Expert Systems With Applications 219 (2023) 119632

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Towards private stylists via personalized compatibility learning

Dongmei Mo, Xingxing Zou, Kaicheng Pang, Wai Keung Wong *

School of Fashion and Textiles, The Hong Kong Polytechnic University, Hong Kong Special Administrative Region of China
The Laboratory for Artificial Intelligence in Design, Hong Kong Special Administrative Region of China

ARTICLE INFO

Keywords:
Personalized outfit compatibility
Fashion compatibility
Multi-label classification
Transformer

ABSTRACT

Personalized outfit compatibility learning is an emerging yet challenging task. Most of the existing methods focus on general outfit compatibility learning. Although a few works have been proposed for personalized fashion compatibility, they either considered user preference on fashion items with specific patterns or design elements or recommended outfits based on the overall visual similarity according to the users' preferred collections. This paper adopts physical and fashion attributes for effective personalized fashion compatibility evaluation and recommendation. The physical attributes are concluded into seven aspects: body shape, skin color, hairstyle, hair color, height, breast size (breasts), and color contrast. The personalized outfit compatibility problem in this paper is a multi-label classification problem and formulated as an optimization function with outfit images, fashion attributes, and physical attributes as input. It is the first attempt to solve the problem by discovering the correlation between visual image features, fashion attributes, and physical attributes. Specifically, the correlation is learned with two transformer encoders by updating attention weights of different embedding pairs during the training process. The model can not only predict the fashion attributes of the outfit's top, bottom, shoes, and bag items, but also predict the incompatible physical attributes of an individual towards the given outfit. It can be used to recommend outfits that best fit an individual and the predicted fashion attributes can be used for result explanation. The O4U dataset, which contains rich annotations of fashion item attributes and human physical attributes of the outfits, is used to evaluate the performance of the proposed method. The quantitative and qualitative results show that the proposed method outperforms state-of-the-art methods for personalized outfit compatibility evaluation.

1. Introduction

Personalized outfit compatibility learning has been recently attracting more attention from the academic and industrial communities. Most research focuses on general fashion compatibility, which is about the compatibility learning among several fashion items within an outfit (Guan et al., 2022, 2021). The methods either learn a compatibility score of the outfit by computing average compatibility probability among different items in the outfit (Cui et al., 2019; Vasileva et al., 2018), or try to reason the evaluation results with inexplicit or predefined fashion factors (Mo et al., 2022; Tan et al., 2019). The existing personalized fashion recommendation works can be divided into two categories: one is to recommend fashion items via highlighting patterns or designs related to the user preference (Chen, Chen, et al., 2019); the other is to recommend fashion outfits based on the visual similarity of the outfits according to the user's preference (Zhan & Lin, 2021).

Despite the effectiveness of these methods, the personalized recommendation for an individual needs to be more consistent with the human sense. For example, when we browse large amounts of outfits on an online shopping platform, we expect to efficiently find the outfit that is not only compatible but also the most suitable for our physical appearance. Thus, this paper proposes considering the outfit compatibility towards individual physical characteristics for personalized outfit recommendations. This is actually an outfit-user compatibility learning problem regarding the compatibility between a set of compatible fashion items and certain personalized physical attributes. The personalized physical attributes are considered from 7 aspects: body figure, skin color, hairstyle, hair color, height, breast size (breast), and color contrast, and each aspect has corresponding detailed attributes. The idea is consistent with human aesthetics on the compatibility of an outfit and an individual. As an example in Fig. 1, given an outfit with certain fashion items (top, bottom, shoes, and bag in this case), fashion lovers who have high requirements on their look usually make a decision on buying the outfit by evaluating the overall compatibility of the outfit with their physical information. We use the body figure as

* Corresponding author at: School of Fashion and Textiles and the Laboratory for Artificial Intelligence in Design, The Hong Kong Polytechnic University, Hong Kong Special Administrative Region of China.

E-mail addresses: dongmei.mo@connect.polyu.hk (D. Mo), xingxingzou@aidlab.hk (X. Zou), kaicpang.pang@connect.polyu.hk (K. Pang), calvin.wong@polyu.edu.hk (W.K. Wong).

https://doi.org/10.1016/j.eswa.2023.119632
Received 15 August 2022; Received in revised form 4 January 2023; Accepted 27 January 2023
Available online 3 February 2023

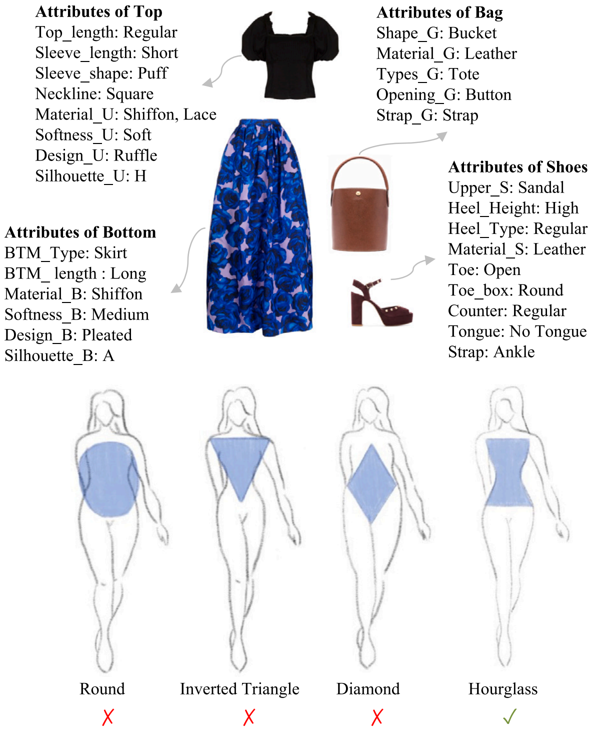

Fig. 1. Illustration of application scenarios of the proposed model.

