Expert Systems With Applications 203 (2022) 117333

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Neural stylist: Towards online styling service

Dongmei Mo [a,b], Xingxing Zou [a,b], WaiKeung Wong [a,b,*]

[a] Institute of Textiles and Clothing, The Hong Kong Polytechnic University, Hong Kong Special Administrative Region of China
[b] Laboratory for Artificial Intelligence in Design, Hong Kong Special Administrative Region of China

ARTICLE INFO

Keywords:
Fashion compatibility
Intelligent evaluation
Fashion recommendation
Neural networks

ABSTRACT

Online stylist service enjoys huge economic potentials due to the trend of transformation of the fashion industry to digitalisation. Existing works either predict the fashion compatibility from the overall aspect or evaluate the compatibility with type-conditional representations. The prediction is hard to interpret due to the abstractive forecast. This paper proposes a visual and semantic representation model for explainable evaluation and recommendation. The model considers fashion compatibility from different factors, such as colour, material and style, by leveraging low to high-level features from former to later layers of CNN. The colour correlation and the pairwise relationship of fashion items in the same outfit are considered during the prediction stage. Instead of just predicting an outfit as compatible or incompatible, the model can classify an outfit as three precise evaluation levels: Good, Normal and Bad. The detailed compatible level is more consistent with the fashion sense of our human brain as Good or Bad outfits may have specific characteristics while Normal outfits tend to be ordinary. Additionally, the model can diagnose and recommend substitutions of the problematic fashion items from overall compatibility or colour-specific aspects by tracking the prediction matrices' backpropagation gradients during the recommendation stage. Experiments in terms of outfit compatibility prediction and fill in the blank are conducted to evaluate the prediction ability of the proposed model. In contrast, fashion substitution recommendation experiments are conducted to assess the compatibility diagnosis and recommendation ability. Quantitative and qualitative results show that the model enables online stylist services with excellent explainability and generalisation on fashion prediction and recommendation.

1. Introduction

What to wear and how to match sets of fashion items to build a good look is ordinary in our daily life (Qasem, 2021). With the fashion industry's shift to digitalisation, online stylist service enjoys increasingly economic potentials (Longo, Padovano, Cimmino, & Pinto, 2021; Seo & Shin, 2019). It would be enjoyable and efficient to have a compatibility assistant that can help to evaluate the compatibility of a given outfit and recommend items to complete an attractive look (Yang & Huang, 2011).

To achieve this goal, this paper designs an intelligent compatibility model to provide online stylist service to improve customer experience. As shown in Fig. 1, the model can be used for the applications of colour-preferred fashion recommendations and outfit complements. A customer can first select the colour she/he likes, and the model will retrieve outfits that fit the requirement. Outfits with similar styles can be recommended and the composition can be revised by changing items with the same or different categories. Meanwhile, the model can help customers make efficient and wise decisions when they are browsing fashion items. As shown in the right of Fig. 1, given a fashion item, the model can recommend other items to complete an outfit. The items can be recommended from both online datasets or personal wardrobes. Customers can pick outfits with high compatibility scores and avoid the cost of buying unnecessary products by being noticed the compatible items that they already have in their wardrobe. The model can also recommend the best outfits to the customers from the shopping cart by evaluating and comparing the compatibility scores. The similarity among items is an important metric to evaluate the compatibility of a given outfit when developing machine learning algorithms (Kuang, Gao, Li, Luo, Chen, Lin, et al., 2019; Liu, Song, Nie, Gan, & Ma, 2019). When assessing the compatibility of an outfit, many factors will be considered. Visually, we analyse the compatibility from the factors like colour, material, design and style, etc.. We expect intelligent models to perceive aesthetics like the human brain and feed them with professional fashion knowledge during the training procedure.

* Corresponding author at: Institute of Textiles and Clothing, The Hong Kong Polytechnic University, Hong Kong Special Administrative Region of China. Laboratory for Artificial Intelligence in Design, Hong Kong Special Administrative Region of China.
E-mail addresses: dongmei.mo@connect.polyu.hk (D. Mo), aemika.zou@connect.polyu.hk (X. Zou), calvin.wong@polyu.edu.hk (W. Wong).

https://doi.org/10.1016/j.eswa.2022.117333
Received 28 January 2022; Received in revised form 12 April 2022; Accepted 25 April 2022
Available online 11 May 2022

Fig. 1. Illustration of application scenarios of the proposed model.

The knowledge like fashion description, category and compatibility label are annotated for each outfit on the existing datasets. The task of fashion compatibility evaluation and recommendation is to extract discriminant features from large amounts of fashion items and learn potential matching patterns from the outfit data. The convolutional neural network (CNN) (Albawi, Mohammed, & Al-Zawi, 2017) is suitable for dealing with such task due to its powerful feature extraction ability. CNN can automatically and effectively detect important features and achieve competitive detection and recognition performance. Although the training procedure of a CNN model usually requires hours to days as there are thousands of parameters to be optimised, the trained model is still efficient in the testing stage and it is practical for real applications on fashion evaluation platforms. There have been some CNN based techniques for solving the evaluation and recommendation problems in fashion area.

Metric learning is a common approach to evaluate the similarity/dissimilarity of any data pairs by applying a distance function to the representations in a high-dimensional space (Kolisnik, Hogan, & Zulkernine, 2021; Tan, Vasileva, Saenko, & Plummer, 2019). The similarity comparison is usually conducted under a unified embedding space where similar data are expected to be close while dissimilar data are widely separated. However, this is not the actual case when a white up item matches both of the shoes 1 and 2 in Fig. 2(a). Metric learning under the unified embedding space will force the two kinds of shoes to be close, but they belong to different styles and should not be of high similarity from our understanding. Thus, the similarity is not a natural way to evaluate the compatibility. Instead, it is a problem telling the model to compare different fashion factors. To this end, several works have been proposed to learn the similarity among fashion items under other conditions (such as fashion category) (Vasileva, Plummer, Dusad, Rajpal, Kumar, & Forsyth, 2018; Veit, Belongie, & Karaletsos, 2017). This kind of method can help to simplify similarity relationships of items from different contexts by learning and comprising the similarity conditioned on only one embedding space at a time. The comparison under type-conditional area thus avoids the problem of forcing incompatible fashion items close with metric learning (Vasileva et al., 2018). The conditioned space learning relies on the labels of the data, such as category or attribute. Thus, it is hard to generalise such learning to unseen cases. For learning the similarity of different contexts, Tan et al. proposed a similarity condition embedding network to learn multiple similarity conditions from the unified embedding space by treating the similarity conditions as latent variables and optimising the problem in a weakly supervised manner (Tan et al., 2019). The conditioned similarity learning methods try to explore the compatibility evaluation, but they fail to explain the reason why an outfit is compatible and how to improve its compatibility (Yang, Song, Feng, Wen, Duan, & Nie, 2021). It is necessary and interesting to explain the compatibility of different fashion factors without using rich fashion information.

The fashion factors of colour, print, material, silhouette and design details are the principal aspects for designers to evaluate the compatibility of a given outfit. As shown in Fig. 2(a), which pair of shoes that best match the given fashion items (i.e. white up, black–white patterned bottom and black crossbody bag)? The first option is leopard-print, peep-toe heels, and the second is black stilettos. Visually, the second is more compatible with the outfit as the factors of colour and print are more harmonious. The second option with solid black is close to the black–white bottom and the black bag, while the first option with leopard-print in three colours of black, brown and off whites makes the outfit dazzling. Additionally, the style of the first option is different from that of the given items. Thus, the second option is recommended due to its compatibility with respecting to the colour correlation, print combination, and style fusion factors. To evaluate and explain the compatibility of outfits from a fashion factor aspect, Zou et al. proposed a compatibility evaluation and reasoning method to include this kind of factor into the design of the network and constructed the Evaluation3 dataset for training and evaluating the performance of the method (Zou, Li, Bai, Lin, & Wong, 2020). The method provides judgement in three compatibility levels: Good, Normal and Bad, and explains the evaluation results with the defined factors of colour and print. Since the method learns explanations based on the training on the dataset with predefined fashion factors, it is hard to generalise to other datasets that have no such annotated factors. Chen et al. proposed a model based on attentive neural networks to capture the discriminative region features, especially in terms of hue, texture and colour, with the

Fig. 2. Evaluation and recommendation samples and analysis.

supervision of both implicit feedback and textual reviews to explain the recommendation (Chen, Zhang, Xu, Cao, Qin, & Zha, 2018b). Hou et al. suggested explaining fashion recommendations with intuitive visual attribute's guidance. Specifically, it first learns a fine-grained interpretable semantic space and then projects users and fashion items to this space to understand users' semantic preferences. Instead of just designing a visual feature for recommendation explanation, Lin et al. considered comments of fashion items and proposed to improve the performance with joint outfit matching learning and comment generation (Lin, Ren, Chen, Ren, Ma, & de Rijke, 2018). Yang et al. proposed to evaluate the outfit compatibility by matching patterns between fashion attributes and developed an explainable solution based on attribute aspect (Yang et al., 2021). Kim et al. also tried to explain the compatibility with visual attribute representations in a self-supervised learning manner. The self-supervised method considers three tasks corresponding to three concepts: colour histograms, discrimination of shapeless local patches and textures of every fashion item (Kim, Saito, Mishra, Sclaroff, Saenko, & Plummer, 2021). Revanur et al. proposed a semi-supervised learning method to help to create large-scale pseudo positive and negative outfits by leveraging unlabelled fashion corpus (Revanur, Kumar, & Sharma, 2021). There are other related methods for solving the outfit compatibility problems with different schemes (Duggal, Zhou, Yang, Xiong, Xia, Tu, et al., 2021; Su, Song, Zheng, Guan, Li, & Nie, 2021; Wang, Cheng, Wang, & Liu, 2021; Zhan, Lin, Ak, Shi, Duan, & Kot, 2021). All these methods tend to explain the recommendation with vague visual overall/region features, or recommend fashion items without considering the relationships among different items within an outfit.

