

Aesthetic-based Clothing Recommendation

Wenhui Yu*

Tsinghua University

Beijing, China

yuwh16@mails.tsinghua.edu.cn

Huidi Zhang*

Tsinghua University

Beijing, China

zhd16@mails.tsinghua.edu.cn

Xiangnan He

National University of Singapore

Singapore, 117417

xiangnanhe@gmail.com

Xu Chen

Tsinghua University

Beijing, China

xu-ch14@mails.tsinghua.edu.cn

Li Xiong

Emory University

Atlanta, USA

lxiong@emory.edu

Zheng Qin†

Tsinghua University

Beijing, China

qingzh@mail.tsinghua.edu.cn

ABSTRACT

Recently, product images have gained increasing attention in clothing recommendation since the visual appearance of clothing products has a significant impact on consumers' decision. Most existing methods rely on conventional features to represent an image, such as the visual features extracted by convolutional neural networks (CNN features) and the scale-invariant feature transform algorithm (SIFT features), color histograms, and so on. Nevertheless, one important type of features, the *aesthetic features*, is seldom considered. It plays a vital role in clothing recommendation since a users' decision depends largely on whether the clothing is in line with her aesthetics, however the conventional image features cannot portray this directly. To bridge this gap, we propose to introduce the aesthetic information, which is highly relevant with user preference, into clothing recommender systems. To achieve this, we first present the aesthetic features extracted by a pre-trained neural network, which is a brain-inspired deep structure trained for the aesthetic assessment task. Considering that the aesthetic preference varies significantly from user to user and by time, we then propose a new tensor factorization model to incorporate the aesthetic features in a personalized manner. We conduct extensive experiments on real-world datasets, which demonstrate that our approach can capture the aesthetic preference of users and significantly outperform several state-of-the-art recommendation methods.

CCS CONCEPTS

- Information systems → Collaborative filtering; Social recommendation; Recommender systems; • Human-centered computing → Social recommendation;

KEYWORDS

Clothing recommendation, side information, aesthetic features, tensor factorization, dynamic collaborative filtering.

School of Software, Tsinghua National Laboratory for Information Science and Technology.

* Both authors contributed equally to this work.

† The corresponding author.

This paper is published under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.

WWW 2018, April 23–27, 2018, Lyon, France

© 2018 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC BY 4.0 License.

ACM ISBN 978-1-4503-5639-8/18/04.

<https://doi.org/10.1145/3178876.3186146>

1 INTRODUCTION

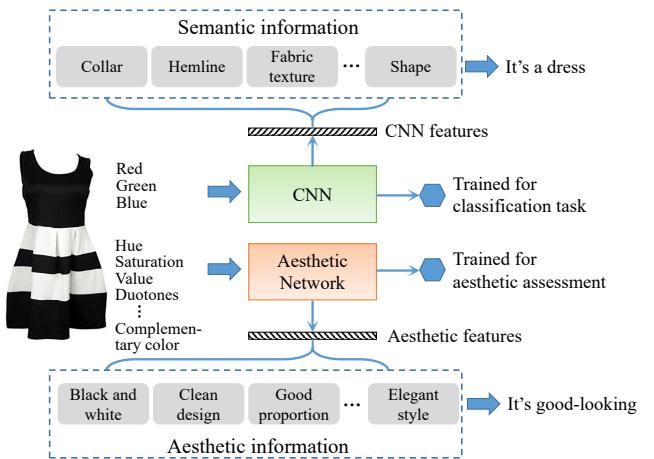


Figure 1: Comparison of CNN features and aesthetic features. The CNN is inputted with the RGB components of an image and trained for the classification task, while the aesthetic network is inputted with raw aesthetic features and trained for the aesthetic assessment task.

When shopping for clothing on the Web, we usually look through product images before making the decision. Product images provide abundant information, including design, color schemes, decorative pattern, texture, and so on; we can even estimate the thickness and quality of a product from its images. As such, product images play a key role in the clothing recommendation task.

To leverage this information and enhance the performance, existing clothing recommender systems use image data with various image features, like features extracted by convolutional neural networks (CNN features) and the scale-invariant feature transform algorithm (SIFT features), color histograms, etc. For example, [8, 12, 15, 30] utilized the CNN features extracted by a deep convolutional neural network. Trained for the classification task, CNN features contain semantic information to distinguish items and have been widely used in recommendation tasks. However, one important factor, aesthetics, has yet been considered in previous research. When purchasing clothing products, what consumers concern is not only “What is the product?”, but also “Is the product good-looking?”.

Taking the product shown in Figure 1 as an example. A consumer will notice that the dress is of colors black and white, of simple but elegant design, and has a delightful proportion. She will purchase it only if she is satisfied with all these aesthetic factors. In fact, for some consumers, especially young females, aesthetic factor could be the primary factor, even more important than others like quality, comfort, and prices. As such, we need novel features to capture this indispensable information. Unfortunately, CNN features do not encode the aesthetic information by nature. [45] used color histograms to portray consumers' intuitive perception about an image while it is too crude and primitive. To provide quality recommendation for the clothing domain, comprehensive and high-level aesthetic features are greatly desired.

In this paper, we leverage the aesthetic network to extract relevant features. The differences between an aesthetic network and a CNN are demonstrated in Figure 1. Recently, [42] proposed a **Brain-inspired Deep Network (BDN)**, which is a deep structure trained for image aesthetic assessment. The inputs are several raw features that are indicative of aesthetic feelings, like hue, saturation, value, duotones, complementary color, etc. It then extracts high-level aesthetic features from the raw features. In this paper, BDN is utilized to extract the holistic features to represent the aesthetic elements of a clothing product (taking Figure 1 as an example, the aesthetic elements can be color, structure, proportion, style, etc.).

It is obvious that the aesthetic preference shows a significant diversity among different people. For instance, children prefer colorful and lovely products while adults prefer those can make them look mature and elegant; women may prefer exquisite decorations while men like concise designs. Moreover, the aesthetic tastes of consumers also change with time, either in short term, or in long term. For example, the aesthetic tastes vary in different seasons periodically—in spring or summer, people may prefer clothes with light color and fine texture, while in autumn or winter, people tend to buy clothes with dark color, rough texture, and loose style. In the long term, the fashion trend changes all the time and the popular color and design may be different by year.

To capture the diversity of the aesthetic preference among consumers and over time, we exploit tensor factorization as a basic model. There are several ways to decompose a tensor [23, 34, 38], however, there are certain drawbacks in existing models. To address the clothing recommendation task better, we first propose a **Dynamic Collaborative Filtering (DCF)** model trained with coupled matrices to mitigate the *sparsity* problem [1]. We then combine it with the additional image features (concatenated aesthetic and CNN features) and term the method as **Dynamic Collaborative Filtering model with Aesthetic Features** (called **DCFA**). We optimize the models with bayesian personalized ranking (BPR) optimization criterion [33] and evaluated their performance on an *Amazon clothing* dataset. Extensive experiments show that we improve the performance significantly by incorporating aesthetic features.

To summarize, our main contributions are as follows:

- We leverage novel aesthetic features in recommendation to capture consumers' aesthetic preference. Moreover, we compare the effect with several conventional features to demonstrate the necessity of the aesthetic features.

- We propose a novel DCF model to portray the purchase events in three dimensions: users, items, and time. We then incorporate aesthetic features into DCF and train it with coupled matrices to alleviate the sparsity problem.
- We conduct comprehensive experiments on real-world datasets to demonstrate the effectiveness of our DCFA method.

2 RELATED WORK

This paper develops aesthetic-aware clothing recommender systems. Specifically, we incorporate the features extracted from the product images by an aesthetic network into a tensor factorization model. As such, we review related work on aesthetic networks, image-based recommendation, and tensor factorization.

