10-2019

Who, where, and what to wear?: extracting fashion knowledge from social media

Yunshan MA

Xun YANG

Lizi LIAO
Singapore Management University, lzliao@smu.edu.sg

Yixin CAO

Tat-Seng CHUA

Who, Where, and What to Wear? Extracting Fashion Knowledge from Social Media

Yunshan Ma
National University of Singapore
yunshan.ma@u.nus.edu

Xun Yang*
National University of Singapore
xunyang@nus.edu.sg

Lizi Liao
National University of Singapore
liaolizi.llz@gmail.com

Yixin Cao
National University of Singapore
caoyixin2011@gmail.com

Tat-Seng Chua
National University of Singapore
dcscts@nus.edu.sg

ABSTRACT

Fashion knowledge helps people to dress properly and addresses not only physiological needs of users, but also the demands of social activities and conventions. It usually involves three mutually related aspects of: occasion, person and clothing. However, there are few works focusing on extracting such knowledge, which will greatly benefit many downstream applications, such as fashion recommendation. In this paper, we propose a novel method to automatically harvest fashion knowledge from social media. We unify three tasks of occasion, person and clothing discovery from multiple modalities of images, texts and metadata. For person detection and analysis, we use the off-the-shelf tools due to their flexibility and satisfactory performance. For clothing recognition and occasion prediction, we unify the two tasks by using a contextualized fashion concept learning module, which captures the dependencies and correlations among different fashion concepts. To alleviate the heavy burden of human annotations, we introduce a weak label modeling module which can effectively exploit machine-labeled data, a complementary of clean data. In experiments, we contribute a benchmark dataset and conduct extensive experiments from both quantitative and qualitative perspectives. The results demonstrate the effectiveness of our model in fashion concept prediction, and the usefulness of extracted knowledge with comprehensive analysis.

CCS CONCEPTS

• **Information systems** → **Specialized information retrieval**.

KEYWORDS

Fashion Knowledge Extraction; Fashion Analysis

ACM Reference Format:
Yunshan Ma, Xun Yang, Lizi Liao, Yixin Cao, and Tat-Seng Chua. 2019. Who, Where, and What to Wear? Extracting Fashion Knowledge from Social Media. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19), October 21–25, 2019, Nice, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3343031.3350889

*Corresponding author.

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10…$15.00
https://doi.org/10.1145/3343031.3350889

1 INTRODUCTION

According to Statista[1], revenue in fashion market amounts to $600 billion dollars in 2019, which demonstrates the great opportunities for various fashion related research and applications. Fashion knowledge plays a critical role in this area. It addresses not only physiological needs of users, but also the demands of social activities and conventions [12, 26, 38]. Take the fashion recommendation as an example, the recommender system should be capable of instructing a man to wear thick clothes in winter rationally with pants instead of shorts, and to dress suits in a formal conference.

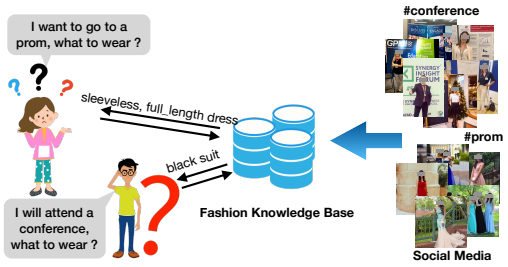


Figure 1: An illustration of fashion knowledge extraction from social media.

Although some existing works focus on recognizing fashion concepts like clothing category and attributes [14, 19, 21], there are few studies at knowledge level in fashion domain, which usually involve three main aspects at the same time: *person*, *clothing*, and *occasion*. As illustrated in Figure 1, it would be better for a young girl to wear a sleeveless full length dress for a prom. To dress properly, people have to consider social conditions like occasion and personal identity, in addition to clothes. Clearly, there exist a large number of such patterns (*e.g.*, dresscode or conventions) guiding people's daily fashion activities. But as we move forward, a key question arises: where and how can we collect such fashion knowledge?

This paper proposes to automatically extract user-centric fashion knowledge from social media, such as Instagram, where massive user-generated multi-modal resources are uploaded every day from all over the world. It is a natural and appropriate source to extract fashion knowledge from general users, because (1) the images posted on social media usually contain the cues to various occasions such as the conference, wedding and travel *etc.*, and also person

[1] https://www.statista.com/outlook/244/100/fashion/worldwide

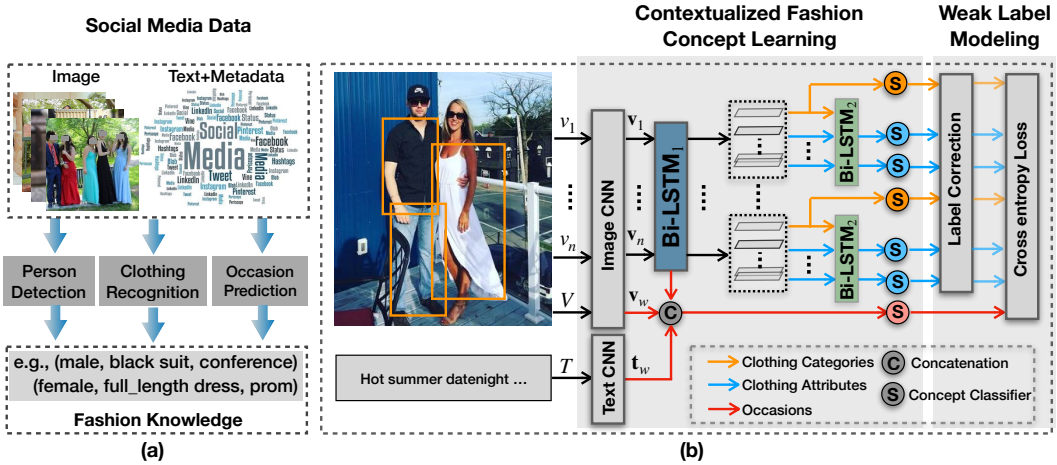


Figure 2: (a) The overall framework of fashion knowledge extraction from social media, and (b) the pipeline of our proposed contextualized fashion concept learning model. Note that the extraction of person attributes (*e.g.*, gender and age) is performed using an off-the-shelf tool for simplicity.

identity information, such as the gender, age *etc*, and (2) there is sufficient up-to-date data to perform the analysis. However, it is a non-trivial task due to the following challenges:

First, the extraction of fashion knowledge from social media content is highly dependant on the performance of fashion concept prediction which still remains unsatisfactory. This is because most of the visual images posted by various users on social media are taken in natural scenes, from which it is hard to detect the fashion concepts (*e.g.*, clothing attributes and occasions). It is also more complex than the fashion product images typically with clean backgrounds, which most existing research studies focus on. Therefore, how to jointly detect the fashion concepts in the natural scene images for knowledge construction is a difficult but critical task.