The task of outfit compatibility evaluation is based on several fashion items, and outfit recommendation is expected to complete or even improve the look based on given one or multiple items (Li, Chen, & Huang, 2021; Lu, Hu, Chen, & Zeng, 2021). The intuitive factors like colour, material and style are potential explanations for the intelligent recommendation (Yu, Hui, & Choi, 2012). Outfit diagnosis is a task that can help to improve compatibility by learning the incompatible factors of a given outfit. Wang et al. proposed conducting compatibility evaluation with a multi-layered comparison network (MCN) and implementing outfit diagnosis by tracking the most problematic items with large backpropagation gradients of the comparison matrix (Wang, Wu, & Zhong, 2019). The recommendation is implemented by substituting the unsuitable items with the most compatible items on the dataset. MCN assumes the former layers in the deep neural network tend to learn low-level features like the colour and texture of images. In contrast, the latter tends to learn more abstracting features like style, and thus, it tries to explain the outfit compatibility with those features in an undefined manner. MCN has been verified effective for evaluating the compatibility of given outfits and providing reasonable explanations of undefined factors like colour, texture and style. However, MCN does not consider the pairwise relationship of colour correlation among different fashion items in the same outfit, which is also potentially crucial for meaningful interpretation. Additionally, MCN and most existing methods evaluate an outfit as compatible or incompatible, ignoring the detailed fashion compatibility levels. The compatibility can be classified into more complex levels: Good, Normal and Bad. It is consistent with the fact that when an outfit has some unique/outstanding features, we consider it good, but most outfits tend to be normal without attractive designs. Bad outfits have inharmonious factors, like colour correlation, patterns or strange designs.

This paper proposes evaluating outfit compatibility and recommending fashion substitutions more generally by exploring potential fashion factors in a unified framework. The proposed model does not assess the compatibility from predefined factors and instead tries to interpret the evaluation from assumed aspects. As shown in Fig. 2(b), the model can provide automatic diagnosis and recommendation service by predicting the compatibility level and interpreting the potential problematic factors of outfits that are captured or uploaded by customers. Substitutions can be recommended based on the predicted potential factors to improve the compatibility level of the outfits. The model aims to provide a general solution for outfit evaluation and recommendation, and it can be easily applied to most of the existing datasets without the requirement of rich annotated fashion factors. Overall, the contributions of this paper can be concluded as:

(1) Different form the existing methods, the proposed method considers the outfit evaluation task in a more generalised and reasonable way. The generalisation results from no requirement of pre-defined factors for the large amounts of fashion images. The method can instead learn the potential compatibility factors with the interpretation from the multi-layered features. The colour-specific compatibility and overall visual-semantic compatibility are jointly learned and the fashion factors are sensed in a way that is more similar to human aesthetics. Thus, the interpretation for the evaluation result is more reasonable.

(2) The diagnosis and recommendation function of the proposed method is different from the existing methods as the compatibility can be improved from Normal/Bad to Good level via recommending more compatible substitutions by comparing the importance of pairwise items with backpropagation gradients.

(3) Experiments on three large-scale outfit datasets are conducted and extensive qualitative and quantitative results demonstrate the effectiveness of the proposed method.

2. Related works

In this section, related works that focus on the problem of outfit compatibility learning and explainable recommendation are introduced.

The similarity of image pairs is usually measured by metric learning, which typically projects two objects to a general embedding space and obtains the distances as the measure of the similarity with a distance function (Chopra, Hadsell, & LeCun, 2005; Hadsell, Chopra, & LeCun, 2006; Wang, Song, Leung, Rosenberg, Wang, Philbin, et al., 2014). Since the distance of any two objects is computed under a shared space, the difference of the objects respecting different contexts is neglected. To overcome this shortcoming, methods that consider the similarity with disentangled representations under other conditions have been proposed (Vasileva et al., 2018; Veit et al., 2017). Vasileva et al. proposed to evaluate the similarity of data pairs by learning different embedding spaces respecting to different type combinations (Vasileva et al., 2018). There are two main drawbacks in these methods for outfit compatibility: (1) the methods measure the compatibility by taking the average of all pairwise similarities. The effect of pairwise similarity for overall compatibility is not considered. Pairwise similarity can be significant for finding the problematic fashion items and providing wise substitutions in the diagnosis and recommendation process; (2) the methods do not offer a concrete judgement of compatibility levels for an outfit, and the compatibility interpretation is not provided.

To deal with the explainability problem, Zou et al. proposed an explainable evaluation network based on Grad-CAM (Selvaraju, Cogswell, Das, Vedantam, Parikh, & Batra, 2017). It first extracts factor-aware features named intra-factor compatibility features with an independent net. All the outfit features are concatenated and fed into an inter-factor compatibility net, and the compatibility judgement in terms of three levels: Good, Normal, Bad, is obtained. The reasons for the judgement are diagnosed by the backpropagation gradients based on the previous concatenation of intra-factor compatibility features (Zou et al., 2020). The explanation is forced to align with annotated reasons in terms of factors of colour and print, which mimics the analysing process of fashion experts. The method focuses on providing reasons for the compatibility judgement from a factor aspect. The recommendation about improving the compatibility is not considered since compatibility tends to be evaluated from a factor aspect. The problematic items in the outfit are hard to be located.

There have been some methods proposed for explainable outfit recommendations. Yang et al. proposed to take advantage of the rich attributes associated with fashion items for interpretable fashion matching. It can predict the compatibility score and provide interpretable patterns of the good matching (Yang, He, Wang, Ma, Feng, Wang, et al., 2019). Feng et al. proposed a partitioned embedding network that uses an attribute partition module to learn attribute embedding within the overall embeddings respecting different parts and applies an adversarial partition module to achieve the independence of other parts. The compatibility is explainable by constructing an attribute matching map, while the outfit compositions can be customised by creating an outfit composition graph (Feng, Yu, Yang, Jing, Jiang, & Song, 2018). Lin et al. proposed a neural outfit recommendation network to provide outfit recommendations with generated abstractive comments by taking visual features, and user comments of fashion items (Lin et al., 2018). First, the method extracts visual features from a convolutional neural network and obtains matching predictions with a rating score. Then, the visible components are transformed into a concise sentence by designing a gated recurrent neural network. The generated comments are then used to explain the recommendations. Most of the existing methods cannot diagnose the problem of the outfit regarding the items, and thus the recommendation cannot guarantee the minor revision with the most considerable compatibility improvement. Wang et al. proposed a multi-layered comparison network (MCN) to take pairwise visual features from former to later layers to construct different comparison matrices for compatibility prediction. The network can diagnose the most problematic fashion items by using backpropagation gradients related to the comparison matrices to approximate the importance of each similarity of pairwise items for the compatibility of the whole outfit (Wang et al., 2019).

The proposed method in this paper also uses multi-layered features to compute the similarity of pairwise fashion items for compatibility prediction. The difference between the proposed method and MCN lies in three folds: first, the proposed method takes the colour correlations of the outfits to enhance the compatibility performance; second, the proposed method first applies Transformer module to learn the relationship among different items in the outfit and then computes the pairwise similarity to predict the compatibility into three levels: Good, Normal and Bad, while MCN directly applies the multi-layered features to compute the pairwise similarity and predict the outfit as compatible or incompatible; third, the proposed method can not only diagnose the problematic items for an incompatible outfit but also recommend items to improve the outfit from low to high levels respecting to colour-specific or overall compatibility.

3. Methodology

Given an outfit with p fashion items with different categories, such as top, bottom, shoes, bag and accessories, without loss of generality, we denote the outfit as $X = [x_1, x_2, \ldots, x_p] \in R^{p \times d}$, where d is the dimensionality of the fashion images. Most existing works evaluate an outfit with a predicted compatibility score and classify it into two levels: *compatible* and *incompatible*. Differently, we develop an end-to-end framework, called a multi-layered features and colour correlation enhancement network (MLCC), to implement the evaluation and recommendation function. MLCC classifies the outfit into three levels: *Good*, *Normal* and *Bad* from overall compatibility and colour-specific aspect. For the outfit which is classified as *Normal* or *Bad*, the model can diagnose the most problematic items and recommend substitutions from overall or colour-specific aspects to improve the outfit compatibility. MLCC comprises multi-layered feature analysis, colour correlation enhancement, and visual-semantic similarity preservation. The architecture of MLCC is illustrated in Fig. 3. More details of each component are described in the following sections.

3.1. Multi-layered feature analysis

Given an outfit, we human brains usually perceive the compatibility from multiple factors, such as fashion colour, print, material, silhouette and design details (Zou et al., 2020). Thus, we expect intelligent evaluation models to perceive fashion aesthetics in the similar way.

When training a model to mimic how our brain evaluates an outfit, two main problems need to be considered. On the one hand, large amounts of annotated data are required for the training of the model. However, most existing outfit datasets are not labelled with rich information, such as compatibility in terms of colour, print, material, etc. Therefore, training a model to perceive such fashion descriptions is not practical. Even if we have this kind of dataset for model training, it is not guaranteed that the model can learn how we perceive aesthetics as some unknown factors can affect our judgement on an outfit. Therefore, predefinition of certain aspects of compatibility is not the only way that can be used to train a model to perceive fashion aesthetics.