2.1 Aesthetic Networks

The aesthetic networks are proposed for image aesthetic assessment. After [14] first proposed the aesthetic assessment problem, many research efforts exploited various handcrafted features to extract the aesthetic information of images [14, 22, 26, 28]. To portray the subjective and complex aesthetic perception, [4, 25, 27, 36, 42] exploited deep networks to emulate the underlying complex neural mechanisms of human perception, and displayed the ability to describe image content from the primitive level (low-level) features to the abstract level (high-level) features.

2.2 Image-based Recommendations

Recommendation has been widely studied due to its extensive use, and many effective methods have been proposed [3, 11, 16–18, 24, 29, 32, 33, 35, 41, 44]. The power of recommender systems lies on their ability to model the complex preference that consumers exhibit toward items based on their past interactions and behavior. To extend their expressive power, various works exploited image data [7–9, 12, 13, 15, 19, 30, 45]. For example, [13] infused product images and item descriptions together to make dynamic predictions, [9, 12] leveraged textual and visual information to recommend tweets and personalized key frames respectively. Image data can also mitigate the sparsity problem and cold start problem. [8, 15, 19, 30] used CNN features of product images while [45] recommended movies with color histograms of posters and frames. [20, 37, 39] recommended clothes by considering the clothing fashion style.

2.3 Tensor Factorization

Time is an important contextual information in recommender systems since the sales of commodities show a distinct time-related succession. In context-aware recommender systems, tensor factorization has been extensively used. For example, [23, 38] introduced two main forms of tensor decomposition, the CANDECOMP/PARAFAC (**CP**) and Tucker decomposition. [21] first utilized tensor factorization for context-aware collaborative filtering. [10, 34] proposed a **Pairwise Interaction Tensor Factorization (PITF)** model to decompose the tensor with a linear complexity. Nevertheless, tensor-based methods suffer from several drawbacks like poor convergence in sparse data [6] and not scalable to large-scale datasets [2]. To address these limitations, [1, 5, 43] formulated recommendation models with the **Coupled Matrix and Tensor Factorization (CMTF)** framework.

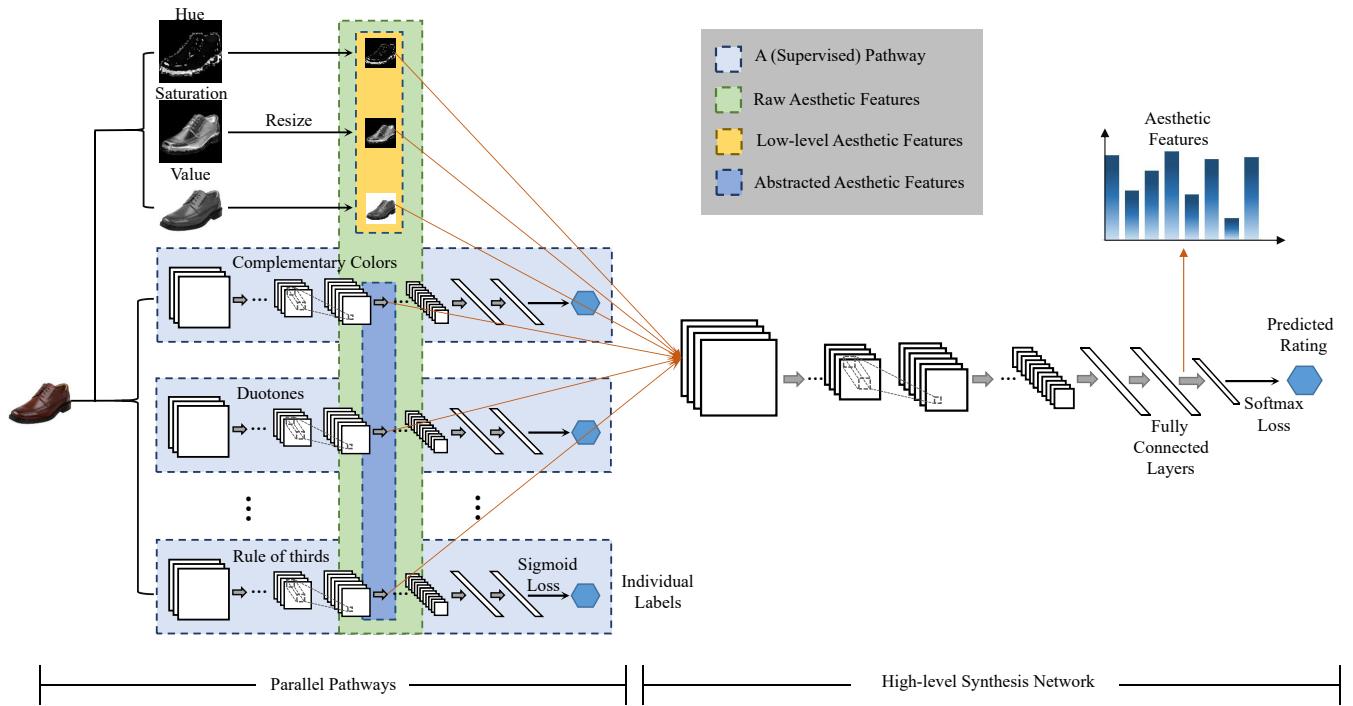


Figure 2: Brain-inspired Deep Network (BDN) architecture.

3 PRELIMINARIES

This section introduces some preliminaries about the aesthetic neural network, which is used to extract the aesthetic features of clothing images. In [42], the authors introduced the Brain-inspired Deep Networks (BDN, shown in Figure 2), a deep CNN structure consists of several parallel pathways (sub-networks) and a high-level synthesis network. It is trained on the *Aesthetic Visual Analysis (AVA)* dataset, which contains 250,000 images with aesthetic ratings and tagged with 14 photographic styles (e.g., complementary colors, duotones, rule of thirds, etc.). The pathways take the form of convolutional networks to exact the abstracted aesthetic features by *pre-trained* with the individual labels of each tag. For example, when training the pathway for complementary colors, the individual label is 1 if the sample is tagged with “complementary colors” and is 0 if not. We input the raw features, which include low-level features (hue, saturation, value) and abstracted features (feature maps of the pathways), into the high-level synthesis network and *jointly tune* it with the pathways for aesthetic rating prediction. Considering that the AVA is a photography dataset and the styles are for photography, so not all the raw features extracted by the pathways are desired in our recommendation task. Thus we only reserve the pathways that are relevant to the clothing aesthetic. Finally, we use the output of the second fully-connected layer of the synthesis network as our aesthetic features.

We then analyze several extensively used features and demonstrate the superiority of our aesthetic features.

CNN Features: These are the most extensively used features due to their extraordinary representation ability. Typically the output of certain fully-connected layer of a deep CNN structure is used.

For example, a common choice is the Caffe reference model with 5 convolutional layers followed by 3 fully-connected layers (pre-trained on the ImageNet dataset); the features are the output of FC7, namely, the second fully-connected layer, which is a feature vector of length 4096.

CNN features mainly contain semantic information, which contributes little to evaluate the aesthetics of an image. Recall the example in Figure 1, it can encode “There is a skirt in the image.” but cannot express “The clothing is beautiful and fits the consumer’s taste.”. Devised for aesthetic assessment, BDN can capture the high-level aesthetic information. As such, our aesthetic features can do better in beauty estimating and complement CNN features in clothing recommendation.

Color Histograms: [45] exploited color histograms to represent human’s feeling about the posters and frames for movie recommendation. Though can get the aesthetic information roughly, the low-level handcrafted features are crude, unilateral, and empirical. BDN can get abundant visual features by the pathways. Also, it is data-driven, since the rules to extract features are learned from the data. Compared with the intuitive color histograms, our aesthetic features are more objective and comprehensive. Recall the example in Figure 1 again, color histograms can tell us no more than “The clothes in the image is white and black”.

4 CLOTHING RECOMMENDATION WITH AESTHETIC FEATURES

In this section, we first introduce the basic tensor factorization model (DCF). We next construct a hybrid model that integrates image features into the basic model (DCFA).