Second, social media data lacks sufficient fashion concept labels which are crucial for fashion knowledge construction. The quality of automatically harvested fashion knowledge highly depends on semantic-level fashion concept learning. However, manually annotating a large amount of data is expensive and time-consuming. Existing datasets are mainly derived from e-commerce site and only focus on a specific set of cloth attributes, which cannot be used to detect the types of occasions or person identities.

To address these challenges, we propose a novel method with two modules that jointly detect the fashion concepts using weakly labeled data. We propose a contextualized fashion concept learning module to effectively capture the dependencies and correlations among different fashion concepts. To alleviate the label insufficiency problem, we enrich the learning procedure with a weak label modeling module that utilizes both the machine-labeled data and clean data. In particular, we incorporate a label transition matrix into this module to enable more robust noise control during the learning process. We then obtain a set of fashion concepts grounded with social media. Finally, through a statistical approach, we obtain the ultimate fashion knowledge. Through extensive evaluations and analyses, we demonstrate that the extracted knowledge is rational and is able to be applied to downstream applications.

The main contributions of this work are as follows:

- We propose a novel method for fashion knowledge extraction with the help of a contextualized fashion concept learning module, which is able to capture the dependencies among occasion, clothing categories and attributes.
- We exploit machine labeled data with weak labels to enrich our learning model with a label correction module for noise control.
- We contribute a benchmark dataset and conduct extensive experiments from both quantitative and qualitative perspectives to demonstrate the effectiveness of our model in fashion concept prediction and the usefulness of extracted knowledge.

2 RELATED WORK

Automatic Knowledge Extraction. In the past few years, researchers from the natural language processing (NLP), data mining, and computer vision communities have conducted extensive studies on automatic knowledge extraction and its applications [3, 4, 31, 32]. In the NLP community, several famous knowledge bases were curated such as YAGO [9], Freebase [1], WikiData [29], DBpedia [17]. These knowledge bases capture large amount of textual facts in the world, which are usually organized into triplets of the form (Subject, Predicate, Object). Most of the facts are about well-known people, places, and things, which were collected whether by crowd sourcing strategies or from large-scale semi-structured web knowledge bases. Nonetheless, all of them were curated only based on textual resources while neglecting the rich information existing in visual data. Thereafter, many efforts have been paid to extracting knowledge from visual data, such as NEIL [7], Visual Genome [16] and VidVRD [24]. Even though many researches targeted at extracting knowledge from both textual and visual data, few works aim to extract knowledge in vertical domains like fashion.

Fashion Concept Prediction. Recently, fashion concept prediction has attracted increasing interests in various tasks such as clothing recognition [5, 21], retrieval [10, 14, 18, 20], parsing [33] and landmark detection [21, 30]. Earlier methods [5, 20] mostly relied on handcrafted features (*e.g.*, SIFT, HOG) to get good clothing representations. However, with the proliferation of deep learning in

computer vision, many deep neural network models have been developed. In particular, Huang *et al.* [14] developed a Dual Attribute-Aware Ranking Network (DARN) for clothing image retrieval. Liu *et al.* [21] proposed a branched neural network FashionNet which learns clothing features by jointly predicting clothing attributes and landmarks. Liao *et al.* [18] introduced a novel data structure of EITree, which organizes the fashion concepts into multiple semantic levels and demonstrates good performance for both fashion image retrieval and clothing attributes prediction. However, most of the models are limited to the fashion concept level, while none of them further extended to fashion knowledge level. Moreover, data from social media lacks high-quality annotations, and weakly supervised methods are usually employed. Corbiere *et al.* [8] learned a model from noisy datasets crawled from e-commerce websites without manual labelling, which demonstrates great generalization capability on DeepFashion [21] dataset. However, it requires a lot of training data (1.3 million images in [8]), which was both time consuming and computationally intensive. In this paper, we also take advantage of weakly-labeled data to enhance our fashion concept prediction model. To counter the noise within the weak labels, we employ a weak label modeling approach, inspired by works on learning with noisy labels [27, 28].

3 PROBLEM FORMULATION

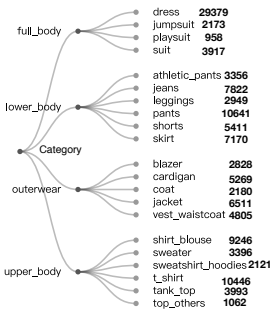
Our goal is to extract user-centric fashion knowledge from social media, such as Instagram, where massive user-centric multimodal resources are uploaded every day. We expect to obtain structured knowledge about *what to wear for a specific occasion* to support downstream fashion applications. We first formally define the user-centric fashion knowledge as triplets of the form $\mathcal{K} = \{\mathcal{P}, C, O\}$, which consists of three aspects defined as follows:

- **Person:** $\mathcal{P}$ refers to a set of person attributes, such as gender, age, body shape, etc. $\mathcal{P}$ should be able to describe a specific type of person, such as *a young woman*.
- **Clothing:** C refers to a set of clothing categories and attributes, such as skirt, a-line, red , etc. C should be able to describe a specific type of clothing, such as *a red a-line skirt*.
- **Occasion:** O refers to a set of occasions, such as conference and dating, and their affiliated metadata, such as location and time.

Given a set of user-generated posts $\mathcal{X} = \{\mathcal{V}, \mathcal{T}, \mathcal{M}\}$ consisting of images $\mathcal{V}$, texts $\mathcal{T}$, and metadata $\mathcal{M}$ (such as time, location) on social media, the problem is to develop a hybrid detection framework which is able to automatically extract the three aspects of fashion knowledge $\{\mathcal{P}, C, O\}$.

Three sub-tasks that need to be tackled are: 1) person attributes detection, 2) clothing categories and attributes detection, and 3) occasion prediction. As existing person detection and analysis methods (such as [23] and [39]) have already achieved satisfactory performance, they can be utilized as off-the-shelf tools for our person information extraction. For simplicity, we only handle the last two sub-tasks in this work.

The main task of this work is to design a fashion concept learning framework as shown in Figure 2 (b), which should jointly detect occasion and clothing categories and attributes from social media, the details of which will be presented in Section 4.



Attribute Name	Attribute Values
color	brown multi gold white yellow green grey purple pink black blue red silver navy orange
neckline	collar round_neck v_neck turtleneck off_shoulder off_shoulder halter_neck hooded
upper_body_length	crop hip_length thigh_length knee_length below_knee
lower_body_length	extra_short thigh_length knee_length calf_length 7_8 full_length
dress_skirt_shape	a_line pencil straight peplum mermaid high_low
closure_type	pullover button zip
sleeve_style	straight shoulder_strap capped
sleeve_length	long three_quarter short sleeveless

(b)

Dating	Daily Work	Travel	Sports	Conference	Prom	Shopping	Party	School	Wedding Guest
9003	7634	7713	6858	9139	9334	6877	8010	7327	8734

(c)

Figure 3: The statistics of the FashionKE dataset, consisting of 21 categories, 8 types of clothing attributes, and 10 common types of occasions.