an example to analyze the compatibility in the figure. Our brain tends to intuitively perceive the outfit compatibility on the fashion attributes like pattern, color, silhouette, or even more abstract elements (design and styles). Even though the process of aesthetic perception is hard to illustrate, we can still find some rules to train an intelligent evaluation model for personalized compatibility evaluation. For the outfit in Fig. 1, we can recognize the fashion attributes for each item and evaluate the compatibility with different body figures. Intuitively, the top item with a square neckline, puff sleeve shape, H silhouette, and ruffle design tends to have high requirements on the upper body, while the bottom item with an A silhouette tends to be more friendly to different types of the lower body. Also, the shoes with round-toe boxes and high heels tend to be more suitable for slim legs. Thus, a person with a body figure of round, inverted triangle, or diamond is suggested to avoid choosing such kinds of outfits.

This paper attempts to evaluate an outfit's compatibility with the consideration of detailed individual physical attributes. To deal with this problem, we need to consider the following challenges: (1) lack of informative supervision. In the literature, there is no outfit dataset that contains rich physical label annotations and fashion attribute annotations on each outfit; (2) difficulties in selecting an effective framework for discovering the correlation among outfit visual features and fashion item attributes towards physical attributes. For a given outfit, an ideal framework is expected to learn the potential correlation between outfit compatibility and individual physical attributes. To address the above issues, we propose to use the outfit dataset (O4U dataset) (Kaicheng Pang, 2022) on which the fashion items of top, bottom, shoes, and bag are annotated with professional fashion attributes, and the outfits are labeled with detailed physical attributes from 7 aspects, namely, body figure, skin color, hairstyle, hair color, height, breast size (breast), color contrast. To discover the compatibility with the learned visual features and fashion attribution embeddings, we propose a new framework based on transformer encode, which has been successfully employed in learning the relationship among different visual regions in general multi-label classification tasks (Han et al., 2021; Yuan et al., 2021). The contributions of the paper can be concluded as below

(1) The proposed method is the first to consider the outfit compatibility towards individual physical attributes and fashion item attributes, which is consistent with human aesthetics on finding the most suitable outfits for personalized information.

(2) The transformer encoder explores the correlation among visual features, fashion attributes and physical attributes. The model can predict the personalized outfit compatibility with fashion attribute prediction as a potential explanation.

(3) Experimental results on the O4U dataset demonstrate the superiority of the proposed method for personalized outfit compatibility prediction compared with the state-of-the-art methods.

2. Related works

In this section, we will introduce more details about personalized fashion compatibility learning and the related applications.

2.1. *Personalized fashion compatibility*

General outfit compatibility learning is to predict if a given outfit with certain fashion items is well matched by giving an explicit compatibility score, while personalized fashion compatibility needs to not only ensure that fashion items in an outfit are compatible with each other but also guarantee that the outfit fits an individual's preference.

The existing personalized fashion compatibility learning methods can be divided into item-level and outfit-level methods. The item-level methods mainly model personalized fashion compatibility by discovering user preferences via visual features or text information supervision. Chen et al. proposed a multi-modal attention network for fashion item recommendation with review information as the guidance of users' interest learning (Chen, Chen, et al., 2019). The method divides a fashion image into regions and employs an attention model to weight the regions with review information as the supervision to discover users' preferences. Despite its effectiveness, the model may perform poorly when the text information is not representative of the users' preferences. Hu et al. proposed applying user–item pairs to discover the users' interest in different items. Unlike many methods that use metric learning to measure the pairwise fashion compatibility, this method used inner product for compatibility calculation (Hu et al., 2015). Furthermore, The item–item and user–item interactions are characterized by a matrix factorization method in Song et al. (2019). Lu et al. proposed a hashing method to learn binary codes for user and item representations to release the efficiency problem in the recommendation process (Lu et al., 2019).

The outfit-level methods conduct personalized fashion compatibility recommendation by treating the outfit as a whole and learning the potential compatibility inexplicitly. The Personalized Outfit Generation (POG) model in Chen, Huang, et al. (2019) suggested that a user should have similar tastes in fashion items and outfits, and user preference should be considered as the connection between the items and the outfits. The model applied transformer architecture to learn the user's interest regarding the user clicked items and outfits. Zhan et al. proposed a Personalized Attention Network (PAN) for personalized outfit recommendation with a user encoder, an item encoder, and a preference predictor as the key component in the framework (Zhan & Lin, 2021). The method also applied the attention network that composed of a sequential user-aware channel-level and spatial-level sub-modules to discover the users' preferences towards fashion items. Additionally, a user-specific ranking loss was proposed to capture the interest of different users in the same outfit.

In summary, the existing personalized compatibility methods mainly learn user preferences from visual fashion preferences on individual fashion items or the whole outfit. They do not take the individual physical information into consideration to recommend the outfits that can fit the individual most.