4.1 Basic Model

Considering the impact of time on aesthetic preference, we propose a context-aware model as the basic model to account for the temporal factor. We use a $P \times Q \times R$ tensor \mathbf{A} to indicate the purchase events among the user, clothes, and time dimensions (where P, Q, R are the number of users, clothes, and time intervals, respectively). If user p purchased item q in time interval r , $A_{pqr} = 1$, otherwise $A_{pqr} = 0$. Tensor factorization has been widely used to predict the missing entries (i.e., zero elements) in \mathbf{A} , which can be used for recommendation. There are several approaches and we introduce the most common ones:

4.1.1 Existing Methods and Their Limitations. In this subsection, we summarize the motivation of proposing our novel tensor factorization model.

Tucker Decomposition: This method [23] decomposes the tensor \mathbf{A} into a tensor core and three matrices,

$$\hat{\mathbf{A}}_{pqr} = \sum_{i=1}^{K_1} \sum_{j=1}^{K_2} \sum_{k=1}^{K_3} \mathbf{a}_{ijk} \mathbf{U}_{ip} \mathbf{V}_{jq} \mathbf{T}_{kr},$$

where $\mathbf{a} \in \mathbb{R}^{K_1 \times K_2 \times K_3}$ is the tensor core, $\mathbf{U} \in \mathbb{R}^{K_1 \times P}$, $\mathbf{V} \in \mathbb{R}^{K_2 \times Q}$, and $\mathbf{T} \in \mathbb{R}^{K_3 \times R}$. Tucker decomposition has very strong representation ability, but it is very time consuming, and hard to converge.

CP Decomposition: The tensor \mathbf{A} is decomposed into three matrices in CP decomposition,

$$\hat{\mathbf{A}}_{pqr} = \sum_{k=1}^K \mathbf{U}_{kp} \mathbf{V}_{kq} \mathbf{T}_{kr},$$

where $\mathbf{U} \in \mathbb{R}^{K \times P}$, $\mathbf{V} \in \mathbb{R}^{K \times Q}$, and $\mathbf{T} \in \mathbb{R}^{K \times R}$. This model has been widely used due to its linear time complexity, especially in Coupled Matrix and Tensor Factorization (CMTF) structure model [1, 2, 5]. However, all dimensions (users, clothes, time) are related by the same latent features. Intuitively, we want the latent features relating users and clothes to contain the information about users' preference, like aesthetics, prices, quality, brands, etc., and the latent features relating clothes and time to contain the information about the seasonal characteristics and fashion elements of clothes like colors, thickness, design, etc.

PITF Decomposition: The Pairwise Interaction Tensor Factorization (PITF) model [34] decomposes \mathbf{A} into three pair of matrices,

$$\hat{\mathbf{A}}_{pqr} = \sum_{k=1}^K \mathbf{U}_{kp}^V \mathbf{V}_{kq}^U + \sum_{k=1}^K \mathbf{U}_{kp}^T \mathbf{T}_{kr}^U + \sum_{k=1}^K \mathbf{V}_{kq}^T \mathbf{T}_{kr}^V,$$

where $\mathbf{U}^V, \mathbf{U}^T \in \mathbb{R}^{K \times P}$, $\mathbf{V}^U, \mathbf{V}^T \in \mathbb{R}^{K \times Q}$, $\mathbf{T}^U, \mathbf{T}^V \in \mathbb{R}^{K \times R}$. PITF has a linear complexity and strong representation ability. Yet, it is not in line with practical applications due to the additive combination of each pair of matrices. For example, in PITF, for certain clothes q liked by the user p but not fitting the current time r , q gets a high score for p and a low score for r . Intuitively it should not be recommended to the user since we want to recommend the right item in the right time. However, the total score can be high enough if p likes q so much that q 's score for p is very high. In this case, q will be returned even it does not fit the time. In addition, PITF model is inappropriate to be trained with coupled matrices.

4.1.2 Dynamic Collaborative Filtering (DCF) Model. To address the limitations of the aforementioned models, we propose a new tensor factorization method. When a user makes a purchase decision on a clothing product, there are two primary factors: if the product fits the user's preference and if it fits the time. A clothing product fits a user's preference if the appearance is appealing, the style fits the user's tastes, the quality is good, and the price is acceptable. And a clothing product fits the time if it is in-season and fashionable. For user p , clothing q , and time interval r , we use the scores S_1 and S_2 to indicate how the user likes the clothing and how the clothing fits the time respectively. $S_1 = 1$ when the user likes the clothing and $S_1 = 0$ otherwise. Similarly, $S_2 = 1$ if the clothing fits the time and $S_2 = 0$ otherwise. The consumer will buy the clothing only if $S_1 = 1$ and $S_2 = 1$, so, $\hat{\mathbf{A}}_{pqr} = S_1 \& S_2$. To make the formula differentiable, we can approximately formulate it as $\hat{\mathbf{A}}_{pqr} = S_1 \cdot S_2$. We present S_1 and S_2 in the form of matrix factorization:

$$S_1 = \sum_{i=1}^{K_1} \mathbf{U}_{ip} \mathbf{V}_{iq}$$

$$S_2 = \sum_{j=1}^{K_2} \mathbf{T}_{jr} \mathbf{W}_{jq},$$

where $\mathbf{U} \in \mathbb{R}^{K_1 \times P}$, $\mathbf{V} \in \mathbb{R}^{K_1 \times Q}$, $\mathbf{T} \in \mathbb{R}^{K_2 \times R}$, and $\mathbf{W} \in \mathbb{R}^{K_2 \times Q}$. The prediction is then given by:

$$\hat{\mathbf{A}}_{pqr} = (\mathbf{U}_{*p}^T \mathbf{V}_{*q}) (\mathbf{T}_{*r}^T \mathbf{W}_{*q}). \quad (1)$$

We can see that in Equation (1), the latent features relating users and clothes are independent with those relating clothes and time. Though K_1 -dimensional vector \mathbf{V}_{*q} and K_2 -dimensional vector \mathbf{W}_{*q} are all latent features of clothing q , \mathbf{V}_{*q} captures the information about users' preference intuitively whereas \mathbf{W}_{*q} captures the temporal information of the clothing. Compared with CP decomposition, our model is more expressive in capturing the underlying latent patterns in purchases. Compared with PITF, combining S_1 and S_2 with $\&$ (approximated by multiplication) is helpful to recommend right clothing in right time. Moreover, our model is efficient and easy to train compared with the Tucker decomposition.

4.1.3 Coupled Matrix and Tensor Factorization. Though widely used to portray the context information in recommendation, tensor factorization suffers from poor convergence due to the sparsity of the tensor. To relieve this problem, [1] proposed a CMTF model, which decomposes the tensor with coupled matrices. In this subsection, we couple our tensor factorization model with restrained matrices during training.

User × Clothing Matrix: We use matrix $\mathbf{B} \in \mathbb{R}^{P \times Q}$ to indicate the purchase activities between users and clothes. $B_{pq} = 1$ if the user p purchased clothing q and $B_{pq} = 0$ if not.

Time × Clothing Matrix: We use matrix $\mathbf{C} \in \mathbb{R}^{R \times Q}$ to record when the clothing was purchased. Since the characteristics of clothing change steadily with time, we do a coarse-grained discretization on time to avoid the tensor from being extremely sparse. Time is divided into R intervals in total. $C_{rq} = 1$ if the clothing q is purchased in time interval r and $C_{rq} = 0$ if not.