Dataset: Currently, there is no available training and evaluation dataset for the task of extracting *user-centric fashion knowledge*. Social media sites such as Instagram provide a huge amount of user generated contents of which a large portion involves people's dressing codes. There are many active fashion influencers who like to share their dressing styles and mix-and-match experience, which provide a valuable and up-to-date source for fashion knowledge extraction. We thus crawl millions of posts from Instagram. Both automated and manual filtering are carried out sequentially to ensure data quality. Finally, we contribute a large and high quality dataset, named FashionKE, which consists of 80,629 eligible images. Each collected image is selected to be easily recognizable and diverse from the three aspects of: person, occasion, and clothing. We first leverage pre-trained object detection model [23] to detect person body and face [39]. Second, we filter out those images without any face or body. Moreover, we align the body and face in the same image and remove those images which do not have normal-sized body or face. Third, to ensure that the images are really user-generated – but not advertisements or posters – we manually check all the images and remove those that cannot reflect any occasion. The ontology of our dataset is shown in Figure 3 and the annotation of the dataset is described as follows:

Occasion annotation. *Occasion* is an important fashion concept that deeply affects people's decisions on dressing. With the help of fashion experts, we manually define 10 common types of occasion concepts as shown in Figure 3 (c). For each occasion, we curate a seed list of hashtags which are widely adopted and highly-correlated with that occasion. We then crawl Instagram posts using such hashtags and each post will have a potential occasion tag corresponding to its seed hashtag. We ask each annotator to give each post a binary label: whether such a post reveals that occasion or not by jointly considering the hashtags, post texts, and image content. It shows that such a process is efficient and effective. Finally, we obtain about 8,000 images for each occasion on average.

Annotation of clothing categories and attributes. This task is very important but time-consuming and expensive, since each image may have multiple persons and multiple sets of clothing. Compared with occasion annotation which requires only one label

for each image, clothing annotation needs about 10 labels (categories and attributes of multiple clothes) for each image. To alleviate this issue, we adopt a two-stage annotation process: we first use a commercial fashion tagging tool[2] to automatically detect and tag the clothes, and then manually check and refine the results by human annotators. The statistic of the categories of all clothes is shown in Figure 3 (a). Note that only 30% of images are carefully refined by human annotators. The rest of data are machine-labeled which are noisy. Therefore, how to exploit the machine-labeled noisy data is also one of our research questions.

4 OUR APPROACH

This paper proposes to develop a hybrid detection framework to extract user-centric fashion knowledge $\mathcal{K} = \{\mathcal{P}, C, O\}$ from social media. As mentioned before, we only focus on detecting clothing categories and attributes, and the occasions. The keys to tackling such a task are: 1) how to design a unified detection framework which is able to effectively capture the correlation among occasions, clothing categories and attributes; and 2) how to effectively utilize the machine labeled data to enhance the fashion concept learning.

For the first question, we design a contextualized fashion concept learning model, as shown in figure 2(b), in which two bidirectional recurrent neural networks are utilized to capture the dependencies and correlations among occasions, clothing attributes and categories, which is presented in section 4.1. For the second question, we introduce a weak label modeling module which estimates a label transition matrix for bridging the gap between weak labels and clean labels, as described in section 4.2.

4.1 Contextualized Fashion Concept Learning

Given a post image V and the affiliated text description T, we first detect a set of clothing regions $\{v_1, \cdots, v_i, \cdots, v_M\}$ in the image V by a clothing detection module. The goal is to predict the occasion label $\hat{y}_{o_V} \in \{y_{o_1}, \cdots, y_{o_{M_o}}\}$ of the posted image V and the clothing category $\hat{y}_{c_{v_i}} \in \{y_{c_1}, \cdots, y_{c_{M_c}}\}$ and attributes $\hat{y}_{a_{v_i}} \in \{y_{a_1}, \cdots, y_{a_{M_a}}\}$ of each clothing region $\{v\}_{i=1}^{M}$. A simple solution is to directly cast it as three independent classification tasks. However, such a straightforward approach may result in sub-optimal performance as it ignores the relations between the occasion, clothing categories, and clothing attributes. For example, it is not likely for a woman to dress shorts or tanks to attend a prom.

We develop a contextualized fashion concept learning framework to capture the correlations among the occasion, clothing categories, and clothing attributes, as shown in Figure 2 (b).

Category Representation: The first step is to learn the contextualized representations of clothing regions for category prediction and the whole image for occasion prediction. We use a bidirectional Long Short-Term Memory (Bi-LSTM) network to encode the dependence among all clothing regions. We first use a pre-trained convolutional neural network (CNN), such as ResNet [11] as our main feature extractor to extract the dense vector representation of the whole image V as $\mathbf{v}_w \in \mathbb{R}^d$, and the vector representation of each clothing region $v_i \in \{v\}_{i=1}^{M}$ as $\mathbf{v}_i \in \mathbb{R}^d$. The final hidden representation for each clothing region is the concatenation of the hidden vectors in both directions:

$$\begin{cases} \overrightarrow{\mathbf{h}}_{v_i} = \overrightarrow{\text{LSTM}_1}(\mathbf{v}_i, \overrightarrow{\mathbf{h}}_{\mathbf{v}_{i-1}}) \\ \overleftarrow{\mathbf{h}}_{v_i} = \overleftarrow{\text{LSTM}_1}(\mathbf{v}_i, \overleftarrow{\mathbf{h}}_{v_{i+1}}) \\ \mathbf{h}_{v_i} = [\overrightarrow{\mathbf{h}}_{v_i}, \overleftarrow{\mathbf{h}}_{v_i}] \end{cases} \tag{1}$$

where $\mathbf{h}_{v_i} \in \mathbb{R}^{2d}$. We then add a fully connected layer $F_c(\cdot)$ parameterized with a weight matrix $\mathbf{W}_c \in \mathbb{R}^{2d \times d}$ and a bias vector $\mathbf{b}_c \in \mathbb{R}^d$ to transform $\mathbf{h}_{v_i}$ as the final category representation of each clothing region $\mathbf{c}_{v_i} = \mathbf{W}_c^T \mathbf{h}_{v_i} + \mathbf{b}_c$.