Deep neural network with the powerful ability of feature extraction can be used to learn representations of fashion items from low to high levels (Feng, Yu, Jing, Wu, Song, Yang, et al., 2019; Vadood & Haji, 2022). The network contains multiple layers and can perceive larger fields with layers that go deeper. The former layers tend to capture low-level features such as colour, texture whilst the later layers tend to learn high-level features such as fashion style and overall compatibility (Wang et al., 2019). Based on this regard, multi-layered

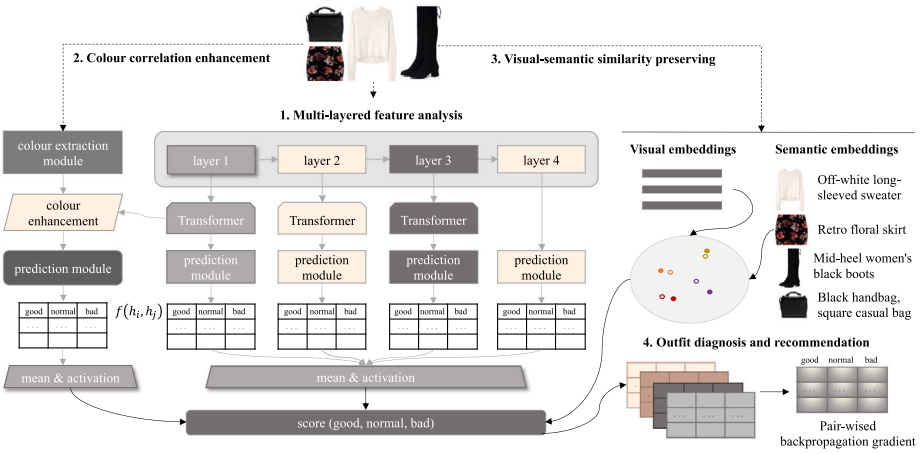


Fig. 3. The architecture of the proposed framework. The framework comprises four components: 1. multi-layer feature analysis; 2. colour-correlation enhancement; 3. visual-semantic similarity preserving, and 4: outfit diagnosis and recommendation. Three compatibility levels (Good, Normal, Bad) are learned from the enumerated pairwise similarity at different layers. The backpropagation gradient is used to approximate the contribution of each item to outfit compatibility. The outfit with levels of Normal or Bad can be improved to the level of Good by recommending more compatible substitutes. The diagnosis and recommendation process is interpretable by tracking the gradient matrix.

representations are considered in the proposed model, and the details of this component are shown in the middle part of Fig. 3. The feature learning layers 1–4 is constructed based on ResNet50 and the construction of each layer is similar to that in Wang et al. (2019). The feature maps obtained from layers 1–3 are reshaped to vectors and fed into Transformer module to learn the pairwise relationship of the same outfit. The features from layer 4 are not fed to Transformer module as the final high-level representations are assumed to be effective and should be directly fed to the prediction module for compatibility prediction.

For the Transformer module, the weight of embedding h_i with respect to h_j in the same outfit is learned through *self-attention*. The attention weight $a_{i,j}$ is defined as

$$a_{i,j} = softmax((W^q h_i)^T (W^k h_j)/\sqrt{d}), \tag{1}$$

$$\bar{h}_i = \sum_{j=1}^{p} a_{i,j} W^v h_j, \tag{2}$$

$$\tilde{h}_i = ReLU(\bar{h}_i W^r + b_1) W^o + b_2, \tag{3}$$

where W^k, W^q, W^v are key, query and value matrices, W^r and W^o are transformation matrices, and b_1, b_2 are bias vectors. The Transformer encoder contains 4 heads and 3 layers, and the weight matrices W^k, W^q, W^v, W^r and W^o at each layer are not shared. The rules of Eq. (1) to Eq. (3) are repeated with updated embeddings $\tilde{h}_i$ fed to the next layer.

As reported by previous works (Chen, Yin, Wang, Wang, Nguyen, & Li, 2018a; Vasileva et al., 2018), evaluating the similarity of fashion items from different types in a shared space will lead to undesired problems. For example, all bottom items compatible with top items will be forced to be close in the shared space. An item like a white T-shirt can be compatible with many kinds of bottoms or bags, such as colourful skirts or bags. However, those skirts and bags can be incompatible, and they should not be forced to be close to each other in the shared space. To avoid this problem, a solution is to define specific space for fair comparison for these items from different types.

In this module, the embeddings after Transformers are projected to the type-aware spaces to obtain masked embeddings by

$$\tilde{h}^{(i->i,j)} = ReLU(\tilde{h}_i \bigotimes m_{i,j}), \tag{4}$$

where $m_{i,j}$ is a learnable mask vector, $\bigotimes$ is element-wise product operation. $m_{i,j}$ provides element-wise gating function and the type-combination relevant elements are selected to feed to the successive prediction modules. The masks are expected to be sparse to select the most relevant features, whilst the masked embeddings should be balanced. To achieve this goal, L_1 and L_2-norms are used to regularise the variables of L_{mask} and L_{emb}:

$$L_{mask} = \|m\|_1, \tag{5}$$

$$L_{emb} = \|x\|_2. \tag{6}$$

The prediction modules of layers 1–4 are independent and work as the following manner. The embedding pairs $(\tilde{h}_i, \tilde{h}_j) \in H = (\tilde{h}_1, \tilde{h}_2, \ldots, \tilde{h}_p)$ from each layer are respectively concatenated and fed to a fully connected layer. The predicted scores $(s_{i,j})$ in terms of three levels (Good, Normal, Bad) are obtained by the linear transformation function, i.e. $s_{i,j} = f(\tilde{h}_i, \tilde{h}_j)$. The prediction matrices $M^l \in R^{p \times p,3}$ are shown as tables in Fig. 3, where l denotes the lth layer. The average prediction matrix is computed based on the four matrices and fed to the activation function to obtain the compatibility scores with respect to three classes of Good, Normal and Bad. The prediction loss is computed with Cross Entropy loss function (De Boer, Kroese, Mannor, & Rubinstein, 2005) as below:

$$L_{tf} = -\frac{1}{N} \sum_{i}^{N} \sum_{c=1}^{C} y_{ic} log(\sigma(h_{tf})), \tag{7}$$

where C is the class number, y_{ic} is 1 when the real label is c, 0 otherwise, $\sigma(h_{tf})$ is the probability function and h_{tf} is the image embeddings after Transformer module. During training, $\sigma(h_{tf})$ is implemented with the average prediction matrix fed to the $softmax$ activation function to obtain the probabilities of the three compatibility levels.

3.2. Colour-correlation enhancement

The colour correlation of an outfit greatly affects compatibility, and more attention should be paid to when designing an evaluation model. In this section, to take advantage of the colours, we first extract colour features for each fashion item and then construct the outfit's colour correlation matrix. To implement this, the FoCo system (Zou, Wong, Gao, & Zhou, 2019) is used to extract 5 main colours of every image and each colour is denoted with a 5-dimensional vector, i.e. $c_i \in R^5$ and $C \in R^{5 \times 5}$, where c_i is ith colour vector and C_j represents colour matrix of jth fashion item in an outfit. Then, we have a colour representation of an outfit as $O_c \in R^{p,5 \times 5}$.

Suppose the batch size as b, the colour correlation is defined as

$$W_{i,j} = \|C_i - C_j\|_F^2, \tag{8}$$

where $W_{i,j} \in R^{b,p\times p}$ indicates the colour difference among different fashion items with given b outfits. We compute the average colour difference of every fashion item concerning others by using

$$\dot{W}_i = \frac{1}{p}\sum_{i=1}^{p} W_{i,j}, \tag{9}$$

where $\dot{W}_i \in R^{b,p\times 1}$. $\dot{W}$ is reshaped as $\bar{W} \in R^{b,dim,p}$ where dim is the dimensionality of embeddings from Transformer module of layer 1. By projecting the embeddings to the colour difference space, we have

$$\tilde{W} = H\bar{W}, \tag{10}$$

where $\tilde{W} \in R^{b,p\times p}$ is the colour correlation representation of b outfits. Further, we project the embeddings to $\tilde{W}$ again and have

$$\tilde{H}_c = \tilde{W}H, \tag{11}$$

where $\tilde{H}_c \in R^{b,p\times dim}$ is the colour-correlation enhanced embeddings. After the transformation, the model is expected to learn the colour correlations of different items in the outfits. Thus, we use the following loss to minimise the difference of colour correlation after the transformation as mentioned above:

$$L_{cw} = \|\tilde{W} - W\|_F^2, \tag{12}$$

where L_{cw} is the colour correlation loss.

The structure of the prediction module is similar to that in the multi-layered feature analysis component. The colour-weighted feature embeddings in Eq. (11) are fed to the prediction module to compute the compatibility score. The prediction score in terms of three levels can be denoted as $y_c \in R^{b\times 3}$ and the loss between the prediction and the ground truth is expected to be minimised with the Cross-Entropy loss function:

$$L_{cf} = -\frac{1}{N}\sum_{i}^{N}\sum_{c=1}^{C} y_{ic} log(\sigma(h_{tcf})), \tag{13}$$

where h_{tcf} is the embeddings after colour-correlation enhancement module and $\sigma(h_{tcf})$ is the prediction function implemented with the prediction matrix fed to the $softmax$ activation function to obtain the probabilities of the three compatibility levels.

3.3. Visual-semantic similarity preserving

Multi-modal information is usually used to describe a fashion item, such as images, texts, etc. The proposed model uses visual Semantic Embedding (VSE) (Kiros, Salakhutdinov, & Zemel, 2014) to make full use of such information for compatibility learning.

For an outfit composed of p fashion items, each item is described by its text description. The text description of ith item can be denoted as e_i, a one-hot vector. The word embedding of e_i is represented as $v_i = W_T e_i$, where W_T is the weight matrix in the word embedding model. Then, we have semantic embeddings of the outfit as $v = \frac{1}{p}\sum_i^p v_i$. Similarly, the visual embeddings of the outfit can be denoted as $\mu = W_I x$, where W_I is a weight matrix for visual embedding transformation.

VSE assumes that the visual and semantic embeddings of the same fashion item should be close while that of different fashion items should be far. To achieve this goal, the visual-semantic loss is formulated as

$$\begin{aligned} &L_{vse}(v,\mu;W_T,W_I) \\ &= \sum_u\sum_k max(0, r - d(\mu,v) + d(\mu,v_k)) \\ &+ \sum_v\sum_k max(0, r - d(v,\mu) + d(v,\mu_k)), \end{aligned} \tag{14}$$

where $d(\cdot)$ is the distance function, r is a margin parameter, (μ, v) denotes visual-semantic embedding pairs while v_k/u_k is the semantic/visual embeddings of different fashion items in the outfit.

3.4. Outfit diagnosis and recommendation

Similar to the work in Wang et al. (2019), the backpropagation gradients are used to approximate the importance of each similarity of every fashion pair on the compatibility of an outfit. Differently, the compatibility is considered from three levels: Good, Normal, Bad, and the procedure of obtaining the similarity importance of each similarity is different.

