DOI: 10.19884/j.1672-5220.202404017

Personal Style Guided Outfit Recommendation with Multi-Modal Fashion Compatibility Modeling

WANG Kexin[1], ZHANG Jie[2*], ZHANG Peng[2], SUN Kexin[3], ZHAN Jiamei[4], WEI Meng[4]
1. School of Computer Science and Technology, Donghua University, Shanghai 201620, China
2. Institute of Artificial Intelligence, Donghua University, Shanghai 201620, China
3. College of Mechanical Engineering, Donghua University, Shanghai 201620, China
4. College of Information Science and Technology, Donghua University, Shanghai 201620, China

Abstract: A personalized outfit recommendation has emerged as a hot research topic in the fashion domain. However, existing recommendations do not fully exploit user style preferences. Typically, users prefer particular styles such as casual and athletic styles, and consider attributes like color and texture when selecting outfits. To achieve personalized outfit recommendations in line with user style preferences, this paper proposes a personal style guided outfit recommendation with multi-modal fashion compatibility modeling, termed as PSGNet. Firstly, a style classifier is designed to categorize fashion images of various clothing types and attributes into distinct style categories. Secondly, a personal style prediction module extracts user style preferences by analyzing historical data. Then, to address the limitations of single-modal representations and enhance fashion compatibility, both fashion images and text data are leveraged to extract multi-modal features. Finally, PSGNet integrates these components through Bayesian personalized ranking (BPR) to unify the personal style and fashion compatibility, where the former is used as personal style features and guides the output of the personalized outfit recommendation tailored to the target user. Extensive experiments on large-scale datasets demonstrate that the proposed model is efficient on the personalized outfit recommendation.

Keywords: personalized outfit recommendation; fashion compatibility modeling; style preference; multi-modal representation; Bayesian personalized ranking (BPR); style classifier
CLC number: Z62 **Document code**: A
Article ID: 1672-5220(2025)02-0156-12

Open Science Identity (OSID)

0 Introduction

In recent years, there has been a growing interest in personalized outfit recommendations within the fashion industry[1-3]. With the advancement of the Internet technology, users can access extensive fashion information online, leading to an ever-expanding choice space. However, this abundance of information also poses challenges for users in filtering out their preferred fashion products. Outfit recommendations are initially generated based on fashion compatibility which quantifies the aesthetic coordination between clothes in terms of colors, styles, fabrics and other multi-modal attributes. A compatible outfit usually conforms to the aesthetic standards of most people. At the same time, personal preferences also influence the recommendation results. Consequently, numerous scholars have dedicated themselves to studying how to recommend outfit combinations that align with both current fashion trends and user style preferences. This trend arises from the increasing emphasis on the personal style and the necessity to tailor recommendation services accordingly. However, most existing fashion outfit recommendations still fall short of fully meeting these personal style needs. Thus, combining fashion styles with user style preferences remains an urgent problem within this field.

One crucial aspect that is often overlooked by current fashion outfit recommendations is users' distinct partiality towards specific styles. A style is a critical element in outfits, represents divergent aesthetic orientations and channels of self-expression, and influences clothing selection. User style preferences are deeply rooted in the individual taste and reflect the desired image projection. By neglecting the personal style, traditional recommendations have shortcomings in providing accurate suggestions that align with user style preferences. Some studies have defined clothing attributes as factors to quantify fashion styles by analyzing the fashion design process, demonstrating a meaningful relationship between clothing attributes and

Received date: 2024-04-30
Foundation items: Shanghai Frontier Science Research Center for Modern Textiles, Donghua University, China; Open Project of Henan Key Laboratory of Intelligent Manufacturing of Mechanical Equipment, Zhengzhou University of Light Industry, China (No. IM202303); National Key Research and Development Program of China (No. 2019YFB1706300)
* Correspondence should be addressed to ZHANG Jie, email: mezhangjie@dhu.edu.cn
Citation: WANG K X, ZHANG J, ZHANG P, et al. Personal style guided outfit recommendation with multi-modal fashion compatibility modeling [J]. *Journal of Donghua University (English Edition)*, 2025, 42(2): 156-167.

fashion styles[4-5]. Despite these advances, the personalization factor of fashion styles remains a challenge. The complexity of capturing user style preferences from a style perspective highlights the necessity of further exploration and improvement in the current fashion outfit recommendation field. Therefore, this study attempts to explore the role of the personal style in the fashion outfit recommendation.

Fashion compatibility is another essential aspect, and fashion representation serves as the foundation for it. Current studies primarily focus on considering either a single visual or textual modality when it comes to fashion representation, as well as methods that simultaneously take both modalities into account[6-7]. The importance of the interaction between these two modalities has been proven by numerous research studies[8-9]. However, existing multi-modal methods that linearly integrate visual and textual modalities, despite recognizing their complementarity, still struggle to reduce redundancy. These linear fusion methods limit the exploration of the diversity of fashion representation and may not fully capture the complex characteristics of fashion trends. Furthermore, the multi-modal information in the fashion representation can be exploited to understand the compatibility between individual fashion items and the whole. Learning fashion compatibility can pave the way for personalized outfit recommendations and enhance the overall consumer experience.

To overcome these challenges, a personal style guided network is proposed that can analyze multi-modal data such as clothing images and textual descriptions related to the outfits favored by users on social media platforms. On the one hand, the multi-modal representation of clothing is modeled to learn the outfit compatibility, that is, the degree of coordination between different items that constitute a set of clothing. On the other hand, the user style preference is extracted and guided to the personalized outfit recommendation.

Primary contributions of this study can be succinctly summarized as follows.

1) A comprehensive personal style guided network for personalized outfit recommendations has been constructed, in which a two-stage style extraction model is used to model the relationship among clothing attributes, outfit styles and personal styles.

2) Multi-modal compatibility of outfits, derived from images and texts, is leveraged to achieve an enhanced representation. Distinct from other studies, the integration of multi-modal features and style features into the Bayesian personalized ranking[10] (BPR) framework is employed to fulfill the guiding function of the personal style.