Occasion Representation: To better represent the whole image V, we augment the CNN feature $\mathbf{v}_w$ with the feature of the post text description $\mathbf{t}_w$ and the final hidden state representation $\mathbf{h}_o = [\overrightarrow{\mathbf{h}}_o, \overleftarrow{\mathbf{h}}_o] \in \mathbb{R}^{2d}$ of Bi-LSTM in Eq. (1):

$$\mathbf{v}'_w = [\mathbf{v}_w, \mathbf{t}_w, \mathbf{h}_o] \tag{2}$$

where $\mathbf{h}_o \in \mathbb{R}^{2d}$ encodes the inter-correlation of different clothing regions extracted from the whole image V. $\mathbf{t}_w \in \mathbb{R}^{d_t \in \mathbb{R}^d}$ denotes the vector representation of the affiliated text description T which usually contains the evidence about the *occasion*. It is extracted by a TextCNN [15]. Both $\mathbf{h}_o$ and $\mathbf{t}_w$ can effectively complement the CNN feature $\mathbf{v}_w$ of the whole image. A fully-connected layer $F_w(\cdot)$, parameterized with a weight matrix $\mathbf{W}_w \in \mathbb{R}^{4d \times d}$ and a bias vector $\mathbf{b}_w \in \mathbb{R}^d$, is added to transform the concatenated representation $\mathbf{v}'_w$ of the whole image to a d-dimensional occasion representation $\mathbf{o}_w = \mathbf{W}_w^T \mathbf{v}'_w + \mathbf{b}_w$ for the whole image.

Attribute Representation: Since each cloth has multiple different types of attributes, such as color, shape, sleeve length, *etc.*, we introduce a multi-branch attribute prediction module, which consists of K fully-connected layers $F_{a_i}(\cdot), i = 1, \cdots, K$, parametrized with weight matrices $\mathbf{W}_{a_i} \in \mathbb{R}^{2d \times d}$ and bias vectors $\mathbf{b}_{a_i} \in \mathbb{R}^d$, to transform each clothing region representation $\mathbf{h}_{v_i} \in \mathbb{R}^{2d}$ into K semantic representations $F_{a_k}(\mathbf{h}_{v_i}), k = 1, \cdots, K$, for attribute prediction. Each branch corresponds to a type of clothing attribute. In this multi-branch structure, the visual representations from the lower-level layers are shared among all attributes. The neuron number in the output-layer of each branch equals to the number of corresponding attribute values.

To capture the dependence among clothing attributes and categories, we introduce the second Bi-LSTM network. For each clothing region, we stack the outputs of K branches $\{F_{a_k}(\mathbf{h}_{v_i})\}_{k=1}^{K}$ and the category representation $\mathbf{c}_{v_i}$ into a sequence of vectors as the inputs to the second Bi-LSTM. The final hidden representation for each attribute is the concatenation of the hidden vectors in both directions:

$$\begin{cases} \overrightarrow{\mathbf{h}}_{a_k}^{v_i} = \overrightarrow{\text{LSTM}_2}(F_{a_k}(\mathbf{h}_{v_i}), \overrightarrow{\mathbf{h}}_{a_{k-1}}^{v_i}) \\ \overleftarrow{\mathbf{h}}_{a_k}^{v_i} = \overleftarrow{\text{LSTM}_2}(F_{a_k}(\mathbf{h}_{v_i}), \overleftarrow{\mathbf{h}}_{a_{k+1}}^{v_i}) \\ \mathbf{h}_{a_k}^{v_i} = [\overrightarrow{\mathbf{h}}_{a_k}^{v_i}, \overleftarrow{\mathbf{h}}_{a_k}^{v_i}] \end{cases} \tag{3}$$

where $\mathbf{h}_{a_k}^{v_i} \in \mathbb{R}^{2d}$ is a contextualized attribute representation which encodes the dependence among clothing attributes and categories. It is further transformed into a d-dimensional attribute representation $\mathbf{a}_{a_k}^{v_i} \in \mathbb{R}^d$ by a fully-connected layer.

[2] www.visenze.com

After obtaining the occasion representation $\mathbf{o}_w$ of the whole image, the category representation $\mathbf{c}_{v_i}$ and attribute representation $\mathbf{a}_{a_k}^{v_i}$ of clothing regions, the prediction scores of occasions, clothing categories, and clothing attributes are obtained by multiple standard classifier layers (i.e., a linear function followed by a softmax layer), respectively. Cross-entropy loss is used to train the model. The training objective is to minimize the following loss function:

$$L = L_o\left(V, y_{o_V}, \Theta\right) + L_c\left(\{v_i\}, \{y_{c_{v_i}}\}, \Theta\right) + L_a\left(\{v_i\}, \{y_{a_{v_i}}\}, \Theta\right) \quad (4)$$

where $L_o(\cdot)$, $L_c(\cdot)$, and $L_a(\cdot)$ denote the cross entropy losses of occasion, category, and attribute, respectively.

4.2 Enhancing Fashion Concept Learning with Weak Label Modeling

As aforementioned, only the occasion label is manually annotated for each image in our dataset. For the annotation of clothing categories and attributes, only a fraction of our dataset is provided with clean clothing category labels and attribute labels, while the rest of data is annotated by a fashion tagging tool. Such machine-labeled data is relatively cheap and easy-to-obtain, but the model would suffer from overfitting due to the label noise in training set. Therefore, how to jointly exploit machine-labeled data and the limited human-labeled data for model training is one of our main research focuses. This paper introduces a weak label modeling strategy to handle the label flip noise [27, 28] in machine-labeled data.

Weak Label Modeling. We first describe the process as a general setting. The goal is to learn a fashion concept learning model, as described in section 4.1, from the machine-labeled data with weak label $y' \in \{1, 2, \cdots, N\}$. The true label is denoted as $y^* \in \{1, 2, \cdots, N\}$. We assume that each weak label y' depends only on the true label y^* and not on the training sample, and further suppose that the weak labels are i.i.d. conditioned on the true labels. Then, we can represent the conditional noise model by a label transition matrix [27, 28] $\mathbf{Q} \in \mathbb{R}^{N \times N}$:

$$p(y' = j | y^* = i) = q_{j,i} \quad (5)$$

where $q_{j,i}$ is the element of label transition matrix $\mathbf{Q}$ at (j, i). The probability of a data sample $\mathbf{x}$ being labeled as a noisy label j can be computed as:

$$p(\hat{y}' = j | \mathbf{x}, \mathbf{Q}, \Theta) = \sum_i q_{j,i} p(\hat{y}^* = i | \mathbf{x}, \theta) \quad (6)$$

Then, the prediction of weak label distribution is:

$$p(\hat{y}' | \mathbf{x}, \mathbf{Q}, \mathbf{\Theta}) = \sum_i p(\hat{y}' | \hat{y}^* = i) p(\hat{y}^* = i | \mathbf{x}, \theta) = \mathbf{Q} p(\hat{y}^* | \mathbf{x}, \theta) \quad (7)$$

where $p(y^* | \mathbf{x}, \theta)$ is used to estimate the true label of testing sample, while $p(\hat{y}' | \mathbf{x}, \mathbf{Q}, \mathbf{\Theta})$ is used for training with a standard cross-entropy loss.

