0.437	0.449	0.240	0.253
Our w	0.244	0.110	0.214	0.357	0.313	0.426	0.472	0.436	0.439	0.600	0.233	0.257
Yamaguchi [33]	0.079	0.029	0.330	0.419	0.209	0.482	0.379	0.521	0.487	0.356	0.153	0.143

Category	jeans	legging	pants	shorts	skirt	socks	stocking	shoe	dress	skin	bk	mean
CAM	0.483	0.475	0.681	0.449	0.508	0.102	0.294	0.183	0.411	0.284	0.273	**0.306**
Our w/o	0.625	0.637	0.630	0.542	0.617	0.158	0.452	0.282	0.602	0.374	0.669	**0.410**
Our w	0.630	0.641	0.619	0.553	0.632	0.174	0.458	0.281	0.604	0.375	0.676	**0.424**
Yamaguchi [33]	0.617	0.616	0.641	0.491	0.589	0.105	0.476	0.238	0.581	0.466	0.920	**0.406**

TABLE V
COMPARISON AMONG CAM, YAMAGUCHI [33] AND OUR METHOD IN TESTING SET

Category	hat	glass	face	hair	scarf	T-shirt	blazer	blouse	sweater	coat	bag	belt
CAM	0.174	0.061	0.177	0.223	0.225	0.271	0.431	0.334	0.388	0.413	0.172	0.162
Our w/o	0.272	0.109	0.200	0.337	0.310	0.401	0.454	0.433	0.501	0.498	0.231	0.242
Our w	0.273	0.107	0.252	0.402	0.318	0.414	0.497	0.440	0.511	0.473	0.245	0.220
Yamaguchi [33]	0.083	0.060	0.320	0.413	0.169	0.462	0.351	0.512	0.441	0.333	0.132	0.158

Category	jeans	legging	pants	shorts	skirt	socks	stocking	shoe	dress	skin	bk	mean
CAM	0.424	0.547	0.474	0.500	0.481	0.109	0.205	0.193	0.476	0.308	0.280	**0.305**
Our w/o	0.634	0.583	0.607	0.481	0.570	0.183	0.475	0.302	0.602	0.361	0.651	**0.410**
Our w	0.623	0.591	0.587	0.525	0.593	0.181	0.496	0.290	0.602	0.367	0.672	**0.421**
Yamaguchi [33]	0.596	0.489	0.563	0.467	0.552	0.070	0.399	0.243	0.573	0.475	0.918	**0.382**

similarly as our model. In the inference stage, we constrain the category label of each patch to take one value from the candidate set defined by category labels of the image. For example, if the image is labeled as {"red skirt", "red T-shirt"}, and we have predicted that a patch belongs to "red", CAM checks the patch's responses to both "skirt" and "T-shirt" and chooses the category producing the higher response. The category labels are further refined by considering the patches' locations. For example, the "T-shirt" is usually worn on the upper body. In summary, CAM considers all the information as our model and thus is very competitive. The pose for the CAM is fixed as the top pose. We choose CAM as the baseline to prove that *joint* inferring of color and category is superior over *sequential* inferring (first inferring color and then category), and that pose selection is dispensable in fashion parsing. The other baseline is from Yamaguchi *et al.* [33]. We re-train their fashion parsing model by using their publicly available source code. When implementing the baseline, we use the same image segmentation and pose estimation results as our methods for fair comparison. In our experiments, we thoroughly evaluate the three methods in two settings: 1) category and color labels (when implementing Yamaguchi *et al.* [33], only category labels) are provided at test time and 2) neither category nor color information is available at test time.

In Tables IV and V, we show the comparison results between our model and two baselines on the training set and testing set, respectively. We report the results for the 23 categories as well as their average value. Results show that our proposed model improves the parsing performance for most categories compared with the baseline methods. We can beat CAM because we can better explore the mutual information of color and category and select the correct human poses for fashion parsing. Note that the comparison between Yamaguchi [33] and our method is not entirely fair, since their method needs pixel-level ground category labels for training, while our method needs image-level color-category labels. Surprisingly, our method achieves better results than Yamaguchi *et al.* [33]. The possible reason is that our method benefits from the image-level color labels, which produces sufficient information as the pixel-level category labels by Yamaguchi *et al.* [33]. Considering the image-level color labels are much easier to annotate, our method is more practical.