Objective Function Formulation: In existing works [1, 5, 21, 43], CMTF models are optimized by minimizing the sum of the squared error of each simulation (MSE_OPT). It is represented as:

$$\text{MSE_OPT} = \frac{1}{2} \|\mathbf{A} - \hat{\mathbf{A}}\|_F^2 + \frac{\lambda_1}{2} \|\mathbf{B} - \hat{\mathbf{B}}\|_F^2 + \frac{\lambda_2}{2} \|\mathbf{C} - \hat{\mathbf{C}}\|_F^2 + \frac{\lambda_3}{2} \|\mathbf{U}\|_F^2 + \frac{\lambda_4}{2} \|\mathbf{V}\|_F^2 + \frac{\lambda_5}{2} \|\mathbf{T}\|_F^2 + \frac{\lambda_6}{2} \|\mathbf{W}\|_F^2, \quad (2)$$

where $\hat{\mathbf{A}}$ is defined in Equation (1), $\hat{\mathbf{B}} = \mathbf{U}^\top \mathbf{V}$, $\hat{\mathbf{C}} = \mathbf{T}^\top \mathbf{W}$, and $\|\cdot\|_F$ is the Frobenius norm of the matrix. The last four terms of Equation (2) are the regularization terms to prevent overfitting. Although the pointwise squared loss has been widely used in recommendation, it is not directly optimized for ranking. To get better top-n performance, we next introduce our hybrid model with BPR [33] optimization criterion.

4.2 Hybrid Model

4.2.1 Problem Formulation. Combined with image features, we formulate the predictive model as:

$$\hat{\mathbf{A}}_{pqr} = \left(\mathbf{U}_{*p}^\top \mathbf{V}_{*q} + \mathbf{M}_{*p}^\top \mathbf{F}_{*q} \right) \left(\mathbf{T}_{*r}^\top \mathbf{W}_{*q} + \mathbf{N}_{*r}^\top \mathbf{F}_{*q} \right), \quad (3)$$

where $\mathbf{F} \in \mathbb{R}^{K \times Q}$ is the feature matrix, \mathbf{F}_{*q} is the image features of clothing q , which is the concatenation of CNN features (\mathbf{f}_{CNN}) and aesthetic features (\mathbf{f}_{AES}), $\mathbf{F}_{*q} = \begin{bmatrix} \mathbf{f}_{CNN} \\ \mathbf{f}_{AES} \end{bmatrix}$ and $K = 8192$. $\mathbf{M} \in \mathbb{R}^{K \times P}$ and $\mathbf{N} \in \mathbb{R}^{K \times R}$ are aesthetic preference matrices. \mathbf{M}_{*p} encodes the preference of user p and \mathbf{N}_{*r} encodes the preference in time interval r . In our model, both the latent features and image features contribute to the final prediction. Though the latent features can uncover any relevant attribute theoretically, they usually cannot in real-world applications on account of the sparsity of the data and lack of information. So the assistance of image information can highly enhance the model. Also, recommender systems often suffer from the *cold start* problem. It is hard to extract information from users and clothes without consumption records. In this case, content and context information can alleviate this problem. For example, for certain “cold” clothing q , we can decide whether to recommend it to certain consumer p in current time r according to if q looks satisfying to the consumer (determined by \mathbf{M}_{*p}) and to the time (determined by \mathbf{N}_{*r}).

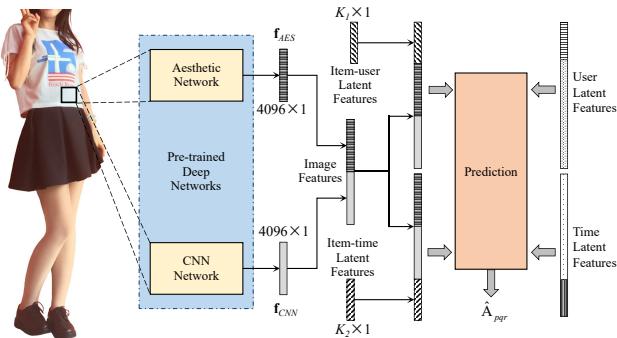


Figure 3: Diagram of our preference predictor.

4.2.2 Model Learning. The model is optimized with BPR optimization criterion from users’ *implicit feedback* (purchase record) with mini-batch gradient descent, which calculates the gradient with a small batch of samples. BPR is a pairwise ranking optimization framework and we represent the training set D into three different forms:

$$D_{pr} = \{(p, q, q', r) | p \in \mathcal{P} \wedge r \in \mathcal{R} \wedge q \in Q_{pr}^+ \wedge q' \in Q \setminus Q_{pr}^+\},$$

$$D_p = \{(p, q, q') | p \in \mathcal{P} \wedge q \in Q_p^+ \wedge q' \in Q \setminus Q_p^+\},$$

$$D_r = \{(r, q, q') | r \in \mathcal{R} \wedge q \in Q_r^+ \wedge q' \in Q \setminus Q_r^+\},$$

where u denotes the user, r represents the time, q represents the positive feedback, and q' represents the non-observed item. The objective function is formulated as:

$$\text{BPR_OPT} = \sum_{(p, q, q', r) \in D_{pr}} \ln \sigma(\hat{\mathbf{A}}_{pqq'r}) + \lambda_1 \sum_{(p, q, q') \in D_p} \ln \sigma(\hat{\mathbf{B}}_{pqq'}) + \lambda_2 \sum_{(r, q, q') \in D_r} \ln \sigma(\hat{\mathbf{C}}_{rqq'}) - \lambda_\Theta \|\Theta\|_F^2, \quad (4)$$

where $\hat{\mathbf{A}}$ is defined in the Equation (3), $\hat{\mathbf{B}} = \mathbf{U}^\top \mathbf{V} + \mathbf{M}^\top \mathbf{F}$, and $\hat{\mathbf{C}} = \mathbf{T}^\top \mathbf{W} + \mathbf{N}^\top \mathbf{F}$; $\hat{\mathbf{A}}_{pqq'r} = \hat{\mathbf{A}}_{pqr} - \hat{\mathbf{A}}_{pq'r}$, $\hat{\mathbf{B}}_{pqq'} = \hat{\mathbf{B}}_{pq} - \hat{\mathbf{B}}_{pq'}$, $\hat{\mathbf{C}}_{rqq'} = \hat{\mathbf{C}}_{rq} - \hat{\mathbf{C}}_{rq'}$; σ is the sigmoid function; $\Theta = \{\mathbf{U}, \mathbf{V}, \mathbf{T}, \mathbf{W}, \mathbf{M}, \mathbf{N}\}$ and $\lambda_\Theta = \{\lambda_3, \dots, \lambda_8\}$ respectively. We then calculate the gradient of Equation (4). To maximize the objective function, we take the first-order derivatives with respect to each model parameter:

$$\nabla_\Theta \text{BPR_OPT} = \sigma(-\hat{\mathbf{A}}_{pqq'r}) \frac{\partial \hat{\mathbf{A}}_{pqq'r}}{\partial \Theta} + \lambda_1 \sigma(-\hat{\mathbf{B}}_{pqq'}) \frac{\partial \hat{\mathbf{B}}_{pqq'}}{\partial \Theta} + \lambda_2 \sigma(-\hat{\mathbf{C}}_{rqq'}) \frac{\partial \hat{\mathbf{C}}_{rqq'}}{\partial \Theta} - \lambda_\Theta \Theta. \quad (5)$$

We use Θ to denote certain column of Θ . For our DCFA model, the derivatives are:

$$\frac{\partial \hat{\mathbf{A}}_{pqq'r}}{\partial \Theta} = \begin{cases} \hat{\mathbf{C}}_{rq} \mathbf{V}_{*q} - \hat{\mathbf{C}}_{rq'} \mathbf{V}_{*q'} & \text{if } \Theta = \mathbf{U}_{*p} \\ \hat{\mathbf{C}}_{rq} \mathbf{U}_{*p} / -\hat{\mathbf{C}}_{rq'} \mathbf{U}_{*p} & \text{if } \Theta = \mathbf{V}_{*q} / \mathbf{V}_{*q'} \\ \hat{\mathbf{C}}_{rq} \mathbf{F}_{*q} - \hat{\mathbf{C}}_{rq'} \mathbf{F}_{*q'} & \text{if } \Theta = \mathbf{M}_{*p} \end{cases} \quad (6)$$

$$\frac{\partial \hat{\mathbf{B}}_{pqq'}}{\partial \Theta} = \begin{cases} \mathbf{V}_{*q} - \mathbf{V}_{*q'} & \text{if } \Theta = \mathbf{U}_{*p} \\ \mathbf{U}_{*p} / -\mathbf{U}_{*p} & \text{if } \Theta = \mathbf{V}_{*q} / \mathbf{V}_{*q'} \\ \mathbf{F}_{*q} - \mathbf{F}_{*q'} & \text{if } \Theta = \mathbf{M}_{*p} \end{cases} \quad (7)$$

Equations (6) and (7) give the derivatives for $\Theta = \{\mathbf{U}, \mathbf{V}, \mathbf{M}\}$, and we can get the similar form for $\Theta = \{\mathbf{T}, \mathbf{W}, \mathbf{N}\}$.