3) A range of experiments have confirmed improvement over established baselines and have validated the effectiveness of various components within the proposed model.

1 Related Works

This study focuses on the personalized outfit recommendation, a field that has been extensively studied[11-13]. These studies can be divided into two tasks: fashion compatibility modeling and personalized outfit matching. In this section, related studies on these two tasks are reviewed. The personal style in this study is highlighted in comparison to that of prior studies.

1.1 Fashion compatibility modeling

Existing studies on fashion compatibility modeling differ in how they evaluate the compatibility score between two or a set of items. One explicit way is to decompose outfit compatibility into pairwise interactions between items. However, these metric learning approaches[14-16] rely on pairwise comparisons between item features alone. In contrast, Cucurull et al.[17] defined the context as a construct that was acknowledged to be compatible with each of these items. They addressed the compatibility prediction problem by employing a graph neural network that learned to generate product embeddings based on their visual features and contextual information. To capture high-order relationships between items, most of the studies treat the outfit as a whole. Lin et al.[18] utilized multi-instance learning to generate correlation embeddings of individual items and acquired users' attention towards different matching items through an attention mechanism. Lu et al.[8] leveraged tensor decomposition of the discrete content and sampled from a set of underlying Bernoulli variables to map items and users into binary codes. Among the existing methods, only a few have studied the problem of personalized outfit recommendation. Moreover, most of these studies utilize either a single data mode with limited information[6] or only superficially combine visual and textual modalities at a low level. They fail to acquire high-level fashion representation knowledge through complementary interaction between the heterogeneous modes. Aligning these complex modes poses a significant challenge for complete fashion representation. Therefore, the proposed model investigates a non-linear multi-modal fusion method to represent fashion items with rich complementary information from images and texts.

1.2 Personalized outfit matching

Another task of personalized outfit recommendations is personalized outfit matching. Some studies have solved this problem from perspectives like multi-instance learning, attention mechanisms and tensor decomposition. Song et al.[7] modeled overall compatibility modeling and user style preference modeling in personalized outfit recommendations as two separate subtasks. They focused on separating the two subtasks and extracted user style preferences from users' historical

outfit data for personalization. Some studies have modeled attribute-based fashion outfit recommendations from three aspects: attribute representation, attribute explanation and attribute preference modeling. Feng et al.[19] partitioned embeddings of different attributes into different regions and adopted adversarial prediction networks to ensure independence between attribute embedding regions. This attribute-region embedding method can enhance the distinction between different attribute semantics. Sagar et al.[9] pre-trained a neural network model for attribute classification and identified harmonious and inharmonious attributes between fashion items to address interpretability in outfit recommendations. This attribute classification-based approach can provide explanations for outfit recommendations. Zhan et al.[20] constructed an attribute-aware fashion knowledge graph and designed a user-specific relational-aware attention mechanism to predict users' fine-grained preferences over different attributes for personalized outfit recommendations. Wang et al.[21] defined conditional preferences by dividing user-item interaction data into preference conditions and constructing the conditional weight branch to learn preference degrees. Ding et al.[22] proposed to integrate coordination knowledge into fashion and define category combinations as templates. Personalization was achieved through learning user style preferences for these templates. Although existing studies have explored personalized outfit recommendations from various perspectives, they overlook the personal style. To this end, the personal style is extracted from users' historical outfit data, and then it is used to guide the outfit recommendation.

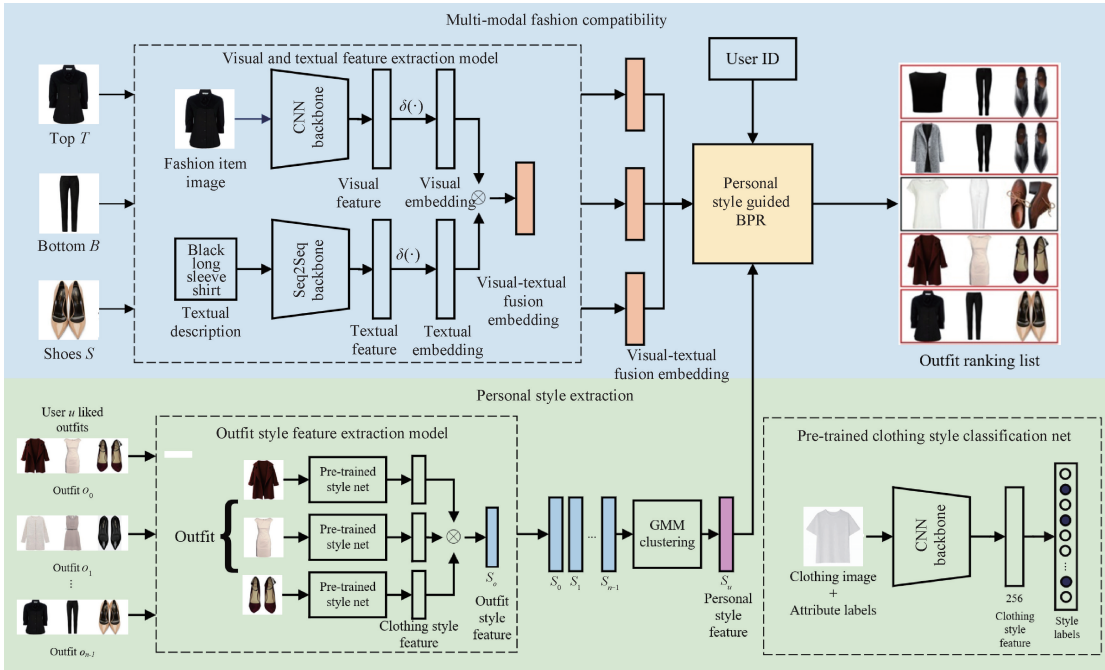
Style is defined as a durable and recognizable pattern of aesthetic choices[23]. De Divitiis et al.[24] trained a universal style classifier based on the color triplets to investigate how specific color combinations were associated with particular emotions and lifestyles. However, solely relying on color combinations presents subjectivity and makes it challenging to identify more complex styles. Sun et al.[25] obtained style-conditioned image patches by randomly cropping from apparel images and trained an image encoder in conjunction with textual descriptions to integrate style information with verbal representations. However, this approach is contingent on manual annotation. An et al.[26] constructed a hybrid style framework to clarify the classification criteria of fashion styles. Banerjee et al.[27] annotated clothing into eight specific styles. Moreover, the style of an outfit is impacted by various attributes, with factors like colors, textures and fabric patterns playing a significant role in determining its style categorization. For instance, soft color collocation represents a gentle style, slim design reveals elegance, and loose version and pockets belong to a casual style. Jeon et al.[4] redefined fashion attributes and proposed 25 new fashion styles, demonstrating the strong association between these styles and fashion attributes. This brings an important breakthrough to the field of fashion and provides us with new perspectives to understand and master fashion styles. Given these complexities surrounding the personalized outfit recommendation encompassing style preferences and attribute considerations, it becomes evident that how to capture user style preferences remains challenging for current outfit recommendations.