In summary, the basic idea is to add a *label correction* layer with an estimated label transition matrix $\mathbf{Q}$ after the prediction layer of our network framework, as shown in figure 2(b), which adapts the prediction to match the weak label distribution.

Estimation of label Transition Matrix: How to effectively estimate the laebl transition matrix $\mathbf{Q}$ is critical to our weak label modeling module. In this work, we implement it as a linear layer to be jointly optimized during training. Since a fraction of our dataset is corrected by human annotators from machine-labeled data, we can first estimate a label transition matrix using the human-corrected labels and the weak labels in the clean part, which is further utilized as an initialization for the linear layer.

Learning: We split the training data into two sets: $\mathcal{X}^*$ and $\mathcal{X}'$ where $\mathcal{X}^*$ denotes the part of our training data with clean labels $\mathcal{Y}^*$, and $\mathcal{X}'$ denotes the rest of our training data with weak labels $\mathcal{Y}'$. Our final objective is to minimize the following fused cross-entropy loss function:

$$L = L^*(\mathcal{X}^*, \mathcal{Y}^*, \Theta^*) + \beta L'(\mathcal{X}', \mathcal{Y}', \mathbf{Q}, \Theta') \quad (8)$$

where $L^*(\cdot)$ is the loss function defined on clean data, as shown in Eq. (4), and $L'(\cdot)$ is the loss function defined on machine-labeled data based on weak label prediction in Eq. (7). β is a trade-off hyperparameter.

5 EXPERIMENTS

To verify the effectiveness of fashion knowledge extraction, we conduct a series of experiments from both the quantitative and qualitative perspectives. We first compare it with three methods for fashion concept prediction on our new benchmark, and then analyze the fashion knowledge based on the extracted concepts. Particularly, we are interested in the following questions:

(1) **RQ1**: Does our fashion knowledge extraction model perform well in the preliminary step of concept prediction?

(2) **RQ2**: Why does our method achieve superior performance?

(3) **RQ3**: Whether the extracted fashion knowledge is reasonable and important to downstream applications?

5.1 Experimental Settings

Experimental Setup. The dataset is constructed in a semi-supervised manner with manual correction (Section 3). For evaluation, we split the dataset into two parts: 90% for training (70% machine-labeled data and 20% clean data) and 10% for testing. Note that all the testing data is randomly selected from the clean part. The evaluation metric is the standard accuracy.

Implementation Details. For visual representations, we use the pretrained ResNet-18 [11] as the feature extractor, which outputs a 512-D dense vector representation. As for the textual information, we utilize the pretrained 300-D word embedding (*i.e.*, Glove [22]) followed by a text CNN architecture [15], which consists of a single channel, four kernels with sizes of {2,3,4,5}, and a max pooling layer, where each kernel has 32 feature maps and uses the rectified linear unit (ReLU) as the activation function. Finally, we obtain a 128-D vector for text representation. The hidden state size of two Bi-LSTM networks is set as 512. In terms of the order of input sequence of the first Bi-LSTM, we sort the clothing regions by their spatial positions (*i.e.*, from left to right, top to bottom). For the second Bi-LSTM, we keep the order of attributes classifiers the same among all training samples. For weak label modeling, we implement the noise transition matrix with a fully-connected layer [27]. We initialize the label transition matrix with a statistical estimation on the part of training data with human-corrected labels. In particular, for each element $q_{j,i}$ of $\mathbf{Q} \in \mathbb{R}^{N \times N}$ (Section 4.2), we count the number of labels whose ground-truth are i while their predictions are label j

by the tagging tool, and then normalize the estimated label transition matrix along each column. We empirically set the trade-off hyperparameter β as 0.5 throughout the experiments.

Our model is implemented with the PyTorch framework. For optimization, we employ the stochastic gradient descent (SGD) [2] with the momentum factor of 0.9. We set the initial learning rate as 0.001 for the text CNN and image CNN, and 10^{-5} for the linear layer of label transition matrix. The learning rate drops by 10 after every 4 epochs. The performance of the model on the testing set is reported until convergence.

Baseline Methods. Since fashion knowledge extraction is a relatively new problem and there are few specific methods for solving it, we choose the following three state-of-the-art baselines that tackle one or more subtasks of fashion concept (*i.e.*, occasion, clothing category and attributes) prediction. The results of predicted concepts play an important role in fashion knowledge analysis.

- **DARN** [14] adopted an attribute-regularized two-stream CNN for cross domain fashion retrieval with a multi-branch fashion concept predictor. We only keep one stream of DARN for our task.
- **FashionNet** [21] is a state-of-the-art model for both clothing landmark detection, and clothing category and attributes prediction, which demonstrates compelling performance in clothing category classification and attribute prediction. We remove the clothing landmark prediction branch in FashionNet in our experiments.
- **EITree** [18] is a state-of-the-art model aiming at multimodal retrieval for fashion products by leveraging a special label structure of EI tree.

5.2 Fashion Concept Prediction (RQ1)

Since there is no occasion classification module in all of the three baselines, we add an additional branch of occasion classifier into these baselines. For fair comparisons, we also remove the textual inputs because the three baselines are not designed to handle textual information. Table 1 shows the accuracy of predicting fashion occasion, category and attributes. We have the following observations:

First, our method outperforms all the baselines on all of the three tasks. This is mainly because: 1) our model takes advantage of machine-labeled data while suppresses the inherent label noise through a weak label modeling module, thus obtaining additional creditable supervisions; and 2) we consider the dependencies and correlations among the occasion, clothing category and attributes, which provide additional discriminating capability for fashion concept prediction. Such dependencies and correlations among multiple fashion concepts implicitly demonstrate the existence of fashion knowledge and its positive impacts on related applications.

Second, our method has been further improved by utilizing textual information, especially on occasion classification. This is because the short texts affiliated with social media posts usually contain rich occasion-aware descriptions, which are important for fashion knowledge extraction.

Table 1: Overall Performance.

setting	occasion	category	attributes
DARN [14]	41.56%	68.04%	66.01%
FashionNet [21]	41.53%	67.33%	65.6%
EITree [18]	39.64%	68.28%	63.95%
our method w/o text	42.61%	73.6%	69.4%
our method	**47.88%**	**73.95%**	**69.59%**

5.3 Ablation Study (RQ2)

5.3.1 Effect of Contextualized Fashion Concept Learning. To evaluate the effect of the proposed contextualized fashion concept learning module, we further compare it with several variants listed below. In addition, we conduct experiments using only the clean data to get rid of the intervention of weak label modeling module, which is the concern of next section. We also remove the textual descriptions because it is modeled by another textCNN module.