Given an outfit $X = [x_1, x_2, \dots, x_p]$, we can obtain four prediction matrices from the multi-layered feature analysis module. As shown in Fig. 3, the compatibility score is obtained by linear projection in the fully connected module at each layer, and the prediction matrix is with the size of $(p \times p, 3)$ as

$$R^l = \begin{bmatrix} s^g_{11} & s^n_{11} & s^b_{11} \\ s^g_{12} & s^n_{12} & s^b_{12} \\ \dots & \dots & \dots \\ s^g_{pp} & s^n_{pp} & s^b_{pp} \end{bmatrix}, \tag{15}$$

where R^l is the prediction matrix at lth layer ($l = 1, 2, 3, 4$), $s^g_{i,j}$, $s^n_{i,j}$, $s^b_{i,j}$ denote the pairwise prediction with respect to the Good, Normal and Bad levels. The score of each compatibility level over the l layers can be expressed as

$$s^g_{i,j} = W^g_{i,j}[\tilde{H}^1_{i,j}; \tilde{H}^2_{i,j}; .., \tilde{H}^l_{i,j};] + b^g, \tag{16}$$

$$s^n_{i,j} = W^n_{i,j}[\tilde{H}^1_{i,j}; \tilde{H}^2_{i,j}; .., \tilde{H}^l_{i,j};] + b^n, \tag{17}$$

$$s^b_{i,j} = W^b_{i,j}[\tilde{H}^1_{i,j}; \tilde{H}^2_{i,j}; .., \tilde{H}^l_{i,j};] + b^b, \tag{18}$$

where $s^g_{i,j}$, $s^n_{i,j}$, $s^b_{i,j}$ are the prediction of the i and jth fashion items in an outfit. $W^g_{i,j}$, $W^n_{i,j}$, $W^b_{i,j}$ are the weight matrices while b^g, b^n and b^b are the biases.

Since the aforementioned equations are linear, the backpropagation gradients of $W^g_{i,j}$, $W^n_{i,j}$, $W^b_{i,j}$ can be used to interpret the importance of each similarity of the pairwise combinations respecting to the compatibility levels of Good, Normal and Bad. The backpropagation gradients are the derivatives of compatibility score s with respect to the embedding combinations $[\tilde{H}^1_{i,j}; \tilde{H}^2_{i,j}; ..; \tilde{H}^l_{i,j}]$ at l layers:

$$w^g_{i,j} = \frac{\partial s^g_{i,j}}{\partial[\tilde{H}^1_{i,j}; \tilde{H}^2_{i,j}; ..; \tilde{H}^l_{i,j}]}, \tag{19}$$

$$w^n_{i,j} = \frac{\partial s^n_{i,j}}{\partial[\tilde{H}^1_{i,j}; \tilde{H}^2_{i,j}; ..; \tilde{H}^l_{i,j}]}, \tag{20}$$

$$w^b_{i,j} = \frac{\partial s^b_{i,j}}{\partial[\tilde{H}^1_{i,j}; \tilde{H}^2_{i,j}; ..; \tilde{H}^l_{i,j}]}. \tag{21}$$

The outfit diagnosis can be implemented by tracking the gradients of the weight matrices of the three compatibility levels. When an outfit is evaluated as Normal or Bad, the compatibility can be improved by outfit diagnosis and recommendation. The similarity importance of each fashion item can be obtained by summing up all the relevant gradients as

$$w_q = \sum_{l=1}^{L}\sum_{i=q,j\neq q}^{p} w^l_{ij}, \tag{22}$$

where w_q is the diagnosed importance of the qth fashion item. The procedure of automatic diagnosis and recommendation on three levels of compatibility (i.e. Good, Normal and Bad) is summarised in Table 1.

The colour-correlation enhancement module can also provide the function of outfit diagnosis and recommendation. The module is expected to focus more on colour when evaluating outfit compatibility or recommending substitutions. The assumption will be verified by the experimental results in the Experiment section.

Table 1
The automatic diagnosis and recommendation algorithm.

Algorithm 1 The procedure of automatic diagnosis and recommendation

Input: $X = [x_1, x_2, \ldots, x_p] \in R^{p \times d}$ is an outfit with p fashion items of d dimensions. T is the maximum iteration of finding substitutions; ϵ is the threshold of the compatibility score of good level ($\epsilon = 90\%$).
Step 1 Prediction procedure: predict the compatibility scores of good, normal and bad levels as $X : (p_g, p_n, p_b)$;
Step 2 Diagnosis procedure: if ($p_g < \epsilon$), find the most problematic items and return the importance order as X_o;
Step 3 Recommendation procedure:
For each $x \in X_o$:
$t = 0$;
do {
(1) Randomly select an outfit (O_r) from the testing set;
(2) Generate new outfit (O_n) by substituting x with item in O_r with the same category;
(3) Predict the compatibility score of O_n;
(4) $t = t + 1$;
}while ($p_g < \epsilon$ && $t \leq T$)
End for
Output: O_n as the final recommendation result.

To the end, the proposed model comprises four components, and the overall optimisation objective is described as

$$L = L_{tf} + \lambda_1 L_{cf} + \lambda_2 L_{cw} + \lambda_3 L_{vse} + \lambda_4 L_{mask} + \lambda_5 L_{emb}, \tag{23}$$

where λ_1, λ_2, λ_3, λ_4, λ_5 are parameters to balance the six terms. The overall parameters of (23) are Θ; W^k, W^q, W^v, W^r and W^o, b_1, b_2; W_T, W_I, M where Θ are the parameters in the CNN network, W^k, W^q, W^v, W^r and W^o, b_1, b_2 are the parameters in the Transformer module in the multi-layered feature analysis component and W_T, W_I, M are parameters in the visual-semantic similarity preserving component. All the parameters can be optimised by back-propagation in a standard manner.

4. Experiments

Experiments on three datasets were conducted to evaluate the performance of the proposed method for outfit compatibility prediction and recommendation. The datasets include the Evaluation3 dataset which provides detailed compatibility levels for each outfit, the Polyvore-T dataset which provides general label annotation as compatible or incompatible, and the POG dataset which contains large amounts of outfit data from the online shopping website (taobao.com) (Chen, Huang, Xu, Guo, Guo, Sun, et al., 2019). The prediction performance of the proposed method is compared with several related methods, and the results in terms of automatic outfit diagnosis and substitution recommendation are presented.

4.1. Datasets

For the Evaluation3 dataset (Zou et al., 2020), the compatibility level is labelled as Good, Normal and Bad. Good outfits possess specific unique designs that make them visually stand out from the others. In contrast, Normal outfits are just visually harmony concerning the predefined factors: colour, print, material, silhouette and design details. Bad outfits contain at least one pair of fashion items incompatible with one factor. The proposed model does not evaluate the compatibility from the predefined factors. Instead, it tries to interpret the evaluation from assumed aspects. We aim to develop a general solution that can be easily applied to most of the existing outfit datasets without rich annotation information.

To evaluate the generalisation ability of the proposed model, experiments on the Polyvore-T dataset were also conducted. The outfits on this dataset are annotated with category, title and compatibility label as compatible or incompatible. More details of the construction of the dataset can be found in Wang et al. (2019). POG dataset comprises data from another domain that is different from Evaluation3 and Polyvore-T datasets. The data source of POG dataset is from taobao.com which is different from Evaluation3 and Polyvore-T datasets. The original POG dataset contains 1.01 million outfits and 583 thousand fashion items with rich annotated context information. We downloaded the first 37,555 outfits and filtered those outfits with less than four items. Finally, we have a subset of POG with 18,961 outfits. The subset was then separated to two partitions: the first partition containing 10,133 outfits was used for training, validation and testing while the second partition containing 8,828 outfits was used for the evaluation of the proposed method and other compared methods at a large-scale level. For clarity, we call the two testing sets in our experiments as Test-1 and Test-2. The attributes of the Evaluation3, Polyvore-T and POG datasets are listed in Table 2.

Positive/Negative Samples. In the experiments, the dataset is divided into training, validation, and testing sets with 7 : 2 : 1. Every outfit on the Evaluation3 dataset is annotated with a class ground truth as class 0, 1, 2 denoting Good, Normal and Bad, respectively. The outfits are considered positive (compatible) and negative (incompatible) for the Polyvore-T dataset, while positive samples come from the ground truth. The negative examples are generated by randomly substituting an item in the positive outfits with another item with the same type from other outfits (Wang et al., 2019). The positive outfits of the POG dataset are from the ground truth while the negative outfits are generated by randomly substituting each item in the outfit from other outfits with the same categories (Chen et al., 2019).

4.2. Tasks

The proposed method can be used to deal with outfit compatibility prediction and fashion diagnosis and recommendation tasks. It can also be used for the fill-in-the-blank task. The tasks and the corresponding evaluation metrics are:

Outfit Compatibility Prediction (AUC): The task of outfit compatibility prediction is to predict a compatibility score of an outfit. On the Evaluation3 dataset, the task is a multi-classes prediction problem, while that on the Polyvore-T dataset is binary classification. The evaluation metric on the Evaluation3 dataset is the Area Under the Receiver Operating Characteristic Curve (ROC AUC) from prediction scores. In our experiments, the AUC under multi-class classification and binary classification is considered and computed by the official implementation in PyTorch. For multi-class, it calculates the ROC AUC scores for each class against all other classes.

Fill-in-the-blank (FITB): Fill-in-the-blank selects the most compatible fashion item from an option list for a given outfit. The list contains four options for the Polyvore-T dataset and the proposed method. Other baselines conduct this task by substituting the blank part in the given outfit with three options and computing the respective compatibility scores. The predicted answer is regarded as the option that can achieve the highest compatibility score. The overall accuracy is computed based on all the predictions and ground truth questions.

Accuracy (ACC): The prediction accuracy concerning multi-class and binary are also computed in our experiments. On the Evaluation3 dataset, the binary prediction accuracy is computed by combining the Good and Normal classes as compatible while the Bad class is regarded as incompatible.

Outfit Compatibility Diagnosis and Recommendation: Compatibility diagnosis and recommendation are to first predict the compatibility level for a given outfit. If the level is Normal or Bad, the model will diagnose the outfit and recommend substitutes that can improve the compatibility level to Good. There is no quantitative metrics for these tasks, and we evaluate the performance by providing samples with human aesthetics.

Table 2
The statistics of the Evaluation3, Polyvore-T and POG datasets.