2 Proposed Methods

2.1 Problem formulation

Given sets of existing users $U = \{u_0, u_1, \cdots, u_{n_U-1}\}$, tops $T = \{t_0, t_1, \cdots, t_{n_T-1}\}$, bottoms $B = \{b_0, b_1, \cdots, b_{n_B-1}\}$ and shoes $S = \{s_0, s_1, \cdots, s_{n_S-1}\}$. For each user u, there is a set of historical outfits $O_u = \{o_0, o_1, \cdots, o_{n_O-1}\}$. Here, n_U, n_T, n_B, n_S and n_O represent the number of elements in sets U, T, B, S and O_u, respectively. Each outfit consists of a top, a bottom and a pair of shoes, denoted as $o = (t_p, b_q, s_r)$, where p, q and r represent indices of items in sets T, B and S, respectively.

The personal style guided network (PSGNet), illustrated in Fig. 1, consists of three main components: a multi-modal fashion compatibility module, a personal style extraction module and a personal style guided BPR framework. The multi-modal fashion compatibility module contains a visual and textual feature extraction model and fashion compatibility evaluation. Visual-textual fusion embeddings are extracted from fashion item images and textual descriptions by the visual and textual feature extraction model. The personal style extraction module is an item-outfit-user three-level structure. By merging item features and personal style features, the personal style guided BPR framework aims to optimize the final outfit preference ranking for users, using positive and negative sample pairs.



CNN—convolutional neural network; GMM—Gaussian mixture model; $\delta(\cdot)$—sigmoid activation function; Seq2Seq—sequence-to-sequence.

Fig. 1 Overview of PSGNet

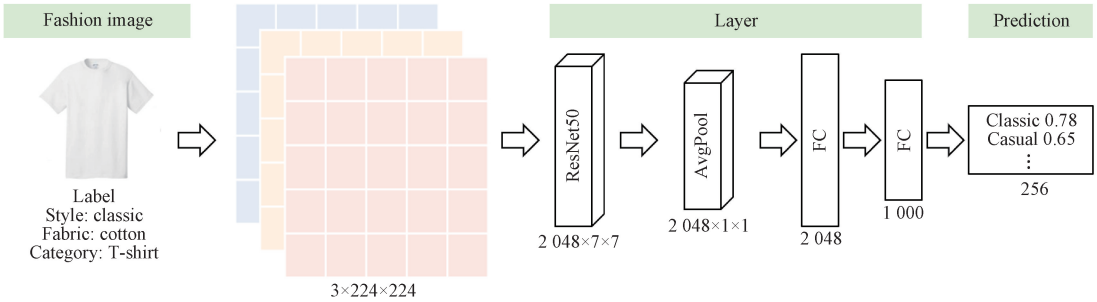
2.2 Personal style extraction

User style preferences refer to types, colors, fabrics and other aspects of clothing they like, which are often contained in users' historical outfit data. The personal style encapsulates a variety of preferences. Specifically, nine personal style categories are delineated across three dimensions: age, line and movement, as detailed in Table 1. By leveraging historical outfit data, it becomes possible to establish a comprehensive understanding of the relationship between users' unique fashion preferences and the personal style. However, the challenge lies in accurately modeling and characterizing the personal style. In response, a method that combines CNNs, specifically the 50-layer residual network[28] (ResNet50), with the GMM[29] clustering algorithm is proposed to tackle this issue. The method leverages the strengths of deep learning for feature extraction and the probabilistic nature of GMM for clustering.

The subtle relationship between fashion style representation and user style preferences is studied by establishing the relationship between clothing attributes and individual preferences. This process entails establishing connections between clothing attributes and individual preferences based on an examination of users' historical outfit data. To extract style features at the item level, CNNs that have been pre-trained on large-scale fashion attribute datasets DeepFashion[5] are employed. Such networks effectively encode subtle style characteristics, such as silhouettes, colors and texture, into vector representations. Specifically, the pre-trained clothing style classification net is used to extract the fashion style. It utilizes ResNet50 to extract fashion features from clothing images and attribute labels as shown in Fig. 2. For instance, the style features of a top t, denoted as $\boldsymbol{S}_t$, a bottom b, denoted as $\boldsymbol{S}_b$ and a pair of shoes s, denoted as $\boldsymbol{S}_s$, are ascertained.

Table 1 Categories and criteria of personal styles

Personal style	Age	Line	Movement
Lively	Youthful	Soft	Dynamic
Sweet	Youthful	Medium	Dynamic
Sporty	Youthful	Sharp	Dynamic
Elegant	Medium	Soft	Medium
Natural	Medium	Medium	Medium
Avant-garde	Medium	Sharp	Medium
Romantic	Mature	Soft	Static
Classic	Mature	Medium	Static
Metropolitan	Mature	Sharp	Static



AvgPool—average pooling; FC—full connection.