Base: We remove the two Bi-LSTM modules without considering any concept dependencies and correlations, leading to a basic version of our model.

Bi-LSTM$_1$: We only keep the first Bi-LSTM network used to encode the dependency among co-occurring clothing regions.

Bi-LSTM$_2$: We only keep the second Bi-LSTM network used to capture the dependencies and correlations among clothing attributes and category.

Final: The proposed contextualized fashion concept learning model with two Bi-LSTM modules.

We can have the following observations from the results presented in Table 2:

First, Bi-LSTM$_1$ outperforms the Base method on all three tasks, especially on category and occasion classification. This is due to two reasons: 1) Bi-LSTM$_1$ improves the prediction of clothing category by modeling the dependencies among co-occurred clothing regions. It is reasonable, since the visual context of clothing regions in the same image is captured in this way, which results in contextualized representation of clothing regions. 2) It improves the prediction of occasions because we augment the CNN representation of occasion with the final hidden state of Bi-LSTM$_1$. It makes sense since the final hidden state encodes the contextualized clothing information which complements the occasion representation significantly.

Second, Bi-LSTM$_2$ again achieves better performance than the Base model, since it models the dependencies among different attribute representations and category representation.

Third, when using both the Bi-LSTM in the Final model, the performances of three predictions are all significantly improved. This demonstrates the necessity of employing both Bi-LSTM$_1$ and Bi-LSTM$_2$ to achieve mutual enhancement.

Table 2: The performance comparison regarding two Bi-LSTM modules. Experiments are conducted on clean data.

setting	occasion	category	attributes
Base	38.86%	69.89%	67.35%
Bi-LSTM$_1$	39.85%	70.82%	67.79%
Bi-LSTM$_2$	38.45%	71.31%	67.57%
Final	**40.06%**	**71.35%**	**67.82%**

5.3.2 Effect of Weak Label Modeling. To verify the utility and robustness of our approach to learning with weak labels, we conduct experiments to compare our model's performance with different weak data ratios. As illustrated in Figure 4, we gradually increase the weak data ratio along the x-axis to compare the performance of

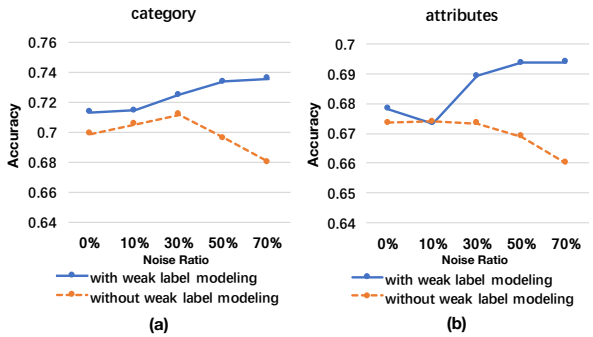


Figure 4: Performance comparison with different ratios of machine-labeled data with weak labels.

our method with/without weak label modeling module in Section 4.2. When using the introduced weak label modeling module, we can observe a clear performance improvement with the increasing weak data ratio. However, if we remove the weak label modeling module, the performance of category prediction would first gradually improve and then degrade rapidly when the ratio is large than 30%. This is because the flip label noise in machine-labeled data has dominated the optimization when the weak data ratio is beyond 30%. It indicates the effectiveness of the weak label modeling module which performs label correction by a label transition matrix to match the distribution of weak labels. With such a weak label modeling module, we can update our model with massive training data with weak labels.

5.4 Obtaining and Analyzing Fashion Knowledge (RQ3)

Given the predicted fashion concepts, the triplet form of fashion knowledge (occasion, person, clothing) will constitute a piece of knowledge that provides guidance for people's dressing in certain occasions. However, the fashion knowledge is subjective even if in the same conventions. Hence it is difficult to formulate a piece of convincing fashion knowledge from a single occurrence. Therefore, in this section, we discuss how to obtain useful fashion knowledge based on concepts from statistical perspective. The basic idea is that a piece of fashion knowledge is useful when it is widely adopted. As far as we know, this is the first work focusing on fashion knowledge rather than the concepts.

5.4.1 From Fashion Concepts to Knowledge. We employ a three-step process to summarize the most popular fashion knowledge from our extracted fashion concepts. First, we distinguish different clothes by their attributes. For example, two clothes are treated as the same only if they have the same category and attributes of interest. For efficiency, we use at most three attributes for two clothes' comparison. Second, we obtain the popular combinations of upper body clothes and lower body clothes, *a.k.a*, outfit, by counting the combinations for upper body clothes and lower body clothes. Third, we count all the combinations of occasion, gender, and clothing (outfits), which are the triplets we defined in Section 3. Finally we sort these triplets by their occurrence in descending order, and those triplets with higher frequencies are kept as useful fashion knowledge. Figure 5 shows some representative pieces of knowledge in different occasions.

occasion	gender	category:attributes
prom	F	dress: sleeveless, full_length, off_shoulder
prom	M	suits: collar, black, button, long_sleeve
conference	F	shirt: collar, long_sleeve, button, pants: full_length
conference	M	blazer: long_sleeve, collar, hip_length; pants: black
sports	F	tank_top: crop_length, athletic_pants: calf_length
sports	M	t-shirt: round_neck, hip_length, shorts: thigh_length
dating	F	dress: thigh_length, pencil, sleeveless
dating	M	t-shirt: short_sleeve, round_neck; jeans: full_length
travel	F	dress: knee_length, v_neck, a_line
travel	M	hoodies: hooded_neck, long_sleeve; jeans: navy
party	F	shirt: crop_length, short_sleeve; skirt: extra_short, a_line
party	M	t-shirt: pullover, round_neck; jeans: 7_8_length
shopping	F	t-shirt: short_sleeve, crop, v_neck; skirt: thigh_length
shopping	M	jacket: zip_closure, hip_length; jeans: full_length

Figure 5: Illustration of some pieces of fashion knowledge we obtained with high popularity.

Figure 6: Some exemplar images in different occasions. The top two rows (a) show the travel occasion in three different areas. The bottom part (b) demonstrates three occasions in three different areas.

5.4.2 Fashion Knowledge Analysis. Figure 6 shows some exemplar images in different occasions, areas and seasons. We can see that images within the same occasion and location or season present some commondition. For example, in the *conference* occasion, people all dress formally with either *dress*, *suit*, or *blazer*. While in the *travel* occasion, even though people wear different clothing, most of them are in casual style. Thus we can see that it is rational and reliable to extract the common dressing patterns under certain occasions.