Finally, we run the three methods in a more challenging setting where no image-level color-category tags are provided in the testing phase. The IOU of our model is 0.156, while CAM only achieves 0.131, which proves the superiority of our model. Yamaguchi [33]'s method can get similar IOU of 0.15 6 as our method due to its pixel-level label annotations.

For in-depth analysis, we also draw the confusion matrix of the testing set in Fig. 9. The most significant observation is that the largest values lie in the diagonal elements. The trousers categories, including "legging", "pants" and "jeans" achieve good performance, because these categories usually have the solid color, such as "blue" or "black". Correct identification of the colors helps localize the category labels. Also, we can see that for "face", "hair" and "hat", one of them is easily misclassified as another. All the three categories belong to the upper body, and possibly have the similar color as "beige". Besides, "bk" is difficult to identify since it usually has very complicated patterns. "skin" is often confused with other categories, since it may appear anywhere in the image, e.g., neck, arms, hand, legs, etc. The most difficult categories are "bag", "belt" and "hat". The possible reason is that "bags" can be at any location around human body, while "belt" and "hat" are usually too small to be identified. Although the confusion matrix of the training set is not given here due to the limited space, similar results can be observed.

Our model is very efficient in parsing an image. It only takes about 0.5 second to infer both the category and color labels of a common 600×400 image on our Intel Xeon CPU 2.93 GHz

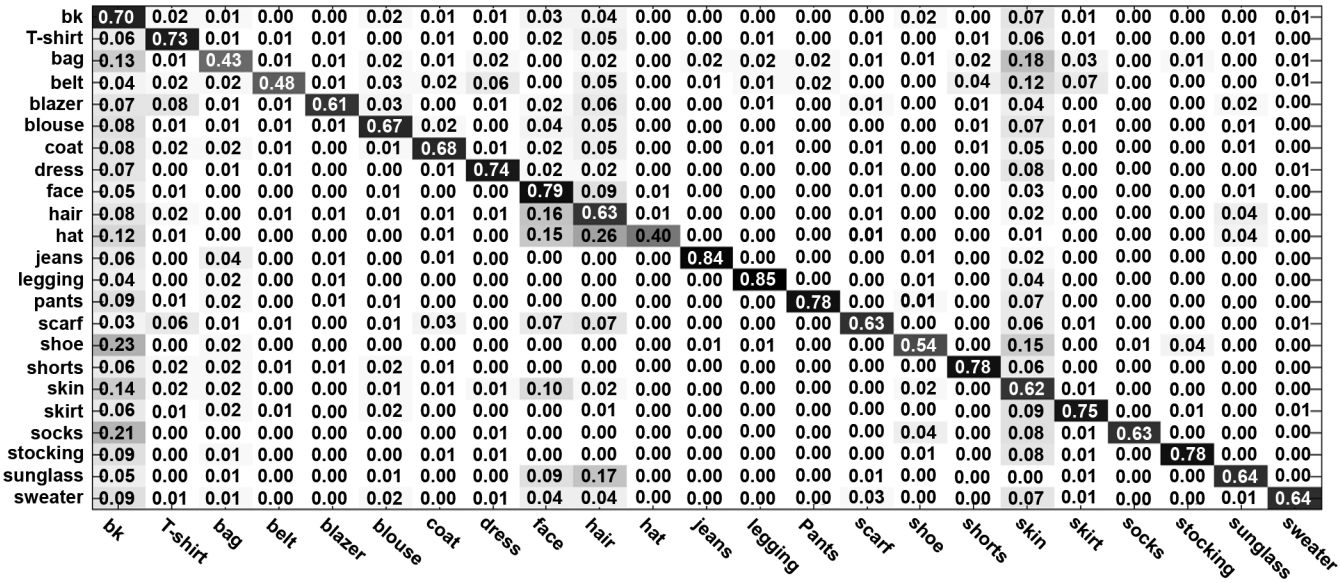


Fig. 9. The confusion matrix of the testing set. The largest values always lie in the diagonal elements, which shows the effectiveness of our model.