2.2. Emerging applications and datasets

Dressing with compatibility of physical characteristics and outfits is more than just the clothing; it is about how the fashion lovers carry themselves to reflect their attitudes and preference. Compatibility learning has been studied in different scenarios in the fashion community. Wang et al. proposed to evaluate the outfit compatibility among items within the same outfit, and diagnose the problematic items via backpropagation gradient comparison, as well as recommend the substitutes with high compatibility score (Wang et al., 2019). Several recent works studied the compatibility of outfits/dresses towards different body shapes. Hidayati et al. proposed the first framework for considering fashion compatibility of clothing styles and body shapes from social big data (Hidayati et al., 2018). The goal is to recommend a user with clothing that fits his/her body shape best. Later, they proposed to learn the correlation between joint deep embeddings of clothing styles and body shapes. Female body shapes can be measured via proportion calculation in the proposed framework, and the fashion knowledge from social big data is utilized to learn the golden styles for an intelligent recommendation of dresses for different body types (Hidayati et al., 2020). All the methods consider fashion compatibility based on a single fashion item or dress and body shape, which is one of the individual physical aspects. More complicated fashion combinations and more complete physical attributes should be considered for practical use in personalized outfit recommendations.

To our knowledge, many types of datasets exist for fashion compatibility learning. They can be categorized as general outfit datasets containing outfit images with or without text descriptions and personalized outfit datasets containing user preference information. The general outfit datasets mainly include WoW (Liu et al., 2012), FashionVC (Song et al., 2017), Maryland Polyvore (Han et al., 2017), UIUC Polyvore (Vasileva et al., 2018), Polyvore-T (Wang et al., 2019) and Evaluation3 (Zou et al., 2020). The personalized outfit datasets mainly include Style4BodyShape (Hidayati et al., 2018), StyleRef. (Hidayati et al., 2020), POG (Chen, Huang, et al., 2019) and Polyvore-U (Polyvore-630, Polyvore-519) (Lu et al., 2019). This paper will use the O4U dataset on which the fashion items are annotated with rich fashion attributes, and the outfits are labeled with complete physical attributes.

3. Methodology

In this section, we first describe the motivation of the paper, then formulate the research problem and present the details of the proposed method.

3.1. Motivation

Previous works consider personalized fashion evaluation or recommendation from different aspects. Some proposed recommending fashion items by taking the image region-level features and reviewing information into a multimodal attention network (Chen, Chen, et al., 2019). The methods can recommend items by highlighting specific regions of the images as the user preference with weak supervision from the user review information. Considering the personalized fashion style, outfit search, and recommendation efficiency issue, Lu et al. proposed learning binary code for efficient and personalized fashion recommendation (Lu et al., 2019). However, the method fails to capture users' interest in details, such as logos and patterns. To solve this problem, Zhan et al. proposed a personalized attention network to integrate user embedding and item representation to compute the user-aware attention maps (Zhan & Lin, 2021). These methods were evaluated on the Polyvore-U benchmark dataset that contains user profiles and outfits. The personalization of the existing methods is not related to a user's physical attributes, while this is general for our human visual justification for evaluating an outfit for a user with a specific body shape, hair color, skin color, etc.

In addition, the relation between fashion attributes and human physical attributes is not explored in the existing methods. Although it may be abstractive when we consider the personalized outfit compatibility with fashion attributes and physical attributes, it must have certain relation as different attributes usually fit different physical attributes. The transformer has been successfully applied to explore the attention between different image regions or image-text information (Pardo-Sixtos et al., 2022). In this paper, we apply the transformer technique and propose a new framework for personalized fashion compatibility learning to explore the relationships among the visual features, fashion attributes, and physical attributes.

3.2. Problem formulation

The personalized fashion compatibility evaluation problem involves the fashion item images from different categories, the attribute labels of fashion items from top, bottom, shoes and bag, and the physical attribute labels. The purpose is to predict the incompatible physical attributes and fashion item attributes with given outfits so that a user can avoid selecting the outfits that do not fit his/her physical characteristics.

Suppose we have a set of outfits from different categories (i.e. top, bottom, shoes, bag, accessories, etc.). Let an outfit denoted as $O = \{x_1, x_2, x_3, x_4, \dots, x_k\}, k \leqslant 9$, where k is the number of items in an outfit. Specifically, the fashion items from the categories of top, bottom, shoes, and bag are annotated with the corresponding number of attributes, i.e. $c_u = 91, c_b = 49, c_s = 47, c_g = 42$, where c_u, c_b, c_s, c_g are the numbers of the attributes of the top, bottom, shoes, and bags. Since the fashion item x_i is attached with a visual image and multi-label description y, we can derive N training outfit samples as $\omega = \{(O_i, y_{ubsg}, y_i) | i = 1, 2, 3, \dots, N\}$, where O_i is the ith outfit and y_i is the ground truth labels that indicate the incompatible physical attributes of the outfit, $y_i \in \{0, 1\}$ where 1 represents incompatibility positive while 0 is incompatibility negative, y_{ubsg} is the fashion attribute labels of the top, bottom, shoes and bag items. Notably, the number of fashion items in the outfit can be changed. Based on these data, we design a personalized transformer scheme which integrates the visual images and fashion attributes for personalized compatibility prediction. Mathematically, we have the following multi-label prediction problem:

$$y_p = \mathcal{F}(o, y_{ubsg} | \Theta), \tag{1}$$

where o is the visual images of the outfit, y_{ubsg} is the attribute labels of the items, Θ is the set of to-be-learned parameters in the model, y_p is the estimated one-hot prediction of the outfit, $y_p = \{y_1, y_2, \dots, y_{c_p}\}$, where $y_i \in \{0, 1\}$ and $c_p = 15$ is the physical attribute number.