Fig. 2 Schematic diagram of pre-trained clothing style classification net

To capture the overall style of the outfit, an aggregation function is implemented to merge these item-level vectors into a unified vector representation, thus encapsulating the holistic style of the ensemble. This unified vector representation $\boldsymbol{S}_o$ serves as a bridge that connects clothing attributes with personalized preferences in a common embedding space. $\boldsymbol{S}_o$ can be expressed as

$$\boldsymbol{S}_o = \boldsymbol{S}_t \otimes \boldsymbol{S}_b \otimes \boldsymbol{S}_s. \tag{1}$$

To realize user-level style modeling, a hierarchical learning approach is pursued. It begins at the item level, progresses to the outfit level and culminates in the modeling of user style preferences. By concatenating outfit style vector representations for each user, clustering algorithms, for example, GMM cluster, can be harnessed to identify user groups exhibiting similar style preferences and patterns. The personal style analysis involves concatenating outfit-level style features for each user and utilizing the user ID as a form of supervision for clustering, as shown in Fig. 3. This process effectively groups users with comparable style preferences, and the cluster assignment for each user characterizes their personalized style signature. GMM cluster groups users with similar fashion styles. Based on the established user clusters, the approach extracts the personal style for each individual.

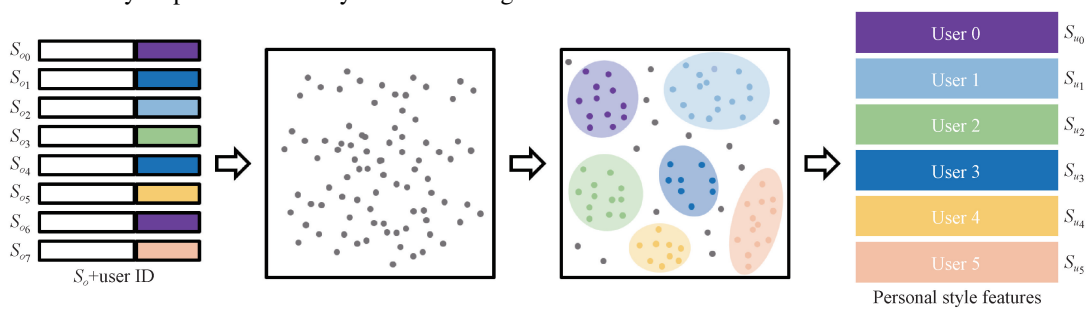


Fig. 3 Process of personal style feature clustering

Consequently, the personal style feature representation $\boldsymbol{S}_u$ for a given user u is aggregated by each outfit o in the users' historical outfit set O_u using adaptive pooling into a user-level style feature:

$$\boldsymbol{S}_u = g(\{\boldsymbol{S}_o \mid o \in O_u\}), \tag{2}$$

where $g(\cdot)$ denotes the GMM clustering mechanism being employed to derive user-level style features from outfit-level style representations. This hierarchical modeling approach allows styles to be learned from clothing attributes at the item level, abstracted to the outfit style and finally personalized to user style preferences. Consequently, a data-driven method emerges whereby the personal style of a user can be learned and represented, thereby facilitating the recommendation of complementary outfits tailored to each individual.

2.3 Multi-modal fashion compatibility

In the visual and textual feature extraction model, CNNs are recognized for their ability to learn visual representations, which are instrumental in extracting the deep features of images to depict the visual information of clothing. Seq2Seq models[30] are adopted in learning textual features, and then are utilized to extract textual features from the textual description of clothing and express semantic information. Using visual and textual embedded features rather than directly visual and textual features is more conducive to learning the potential correlation between the two modalities. Therefore, multilayer perceptron (MLP) is used to convert visual and textual features into latent embeddings. Here, in a given example, the upper clothing is denoted as top t and an MLP is employed to derive the latent visual embedding $\boldsymbol{v}_t$:

$$\boldsymbol{v}_t = \delta(\boldsymbol{\omega}_t \tilde{\boldsymbol{v}}_t + \boldsymbol{b}_t), \tag{3}$$

where $\boldsymbol{\omega}_t$ is the weight parameter; $\tilde{\boldsymbol{v}}_t$ is the visual feature

extracted by the CNN; $\boldsymbol{b}_t$ is the bias parameter. Similarly, latent textual embeddings $\boldsymbol{c}_t$ can be obtained. Visual-textual fusion embedding for the top t is denoted as $\boldsymbol{z}_t$:

$$\boldsymbol{z}_t = \boldsymbol{v}_t \otimes \boldsymbol{c}_t, \tag{4}$$

where $\boldsymbol{v}_t$ denotes the latent visual embeddings for the top t. Likewise, the latent visual, textual and visual-textual fusion embeddings for b and s are represented by $\boldsymbol{v}_b$ and $\boldsymbol{v}_s$, $\boldsymbol{c}_b$ and $\boldsymbol{c}_s$, $\boldsymbol{z}_b$ and $\boldsymbol{z}_s$, respectively.

Understanding the compatibility relationships between clothing items for effective collocation is essential for fashion recommendation systems. A method that distinguishes positive and negative outfit samples is proposed, enabling us to capture the essence of compatibility in clothing matching. The similarity and complementarity of visual, textual and visual-textual fusion embeddings are leveraged to design the pairwise compatibility scoring function:

$$\begin{cases} C_{tb} = \boldsymbol{v}_t^{\mathrm{T}}\boldsymbol{v}_b + \boldsymbol{c}_t^{\mathrm{T}}\boldsymbol{c}_b + \boldsymbol{z}_t^{\mathrm{T}}\boldsymbol{z}_b, \\ C_{ts} = \boldsymbol{v}_t^{\mathrm{T}}\boldsymbol{v}_s + \boldsymbol{c}_t^{\mathrm{T}}\boldsymbol{c}_s + \boldsymbol{z}_t^{\mathrm{T}}\boldsymbol{z}_s, \\ C_{bs} = \boldsymbol{v}_b^{\mathrm{T}}\boldsymbol{v}_s + \boldsymbol{c}_b^{\mathrm{T}}\boldsymbol{c}_s + \boldsymbol{z}_b^{\mathrm{T}}\boldsymbol{z}_s, \end{cases} \tag{5}$$

where C_{tb}, C_{ts} and C_{bs} represent the pairwise compatibility scores between t and b, t and s, and b and s, respectively. Equation (5) allows the proposed model to capture the intricate dynamics of clothing compatibility and refine the recommendations that it provides to users.