Fig. 10. Training fashion parsing results: 16 (4 rows $\times$4 columns) fashion parsing results are shown. The original images, the parsing ground truth and our parsing results are shown sequentially. Better view in color.

PC. Note that although in this paper, we only illustrate using one kind of attribute, i.e., color to assist image parsing, our framework can be easily extended to other kinds of fashion attributes, such as clothes material, pattern, etc.

F. Qualitative Fashion Parings Results

Twenty exemplar parsing results of the training set and testing set are shown in Figs. 10 and 11 respectively. We can observe that our method achieves great performance. For example, in the third image of the top row in Fig. 10, although the color of human clothes is similar as that of the background, which makes the parsing task very challenging, our method still well segments the human body from the background. Besides, the lady wears "T-shirt" inside "blazer", which are difficult to be separated from each other, but our model still correctly infers the labels. Another example is the first image of the third row

Fig. 11. Testing fashion parsing results: 16 (4 rows $\times$4 columns) fashion parsing results are shown. The original images, the parsing ground truth and our parsing results are shown sequentially. Better view in color.

in Fig. 11. Although the girl is not in the exact frontal view, our system successfully handles this case.

In some scenarios, our method does not achieve satisfactory performance. We show a few such cases in the fourth row of Figs. 10 and 11. For example, in the last image of the fourth row in Fig. 10, "hair" is missing. This is because in our current inference model, we do not constrain the labels of the image to be assigned to at least one patch. We will try to address this problem in future. Take the first image of the fourth row in Fig. 11 as another example, where "sweater" mistakenly expands to the background area. This is caused by the inaccurate human pose estimation. In the future, we will also focus on how to improve the pose estimation.

We further show more patterned and cartoon clothing parsing results in the last row of Figs. 10 and 11. Generally speaking, our method can well handle the patterned and cartoon clothes. One example is the last image of Fig. 10. Even the girl wears a spotted "shorts", our algorithm still correctly classify all the "shorts" patches. The other example is the second image of the last row of Fig. 10. Our algorithm can still correctly tell which pixel belong to "T-shirt", although the girl wears a white "T-shirt" with a black cartoon logo. However, we acknowledge that our method achieves better performance when parsing plain clothes, where the color prediction is more reliable thus can better assist category label estimation. For example, in the first image of the last row of Fig. 10, our algorithm wrongly predict several "skirt" patches as "bag". Take a closer look at the errors, we can find that all confusing patches share similar color with the red "bag". In this case, color information cannot help the imperfect category classifiers, which produces the parsing errors.

TABLE VI
THE IOU WHEN INCREASING PERCENT OF NOISES ARE ADDED

Noises_level	0	12%	24%
IOU	0.421	0.325	0.256

G. Sensitive to Tags Noises

In this part, we briefly discuss how our system performs when the color-category tags are contaminated. In real scenarios, people are very likely to input wrong tags. We simulate the tagging errors by altering the ground truth color-category tags on the testing set. For example, the original "blue legging" may be changed to "blue jeans", "green legging" or even "green jeans". In the experiment, we replace one or two tags of each image with corresponding one or two noisy tags, which causes a 12% or 24% noise level for the whole dataset. From the results in Table VI, we can see that our system has achieved an IOU of 0.325 when 12% tags are wrong, which is still higher than baseline. It shows the robustness of our system to noise.