We exploit the mini-batch gradient descent to maximize the objective function. For each iteration, all positive samples are enumerated (lines 3-12). We compute the gradients with a batch, including b positive samples (line 5) and $5b$ negative samples (lines 7-9) to construct $5b$ preference pairs, and update the parameters (line 11). To calculate the gradients (line 10), we combine Equations (5) with (6) and (7). Of special note is that $\frac{\partial \hat{\mathbf{A}}_{pqq'r}}{\partial \Theta}$ in Equation (6) is certain column of $\frac{\partial \hat{\mathbf{A}}_{pqq'r}}{\partial \Theta}$ in Equation (5), for example, the p -th column when $\Theta = \mathbf{U}_{*p}$.

Algorithm 1: Mini-batch gradient descent based algorithm.

Input: sparse tensor A , coupled matrices B and C , image features F , regularization coefficients λ_Θ , batch size b , learning rate η , maximum number of iterations $iter_max$, and convergence criteria.

Output: top- n prediction given by the complete tensor \hat{A} .

```

1 initialize  $\Theta$  randomly;
2  $iter = 0$ ;
3 while not converged  $\&\&$   $iter < iter\_max$  do
4    $iter+ = 1$ ;
5   split all purchase records into  $b$ -size batches;
6   for each batch do
7     for each record in current batch do
8       select 5 non-observed items  $q'$  randomly from
9          $Q \setminus (Q_p^+ \cup Q_r^+)$ ;
10      add these negative samples to the current batch;
11      calculate  $\nabla_\Theta BPR\_OPT$  with current batch;
12       $\Theta = \Theta + \eta \nabla_\Theta BPR\_OPT$ ;
13   calculate  $\hat{A}$  and predict the top- $n$  items;
14 return the top- $n$  items;

```

5 EXPERIMENT

In this section, we conduct experiments on real-world datasets to verify the feasibility of our proposed model. We then analyze the experiment results and demonstrate the precision promotion by comparing it with various baselines. We focus on answering the following three key research questions:

RQ1: How is the performance of our final framework for the clothing recommendation task?

RQ2: What are the advantages of the aesthetic features compared with conventional image features?

RQ3: Is it reasonable to transfer the knowledge gained from AVA, which is a dataset of photographic competition works, to the clothing aesthetics assessment task?

5.1 Experimental Setup

5.1.1 Datasets. We use the AVA dataset to train the aesthetic network and use the Amazon dataset to train the recommendation models.

- **Amazon clothing:** The Amazon dataset [15] is the consumption records from *Amazon.com*. In this paper, we use the *clothing shoes and jewelry* category filtered with 5-score (remove users and items with less than 5 purchase records) to train all recommendation models. There are 39,371 users, 23,022 items, and 278,677 records in total (after 2010). The sparsity of the dataset is 99.969%.
- **Aesthetic Visual Analysis (AVA):** We train the aesthetic network with the AVA dataset [31], which is the collection of images and meta-data derived from *DPCChallenge.com*. It contains over 250,000 images with aesthetic ratings from 1 to 10, 66 textual tags describing the semantics of images, and 14 photographic styles (complementary colors, duotones, high dynamic range, image grain, light on white, long exposure,

macro, motion blur, negative image, rule of thirds, shallow DOF, silhouettes, soft focus, and vanishing point).

5.1.2 Baselines. To demonstrate the effectiveness of our model, we adopt the following methods as baselines for performance comparison:

- **Random (RAND):** This baseline ranks items randomly for all users.
- **Most Popular (MP):** This baseline ranks items according to their popularity and is non-personalized.
- **MF:** This Matrix Factorization method ranks items according to the prediction provided by a singular value decomposition structure. It is the basis of many state-of-the-art recommendation approaches.
- **VBPR:** This is a stat-of-the-art visual-based recommendation method [15]. The image features are pre-generated from the product image using the Caffe deep learning framework.
- **CMTF:** This is a stat-of-the-art context-aware recommendation method [1]. The tensor factorization is jointly learned with several coupled matrices.

5.1.3 Experiment Settings. In the *Amazon* dataset, we remove the record before 2010 and discretize the time by weeks. There are 237 time intervals, the sparsity of the tensor is 99.99987%. We randomly split the dataset into training (80%), validation (10%), and test (10%) sets. The validation set was used for tuning hyper-parameters and the final performance comparison was conducted on the test set. We do the prediction and recommend the top- n items to consumers. The Recall and the normalized discounted cumulative gain (NDCG) are calculated to evaluate the performance of the baselines and our model. When n is fixed, the Precision is only determined by true positives whereas the Recall is determined by both true positives and positive samples. To give a more comprehensive evaluation, we exhibit the Recall rather than the Precision and F_1 -score (F_1 -score is almost determined by the Precision since the Precision is much smaller than the Recall in our experiments). Our experiments are conducted by predicting Top-5, 10, 20, 50, and 100 favorite clothing.

5.2 Performance of Our Model (RQ1)

We iterate 200 times to train all models (except RAND and MP). In each iteration, we enumerate all positive records to optimize models and select 1000 users in test (or validation) set to calculate evaluation metrics, then show the best performance every 10 iterations. Figure 4(a) shows the Recall and Figure 4(b) shows the NDCG during training. We set $n = 50$ when representing the Recall and $n = 5$ when representing the NDCG, due to the relatively large value respectively (represented in Figure 5). We can see that NDCG@5 shows a heavier fluctuation than Recall@50 (Figure 4 and Figure 7) since a smaller n leads to a more random prediction. Compared with MP, personalized methods show stronger ability to represent the preference of users and outperform MP several times. By recommending clothes that fit the current season, CMTF can outperform MF on both Recall and NDCG. Enhanced by side information, VBPR performs the best among all baselines. The proposed DCFA model outperforms VBPR about 8.53% on Recall@50 and 8.73% on NDCG@5.