To determine the outfits with higher compatibility scores, the optimization goal C can be formulated:

$$C \equiv \{(t, b_i, b_j) \mid C_{tb_i} > C_{tb_j}\}, \tag{6}$$

where the subscripts i and j stand for positive samples and negative samples, respectively.

Pairwise compatibility scores among tops, bottoms and shoes are considered, along with the corresponding weights for each type, to calculate the overall compatibility score C_{tbs} of the outfit consisting of t, b and s:

$$C_{tbs} = \rho C_{tb} + \sigma C_{ts} + \tau C_{bs}, \tag{7}$$

where $\rho, \sigma, \tau \in [0, 1]$, and they are non-negative trade-off parameters that control the importance of different pairwise compatibilities for modeling the overall compatibility. By considering these factors, the proposed model can provide more accurate and refined compatibility assessments for users to seek outfit recommendations.

2.4 Personal style guided BPR

At this point, the multi-modal fashion compatibility module provides the fashion representation and learns the compatibility of different modalities, and the personal style extraction module captures the personal style. Naturally, Eqs. (2) and (7) are combined for the complete fashion representation. The performance score $F_{u,o}$ of o consisting of t, b and s for a user u is

$$F_{u,o} = \boldsymbol{S}_u \vartheta C_{tbs}, \tag{8}$$

where ϑ is a tensor with a fixed dimension.

Through positive and negative sample classification and multi-modal feature extraction, the proposed model can effectively learn the compatibility law of clothing matching to model the compatibility relationship. On this basis, the BPR framework has been modified to optimize personalized outfit recommendation. Given tuples (u, o_i, o_j), where o_i is the outfit that the user u prefers and o_j is the outfit that the user u does not prefer. Therefore, the optimization goal P is

$$P \equiv \{(u, o_i, o_j) \mid F_{u,o_i} > F_{u,o_j}\}. \tag{9}$$

A rank loss function $\mathcal{L}_{\mathrm{rank}}$ can be expressed as

$$\mathcal{L}_{\mathrm{rank}} = \sum_{(u,o_i,o_j)\in P} \log\{1 + \exp[-(F_{u,o_i} - F_{u,o_j})]\}. \tag{10}$$

By building the compatibility model and the personalized model, the proposed model can effectively leverage the advantages of both to recommend outfits that are personalized and compatible. The personalized sorting mechanism and compatibility constraint learning are the keys to realizing personalized and compatible clothing-matching recommendations.

3 Experiments

The performance of the proposed PSGNet is evaluated through a comparative analysis with state-of-the-art models. Additionally, ablation experiments are conducted to assess the individual contributions of each module.

3.1 Dataset and experiment settings

3.1.1 Outfit dataset

The Polyvore-630 dataset employed in these experiments is a large-scale fashion outfit dataset constructed by Lu et al.[8], and mines user data from the Polyvore website. This dataset comprises 150 380 outfits created by 630 users, with each outfit consisting of a top, a bottom and a pair of shoes. Unlike the original Polyvore dataset[31], the Polyvore-630 dataset incorporates abundant historical collocation information from users, making it amenable to personalized outfit recommendations. The dataset includes positive and negative samples for users. Negative samples are generated by replacing an item in a positive outfit with a randomly selected item from a different category. A ratio of one positive outfit to 10 negative outfits is maintained for each user. With its abundance of user preference cues, the Polyvore-630 dataset offers a realistic and comprehensive validation platform for personalized outfit recommendations. The statistics of this dataset are shown

in Table 2.

Table 2 Statistics of Polyvore-630 dataset

Polyvore-630	Number of outfits	Number of items
Training dataset	127 326	159 729
Test dataset	23 054	45 505

3.1.2 Experiment settings

In the experiments, AlexNet[32] is utilized as the default backbone network for image feature extraction. Following the approach presented in Ref. [8], in order to handle images of arbitrary sizes, the FC layers of AlexNet are replaced with convolutional layers, accompanied by average pooling layers that result in 4 096-dimensional feature vectors. For textual feature extraction from item descriptions, the Seq2Seq model is employed, and yields textual features of dimensionality-2 400. To train the style classifier, ResNet50 is used as the backbone network, augmented with an additional average pooling layer and two FC layers to extract style embeddings of dimensionality-256. The proposed model is implemented using the PyTorch framework. To enhance robustness, the models resample the negative samples in each epoch instead of fixing them during training. The experiments are performed on the tasks of personalized outfit recommendations, and test outfits are ranked for each user based on descending compatibility and personalized scores. The ranking performance is evaluated by using metrics such as the area under the receiver operating characteristic curve (AUC) and normalized discounted cumulative gain (NDCG).

3.2 Personalized outfit recommendation

The proposed model is compared with the following state-of-the-art models.

1) Visual BPR (VBPR)[6]: apart from capturing the latent factors of user-item interactions, VBPR specifically models the user style preference towards visual factors.

2) General compatibility and personal preference-BPR (GP-BPR)[7]: GP-BPR incorporates visual and textual modalities of fashion items and user-item data, utilizing both visual and textual information, to create a personalized compatibility model for clothing matching in a linear manner.

3) Personalized attribute-wise interpretable-BPR (PAI-BPR)[9]: PAI-BPR employs fashion attributes and user-item data to develop an interpretable personalized fashion recommendation scheme that considers attributes on an individual basis.

4) Fashion hash net (FHN)[8]: FHN proposes a discrete content-based tensor factorization model that maps items and users to binary codes for efficient fashion recommendations.

5) Multi-modal fashion compatibility and conditional preference model (MCCP)[21]: MCCP leverages multi-modal features to partition the user-item data into preference conditions and establishes a conditional preference model.