VII. CONCLUSION

In this paper, we propose a general weakly-supervised fashion parsing framework. Given an image set and its cor-

responding image-level color-category tags, our system can efficiently parse all the images, i.e., assign each pixel with a color-category label. During the fashion parsing process, we also learn color classifiers and category classifiers, which can be used to parse a testing image. Extensive experiments on our Colorful-Fashion dataset validate the effectiveness of the proposed framework. Currently, we have proved that the color-category tags can assist the fashion parsing. In future, we will explore the effectiveness of other structured tags. For example, we will investigate whether pattern-category tags (e.g., stripped T-shirt, plaid skirt, etc.) can facilitate clothes paring.

REFERENCES

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[2] T. Berg, A. Berg, and J. Shih, "Automatic attribute discovery and characterization from noisy web data," in *Proc. ECCV*.

[3] L. Bossard, M. Dantone, C. Leistner, C. Wengert, T. Quack, and L. Van Gool, "Apparel classification with style," in *Proc. ACCV*.

[4] J. B. Gabriel, S. Jamie, F. Julien, and C. Roberto, "Segmentation and recognition using structure from motion point clouds," in *Proc. ECCV*.

[5] C. Joao and S. Cristian, "CPMC: Automatic object segmentation using constrained parametric min-cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1312–1328, Jul. 2012.

[6] H. Chen, A. Gallagher, and B. Girod, "Describing clothing by semantic attributes," in *Proc. ECCV*.

[7] H. Chen, Z. J. Xu, Z. Q. Liu, and S. C. Zhu, "Composite templates for cloth modeling and sketching," in *Proc. CVPR*.

[8] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. CVPR*.

[9] E. Ian and H. Derek, "Category independent object proposals," in *Proc. ECCV*.

[10] E. Mark, V. G. Luc, K. W. Christopher, W. John, and Z. Andrew, "The pascal visual object classes (VOC) challenge," *Int. J. Comput. Vision*, vol. 88, pp. 303–338, 2010.

[11] F. Rong-En, C. Kai-Wei, H. Cho-Jui, W. Xiang-Rui, and L. Chih-Jen, "Liblinear: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, 2008.

[12] B. Fulkerson, A. Vedaldi, and S. Soatto, "Class segmentation and object localization with superpixel neighborhoods," in *Proc. ICCV*, 2009.

[13] Y. Han, F. Wu, J. Shao, Q. Tian, and Y. Zhuang, "Graph-guided sparse reconstruction for region tagging," in *Proc. CVPR*, 2012.

[14] B. Hasan and D. Hogg, "Segmentation using deformable spatial priors with application to clothing," in *Proc. BMVC*, 2010.

[15] H. Derek, A. E. Alexei, and H. Martial, "Recovering occlusion boundaries from an image," *Int. J. Comput. Vision*, vol. 91, pp. 328–346, 2011.

[16] S. K. Sathiya, S. Sellamanickam, C. Kai-Wei, H. Cho-Jui, and L. Chih-Jen, "A sequential dual method for large scale multi-class linear SVMs," in *Proc. KDD*, 2008.

[17] P. Kohli *et al.*, "Robust higher order potentials for enforcing label consistency," *Int. J. Comp. Vision*, vol. 82, pp. 302–324, 2009.

[18] K. Daniel and F. Vittorio, "Figure-ground segmentation by transferring window masks," in *Proc. CVPR*.

[19] C. Liu, J. Yuen, and A. Torralba, "Nonparametric scene parsing: Label transfer via dense scene alignment," in *Proc. CVPR*.

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

[21] 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 *Proc. CVPR*.

[22] X. Liu, B. Cheng, S. Yan, J. Tang, T. S. Chua, and H. Jin, "Label to region by bi-layer sparsity priors," in *Proc. MM*.

[23] X. Liu, S. Yan, J. Luo, J. Tang, Z. Huango, and H. Jin, "Nonparametric label-to-region by search," in *Proc. CVPR*.

[24] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[25] D. Park and D. Ramanan, "N-best maximal decoders for part models," in *Proc. ICCV*.

[26] R. Carsten, K. Vladimir, and B. Andrew, "Grabcut: Interactive foreground extraction using iterated graph cuts," in *Proc. TOG*.