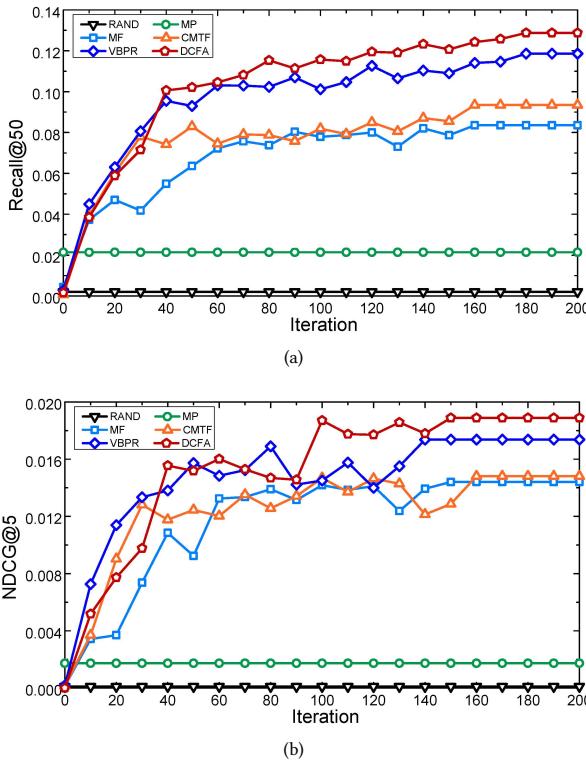
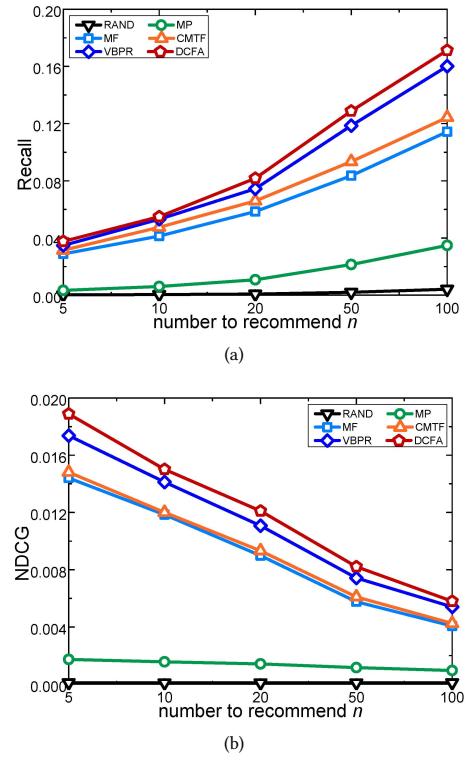
**Figure 4: Performance with training iterations (test set)**

Figure 5 represents the variation of the Recall and the NDCG with different n . In Figure 5(a), we can see that the Recall increases almost linearly with the increasing of n while in Figure 5(b), for most methods (except RAND), the NDCG decreases with the increasing of n . Since for most models (except RAND), the higher-rated clothing is with more possibility to be chosen by consumers. So the ordering quality decreases with the increasing of n . To the contrary, since RAND orders all items randomly, its ordering quality keeps constant.

In our experiments, we tune all hyperparameters sequentially on the validation set (include those in our model and in baselines). There are 8 hyperparameters in Equation (4) and the sensitivity analysis is shown in Figure 6. We can see that when $\lambda_1 = 0.1$, $\lambda_2 = 0.1$, $\lambda_3 = 0.3$, $\lambda_4 = 0.3$, $\lambda_5 = 0.5$, $\lambda_6 = 0.2$, $\lambda_7 = 0.5$, $\lambda_8 = 0.5$, DCFA can achieve the best performance. Influences of hyperparameters in baselines are also shown in Figure 6. For all models, λ_1 and λ_2 are used to represent weights of the coupled user-item matrix and time-item matrix. λ_3 to λ_8 are regularization coefficients of the user matrix, item matrix (connecting with user), time matrix, item matrix (connecting with time), aesthetic preference matrix of consumers, and aesthetic preference matrix of time respectively. For example, we can see that the performance of MF varies with regularization coefficients of the user matrix (λ_3) and the item matrix (connecting with user, λ_4), while keeps constant with the variation of λ_5 because there is no time matrix in MF. Specially, in CMTF, the item matrix connects both the user and time matrices, we use λ_3 , λ_4 , λ_5 to represent the regularization coefficients of the user, item, and time matrices respectively.

**Figure 5: Performance with different n (test set)**

5.3 Necessity of the aesthetic features (RQ2)

In this subsection, we discuss the necessity of the aesthetic features. We combine various widely used features to our basic model and compare the effect of each features by constructing five models:

- **DCF:** This is our basic **D**ynamic **C**ollaborative **F**iltering model without any image features, which is represented in the subsection 4.1.
- **DCFH:** This is a **D**ynamic **C**ollaborative **F**iltering model with **C**olor **H**istograms.
- **DCFCo:** This is a **D**ynamic **C**ollaborative **F**iltering model with **C**NN **F**eatures **o**nly.
- **DCFAo:** This is a **D**ynamic **C**ollaborative **F**iltering model with **A**esthetics **F**eatures **o**nly.
- **DCFA:** This is our proposed model represented in the subsection 4.2, utilizing both CNN features and aesthetic features.

Figures 7(a) and 7(b) show the distribution of 10 maximum on Recall@50 and the NDCG@5 of each model during the 200 iterations. As shown in Figure 7, DCF performs the worst since no image features are involved to provide the extra information. With the information of color distribution, DCFH performs better, though still worse than DCFCo and DCFAo, because the low-level features are too crude and unilateral, and can provide very limited information about consumers' aesthetic preference. DCFCo and DCFAo show the similar performance because both CNN features and aesthetic features have strong ability to mine the user's preference. Our DCFA model, capturing both semantic information and aesthetic information, performs the best on the *Amazon* dataset since those

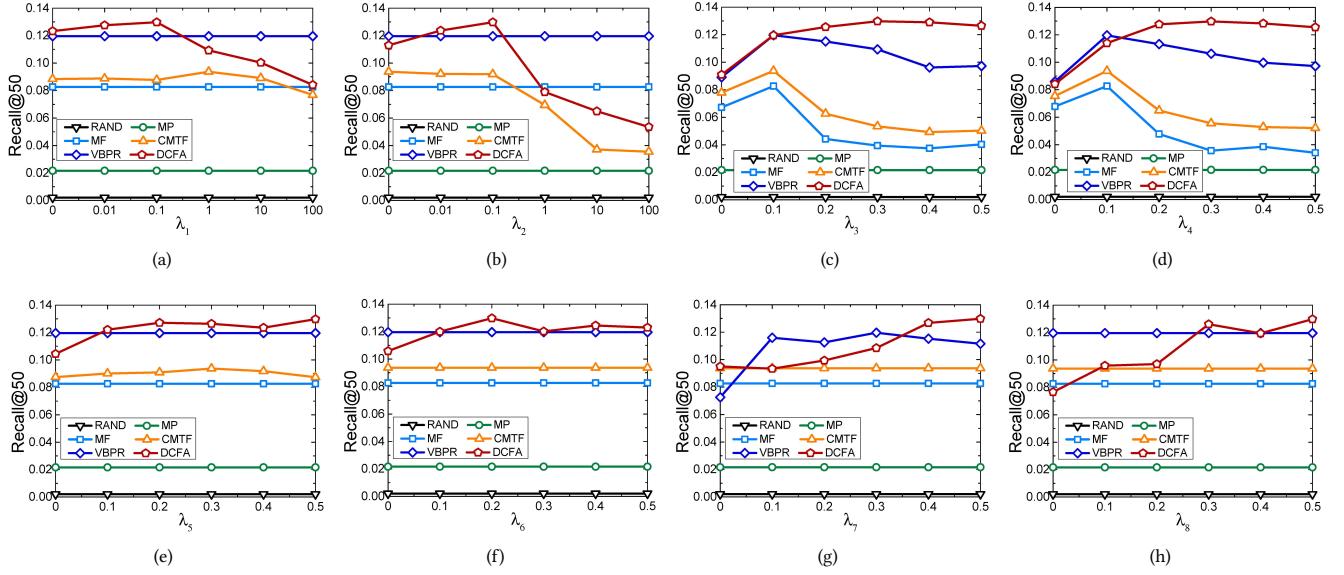


Figure 6: Impacts of hyperparameters (validation set)