On the other hand, different pieces of knowledge present rich and diverse information in terms of occasion, person, and clothing. First, occasion hugely influences the clothing distribution. For example, *dress* is popular in *wedding_guest* and *prom* occasions, *blazer* is popular in *conference* occasion, *t_shirt* and *tank_top* are popular in *sports* and *travel* occasions. However *shorts* scarcely appear in the *wedding_guest* and *prom* occasions, and *dress* rarely appears in *sports* occasion. This observation is in harmony with common sense and demonstrates that our extracted knowledge captures the

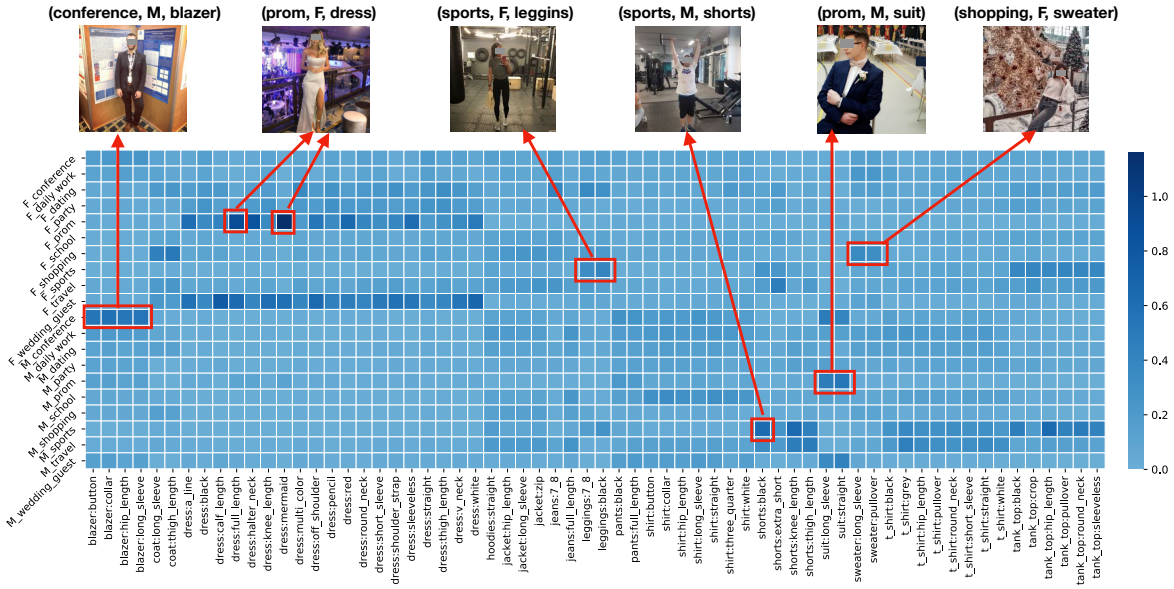


Figure 7: The correlation between occasions and popular clothing. The horizontal axis demonstrates various popular clothes. The vertical axis shows the occasions for male (M-) and female (F-). The darker color means there are more images satisfying the condition of triplet at that point. The result is based on our model's prediction of testing data.

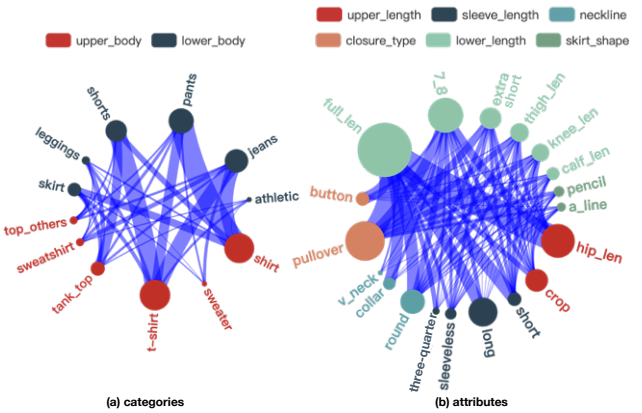


Figure 8: The cross-category matching relationships between categories (a) and attributes (b). The size of each node shows the number of clothing and the strength of edges represents the number of connections.

correlation between occasions and clothing. Second, people with different genders have distinctive dressing styles. The top left part of Figure 7 indicates that females are more likely to wear *dress* and the bottom right of Figure 7 illustrates that males are more likely to wear *t-shirt* and *pants*. This phenomenon seamlessly verifies our claim in the very beginning of this paper that what to wear and how to wear are hugely affected by human identity.

Interestingly, we can also discover some insightful points by fine-grained comparisons. For example, in terms of the dresses in occasion *F_prom* and *F_wedding_guest*, the attributes *full_length*, *sleeveless*, and *a_line* are most popular, while *thigh_length* and *pencil* are less popular. It makes sense that in a formal occasion like *prom*, a *dress* of *full_length*, *sleeveless*, and *a_line* are much more formal than that of thigh_length and *pencil*.

Cross Category Matching. Without considering the occasions, the cross category matching reveals the relationship among clothes themselves, which will benefit many downstream applications. For example, in fashion recommendation [6, 25, 34, 35], such cross-category matching rules can guide the model to recommend clothes that are functionally compatible with the given clothes. We illustrate the matching popularity between *upper_body* clothes and *lower_body* clothes in Figure 8 (a). For example, *shirts* are more likely to match with *pants* other than *shorts*, and *tank_top* is linked to shorts with the highest weight among all the connections. Similar matching patterns also exist between attributes. As shown in Figure 8 (b), *long_sleeves*'s connection with *full_length* has the highest strength but its connections with *thigh_length* are weaker.

6 CONCLUSION

In this paper, we explored a new task of automatically extracting fashion knowledge from social media. To build an effective model for fashion concept prediction, we designed a contextualized fashion concept learning model and enhanced it with weak label modeling. And the extracted knowledge verifies our hypothesis that fashion are affected by person, occasion, and clothing.

There are several research directions that can be done in the future: 1) The use of the extracted fashion knowledge into various downstream applications. 2) The extraction of fine-grained knowledge. In the future, we will try to apply the proposed visual concept learning method to enhance the general visual retrieval tasks, such as cross-modal retrieval [13] and person retrieval [36, 37].

ACKNOWLEDGEMENT

This research is part of NExT++ project, which is supported by the National Research Foundation, Prime Minister's Office, Singapore under its IRC@SG Funding Initiative.

REFERENCES

[1] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. AcM, 1247–1250.

[2] Léon Bottou. 1991. Stochastic gradient learning in neural networks. *Proceedings of Neuro-Nımes* 91, 8 (1991), 12.