Dataset	Split	Top	Bottom	Shoes	Bag	Accessory	Item	Outfit
Evaluation3	Train	12,080	9,930	13,885	10,335	–	46,230	22,035
	Val	3,451	2,837	3,967	2,953	–	13,208	6,295
	Test	1,725	1,419	1,984	1,476	–	6,604	3,147
Polyvore-T	Train	13,764	14,849	15,268	12,640	12,093	68614	16,176
	Val	962	1,052	1,124	948	823	4,904	1,196
	Test	2,000	2,153	2,314	1,994	1,712	10,173	2,463
POG	Train	7,996	5,080	6,594	4,709	5,923	30,302	7,093
	Val	2,307	1,419	1,875	1,357	1,692	8,650	2,026
	Test-1	1,139	705	949	678	865	4,336	1,014
	Test-2	9,980	6,239	8,270	5,861	7,357	37,707	8,828

4.3. Baselines

The proposed method is compared with several state-of-the-art baselines designed for fashion compatibility evaluation. These methods include:

Pooling (Li, Cao, Zhu, & Luo, 2017): Pooling operation takes a fashion outfit and its category and title as input and encodes the visual as well as semantic features with a deep convolutional network, and Word2vec (Pennington, Socher, & Manning, 2014). It obtains the quality score with multi-modal fusion and multi-instance pooling.

Self-Attention (Wang, Girshick, Gupta, & He, 2018): The self-attention mechanism computes the representation at a position as the weighted representation of the features at all positions. For outfit compatibility evaluation, it learns representations by considering the relation of all items in an outfit. In our experiment, the scaled dot-product attention (Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, et al., 2017) was used in which the query, key, and value are item features in the same outfit.

BiLSTM+VSE (Han, Wu, Jiang, & Davis, 2017): BiLSTM learns outfit compatibility relationships by sequentially predicting the following item with the condition of previous ones. It regresses image features to the corresponding semantic representations to learn a visual-semantic space during the training, and the forward, backward LSTM and VSE losses are jointly optimised. The cross-entropy between the prediction and the ground truth is the compatibility score.

CSN (Vasileva et al., 2018): It learns image embeddings concerning corresponding type-aware space by conditioning different type combinations. The compatibility score is obtained based on the average of all pairwise compatibility.

MCN (Wang et al., 2019): It considers image features at different layers and corresponding semantic embeddings to learn and interpret the fashion compatibility from various potential aspects. The overall compatibility score is computed by type-specified pairwise similarities within a given outfit. MCN diagnoses the problematic items in an outfit by tracking the backpropagation gradients with the comparison matrix.

OCM-CF (Su et al., 2021): OCM-CF is outfit compatibility modelling scheme via complementary factorisation that consists of two components: context-aware outfit representation learning and hidden complementary factors modelling. The first component learns outfit representation with graph convolutional networks and multi-head attention mechanism, while the second component applies multiple parallel networks to discover the latent complementary factors. The final compatibility score is obtained by summing the scores that derived from the outfit representations by the parallel networks.

CSA (Lin, Tran, & Davis, 2020): CSA is a category-based subspace attention network to capture similarities of different outfits for complementary fashion item recommendation. In addition to the category-based attention mechanism, it utilises a new outfit ranking loss to learn the item relationships among the whole outfit. The loss needs triples as input, i.e. an outfit with a set of compatible fashion items, a positive fashion item that goes well to the outfit, and a set of negative items that is incompatible to the positive outfit.

4.4. Experiment settings

The experiments were completed on a desktop PC with NVIDIA GTX2080 GPU with 8 GB memory. The OS is Ubuntu 20.04.2 LTS, the CPU is Intel(R) Core(TM) i7-8700K @3.70 GHz with 11 processors, and the memory size is 32 GB.

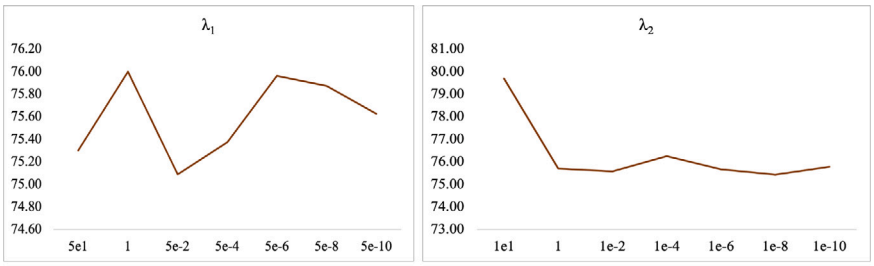
Training Settings. In all experiments, the ResNet50 pre-trained on ImageNet is used as the backbone, and the size of the training images on the Polyvore-T and POG datasets is 224×224 while that on the Evaluation3 dataset is 112×112. The lengths of fashion items of each outfit on the Evaluation3 dataset is 4 including top, bottom, shoes and bag, while that on the Polyvore-T and POG datasets is 5 including top, bottom, shoes, bag and accessories. On the Polyvore-T dataset, the input length can be from 3 to 5 with the missing part filled with the mean image while POG dataset has at least 4 fashion items in each outfit. Unlike the Evaluation3 and Polyvore-T datasets that composed of fashion items labelled with 4 or 5 types (i.e. top, bottom, shoes, bag and accessories), the fashion items on the POG dataset are labelled with specific categories. For simplicity, the type-aware learning strategy is not used for training the proposed method on this dataset, and instead, the embeddings after Transformer are directly fed to the prediction module. The batch size for the experiments on the Evaluation3 and POG is 16 while that on the Polyvore-T dataset is 24. SGD optimising strategy with a momentum of 0.9 is used during the training process. All methods are trained with 50 epochs during the training process, and the model parameters with the best performance on the validation set will be saved for testing.

Parameter Analysis. There are five parameters (i.e. λ_{1-5}) in the objective function and their values should be selected properly. Figs. 4 and 5 show the performance of the proposed method at different parameter values in which the variation curves are obtained by changing the values of one parameter while fixing the others. As we can see, the parameters can affect the performance to a certain degree. Specially, different values of $\lambda_{1,2,3}$ can lead to fluctuation of the performance. In our experiments, it is reasonable to set the values of λ_1 and λ_2 on all datasets as $5e^{-3}$ and e^{-2} for simplicity, while the values of $\lambda_{3,4,5}$ are set as 1, $5e^{-4}$ and $5e^{-3}$, respectively (the settings of $\lambda_{3,4,5}$ is similar to Wang et al. (2019)).

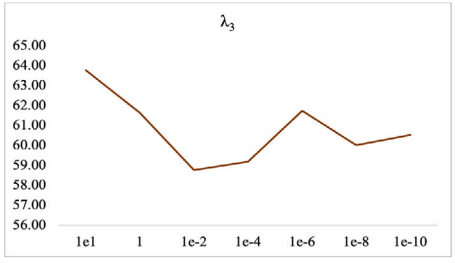
4.5. Prediction performance

The prediction performance of the proposed method and the compared methods are evaluated in this section. All methods are trained with multi-class cross-entropy during the training process and evaluated with three and binary classification performance during the testing stage.

The classification results on Evaluation3, Polyvore-T and POG datasets are shown in Tables 3–5, respectively. The result shows that the proposed method can obtain the best performance compared with other methods. The task of FITB is challenging as filling in the blank in the given outfit with a substitution may have little effect on the overall compatibility, and the compatibility difference among the four

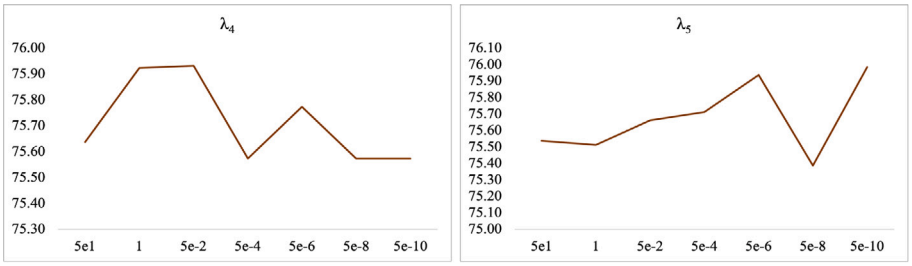


(a) The ACC-λ_1 on Evaluation3 dataset. (b) The ACC-λ_2 on Evaluation3 dataset.



(c) The AUC-λ_3 on Polyvore-T dataset.

Fig. 4. Parameters analysis of λ_1, λ_2 and λ_3.



(a) The ACC-λ_4 on Evaluation3 dataset. (b) The ACC-λ_5 on Evaluation3 dataset.

Fig. 5. Parameters analysis of λ_4 and λ_5.

Table 3
Prediction performance of different methods on the Evaluation3 dataset.

Method	Good/normal/bad		Compatible/incompatible	
	AUC (%)	ACC (%)	AUC (%)	ACC (%)
Pooling	79.89 ± 0.00	78.58 ± 0.00	83.22 ± 0.00	90.21 ± 0.00
Self-Att	77.99 ± 0.00	77.18 ± 0.00	81.64 ± 0.00	89.89 ± 0.00
BiLSTM	49.70 ± 2.06	25.17 ± 27.42	49.86 ± 1.35	57.67 ± 42.03
CSN	49.80 ± 1.00	37.67 ± 33.33	50.32 ± 1.62	73.03 ± 34.29
SCE	48.72 ± 0.88	25.68 ± 27.13	50.13 ± 0.19	73.02 ± 34.32
MCN	82.45 ± 0.00	77.95 ± 0.00	87.31 ± 0.00	90.37 ± 0.00
CSA	52.71 ± 0.00	74.16 ± 0.00	52.24 ± 0.00	88.36 ± 0.00
OCM-CF	45.78 ± 0.00	74.17 ± 0.00	47.81 ± 0.00	88.37 ± 0.00
Ours	**85.21 ± 0.00**	**79.00 ± 0.00**	**89.84 ± 0.00**	**91.26 ± 0.00**

conditional outfits are too trivial to classify. BiLSTM obtains poor performance because of no repeating items of the same type in the outfit, which is similar to the experiments in Vasileva et al. (2018), Wang et al. (2019). MCN and the proposed method obtain better performance than most of the compared methods as they both consider multi-layered features and learn compatibility with pairwise comparison in the outfits. However, the proposed method outperforms MCN as it weights feature relationships at different layers with the Transformer module and incorporates colour correlation of different items in the outfit to highlight the interaction in the colour aspect. The advantage of the proposed method can be more distinct for 3-class classification on the Evaluation3 dataset as detailed compatibility level leads to a clearer distinction on potential fashion factors like colour, material and pattern. As shown in Fig. 6, the outfits on the left-hand side belong to bad compatibility, and they tend to be inharmonious in terms of colour and pattern combination. In contrast, the outfits from the middle and right-hand sides look more compatible with harmonious colour combinations and consistent styles. The visualisation of binary classification on the Polyvore-T dataset is shown in Fig. 7, where the sum of prediction scores of compatible and incompatible is 100% and the distribution of the data points is linear. Then we can see that the incompatible and compatible outfits have a large gap respecting to fashion factors like colour, material, pattern and style. Additionally, the gap among the outfits that are classified as compatible is still distinct

Fig. 6. The visualisation of evaluating results of the proposed method on the Evaluation3 dataset.