3.3. The proposed method

The proposed method takes outfit images and the corresponding attribute annotations as input and explores the relation between fashion attributes and physical attributes. The personalized compatibility of outfits towards physical attributes is learned and evaluated as a multi-label classification task as the following.

3.3.1. The proposed framework

The personalized compatibility learning framework is presented in Fig. 2. The framework comprises three parts: a visual feature extractor, the first transformer encoder for fashion attribute prediction, and the second transformer encoder for physical attribute prediction. The visual feature extractor is based on the Resnet (He et al., 2016) which was pretrained on ImageNet (Deng et al., 2009). As shown on the left side of Fig. 2, the fashion items in the outfit are input to the Resnet, and the feature embeddings are obtained as the representation of the outfit. Then the attribute label embedding together with the visual feature embedding are input to the first transformer, and their correlation is learned for item attribute prediction. The learned visual feature and

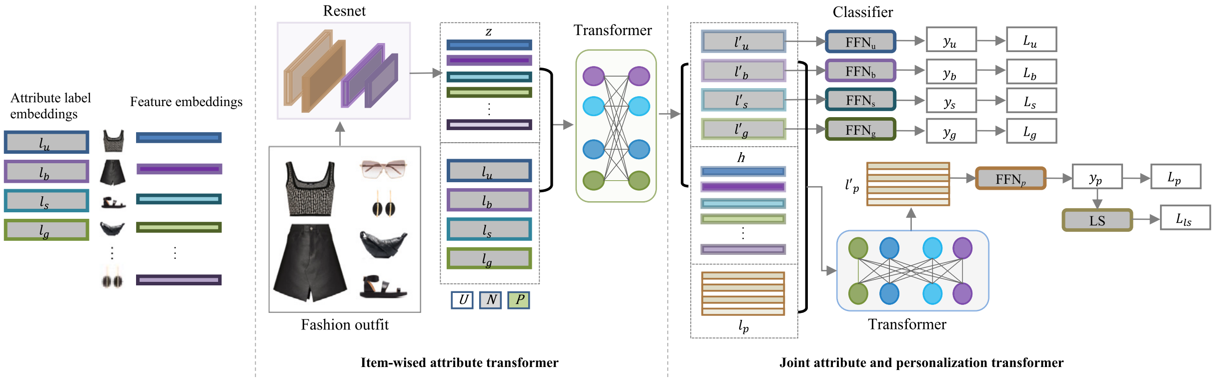

Fig. 2. The personalized compatibility evaluation framework. The framework composes of three parts: The image feature extractor, the first transformer encoder which is for fashion attribute prediction, and the second transformer encoder which is for physical attribute prediction. $\{U, N, P\}$ denote different states of the label with negative (N), positive (P), and unknown (U). The multi-label classifier is the feedforward network (FFN). z, h denote the visual image features learned by the Resnet feature extractor and the first transformer encoder. l_u, l_b, l_s, l_g denote the initial fashion attribute embeddings of top, bottom, shoes and bag while l'_u, l'_b, l'_s, l'_g denote the fashion attribute embeddings after the first transformer encoder. l_p is the initial physical attribute embeddings while l'_p is the physical attribute embeddings after the second transformer encoder. y_u, y_b, y_s, y_g are the multi-label prediction of fashion attributes of top, bottom, shoes and bag, y_p is the multi-label prediction of physical attributes. LS is the label smoothing loss function. $L_u, L_b, L_s, L_g, L_p, L_{ls}$ correspond to the loss components of the objective function in Eq. (25).

attribute embeddings from the first transformer encoder as well as the initial physical attribute embeddings are fed to the second transformer for personalized compatibility prediction.

The proposed framework can be applied to many practical scenarios. However, directly applying it without pre-training to a specific case may obtain unexpected performance due to the domain gap. Thus it is suggested to finetune the framework with related data first. In the scenario of recommendation, given a set of outfit images, the framework can automatically recognize the fashion attributes of different items and the physical attributes that are compatible with the outfit. So when an individual inputs his/her personal physical information to the system, the framework can recommend a set of outfits that are compatible with the personalized information. Further, the individual can select his/her preferred fashion attributes, then the system can recommend the outfits that fulfill the requirement.

Fashion item embedding Given an outfit with four fashion items (i.e., x_u, x_b, x_s, x_g corresponding to the items of top, bottom, shoes, bag), the feature extractor can represent each item with a tensor $Z \in R^{w\times h\times d}$, where w, h, d are the width, height, and channel of the output. The 3D tensor can be transformed to 2D embeddings with the size of $p \times d$, where $p = w \times h$. Then we have outfit embeddings as $o \in R^{4p\times d}$, where each 1D vector $v \in R^{1\times d}$ from o is the representation of a sub region that can map back to the patches in the original image space.