[3] Yixin Cao, Lei Hou, Juanzi Li, and Zhiyuan Liu. 2018. Neural Collective Entity Linking. In *Proceedings of the 27th International Conference on Computational Linguistics*. 675–686.

[4] Yixin Cao, Zikun Hu, Tat-Seng Chua, Zhiyuan Liu, and Heng Ji. 2019. Low-Resource Name Tagging Learned with Weakly Labeled Data. *arXiv preprint arXiv:1908.09659* (2019).

[5] Huizhong Chen, Andrew Gallagher, and Bernd Girod. 2012. Describing clothing by semantic attributes. In *European conference on computer vision*. Springer, 609–623.

[6] Xu Chen, Hanxiong Chen, Hongteng Xu, Yongfeng Zhang, Yixin Cao, Zheng Qin, and Hongyuan Zha. 2019. Personalized Fashion Recommendation with Visual Explanations based on Multimodal Attention Network: Towards Visually Explainable Recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 765–774.

[7] Xinlei Chen, Abhinav Shrivastava, and Abhinav Gupta. 2013. NEIL: Extracting Visual Knowledge from Web Data. In *The IEEE International Conference on Computer Vision (ICCV)*.

[8] Charles Corbiere, Hedi Ben-Younes, Alexandre Ramé, and Charles Ollion. 2017. Leveraging weakly annotated data for fashion image retrieval and label prediction. In *Proceedings of the IEEE International Conference on Computer Vision*. 2268–2274.

[9] MS Fabian, K Gjergji, WEIKUM Gerhard, et al. 2007. Yago: A core of semantic knowledge unifying wordnet and wikipedia. In *16th International World Wide Web Conference, WWW*. 697–706.

[10] M Hadi Kiapour, Xufeng Han, Svetlana Lazebnik, Alexander C Berg, and Tamara L Berg. 2015. Where to buy it: Matching street clothing photos in online shops. In *Proceedings of the IEEE international conference on computer vision*. 3343–3351.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[12] Shintami Chusnul Hidayati, Cheng-Chun Hsu, Yu-Ting Chang, Kai-Lung Hua, Jianlong Fu, and Wen-Huang Cheng. 2018. What Dress Fits Me Best?: Fashion Recommendation on the Clothing Style for Personal Body Shape. In *2018 ACM Multimedia Conference on Multimedia Conference*. ACM, 438–446.

[13] Richang Hong, Lei Li, Junjie Cai, Dapeng Tao, Meng Wang, and Qi Tian. 2017. Coherent semantic-visual indexing for large-scale image retrieval in the cloud. *IEEE Transactions on Image Processing* 26, 9 (2017), 4128–4138.

[14] Junshi Huang, Rogerio S Feris, Qiang Chen, and Shuicheng Yan. 2015. Cross-domain image retrieval with a dual attribute-aware ranking network. In *Proceedings of the IEEE international conference on computer vision*. 1062–1070.

[15] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1746–1751.

[16] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* 123, 1 (2017), 32–73.

[17] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web* 6, 2 (2015), 167–195.

[18] Lizi Liao, Xiangnan He, Bo Zhao, Chong-Wah Ngo, and Tat-Seng Chua. 2018. Interpretable multimodal retrieval for fashion products. In *2018 ACM Multimedia Conference on Multimedia Conference*. ACM, 1571–1579.

[19] Lizi Liao, Yunshan Ma, Xiangnan He, Richang Hong, and Tat-seng Chua. 2018. Knowledge-aware Multimodal Dialogue Systems. In *2018 ACM Multimedia Conference on Multimedia Conference*. ACM, 801–809.

[20] Si Liu, Zheng Song, Guangcan Liu, Changsheng Xu, Hanqing Lu, and Shuicheng Yan. 2012. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3330–3337.

[21] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. 2016. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1096–1104.

[22] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[23] Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (2018).

[24] Xindi Shang, Tongwei Ren, Jingfan Guo, Hanwang Zhang, and Tat-Seng Chua. 2017. Video visual relation detection. In *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 1300–1308.

[25] Xuemeng Song, Fuli Feng, Xianjing Han, Xin Yang, Wei Liu, and Liqiang Nie. 2018. Neural compatibility modeling with attentive knowledge distillation. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 5–14.

[26] Xuemeng Song, Fuli Feng, Jinhuan Liu, Zekun Li, Liqiang Nie, and Jun Ma. 2017. Neurostylist: Neural compatibility modeling for clothing matching. In *Proceedings of the 25th ACM international conference on Multimedia*. ACM, 753–761.

[27] Sainbayar Sukhbaatar, Joan Bruna, Manohar Paluri, Lubomir Bourdev, and Rob Fergus. 2014. Training convolutional networks with noisy labels. *arXiv preprint arXiv:1406.2080* (2014).

[28] Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. 2018. Joint optimization framework for learning with noisy labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5552–5560.

[29] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM* 57, 10 (2014), 78–85.

[30] Wenguan Wang, Yuanlu Xu, Jianbing Shen, and Song-Chun Zhu. 2018. Attentive fashion grammar network for fashion landmark detection and clothing category classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4271–4280.

[31] Xiang Wang, Xiangnan He, Yixin Cao, Meng Liu, and Tat-Seng Chua. 2019. KGAT: Knowledge Graph Attention Network for Recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*. 950–958.

[32] Xiang Wang, Dingxian Wang, Canran Xu, Xiangnan He, Yixin Cao, and Tat-Seng Chua. 2019. Explainable Reasoning over Knowledge Graphs for Recommendation. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. 5329–5336.

[33] Wei Yang, Ping Luo, and Liang Lin. 2014. Clothing co-parsing by joint image segmentation and labeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3182–3189.

[34] Xun Yang, Xiangnan He, Xiang Wang, Yunshan Ma, Fuli Feng, Meng Wang, and Tat-Seng Chua. 2019. Interpretable Fashion Matching with Rich Attributes. SIGIR.

[35] Xun Yang, Yunshan Ma, Lizi Liao, Meng Wang, and Tat-Seng Chua. 2019. TransNFCM: Translation-Based Neural Fashion Compatibility Modeling. *AAAI* (2019).

[36] Xun Yang, Meng Wang, Richang Hong, Qi Tian, and Yong Rui. 2017. Enhancing person re-identification in a self-trained subspace. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 13, 3 (2017), 27.

[37] Xun Yang, Meng Wang, and Dacheng Tao. 2017. Person re-identification with metric learning using privileged information. *IEEE Transactions on Image Processing* 27, 2 (2017), 791–805.

[38] Wenhui Yu, Huidi Zhang, Xiangnan He, Xu Chen, Li Xiong, and Zheng Qin. 2018. Aesthetic-based clothing recommendation. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 649–658.

[39] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters* 23, 10 (2016), 1499–1503.