According to the experimental results shown in Table 3, PSGNet performs better than the baseline models in recommending outfits that conform to user style preferences and fashion compatibility. Compared with VBPR only using visual features, GP-BPR performs better, which proves the effectiveness of combining the visual and textual modalities. However, since GP-BPR can only linearly fuse visual and textual features, its performance is inferior to FHN, MCCP and PSGNet. Among these models, MCCP incorporating user conditional preferences performs the second best, highlighting the importance of modeling user style preferences for personalized outfit recommendations. The proposed PSGNet combines the multi-modal and personalized advantages of other models and focuses on learning personal styles, achieving the highest on both evaluation metrics.

Table 3 Results of personalized outfit recommendation performance comparison

Model	AUC	NDCG
VBPR	0.779 6	0.698 6
GB-BPR	0.838 8	0.727 1
PAI-BPR	0.827 8	0.756 6
FHN	0.846 5	0.764 8
MCCP	0.861 2	0.802 2
PSGNet	0.887 6	0.823 1

In Fig. 4, supplementary top-10 recommendation results are presented for a specific user, showcasing the performance of different models. Among them, PSGNet exhibits the best performance, as evidenced by its ability to recommend outfits that are consistent in the style while maintaining a high ranking of positive samples. While VBPR, GP-BPR, PAI-BPR and FHN are capable of recommending partially preferred outfits for users, they also tend to rank negative samples highly. The recommendation results of MCCP and PSGNet align more closely with user style preferences. However, it is worth noting that the recommended outfits by PSGNet exhibit a consistent personal style. These findings underscore the superiority and reliability of the proposed model in capturing and modeling the personal style of the user, thus highlighting its significant potential in the field of personalized outfit recommendations.

In summary, the experimental results validate that the proposed model can promote personalization and fashion compatibility in outfit recommendations by multi-modal fusion feature learning and personal style modeling.

Fig. 4 Top 10 recommendations for user 1: (a) VBPR; (b) GB-BPR; (c) PAI-BPR; (d) FHN; (e) MCCP; (f) PSGNet

3.3 Fashion compatibility matching

Previous experiments have demonstrated the performance of the proposed model on the personalized outfit recommendation task. The results on another task, fashion compatibility matching, are expanded. To demonstrate the ability of the proposed model in fashion compatibility matching, a fill-in-the-blank (FITB) fashion recommendation experiment was conducted by randomly selecting an item as a blank and setting three negative candidates for each clothing item within the test dataset. The categories of the negative candidates are the same as the ground-truth (GT). The accuracy of the proposed model is compared with the following baselines.

1) Bidirectional long short term memory (Bi-LSTM)[31]: Bi-LSTM treats a fashion outfit as a sequence conditioned on the previous items and sequentially predicts the next item to learn their compatibility relationship.

2) Compatibility scoring network (CSN)[33]: CSN learns image embeddings that respect item types, and jointly learns notions of item similarity and compatibility in an end-to-end model.

3) Multi-layered comparison network (MCN)[34]: MCN leverages feature mappings from diverse layers of CNN and global average pooling to construct representations in various aspects. Subsequently, it derives an overall compatibility score through pairwise similarity enumerations across different layers.

4) FHN: FHN utilizes type-dependent hashing modules to generate binary codes for outfits, while employing visual semantic embedding to ensure consistent representation across visual and textual modalities.

As presented in Table 4, superior performance of PSGNet over the baselines highlights the effectiveness of the proposed model in addressing the fashion compatibility matching task. Upon a thorough examination of Table 4, the following accurate results for the evaluation are observed.

1) PSGNet exhibits the highest accuracy in the FITB task. This notable improvement stems from the utilization of both visual and textual features to effectively represent fashion items, resulting in the enhanced accuracy in the compatibility modeling. Conversely, Bi-LSTM and CSN mainly rely on visual features to analyze the compatibility between fashion items, failing to exploit the advantage of integrating visual and textual modalities, thereby leading to subpar performance.

2) While MCN and FHN incorporate visual and textual modalities for modeling fashion compatibility, they fall short in deeply integrating multi-modal information and neglect the compatibility relationships between individual items. Consequently, MCN and FHN exhibit limited predictive performance regarding compatibility relative to PSGNet.

Table 4 Results of fashion compatibility matching performance comparison

Model	FITB accuracy
Bi-LSTM	0.462 7
CSN	0.583 0
MCN	0.641 5
FHN	0.658 3
PSGNet	0.675 9

In Fig. 5, a comparative analysis of individual models is presented by visually examining various test examples of the FITB task. The purpose of including different categories as blank options in these examples is to evaluate the degree to which the model's choice is influenced by the category. The GT option is A, while the negative candidates are options B, C and D. In example 1, it becomes apparent that Bi-LSTM incorrectly selects option B which resembles option A. This outcome suggests that Bi-LSTM is susceptible to visual similarities and disregards the crucial aspect of compatibility. Subsequently, in example 2, it can be observed that the erroneous selection of CSN could be primarily attributed to its susceptibility to color influences. This observation intimates that CSN might lack the necessary capacity to effectively capture color compatibility due to its limited expression of visual features. Example 3 demonstrates that MCN can successfully identify options compatible with the candidates. However, these choices do not align with the personal style. Example 4 exemplifies the capability of the proposed model to select options that align with the personal style rather than solely focusing on fashion compatibility. These examples underscore the capacity of the proposed model to holistically consider both fashion compatibility and personalized preferences.

Example	Query	Candidate A (GT)	Candidate B	Candidate C	Candidate D	Result
Example 1						Bi-LSTM: B CSN: A MCN: A FHN: A PSGNet: A
Example 2						Bi-LSTM: D CSN: B MCN: A FHN: A PSGNet: A
Example 3						Bi-LSTM: B CSN: C MCN: C FHN: A PSGNet: A
Example 4						Bi-LSTM: D CSN: C MCN: D FHN: D PSGNet: A

Fig. 5 Four examples of FITB query results

It can be affirmed that the proposed model exhibits effective performance in the FITB task, effectively circumventing the impact of diverse categories while simultaneously incorporating compatibility and user style preferences. Consequently, the findings further substantiate the efficacy of the proposed model in fashion compatibility modeling. To summarize, PSGNet achieves optimized fashion compatibility modeling by extensively integrating multi-modal representations and simultaneously addressing the compatibility relationships between fashion items.