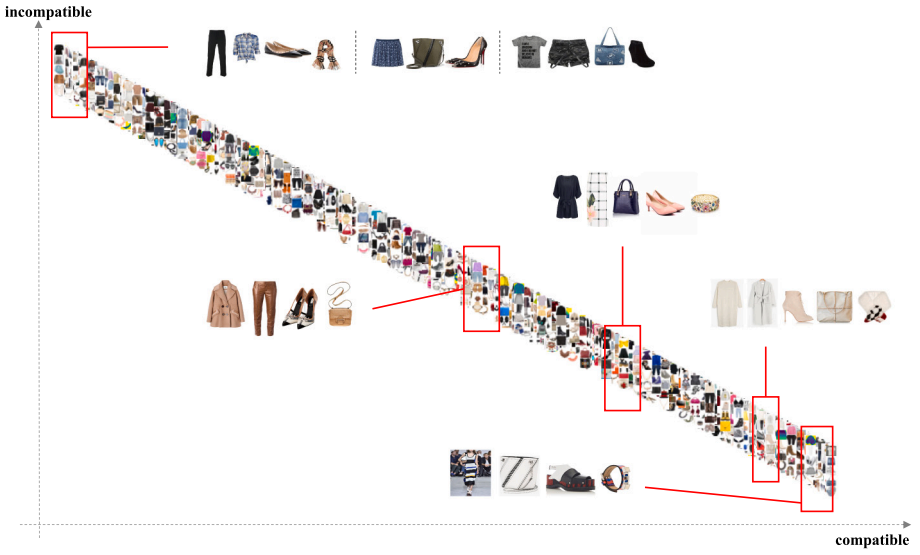


Fig. 7. The visualisation of compatibility on Polyvore-T dataset.

Table 4
Prediction performance of different methods on Polyvore-T dataset.

Method	AUC (%)	FITB (%)
Pooling	88.35 ± 0.26	57.28 ± 0.31
Self-Attention	79.65 ± 0.68	48.60 ± 0.70
BiLSTM	74.44 ± 0.95	45.41 ± 0.40
CSN	84.90 ± 0.52	57.06 ± 1.70
SCE	67.77 ± 1.04	14.04 ± 0.38
MCN	91.90 ± 0.40	64.35 ± 0.92
CSA	91.00 ± 0.00	63.73 ± 0.00
OCM-CF	92.00 ± 0.00	63.62 ± 0.00
Ours	**92.16 ± 0.25**	**65.65 ± 0.37**

Table 5
Prediction performance of different methods on the POG dataset.

Method	Test-1		Test-2	
	AUC (%)	ACC (%)	AUC (%)	ACC (%)
Pooling	50.25	49.41	50.07	50.24
Self-Att	55.58	55.58	53.29	49.91
BiLSTM	57.96	53.55	53.54	52.21
CSN	44.60	52.27	45.77	49.83
SCE	52.95	49.31	50.75	49.90
MCN	74.67	68.05	76.28	68.03
CSA	48.20	48.72	50.67	49.58
OCM-CF	77.59	66.17	76.69	66.74
Ours	**79.02**	**69.53**	**77.01**	**69.03**

from human aesthetics as the outfits in the middle and the right-hand side can be perceived as compatible at different levels.

4.6. Outfit diagnosis and recommendation

In this section, we explore the performance of the proposed method in terms of compatibility diagnosis and recommendation by compatibility improvement from low to the high level, the diagnosis at different layers and recommendation comparison between colour enhancement and overall compatibility.

4.6.1. Compatibility improvement from low to high level

This section shows examples of the proposed method in diagnosing problematic fashion items and recommending substitutions for given outfits. Figs. 8 and 9 show the results of compatibility improvement on the Evaluation3 and the Polyvore-T dataset, respectively.

As shown in Fig. 8, the first outfit on the left column is predicted as Normal, and the trousers are considered as the problematic item and suggested to be replaced with a skirt that can improve the overall compatibility level. The prediction and recommendation are consistent

Fig. 8. The compatibility evaluation and recommendation result of the proposed method on Evaluation3 dataset.

Fig. 9. The compatibility evaluation and recommendation result of the proposed method on Polyvore-T dataset.

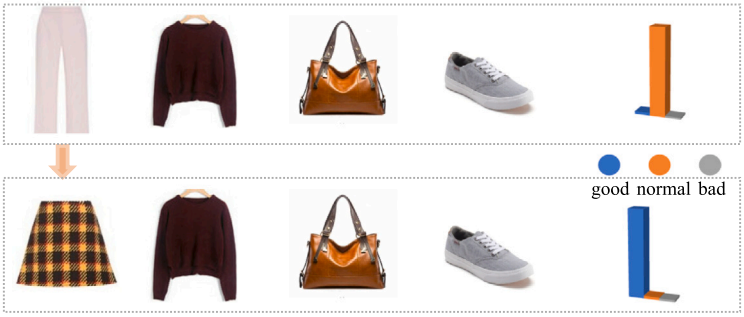


Fig. 10. The overall compatibility diagnosis and improvement result of the proposed method on Evaluation3 dataset. The ground truth is Normal, the prediction is Normal.

with human aesthetics. The original outfit has no outstanding aspects as the four items' colour, material, and design are compatible but not attractive. The floral-patterned skirt with harmonious colour makes the outfit more appealing and syllable for the improved outfit. For the second outfit, on the one hand, the up item with highly saturated colour make the whole look too dazzling, and the styles of the four items are inconsistent. On the other hand, the primary colours: red, blue, baby blue, lake blue and grey, are incompatible. The proposed method tries to improve the overall compatibility with as few as possible substitutions. Finally, the bottom, bag, and shoes are replaced with more compatible items in terms of colour and style. The improved outfit is more compatible, especially the colour combination, and the style is more consistent. The potential reason for this result is that the model can learn some hidden rules for obtaining a high compatibility score, such as the colour correlation, style preference from the training data, and the recommended items consistent with the rules. The potential rules are compatible with the illustration in Fig. 6 as the colours, prints and styles within the same outfit from Bad to Good levels are becoming more and more compatible.

4.6.2. Comparison between colour enhancement and overall compatibility

To further explore the potential rules of how the proposed method improves outfit compatibility from low to the high level, we show the diagnosis and improvement procedure of examples on the Evaluation3 dataset. In particular, Figs. 10 and 12 show the procedures of diagnosis and recommendation from overall compatibility while Figs. 11 and 13 show the procedures from the colour enhancement module. In Fig. 10, the model predicts the bottom as the most difficult part and recommends a brown patterned skirt to improve the compatibility. This is visually consistent with human aesthetics as the skirt makes the outfit more outstanding. In Fig. 11, the colour module tends to recommend items with similar colours or tones to improve the outfit as the colour similarity among different items are getting high while the recommendation step increases. However, the style between the previous item and the substitute is slightly different (first vs the second row). The improved outfit has a different style from the original outfit (first vs the last row). This can also be verified by Fig. 13. Unlike the colour module that tends to focus more on colour or other low-level aspects, the overall compatibility tends to consider not only colour, material but also the style and other potential factors, and the style of the improved outfit is more consistent with the original one. The comparison between Figs. 10, 12, 11 and 13 indicates that the proposed method can diagnose the most problematic items and recommend compatible items to improve the compatibility level from low to high. On the other hand, the colour module tends to focus more on colour or other low-level features, and the recommendation results may change

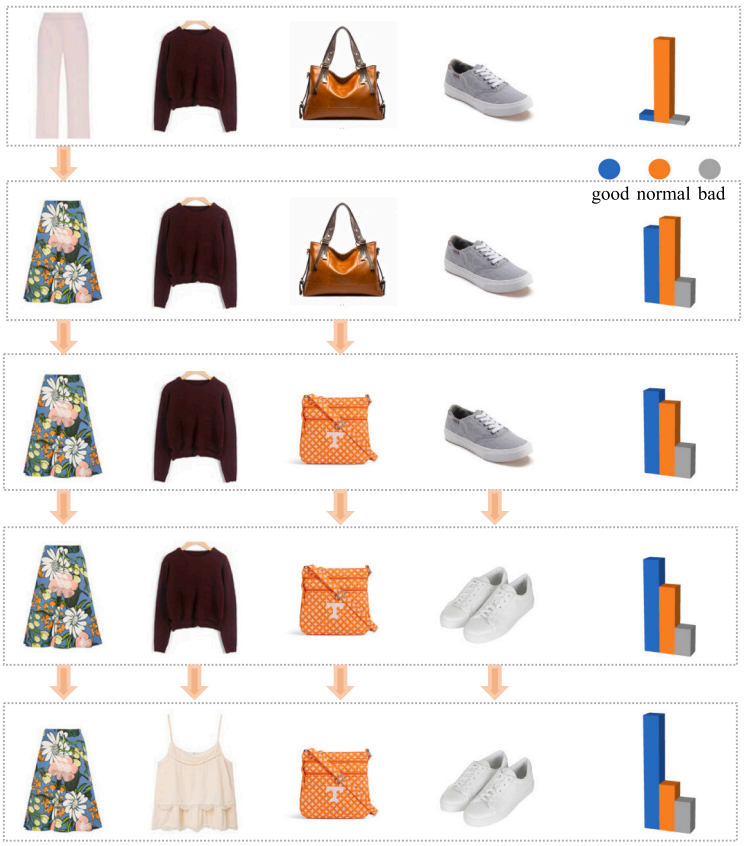


Fig. 11. The colour module compatibility diagnosis and improvement result of the proposed method on Evaluation3 dataset. The ground truth is Normal, the prediction is Normal.