[27] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation," in *Proc. ECCV 2006*.

[28] V. D. W. Joost, S. Cordelia, and V. Jakob, "Learning color names from real-world images," in *Proc. CVPR*.

[29] A. Vezhnevets and J. M. Buhmann, "Towards weakly supervised semantic segmentation by means of multiple instance and multitask learning," in *Proc. CVPR*.

[30] A. Vezhnevets, J. M. Buhmann, and V. Ferrari, "Active learning for semantic segmentation with expected change," in *Proc. CVPR*.

[31] A. Vezhnevets, V. Ferrari, and J. M. Buhmann, "Weakly supervised semantic segmentation with a multi-image model," in *Proc. ICCV*.

[32] G. Wang and D. Forsyth, "Joint learning of visual attributes, object classes and visual saliency," in *Proc. CVPR*.

[33] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg, "Parsing clothing in fashion photographs," in *Proc. CVPR*.

[34] Y. Ming and Y. Kai, "Real-time clothing recognition in surveillance videos," in *Proc. ICIP*.

[35] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *Proc. CVPR*.

[36] Y. Yang, Y. Yang, Z. Huang, H. T. Shen, and F. Nie, "Tag localization with spatial correlations and joint group sparsity," in *Proc. CVPR*.

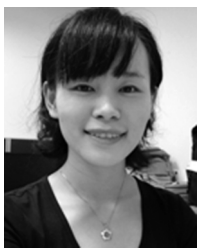
Si Liu is currently a research fellow at Learning and Vision Group of National University of Singapore. She received her Ph.D. degree from National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences in 2012. Her research interests includes computer vision and multimedia.

Jiashi Feng received the B.S. degree from the University of Science and Technology of China (USTC), Hefei, China, in 2007. He is currently pursuing the Ph.D. degree from Department of Electrical and Computer Engineering, National University of Singapore (NUS), Singapore. His research include computer vision and machine learning.

Csaba Domokos received the MS degree in Computer Science and the MS degree in Mathematics from the University of Szeged, Hungary in 2006 and 2010, respectively. He obtained his Ph.D. degree from University of Szeged in 2011. He is currently a Research Fellow at the Learning and Vision Research Group, National University of Singapore. His current research interests include semantic segmentation, low-rank approximation, image registration and shape matching. He received the winner prize of the segmentation task in PASCAL VOC 2012.

Hui Xu is currently a Researcher Assistant in Chongqing Institute of Green and Intelligent Technology, Chinese Academy of China. From September 2005 to March 2012, she received bachelor degree and master degree in Beijing University of Posts and Telecommunications.

Junshi Huang received his bachelor and master degree in Beijing Institute of Technology in 2009 and 2012 respectively. He currently is a Ph.D. student in the Department of Electrical and Computer Engineering at the National University of Singapore. His research interests are in the computer vision.



Zhenzhen Hu is a Ph.D. candidate of Hefei University of Technology. She received the M.Sc. degrees from the School of Computer and Information Science, Hefei University of Technology in 2011.

Shuicheng Yan is currently an Assistant Professor in the Department of Electrical and Computer Engineering at National University of Singapore, and the founding lead of the Learning and Vision Research Group (http://www.lv-nus.org). Dr. Yan's research areas include computer vision, multimedia and machine learning, and he has authored or co-authored over 200 technical papers over a wide range of research topics. He is an associate editor of IEEE Transactions on Circuits and Systems for Video Technology, and has been serving as the guest editor of the special issues for TMM and CVIU. He received the Best Paper Awards from ACM MM'10, ICME'10 and ICIMCS'09, the winner prize of the classification task in PASCAL VOC'10, the honorable mention prize of the detection task in PASCAL VOC'10, 2010 TCSVT Best Associate Editor (BAE) Award, 2010 Young Faculty Research Award, 2011 Young Scientist Award Singapore, and the co-author of the best student paper awards of PREMIA'09 and PREMIA'11.