3.4 Ablation study of PSGNet

In order to validate the effectiveness of each component in the proposed model, ablation experiments are conducted by involving the removal of specific modules. Based on PSGNet (the full model), various architecture variants have been developed by ablating components, including the personal style feature extraction module, visual feature extraction module and textual feature extraction module. The experiments have been performed on the Polyvore-630 dataset, with the accuracy, AUC and NDCG serving as the metrics.

To verify the performance improvements facilitated by the personal style guidance and multi-modal information for personalized outfit recommendations, four comparative experiments are designed.

1) PSGNet: training a multi-modal clothing matching model using personal style features to guide both visual and textual inputs.

2) PSGNet−style: remove personal style features. Calculate compatibility directly, without a personal style feature extraction module.

3) PSGNet−text: remove textual features. Train a recommendation model exclusively based on users' historical images and pre-trained style features, while disregarding textual features.

4) PSGNet−image: remove visual features. Train a recommendation model solely using users' historical

textual and personal style features without leveraging visual features.

Results of the ablation study are listed in Table 5. After removing the personal style feature extraction module, the overall performance metric of PSGNet-style decreases by about 0.03, indicating that the personal style is important for representing personalized user preferences. Furthermore, the results also demonstrate that the multi-modal module can substantially enhance the accuracy, AUC and NDCG metrics for personalized outfit recommendations, as compared to using only a single modality (either visual or textual modality). This verifies the complementary role of multi-modal information which can provide richer expressions of user style preferences than a unimodal model in personalized scenarios.

Table 5 Results of ablation experiments

Model	Accuracy	AUC	NDCG
PSGNet	0.881 3	0.887 6	0.823 1
PSGNet-style	0.853 4	0.854 3	0.796 4
PSGNet-text	0.865 6	0.861 8	0.802 5
PSGNet-image	0.832 8	0.819 4	0.768 0

Additionally, ablation studies quantify the contribution of each component. Style features encode personalized preferences, while the multi-modal module generates aesthetically compatible matchings. By incorporating these factors in a balanced manner, PSGNet exhibits strong representational power and potential for expansion. The modularized framework also allows convenient ablation of components to isolate their unique contributions.

In summary, the comparative experiments and ablation studies provide compelling quantitative evidence for the utility of the proposed model in improving personalized outfit recommendations.

4 Conclusions

In summary, the proposed model effectively overcomes the limitations of existing personalized outfit recommendations, making significant progress towards more personalized and accurate outfit recommendations. By modeling personal style features, fashion compatibility relationships and integrating multi-modal fashion item representations, these innovations benefit both individuals and fashion businesses. For individuals, it would improve satisfaction with tailored outfit recommendations; for fashion businesses, it would boost customer engagement, clicks and sales. Based on strong empirical results, integrating personal style guidance and multi-modal learning offers promising directions for further enhancing personalized outfit recommendations.

In subsequent research, more factors like target occasions, environmental conditions and style preferences will be explored to better understand user needs. Furthermore, the proposed model currently falls short in addressing personalized outfit recommendations for new users. To enhance the model's applicability, efforts will focus on solving the user cold-start problem, perhaps by incorporating objective factors and devising targeted strategies to ensure effective generalization.

References

[1] CHEN H J, SHUAI H H, CHENG W H. A survey of artificial intelligence in fashion[J]. *IEEE Signal Processing Magazine*, 2023, 40(3): 64-73.

[2] CHEN X, CHEN H, XU H, et al. Personalized fashion recommendation with visual explanations based on multi-modal attention network: towards visually explainable recommendation[C]// Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM, 2019: 765-774.

[3] CHENG W H, SONG S J, CHEN C Y, et al. Fashion meets computer vision: a survey[EB/OL]. (2021-01-28)[2024-04-01]. https://arxiv.org/abs/2003.13988.

[4] JEON Y, JIN S, HAN K. FANCY: human-centered, deep learning-based framework for fashion style analysis[C]//Proceedings of the Web Conference 2021. New York, USA: ACM, 2021: 2367-2378.

[5] LIU Z, LUO P, QIU S, et al. DeepFashion: powering robust clothes recognition and retrieval with rich annotations[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2016: 1096-1104.

[6] HE R N, MCAULEY J. VBPR: visual Bayesian personalized ranking from implicit feedback[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2016, 30(1): 144-150.

[7] SONG X M, HAN X J, LI Y K, et al. GP-BPR: personalized compatibility modeling for clothing matching[C]//Proceedings of the 27th ACM International Conference on Multimedia. New York, USA: ACM, 2019: 320-328.

[8] LU Z, HU Y, YU C, et al. Personalized fashion recommendation with discrete content-based tensor factorization[J]. *IEEE Transactions on Multimedia*, 2023, 25: 5053-5064.

[9] SAGAR D, GARG J, KANSAL P, et al. PAI-BPR: personalized outfit recommendation scheme with attribute-wise interpretability[C]//2020 IEEE Sixth International Conference on Multimedia Big Data. New York, USA: IEEE, 2020: 221-230.

[10] RENDLE S, FREUDENTHALER C, GANTNER Z, et al. BPR: Bayesian personalized ranking