Fashion attribute embedding We denote the attribute embeddings of top, bottom, shoes, and bag as $l_u = \{l_1, l_2, \dots, l_{c_u}\}$, $l_b = \{l_1, l_2, \dots, l_{c_b}\}$, $l_s = \{l_1, l_2, \dots, l_{c_s}\}$ and $l_g = \{l_1, l_2, \dots, l_{c_g}\}, l_i \in R^d$. The label embeddings represent the possible labels of the actual one-hot labels that are learned from the corresponding embedding layers of the size of $d \times c_u, d \times c_b, d \times c_s, d \times c_g$, respectively.

Similar to the mask embedding strategy in Lanchantin et al. (2021), the state embeddings are incorporated into the label embeddings to obtain the masked label embeddings as

$$\tilde{l}_i = l_i + s_i, \tag{2}$$

where $s_i \in R^d$ is the state embedding of the label l_i, and s_i comes from the possible states: negative (N), positive (P), and unknown (U). The state embeddings are learned from the learnable embedding layer with a size of $d \times 3$. With the state embedding, the label information can be fully or partially considered during the training stage by controlling the ratio of the labels with P or N state via label mask training. The known labels are randomly selected from the training sets, and the ratio is within $\{0, 0.75\}$ while the unknown labels are within $\{0.25, 1\}$. The masked label embeddings are obtained for the known labels based on the initial label embeddings and the corresponding ground truth state embeddings. The binary cross entropy is employed to compute the difference between the ground truths and the predicted fashion attributes. Then, we can have the binary cross entropy losses of $\mathcal{L}_u, \mathcal{L}_b, \mathcal{L}_s, \mathcal{L}_g$ as

$$\mathcal{L}_u = \sum_{n=1}^{N} E^u_{(y_k)}\{CE(l_n, y_n)|y_k\}, \tag{3}$$

$$\mathcal{L}_b = \sum_{n=1}^{N} E^b_{(y_k)}\{CE(l_n, y_n)|y_k\}, \tag{4}$$

$$\mathcal{L}_s = \sum_{n=1}^{N} E^s_{(y_k)}\{CE(l_n, y_n)|y_k\}, \tag{5}$$

$$\mathcal{L}_g = \sum_{n=1}^{N} E^g_{(y_k)}\{CE(l_n, y_n)|y_k\}, \tag{6}$$

where $E^u_{(y_k)}\{\cdot|y_k\}$, $E^b_{(y_k)}\{\cdot|y_k\}$, $E^s_{(y_k)}\{\cdot|y_k\}$, $E^g_{(y_k)}\{\cdot|y_k\}$ denote the expectation of fashion attributes regarding the probability distribution of known labels y_k of the fashion items of top, bottom, shoes and bag. $CE(\cdot)$ is the binary cross entropy loss function as

$$\begin{aligned} CE(l, y) = &-w_{n,c}[p_c y_{n,c} \cdot log\sigma(l_{n,c}) \\ &+(1 - y_{n,c}) \cdot log(1 - \sigma(l_{n,c}))], \end{aligned} \tag{7}$$

where $w_{n,c}$ is the weight for the nth image/outfit corresponding to the cth class, p_c is the weight of the positive prediction of the class, and $\sigma(\cdot)$ is the probability function. Training the model with randomly masked label embeddings can help the model discover potential correlations among different label combinations and generalize to the prediction cases with an arbitrary amount of known information.

Physical attribute embedding For an outfit, personalized fashion compatibility regarding physical attributes predicts a set of labels that indicate the incompatible physical attributes. The physical attribute embeddings are denoted as $l_p = \{l_1, l_2, \dots, l_{c_p}\}, l_i \in R^d$ with the size of $d \times c_p$. Similar to the binary cross entropy losses of $\mathcal{L}_u, \mathcal{L}_b, \mathcal{L}_s, \mathcal{L}_g$, the optimization loss of physical attribute embeddings is

$$\mathcal{L}_p = \sum_{n=1}^{N} E^p_{(y_k)}\{CE(l_n, y_n)|y_k\}, \tag{8}$$

where $E^p_{(y_k)}\{\cdot|y_k\}$ is the expectation of physical attributes regarding the probability distribution of the corresponding known labels.

Transformer encoder Given feature embeddings that are learned from the feature extractor and the label embeddings that are randomly initialized, the transformer encoder similar to Lanchantin et al. (2021) is applied to learn the correlations between the features and label embeddings. The transformer is proved effective and suitable for capturing the dependencies between the general image features and the label embeddings as it is ordered invariant and the weight of pairwise feature-label embeddings can be learned by the self-attention mechanism inside (Vaswani et al., 2017).

For the fashion items of top, bottom, shoes and bag in the outfit, we have the corresponding combined embeddings as

$$H_u = \{z_1^u, z_2^u, \dots, z_{w\times h}^u, l_1^u, l_2^u, \dots, l_{c_u}^u\}, \tag{9}$$

$$H_b = \{z_1^b, z_2^b, \dots, z_{w\times h}^b, l_1^b, l_2^b, \dots, l_{c_b}^b\}, \tag{10}$$

$$H_s = \{z_1^s, z_2^s, \dots, z_{w\times h}^s, l_1^s, l_2^s, \dots, l_{c_s}^s\}, \tag{11}$$

$$H_g = \{z_1^g, z_2^g, \dots, z_{w\times h}^g, l_1^g, l_2^g, \dots, l_{c_g}^g\}. \tag{12}$$

For the outfit towards physical attributes, we have

$$H_p = \{z_1^p, z_2^p, \dots, z_{w\times h}^p, l_1^p, l_2^p, \dots, l_{c_p}^p\}. \tag{13}$$