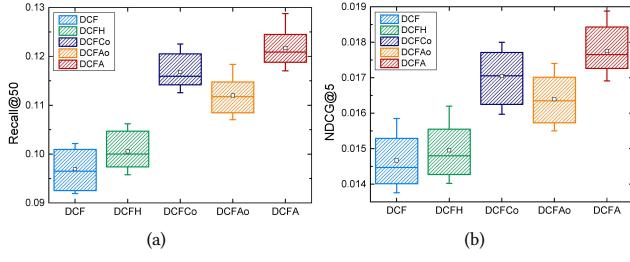


Figure 7: Performance of various features (test set)

two kinds of information mutually enhance each other to a certain extent. Give an intuitive example, if a consumer want to purchase a skirt, she needs to tell whether there is a skirt in the image (semantic information) when look through products, and then she needs to evaluate if the skirt is good-looking and fits her tastes (aesthetic information) to make the final decision. We can see that in the actual scene, semantic information and aesthetic information are both important for decision making and the two kinds of features complement each other in modeling this procedure. Though CNN features also contain some aesthetic information (like color, texture, etc.), it is far from a comprehensive description, which can be provided by the aesthetic features on account of the abundant raw aesthetic features inputted and training for aesthetic assessment tasks. Also, aesthetic features contain some holistic information (like structure and portion), while cannot provide a complete semantic description. So, these two kind of features cannot replace each other and are supposed to model users' preference collaboratively. In our experiments, DCFA outperforms DCFCo and DCFAo about 5.06% and 8.79% on Recall@50, 4.89% and 8.51% on NDCG@5 respectively. We can see that though the aesthetic features and CNN features do not perform the best separately, they mutually enhance each other and achieve improvement together.

Several purchased and recommended items are represented in the Figure 8. Items in the first row are purchased by certain consumer (training data, the number is random). To illustrate the effect of the aesthetic features intuitively, we choose the consumers with explicit style preference and single category of items. Items in the second row and third row are recommended by DCFCo and DCFA respectively. For these two rows, we choose five best items from the 50 recommendations to exhibit. Comparing the first and the second row, we can see that leveraging semantic information, DCFCo can recommend the congeneric (with the CNN features) and relevant (with tensor factorization) commodities. Though it can recommend the pertinent products, they are usually not in the same style with what the consumer has purchased. Capturing both aesthetic and semantic information, DCFA performs much better. We can see that items in the third row have more similar style with the training samples than items in the second row. Take Figure 8(f) as an example, we can see that what the consumer likes are vibrant watches for young men. However, watches in the second row are in pretty different styles, like digital watches for children, luxuriantly-decorated ones for ladies, old-fashioned ones for adults. Evidently, watches in the third row are in similar style with the train samples. They have similar color schemes and design elements, like the intricate-designed dials, nonmetallic watchbands, small dials, and tachymeters. It is also obvious in Figure 8(c), we can see that the consumer prefers boots, ankle boots or thigh boots. However, products recommended by DCFCo are some different type of women's shoes, like high heels, snow boots, thigh boots, and cotton slippers. Though there is a thigh boot, it is not in line with the consumer's aesthetics due to the gaudy patterns and stumpy proportion, which rarely appears in her choices. Products recommended by DCFA are better. First, almost all recommendations are boots. Then, thigh boots in the third row are in the same style with the training samples, like leather texture, slender proportions, simple design and



Figure 8: Items purchased by consumers and recommended by different models.

some design elements of detail like straps and buckles (the second and third ones). Though the last one seems a bit different with the training samples, it is in the uniform style with them intuitively, since they are all designed for young ladies. As we can see, with the aesthetic features and the CNN features complementing each other, DCFA performs much better than DCFCo.

5.4 Rationality of using the AVA dataset (RQ3)

The BDN is trained on the AVA dataset, which contains photographic works labeled with aesthetic ratings, textual tags, and photographic styles. We utilize aesthetic ratings and photographic styles to train the aesthetic network. In this subsection, we simply discuss if it is reasonable to estimate clothing by the features trained for photographic assessment.

With no doubt that there are many similarities between aesthetic photographs and well-designed clothing, like delightful color combinations, saturation, brightness, structures, proportion, etc. Of course, there are also many differences. To address this gap, we modify the BDN. In [42], there are 14 pathways to captures all photographic styles. In this paper, we remove several pathways for the photographic styles which contribute little in clothing estimation, like high dynamic range, long exposure, macro, motion blur, shallow DOF, and soft focus. These features mainly describe the camera parameters setting or photography skills but not the image, so they help little in our clothing aesthetic assessment task. Experiments show that our proposed model can uncover consumers' aesthetic preference and recommend the clothing that are in line with their aesthetics, and the performance is obviously promoted.

There are many works recommending clothing or garments with fashion information [20, 37, 39] and there are several datasets for clothing fashion style. [20] utilized three datasets containing street

fashion images and annotations by fashionistas to train phase, input queries, and return ranked list respectively. [39] proposed a novel dataset crawled from *chictopia.com* containing photographs, text in the form of descriptions, votes, and garment tags. However, these datasets are mainly for fashion style and not appropriate for BDN training because of the lack of aesthetic ratings and style tags, so we choose AVA. There are abundant images and tags to provide raw aesthetic features. Though not all raw features are needed due to the gap of photographic works and clothing, many of them are important in clothing aesthetic assessment. Beyond that, our model should have ability to extend to a wider range of application scenarios, like the recommendation of electronic products, movies, toys, etc., so a general dataset for aesthetic network training is important.

6 CONCLUSION

In this paper, we investigated the usefulness of aesthetic features for personalized recommendation on implicit feedback datasets. We proposed a novel model that incorporates aesthetic features into a tensor factorization model to capture the aesthetic preference of consumers at a particular time. Experiments on challenging real-word datasets show that our proposed method dramatically outperforms state-of-the-art models, and succeeds in recommending items that fit consumers' style.

For future work, we will establish a large dataset for product aesthetic assessment, and train the networks to extract the aesthetic information better. Moreover, we will investigate the effectiveness of our proposed method in the setting of explicit feedback. Lastly, we are interested in integrating the domain knowledge about aesthetic assessment, e.g., in the form of decision rules [40], into the recommender model.