Fig. 12. The overall compatibility diagnosis and improvement result of the proposed method on Evaluation3 dataset. The ground truth is Bad, the prediction is Bad.

the original style. However, the proposed method tends to consider low-level features and complex and abstract features when evaluating the compatibility, and it can achieve high compatibility in a few steps.

4.6.3. Diagnosis and recommendation at different layers

The importance of pairwise similarity approximated with the backward gradients at different layers on the Polyvore-T dataset are shown in Figs. 14–16. In Fig. 14, the orange bottom has the greatest effect on the compatibility at all layers, which is consistent with the human diagnosis. The bag and shoes are more harmonious on colour or style while they are not very compatible with the causal bottom no matter on colour or style aspect.

In Fig. 15, the up and bag with high saturated colours tend to be incompatible in terms of low-level features, which can be reflected by the first two layers. Visually, the bottom and up colours are harmonious, but the styles are different. This is diagnosed by the last two layers that focus more on high-level features that reflect the abstract factors. In Fig. 16, we can find that the up-bottom or up-accessory have the most significant impact on the overall compatibility at all layers. This is because the material and style of the up are visually inharmonious to other items.

4.7. Study of explainability and generalisation

Explainability. Fashion is extremely subjective and every user may have their own opinion on the same outfit. Explainability makes the results provided by the model more convincing which is crucial to enable users to trust the online stylist service. The proposed method evaluates compatibility by taking advantage of multi-layered features. The backpropagation gradients of the prediction matrices at different layers can be used to approximate the potential factors of the incompatibility. The maximum of gradients at different layers are listed, and the possible factors are predicted. As shown in Figs. 17 and 18, the outfits on the left columns are with low compatibility while those on the right columns are revised to have high compatibility. For example, it is obvious that the up at the fourth row in Fig. 17 has a different

Fig. 13. The colour module compatibility diagnosis and improvement result of the proposed method on Evaluation3 dataset. The ground truth is Bad, the prediction is Normal.

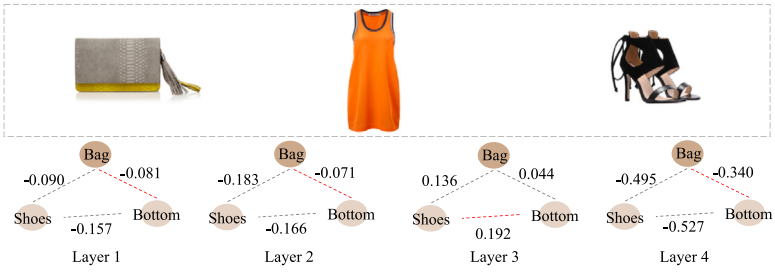


Fig. 14. Example of diagnosing incompatible outfit at different layers.

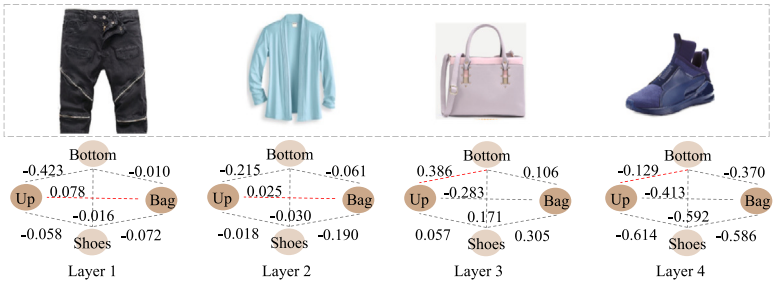


Fig. 15. Example of diagnosing incompatible outfit at different layers.

style with the bottom as the up is casual while the bottom is elegant. The original outfit is suggested to substitute the up and bag with the more compatible items in style and colour aspects. The revised outfit enjoys a Good level of compatibility, which is consistent with our human aesthetics. At the first row of Fig. 18, Layer 1 holds the most significant gradient, which indicates that the low-level features, such as colour and material, have a strong impact on the compatibility. The grey up is not as consistent as the brown up with the black bag and dark brown shoes. The down jacket in the original outfit at the second row is not compatible with the bag while substituting the jacket with the black–white patterned sweater makes the outfit more harmonious and neat.

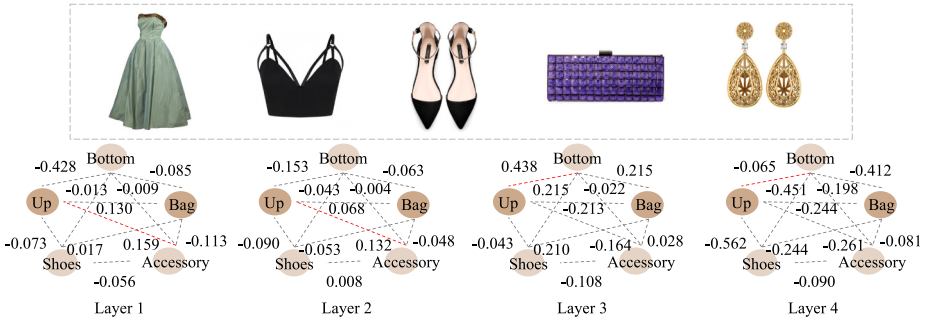


Fig. 16. Example of diagnosing incompatible outfit at different layers.

Fig. 17. Potential factor prediction based on comparison of different layers on Evaluation3 dataset.

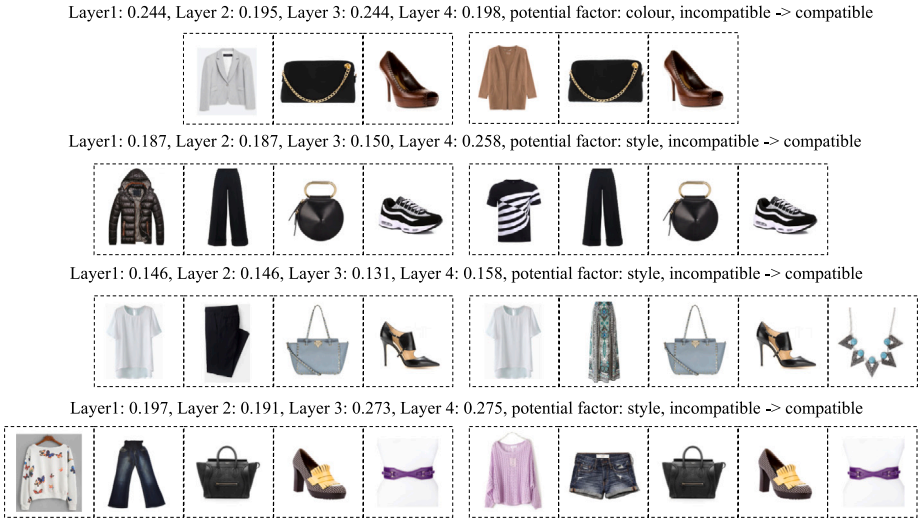


Fig. 18. Potential factor prediction based on comparison of different layers on Polyvore-T dataset.

Generalisation. Apart from being explainable, generalisation ability is vital for online stylist service as well. Currently, all outfit datasets proposed for fashion compatibility learning are collected from Polyvore, on which the outfits are made and uploaded by different fashion lovers. The compatibility annotations of the outfits actually reflect subjective fashion senses. Thus, we expect the model trained on these data can

Fig. 19. Compatibility prediction based on downloaded images.

Table 6
Statistical significance test for comparing relevant methods.

p-value/t-staticts	Ours	MCN	Pooling	Self-Att
Ours	–	0.0169/3.5237	0.0199/3.3686	0.0411/2.7346
MCN	0.0169/3.5237	–	0.0036/5.1591	0.0802/2.1893
Pooling	0.0199/3.3686	0.0036/5.1591	–	0.0369/2.8247
Self-Att	0.0411/2.7346	0.0802/2.1893	0.0369/2.8247	–

perceive the general essentials of fashion matching instead of only learning the distribution of the Polyvore data. To test the generalisation ability of the proposed model, we have conducted compatibility evaluation experiment on other source domain. Specifically, a set of outfits that captured from a TV drama [1] are used to test the performance of the proposed model for cross-domain compatibility prediction. The compatibility of the original items in an outfit from the drama plus two different items with the same category are evaluated, and the result is shown in Fig. 19. As we can see, the shoes in the first outfit in the first row can lead to a large compatibility gap, which indicates that an item can improve or even deteriorate an outfit. For the first outfit in the second row, the two shoes with the same designs, materials but different colours can lead to other compatibility scores. It indicates that colour is an essential factor when conducting compatibility evaluation with the proposed model, which is consistent with our initial assumption. Overall, the prediction result is reasonable as the outfits from the drama are commented on with high compatibility by the audience.

4.8. Statistical significance test

The comparison of prediction performance is an effective way to evaluate an model, but there may exist statistical fluke. To improve the confidence in the interpretation and presentation of the experimental results of different methods, in this section, we apply statistical hypothesis testing (Benjamin, Berger, Johannesson, Nosek, Wagenmakers, Berk, et al., 2018) to conduct a deeper comparative study.

The statistical hypothesis testing is to evaluate the mean performance difference that comes from the cross-validation procedure. In our experiments, we first state the null hypothesis that the difference does not exist between the two methods and they have the same performance. The threshold of significance level is set as $\alpha = 0.05$ for rejecting the null hypothesis. Then we use 5×2cv paired t test (Dietterich, 1998) to compare the two models and obtain the "p-value" that computed from "t-statistic" as difference. For every two methods, if the "p-value" is below the significance threshold, the null hypothesis can be rejected, that is, the difference is statistically significant and the two methods have different performance.

Since MCN, Pooling, Self-Att and the proposed method are directly convolutional based methods and they require similar data structure as input, we use these methods in the statistical hypothesis testing. For the other compared methods, since they are either based on conditional networks or require triples or graph data as input, they can be considered distinctly different and they are not necessary for comparison. The testing result of the pairwise methods is listed in Table 6. As we can see, the "p-values" of the proposed method are all below 0.05, which indicates that the null hypothesis can be rejected and the proposed method is statistically significant.