The pair-wise weight of $\{h_i, h_j\}$ denoted as $a_{i,j}$ can be learned by self-attention in the transformer encoder. Specifically, the attention weight $a_{i,j}^t$ of the ith embedding and the jth embedding at the t-step can be obtained with the following procedure:

(1) compute the normalized scalar attention coefficient with

$$a_{i,j} = softmax((W^q h_i)^T (W^k h_j)/\sqrt{d}); \tag{14}$$

(2) update each embedding h_i to $\bar{h}_i$ with the weighted sum of all related embeddings via

$$\bar{h}_i = \sum_{j=1}^{M} a_{i,j} W^v h_j; \tag{15}$$

(3) finally obtain the activated embedding $\widetilde{h}_i$ with

$$\widetilde{h}_i = ReLU(\bar{h}_i W^r + b_1)W^o + b_2, \tag{16}$$

where M is the number of data pairs related to h_i, W^q, W^k, W^v is the query, key, and value matrices respectively, W^r, W^o are transformation matrices, and b_1, b_2 are bias vectors. The transformer encoder is composed of L layers with the same structures, and each layer has its respective variables.

The fashion item feature embeddings $\{z_u, z_b, z_s, z_g\}$ and the corresponding label embeddings $\{l_u, l_b, l_s, l_g\}$ are fed to the successive layers of the transformer encoder. The first transformer encoder in the proposed framework learns the correlation between the fashion items and the corresponding fashion attributes, and the learned attribute label embeddings are then fed to the successive classifier for attribute prediction.

The initial physical label embeddings l_p and the learned feature embeddings $\{h_u,\ h_b,\ h_s,\ h_g\}$ as well as the learned attribute label embeddings $\{l'_u, l'_b, l'_s, l'_g\}$ from the first transformer encoder are stacked and fed to the second transformer in the proposed framework to learn the potential dependencies. Finally, the transformed physical label embeddings are input to the successive classifier for physical label prediction.

Multi-label classifier As shown in the right-hand side of Fig. 2, the obtained label embeddings $\{l'_u, l'_b, l'_s, l'_g, l'_p\}$ are fed to the feedforward network (FFN) for attribute prediction, which can be formulated as

$$y_i^u = \mathrm{FFN}_u(l_i'^u) = s(w_i^u \cdot l_i'^u + b_i^u), \tag{17}$$

$$y_i^b = \mathrm{FFN}_b(l_i'^b) = s(w_i^b \cdot l_i'^b + b_i^b), \tag{18}$$

$$y_i^s = \mathrm{FFN}_s(l_i'^s) = s(w_i^s \cdot l_i'^s + b_i^s), \tag{19}$$

$$y_i^g = \mathrm{FFN}_g(l_i'^g) = s(w_i^g \cdot l_i'^g + b_i^g), \tag{20}$$

$$y_i^p = \mathrm{FFN}_p(l_i'^p) = s(w_i^p \cdot l_i'^p + b_i^p), \tag{21}$$

where $s(\cdot)$ is a sigmoid function and w_i, b_i is the weight of label i with size of $1 \times d$ and the corresponding bias.

3.3.2. Online multi-label smoothing

Label smoothing utilizes soft labels generated from a uniform distribution to take the place of hard labels to reduce the overfitting problem for model training. It can be used to improve classification performance, especially under the case when the class label is imbalanced (Szegedy et al., 2016; Tzelepi et al., 2021). Unlike label smoothing, which uses a static soft label, Zhang et al. proposed to use model predictions to continuously update the soft labels during the training process for single class prediction (Zhang et al., 2021). In this paper, we propose to exploit model predictions for online label smoothing for the multi-label prediction case. If the prediction is correct for a given image, the soft labels corresponding to the target classes will be updated, and the updated soft labels will be applied to supervise the training of the model.

For the physical labels, we denote the class number as c_p, suppose the total training epochs is T, then we have the soft labels $S = \{S^0, S^1, \dots, S^T\}$, where $S^t \in R^{c_p \times c_p}$ is the soft label for the tth epoch. When $t = 0$, the soft label S^t is initialized as zero matrices. Given the ith outfit ($i = 1, 2, \dots, N$), $S^{t-1}_{y_i^p = y_i^{gt}}$ is denoted as the soft label that is correctly predicted regarding the ground truth label of the outfit, the training loss of the model supervised by $S^{t-1}_{y_i}$ can be formulated as

$$L_{ls} = -\sum_{i=1}^{N} S^{t-1}_{y_i^p = y_i^{gt}} \cdot \log p(i|(y_i^p = y_i^{gt}, o)), \tag{22}$$

where $p(i|(y_i^p = y_i^{gt}, o))$ is the prediction score corresponding to the correctly predicted labels regarding the ground truth labels with the outfit feature o. $S^t_{y_i^p = y_i^{gt}}$ forms a temporary label distribution to supervise the model training, and it can be updated by

$$S^t_{y_i^p = y_i^{gt}} = S^{t-1}_{y_i^p = y_i^{gt}} + p(i|(y_i^p = y_i^{gt}, o)). \tag{23}$$

At the end of the tth training epoch, the cumulative S^t is normalized as

$$S^t_{y_i^p = y_i^{gt}} = \frac{S^{t-1}_{y_i^p = y_i^{gt}}}{\sum_{i=1}^{N} S^{t-1}_{y_i^p = y_i^{gt}}}. \tag{24}$$

The normalized soft label $S^t_{y_i^p = y_i^{gt}}$ over N outfits is obtained and will be used for supervising the training of the model at the next epoch.