REFERENCES

- [1] Evrin Acar, Tamara G. Kolda, and Daniel M. Dunlavy. 2011. All-at-once Optimization for Coupled Matrix and Tensor Factorizations. *Computing Research Repository - CORR* abs/1105.3422 (2011). arXiv:1105.3422
- [2] Evrin Acar, Tamara G Kolda, Daniel M Dunlavy, and Morten Mørup. 2010. Scalable Tensor Factorizations for Incomplete Data. *Chemometrics and Intelligent Laboratory Systems* 106, 1 (2010), 41–56.
- [3] Immanuel Bayer, Xiangnan He, Bhargav Kanagal, and Steffen Rendle. 2017. A Generic Coordinate Descent Framework for Learning from Implicit Feedback. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*. 1341–1350.
- [4] Yoshua Bengio. 2009. Learning Deep Architectures for AI. *Found. Trends Mach. Learn.* 2, 1 (Jan. 2009), 1–127.
- [5] Preeti Bhargava, Thomas Phan, Jiayu Zhou, and Juhan Lee. 2015. Who, What, When, and Where: Multi-Dimensional Collaborative Recommendations Using Tensor Factorization on Sparse User-Generated Data. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15)*. 130–140.
- [6] A. M. Buchanan and A. W. Fitzgibbon. 2005. Damped Newton algorithms for matrix factorization with missing data. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, Vol. 2. 316–322 vol. 2.
- [7] Da Cao, Liqiang Nie, Xiangnan He, Xiaochi Wei, Shunzhi Zhu, and Tat-Seng Chua. 2017. Embedding Factorization Models for Jointly Recommending Items and User Generated Lists. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17)*. 585–594.
- [8] Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat-Seng Chua. 2017. Attentive Collaborative Filtering: Multimedia Recommendation with Item- and Component-Level Attention. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17)*. 335–344.
- [9] Tao Chen, Xiangnan He, and Min-Yen Kan. 2016. Context-aware Image Tweet Modelling and Recommendation. In *Proceedings of the 2016 ACM on Multimedia Conference (MM '16)*. 1018–1027.
- [10] Xu Chen, Zheng Qin, Yongfeng Zhang, and Tao Xu. 2016. Learning to Rank Features for Recommendation over Multiple Categories. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '16)*. 305–314.
- [11] Xu Chen, Pengfei Wang, Zheng Qin, and Yongfeng Zhang. 2016. HLBPR: A Hybrid Local Bayesian Personal Ranking Method. In *Proceedings of the 25th International Conference Companion on World Wide Web (WWW '16 Companion)*. 21–22.
- [12] Xu Chen, Yongfeng Zhang, Qingyao Ai, Hongteng Xu, Junchi Yan, and Zheng Qin. 2017. Personalized Key Frame Recommendation. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17)*. 315–324.
- [13] Qiang Cui, Shu Wu, Qiang Liu, and Liang Wang. 2016. A Visual and Textual Recurrent Neural Network for Sequential Prediction. *arXiv preprint arXiv:1611.06668* (2016).
- [14] Ritendra Datta, Dhira Joshi, Jia Li, and James Z Wang. 2006. Studying Aesthetics in Photographic Images Using a Computational Approach. In *Proceedings of the 9th European Conference on Computer Vision (ECCV '06)*. 288–301.
- [15] Ruining He and Julian McAuley. 2016. VBPR: Visual Bayesian Personalized Ranking from Implicit Feedback. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI '16)*. 144–150.
- [16] Xiangnan He and Tat-Seng Chua. 2017. Neural Factorization Machines for Sparse Predictive Analytics. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17)*. 355–364.
- [17] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*. 173–182.
- [18] Xiangnan He, Hanwang Zhang, Min-Yen Kan, and Tat-Seng Chua. 2016. Fast Matrix Factorization for Online Recommendation with Implicit Feedback. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '16)*. 549–558.
- [19] Balázs Hidasi, Massimo Quadrana, Alexandros Karatzoglou, and Domonkos Tikk. 2016. Parallel Recurrent Neural Network Architectures for Feature-rich Session-based Recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys '16)*. 241–248.
- [20] Vignesh Jagadeesh, Robinson Piramuthu, Anurag Bhardwaj, Wei Di, and Neel Sundaresan. 2014. Large Scale Visual Recommendations from Street Fashion Images. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '14)*. 1925–1934.
- [21] Alexandros Karatzoglou, Xavier Amatriain, Linas Baltrunas, and Nuria Oliver. 2010. Multiverse Recommendation: N-dimensional Tensor Factorization for Context-aware Collaborative Filtering. In *Proceedings of the Fourth ACM Conference on Recommender Systems (RecSys '10)*. 79–86.
- [22] Yan Ke, Xiaoou Tang, and Feng Jing. 2006. The Design of High-Level Features for Photo Quality Assessment. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*, Vol. 1. 419–426.
- [23] Tamara G. Kolda and Brett W. Bader. 2009. Tensor Decompositions and Applications. *Siam Review* 51, 3 (2009), 455–500.
- [24] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix Factorization Techniques for Recommender Systems. *Computer* 42, 8 (2009), 30–37.
- [25] Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James Zijun Wang. 2014. RAPID: Rating Pictorial Aesthetics using Deep Learning. In *Proceedings of the ACM International Conference on Multimedia (MM '14)*. 457–466.
- [26] Wei Luo, Xiaogang Wang, and Xiaoou Tang. 2013. Content-Based Photo Quality Assessment. *IEEE Transactions on Multimedia* 15, 8 (Dec 2013), 1930–1943.
- [27] Shuang Ma, Jing Liu, and Chang Wen Chen. 2017. A-Lamp: Adaptive Layout-Aware Multi-patch Deep Convolutional Neural Network for Photo Aesthetic Assessment. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '17)*. 722–731.
- [28] Luca Marchesotti, Florent Perronnin, Diane Larlus, and Gabriela Csurka. 2011. Assessing the aesthetic quality of photographs using generic image descriptors. In *2011 International Conference on Computer Vision (ICCV '06)*. 1784–1791.
- [29] Benjamin M. Marlin. 2003. Modeling User Rating Profiles For Collaborative Filtering. In *International Conference on Neural Information Processing Systems (NIPS '03)*. 627–634.
- [30] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. Image-Based Recommendations on Styles and Substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '15)*. 43–52.
- [31] Naila Murray, Luca Marchesotti, and Florent Perronnin. 2012. AVA: A large-scale database for aesthetic visual analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '12)*. 2408–2415.
- [32] Dmitry Pavlov and David M. Pennock. 2002. A Maximum Entropy Approach to Collaborative Filtering in Dynamic, Sparse, High-Dimensional Domains. In *International Conference on Neural Information Processing Systems (NIPS '02)*. 1441–1448.
- [33] Steffen Rendle, Christoph Freudenthaler, Zeno Ganter, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Conference on Uncertainty in Artificial Intelligence (UAI '09)*. 452–461.
- [34] Steffen Rendle and Lars Schmidt-Thieme. 2010. Pairwise Interaction Tensor Factorization for Personalized Tag Recommendation. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining (WSDM '10)*. 81–90.
- [35] Ruslan Salakhutdinov and Andriy Mnih. 2007. Probabilistic Matrix Factorization. In *Proceedings of the 20th International Conference on Neural Information Processing Systems (NIPS'07)*. 1257–1264.
- [36] Katharina Schwarz, Patrick Wieschollek, and Hendrik P. A. Lensch. 2016. Will People Like Your Image? *CoRR* abs/1611.05203 (2016). arXiv:1611.05203
- [37] Dandan Sha, Daling Wang, Xiangmin Zhou, Shi Feng, Yifei Zhang, and Ge Yu. 2016. An Approach for Clothing Recommendation Based on Multiple Image Attributes. In *Web-Age Information Management: 17th International Conference (WAIM '16)*. 272–285.
- [38] Nicholas Sidiropoulos, Lieven De Lathauwer, Xiao Fu, Kejun Huang, Evangelos Papalexakis, and Christos Faloutsos. 2017. Tensor Decomposition for Signal Processing and Machine Learning. *IEEE Transactions on Signal Processing* 65, 13 (July 2017), 3551–3582.
- [39] Edgar Simoserra, Sanja Fidler, Francesc Morenónoguer, and Raquel Urtasun. 2015. Neuroaesthetics in fashion: Modeling the perception of fashionability. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '15)*. 869–877.
- [40] Xiang Wang, Xiangnan He, Fuli Feng, Liqiang Nie, and Tat-Seng Chua. 2018. TEM: Tree-enhanced Embedding Model for Explainable Recommendation. In *Proceedings of the 27th International Conference on World Wide Web (WWW '18)*.
- [41] Xiang Wang, Xiangnan He, Liqiang Nie, and Tat-Seng Chua. 2017. Item Silk Road: Recommending Items from Information Domains to Social Users. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17)*. 185–194.
- [42] Zhangyang Wang, Shiyu Chang, Florin Dolcos, Diane Beck, Ding Liu, and Thomas S. Huang. 2016. Brain-Inspired Deep Networks for Image Aesthetics Assessment. *Michigan Law Review* 52, 1 (2016), 123–128.
- [43] Liang Xiong, Xi Chen, Tzu Kuo Huang, Jeff G. Schneider, and Jaime G. Carbonell. 2010. Temporal Collaborative Filtering with Bayesian Probabilistic Tensor Factorization. In *Siam International Conference on Data Mining (SDM '10)*. 211–222.
- [44] Yongfeng Zhang, Min Zhang, Yiqun Liu, Shaoping Ma, and Shi Feng. 2013. Localized Matrix Factorization for Recommendation Based on Matrix Block Diagonal Forms. In *Proceedings of the 22Nd International Conference on World Wide Web (WWW '13)*. 1511–1520.
- [45] Lili Zhao, Zhongqi Lu, Sinnen Jialin Pan, and Qiang Yang. 2016. Matrix Factorization+ for Movie Recommendation. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI '16)*. 3945–3951.