Table 7
Ablation study on Evaluation3 dataset.

Method	Good/normal/bad		Compatible/incompatible	
	AUC (%)	ACC (%)	AUC (%)	ACC (%)
Ours+1FC	83.48	79.69	89.01	91.74
Ours+2FC	75.76	78.20	89.72	91.36
Ours+CP	82.68	80.04	88.98	91.99
Ours_noC	84.05	78.90	89.15	91.07
Ours	85.21	79.00	89.84	91.26

[1] <<Now, We Are Breaking Up>>; image source: https://wakwb.com/.

4.9. Ablation study

To explore the performance of the proposed method with different modules and predictors, an ablation study was conducted on the Evaluation3 dataset, and the result is shown in Table 7, where Ours+1FC and Ours+2FC indicate that the prediction module obtains the compatibility score with 1 or 2 fully connected layers, respectively, Ours+CP represents the structure of prediction module which is similar to the comparison procedure in MCN, Ours_noC denotes our model without the colour-correlation enhancement module. The performance of Ours+1FC and Ours+2FC is inferior to Ours, which indicates the advantage of the prediction module of the proposed method. Based on the observation, we also use the same prediction module for MCN for fair comparison on the Evaluation3 dataset. The prediction module of our method on the Polyvore-T dataset is similar to that in MCN for simplicity. The comparison between Ours_noC and Ours indicates that our method without the colour-correlation enhancement module is still competitive, and the incorporation of colour-correlation can further improve the prediction performance.

5. Conclusion

This paper targets on fashion compatibility evaluation via jointly learning colour and visual semantic features based on multi-layered convolutional networks and Transformer scheme. Experimental results on three mainstream fashion outfit datasets, i.e., Evaluation3, Polyvore-T, and POG, show the advance of the proposed approach. Although the generalisation enables the easy embedding of the method on fashion evaluation platform, and the explainability of our method can help to achieve user's trust, the lack of personalisation information makes the method hard to fulfil customers' expectation. Actually, personalised outfit evaluation and recommendation can help fashion lovers to present their fashion style and personality. Therefore, developing an algorithm with personalised service towards a specific customer group will be our next step.

CRediT authorship contribution statement

Dongmei Mo: Conceptualisation, Methodology, Formal analysis, Writing – original draft. **Xingxing Zou:** Conceptualisation, Writing – review & editing. **WaiKeung Wong:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The research was funded by Laboratory for Artificial Intelligence in Design (Project Code: RP3-2) under the InnoHK Research Clusters, Hong Kong SAR Government.

References

Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017). Understanding of a convolutional neural network. In *2017 international conference on engineering and technology (ICET)* (pp. 1–6). IEEE.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., et al. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10.

Chen, W., Huang, P., Xu, J., Guo, X., Guo, C., Sun, F., et al. (2019). POG: personalized outfit generation for fashion recommendation at Alibaba iFashion. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 2662–2670).

Chen, H., Yin, H., Wang, W., Wang, H., Nguyen, Q. V. H., & Li, X. (2018a). PME: projected metric embedding on heterogeneous networks for link prediction. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 1177–1186).

Chen, X., Zhang, Y., Xu, H., Cao, Y., Qin, Z., & Zha, H. (2018b). Visually explainable recommendation. arXiv preprint arXiv:1801.10288.

Chopra, S., Hadsell, R., & LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. *vol. 1*, In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)* (pp. 539–546). IEEE.

De Boer, P.-T., Kroese, D. P., Mannor, S., & Rubinstein, R. Y. (2005). A tutorial on the cross-entropy method. *Annals of Operations Research*, *134*(1), 19–67.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, *10*(7), 1895–1923.

Duggal, R., Zhou, H., Yang, S., Xiong, Y., Xia, W., Tu, Z., et al. (2021). Compatibility-aware heterogeneous visual search. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10723–10732).

Feng, Z., Yu, Z., Jing, Y., Wu, S., Song, M., Yang, Y., et al. (2019). Interpretable partitioned embedding for intelligent multi-item fashion outfit composition. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *15*(2s), 1–20.

Feng, Z., Yu, Z., Yang, Y., Jing, Y., Jiang, J., & Song, M. (2018). Interpretable partitioned embedding for customized multi-item fashion outfit composition. In *Proceedings of the 2018 ACM on international conference on multimedia retrieval* (pp. 143–151).

Hadsell, R., Chopra, S., & LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. *vol. 2*, In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)* (pp. 1735–1742). IEEE.

Han, X., Wu, Z., Jiang, Y.-G., & Davis, L. S. (2017). Learning fashion compatibility with bidirectional lstms. In *Proceedings of the 25th ACM international conference on multimedia* (pp. 1078–1086).

Kim, D., Saito, K., Mishra, S., Sclaroff, S., Saenko, K., & Plummer, B. A. (2021). Self-supervised visual attribute learning for fashion compatibility. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1057–1066).

Kiros, R., Salakhutdinov, R., & Zemel, R. S. (2014). Unifying visual-semantic embeddings with multimodal neural language models. arXiv preprint arXiv:1411.2539.

Kolisnik, B., Hogan, I., & Zulkernine, F. (2021). Condition-CNN: A hierarchical multi-label fashion image classification model. *Expert Systems with Applications*, *182*, Article 115195.

Kuang, Z., Gao, Y., Li, G., Luo, P., Chen, Y., Lin, L., et al. (2019). Fashion retrieval via graph reasoning networks on a similarity pyramid. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3066–3075).

Li, Y., Cao, L., Zhu, J., & Luo, J. (2017). Mining fashion outfit composition using an end-to-end deep learning approach on set data. *IEEE Transactions on Multimedia*, *19*(8), 1946–1955.

Li, Y., Chen, T., & Huang, Z. (2021). Attribute-aware explainable complementary clothing recommendation. *World Wide Web*, *24*(5), 1885–1901.

Lin, Y., Ren, P., Chen, Z., Ren, Z., Ma, J., & de Rijke, M. (2018). Explainable fashion recommendation with joint outfit matching and comment generation. CoRR abs/1806.08977, arXiv preprint arXiv:1806.08977.

Lin, Y.-L., Tran, S., & Davis, L. S. (2020). Fashion outfit complementary item retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3311–3319).

Liu, J., Song, X., Nie, L., Gan, T., & Ma, J. (2019). An end-to-end attention-based neural model for complementary clothing matching. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *15*(4), 1–16.

Longo, F., Padovano, A., Cimmino, B., & Pinto, P. (2021). Towards a mass customization in the fashion industry: An evolutionary decision aid model for apparel product platform design and optimization. *Computers & Industrial Engineering*, *162*, Article 107742.

Lu, Z., Hu, Y., Chen, Y., & Zeng, B. (2021). Personalized outfit recommendation with learnable anchors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12722–12731).

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543).

Qasem, Z. (2021). The effect of positive TRI traits on centennials adoption of try-on technology in the context of E-fashion retailing. *International Journal of Information Management*, *56*, Article 102254.

Revanur, A., Kumar, V., & Sharma, D. (2021). Semi-supervised visual representation learning for fashion compatibility. In *Fifteenth ACM conference on recommender systems* (pp. 463–472).

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618–626).

Seo, Y., & Shin, K.-s. (2019). Hierarchical convolutional neural networks for fashion image classification. *Expert Systems with Applications*, *116*, 328–339.

Su, T., Song, X., Zheng, N., Guan, W., Li, Y., & Nie, L. (2021). Complementary factorization towards outfit compatibility modeling. In *Proceedings of the 29th ACM international conference on multimedia* (pp. 4073–4081).

Tan, R., Vasileva, M. I., Saenko, K., & Plummer, B. A. (2019). Learning similarity conditions without explicit supervision. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10373–10382).

Vadood, M., & Haji, A. (2022). A hybrid artificial intelligence model to predict the color coordinates of polyester fabric dyed with madder natural dye. *Expert Systems with Applications*, Article 116514.

Vasileva, M. I., Plummer, B. A., Dusad, K., Rajpal, S., Kumar, R., & Forsyth, D. (2018). Learning type-aware embeddings for fashion compatibility. In *Proceedings of the european conference on computer vision (ECCV)* (pp. 390–405).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).

Veit, A., Belongie, S., & Karaletsos, T. (2017). Conditional similarity networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 830–838).

Wang, J., Cheng, X., Wang, R., & Liu, S. (2021). Learning outfit compatibility with graph attention network and visual-semantic embedding. In *2021 IEEE international conference on multimedia and expo (ICME)* (pp. 1–6). IEEE.

Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*.

Wang, J., Song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., et al. (2014). Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1386–1393).

Wang, X., Wu, B., & Zhong, Y. (2019). Outfit compatibility prediction and diagnosis with multi-layered comparison network. In *Proceedings of the 27th ACM international conference on multimedia* (pp. 329–337).

Yang, X., He, X., Wang, X., Ma, Y., Feng, F., Wang, M., et al. (2019). Interpretable fashion matching with rich attributes. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval* (pp. 775–784).

Yang, C.-L., & Huang, R.-H. (2011). Key success factors for online auctions: Analysis of auctions of fashion clothing. *Expert Systems with Applications*, *38*(6), 7774–7783.

Yang, X., Song, X., Feng, F., Wen, H., Duan, L.-Y., & Nie, L. (2021). Attribute-wise explainable fashion compatibility modeling. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *17*(1), 1–21.

Yu, Y., Hui, C.-L., & Choi, T.-M. (2012). An empirical study of intelligent expert systems on forecasting of fashion color trend. *Expert Systems with Applications*, *39*(4), 4383–4389.

Zhan, H., Lin, J., Ak, K. E., Shi, B., Duan, L.-Y., & Kot, A. C. (2021). A3-fKG: Attentive attribute-aware fashion knowledge graph for outfit preference prediction. *IEEE Transactions on Multimedia*.

Zou, X., Li, Z., Bai, K., Lin, D., & Wong, W. (2020). Regularizing reasons for outfit evaluation with gradient penalty. arXiv preprint arXiv:2002.00460.

Zou, X., Wong, W. K., Gao, C., & Zhou, J. (2019). Foco system: A tool to bridge the domain gap between fashion and artificial intelligence. *International Journal of Clothing Science and Technology*.