The overall loss is derived from Eqs. (4), (5), (6), (8) and (22), which is summarized as

$$\mathcal{L} = \mathcal{L}_u + \mathcal{L}_b + \mathcal{L}_s + \mathcal{L}_g + \mathcal{L}_p + \lambda \mathcal{L}_{ls}, \tag{25}$$

where λ is a parameter to balance the online multi-label smoothing loss towards other losses. The parameters of the proposed model can be optimized via backpropagation in an end-to-end manner.

Although the classification transformer (C-Tran) (Lanchantin et al., 2021) also uses transformer encoder for class prediction, it is different from the proposed method. First, C-Tran only considers visual features for physical compatibility learning while the proposed method jointly considers the correlation of the visual features, the outfit attributes and the physical attributes for comprehensive compatibility learning. Meanwhile, the fashion attribute recognition losses are optimized during the training process, thus the model can learn more effective attribute featrues for the later physical compatibility learning. Second, the proposed model applies the online multi-label smoothing strategy to deal with the physical label imbalance problem, by which the prediction performance can be improved.

4. Experiments

In this section, we conduct extensive experiments to evaluate the performance of the proposed method by answering the following research questions:

RQ1: Does the proposed method achieve better performance compared with state-of-the-art methods on personalized compatibility evaluation?

RQ2: How is the performance of the proposed method for attribute prediction?

RQ3: How does the proposed method recognize the fashion attributes with the attention mechanism?

RQ4: How does the proposed method perceive the outfit compatibility towards the physical attributes?

RQ5: Can we explore some potential correlations between the fashion attributes and the physical attributes with the proposed method?

RQ6: How is the performance when the proposed method conducts compatibility evaluation in different cases?

4.1. Experimental design

In this section, the experimental design including implementation details, the evaluation metrics, the dataset, and the comparison methods will be introduced.

4.1.1. Implementation details

For a fair comparison, the proposed and compared methods use the same feature extractor, i.e. Resnet18, Resnet50, and Resnet101 that were pretrained on ImageNet with corresponding output dimension $d = 512, 2048, 2048$. The size of the input images was cropped and resized to 112×112. The output 3D tensor from the feature extractor is $7 \times 7 \times d$, and we can have feature embeddings of each fashion item as $z \in R^{49\times d}$. The label embeddings also have the same dimension as the feature embeddings. Specifically, the input embeddings for the first transformer encoder for top, bottom, shoes, bag fashion attribute prediction are $H'_u \in R^{(49+c_u)\times d}$, $H'_b \in R^{(49+c_b)\times d}$, $H'_s \in R^{(49+c_s)\times d}$, $H'_g \in R^{(49+c_g)\times d}$ respectively, where c_u, c_b, c_s, c_g is the corresponding attribute number. The input embedding for the second transformer encoder is $H'_p \in R^{(49\times k+C_p)\times d}$, where k is the number of fashion items in the outfit and c_p is the number of the physical attributes.

The number of the attention head and layer of the transformer encoder is 4 and 3, respectively, which is similar to the previous works (Lanchantin et al., 2021). The proposed model can be trained end-to-end, and the variables except the parameter λ can be optimized by the backpropagation. In the experiments, the optimizer of the proposed method is SGD with weight decay of 4×10^{-4}. The learning rate is 2×10^{-4} and the dropout with $p = 0.1$ is used for avoiding overfitting. The batch size is 16, and the maximum training epoch of all methods is 50. The parameters of the comparison methods are set as default according to the original papers. We conducted the experiments on a desktop PC with OS as Ubuntu 20.04.2 LTS, CPU as Intel(R) Core(TM) i7-8700K @3.70 GHz with 11 processors and 32 GB memory, GPU as NVIDIA GTX2080 with 8 GB memory.

4.1.2. Baselines and evaluation metrics

The task of compatibility prediction of an outfit regarding the individual physical attributes can be considered as a multi-label classification problem. Several relevant state-of-the-art multi-label classification methods are used for comparison in the experiments. The methods include classification transformer (C-Tran) (Lanchantin et al., 2021), Modular Graph Transformer Networks (MGTN) (Nguyen et al., 2021), Class-specific Residual Attention (CSRA) (Zhu & Wu, 2021), Multi-class Attentional Regions (MCAR) (Gao & Zhou, 2021), Multi-modal multi-label recognition (M3TR) (Zhao, Zhao, & Li, 2021), Asymmetric loss for multi-label classification (ASL) (Ridnik et al., 2021), and Transformer-based dual relation graph (TDRG) (Zhao et al., 2021).

Table 1
The physical label distribution on O4U dataset.