from implicit feedback[C]//Proceedings of the International Conference on Uncertainty in Artificial Intelligence. [S. l.]: AUAI Press, 2009, 452-461.
[11] GONG W, KHALID L. Aesthetics, personalization and recommendation: a survey on deep learning in fashion[EB/OL]. (2021-01-20)[2024-04-01]. https://arxiv.org/abs/2101.08301v1.
[12] GU X L, GAO F, TAN M, et al. Fashion analysis and understanding with artificial intelligence [J]. *Information Processing & Management*, 2020, 57(5): 102276.
[13] DELDJOO Y, NAZARY F, RAMISA A, et al. A review of modern fashion recommender systems[EB/OL]. (2022-02-06)[2024-04-01]. https://arxiv.org/abs/2202.02757v3.
[14] HAN X T, WU Z X, JIANG Y G, et al. Learning fashion compatibility with bidirectional LSTMs[EB/OL]. (2017-07-08)[2024-04-01]. http://arxiv.org/abs/1707.05691.
[15] MCAULEY J, TARGETT C, SHI Q F, et al. Image-based recommendations on styles and substitutes[EB/OL]. (2015-06-15)[2024-04-01]. https://arxiv.org/abs/1506.04757v1.
[16] SONG X M, FENG F L, LIU J H, et al. NeuroStylist: neural compatibility modeling for clothing matching[C]//Proceedings of the 25th ACM International Conference on Multimedia. New York, USA: ACM, 2017: 753-761.
[17] CUCURULL G, TASLAKIAN P, VAZQUEZ D. Context-aware visual compatibility prediction[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2019: 12609-12618.
[18] LIN Y S, MOOSAEI M, YANG H. OutfitNet: fashion outfit recommendation with attention-based multiple instance learning [C]// Proceedings of the Web Conference 2020. New York, USA: ACM, 2020: 77-87.
[19] FENG Z L, YU Z Y, YANG Y Z, et al. Interpretable partitioned embedding for customized fashion outfit composition[EB/OL]. (2018-06-21)[2024-03-28]. https://arxiv.org/abs/1806.04845.
[20] ZHAN H J, LIN J, AK K E, et al. A^3-FKG: attentive attribute-aware fashion knowledge graph for outfit preference prediction [J]. *IEEE Transactions on Multimedia*, 2022, 24: 819-831.
[21] WANG Y Z, LIU L, FU X D, et al. MCCP: multi-modal fashion compatibility and conditional preference model for personalized clothing recommendation [J]. *Multimedia Tools and Applications*, 2024, 83(4): 9621-9645.
[22] DING Y J, MOK P Y, MA Y S, et al. Personalized fashion outfit generation with user coordination preference learning[J]. *Information Processing & Management*, 2023, 60(5): 103434.
[23] GODART F C. Why is style not in fashion? Using the concept of 'style' to understand the creative industries[M]//JONES C, MAORET M, eds. Research in the Sociology of Organizations. [S. l.]: Emerald Publishing Limited, 2018: 103-128.
[24] DE DIVITIIS L, BECATTINI F, BAECCHI C, et al. Style-based outfit recommendation[C]// 2021 International Conference on Content-Based Multimedia Indexing. New York, USA: IEEE, 2021: 1-4.
[25] SUN Z, ZHOU Y H, HE H H, et al. SGDiff: a style guided diffusion model for fashion synthesis [EB/OL]. (2023-08-15)[2023-11-14]. https://arxiv.org/abs/2308.07605v1.
[26] AN H, LEE K Y, CHOI Y, et al. Conceptual framework of hybrid style in fashion image datasets for machine learning[J]. *Fashion and Textiles*, 2023, 10(1): 18.
[27] BANERJEE D, DHAKAD L, MAHESHWARI H, et al. Recommendation of compatible outfits conditioned on style[EB/OL]. (2022-03-30)[2023-11-14]. https://arxiv.org/abs/2203.16161v1.
[28] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition [EB/OL]. (2015-12-10)[2023-11-14]. https://arxiv.org/abs/1512.03385v1.
[29] FRALEY C. How many clusters? Which clustering method? Answers via model-based cluster analysis [J]. *The Computer Journal*, 1998, 41(8): 578-588.
[30] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[EB/OL]. (2014-12-14)[2023-11-06]. https://arxiv.org/abs/1409.3215.
[31] HAN X, WU Z, JIANG Y G, et al. Learning fashion compatibility with bidirectional LSTMs[C]//Proceedings of the 25th ACM International Conference on Multimedia. New York, USA: ACM, 2017: 1078-1086.
[32] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks [J]. *Communications of the ACM*, 2017, 60(6): 84-90.
[33] VASILEVA M I, PLUMMER B A, DUSAD K, et al. Learning type-aware embeddings for fashion compatibility[EB/OL]. (2018-03-25)[2024-01-22]. https://arxiv.org/abs/1803.09196v2.
[34] WANG X, WU B, ZHONG Y. Outfit compatibility prediction and diagnosis with multi-layered comparison network[C]//Proceedings of the 27th ACM International Conference on Multimedia. Nice, France: ACM, 2019: 329-337.

结合多模态时尚兼容性建模和个人风格引导的服装搭配推荐

汪可欣[1]，张　洁[2*]，张　朋[2]，孙可芯[3]，占家美[4]，魏　濛[4]
1. 东华大学 计算机科学与技术学院，上海 201620
2. 东华大学 人工智能研究院，上海 201620
3. 东华大学 机械工程学院，上海 201620
4. 东华大学 信息科学与技术学院，上海 201620

摘　要：个性化服装搭配推荐已经成为时尚领域的一个研究热点。然而，现有的推荐方法尚未充分挖掘用户风格偏好。通常情况下，用户在选择服装时，不仅会倾向于特定风格，如休闲风格或运动风格，还会关注服装的颜色、质地等细节特征。为了推荐符合用户风格偏好的个性化服装搭配，该文提出了一种结合多模态时尚兼容性建模和个人风格引导的服装搭配推荐方法，简称 PSGNet。首先，设计一个风格分类器，将不同服装类型和属性的时尚图像划分到不同的风格类别中；其次，建立个人风格预测模块，通过分析历史数据提取用户风格偏好；再次，为了克服单模态表示的局限性并增强时尚兼容性，利用时尚图像和文本数据来同时提取服装的多模态特征；最后，通过贝叶斯个性化排序（Bayesian personalized ranking，BPR）算法来整合这些模块以统一个人风格和时尚兼容性，其中个人风格特征可引导输出推荐结果，为每位目标用户提供量身定制的个性化服装搭配推荐。在大规模数据集上进行广泛实验。结果表明，该文所提出的方法可有效推荐个性化服装搭配。

关键词：个性化服装搭配推荐；时尚兼容性建模；风格偏好；多模态表示；贝叶斯个性化排序；风格分类器