Aspect	Attribute name	Frequency of occurrence
Body figure	triangle	10,710
	spoon	9,058
	bottom_hourglass	8,913
	top_hourglass	1,486
	inverted_triangle	4,620
	round	10,116
	diamond	9,872
	hourglass	20
	rectangle	41
	athietic	4,407
Skin color	yellow	1,886
	dark	1,940
	fair	247
	brown	2,273
Hair style	long_curls	11
	long_straight_hair	19
	middle_long_curls	9
	middle_long_straight_hair	18
	short_curls	21
	short_straight_hair	33
Hair color	ginger	721
	black	766
	dark_brown	239
	light_brown	1,404
	gray/silver	2,098
	golden	36
Height	high	2,538
	middle	304
	low	1,927
Breasts	big	4,362
	average	4
	small	348
Color contrast	high	301
	low	1,805

C-Tran is a general framework that exploits the dependencies between visual features and the labels for multi-label classification problems. The method first uses a label mask training objective with label states of positive, negative, or unknown for model training. The method is proved to be more general and robust for multi-label classification when only partial or extra-label annotations are available.

MGTN employs multiple backbones for different sub-graphs derived from graph transformers and convolutions to learn better representation for classification performance improvement. The method applies several strategies to integrate object labels' semantic and network properties to the multi-label classification problem.

CSRA applies class-specific residual attention to obtain class-specific features of each category via combining the spatial attention scores with the class-agnostic average pooling features. The method can effectively learn different spatial regions corresponding to objects from different categories for effective multi-label image classification.

MCAR uses a two-stream framework to distinguish multi-category objects of local regions from global images. It aims to use attentional regions as few as possible and simultaneously keep the diversity of these regions as high as possible for efficient and effective multi-class object recognition.

M3TR considers the relations of visual structures and multi-modality information for multi-label classification. It combines CNNs and Transformers to learn the semantic cross-attention to embed visual structures into the high-level features for intra-modal relationship learning. It also uses a linguistic cross-attention to obtain high-level semantic representation with a linguistic-guided enhancement module to learn the interactions between the visual and linguistic modalities.

ASL utilizes an asymmetric loss scheme to treat the positive and negative samples differently to balance the probabilities of different samples for dealing with the high negative-positive imbalance and ground-truth mislabeling challenges in the multi-label classification task.

TDRG proposes a Transformer-based Dual Relation Graph to construct complementary relationships with structural and semantic relation graphs. It incorporates the relationship into the semantic graph to construct a joint relation graph for feature representation learning.

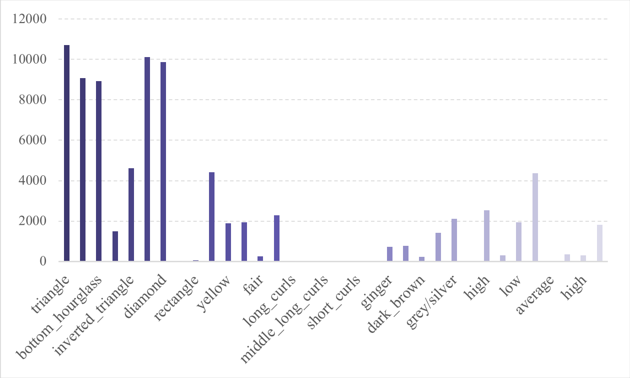

Fig. 3. Original physical label distribution.

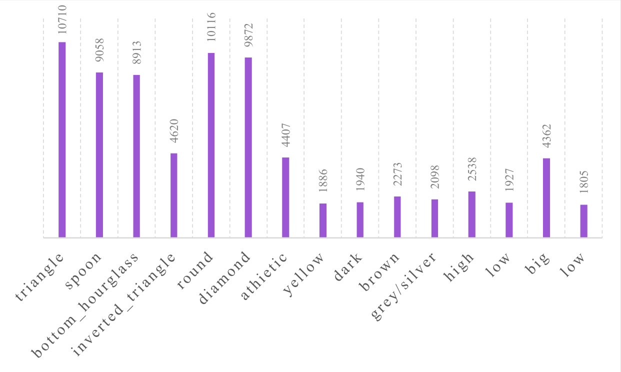

Fig. 4. Selected physical label distribution with rate≥ 0.1.

For a fair comparison, the general evaluation metrics (Chen, Wei, et al., 2019; Ge et al., 2018; Zhu et al., 2017), such as mean average precision (mAP), average per-class precision (CP), recall (CR), F1 (CF1), and the average overall precision (OP), recall (OR), F1 (OF1) and the average weighted precision (WP), weighted recall (WR), and weighted F1 (WF1), are employed to evaluate the performance of the proposed method and the comparison methods.

4.1.3. Dataset

The O4U dataset (Kaicheng Pang, 2022) is used to evaluate the performance of the proposed method on incompatible physical label prediction regarding a giving outfit. The outfit numbers of training, validation, and testing sets in the experiment are 11,023, 3149, and 1575, respectively. The physical attributes on the dataset are categorized as body figure, skin color, hairstyle, hair color, height, breasts, and color contrast. Each aspect has corresponding detailed attributes. For clarity, the labels of each aspect are listed in Table 1 and the corresponding label distribution is presented in Fig. 3.

From Table 1, we can see that the frequency of occurrence of some attributes is low while others are relatively high. Fig. 3 indicates that the label distribution of the dataset is imbalanced, and directly conducting multi-label classification on the dataset will be very challenging. More importantly, the low frequency of some labels indicates that the labels are not significant to the compatibility of an outfit for an individual with specific physical information. For example, the labels from the aspect of hair style rarely occur, which indicates that the probability of individual incompatibility tends to have little connection with this kind of label. Thus, it is reasonable to ignore such labels and train a model focusing on the important labels for the evaluation of physical incompatibility. To this end, we calculate the occurrence rate of each label and select the labels with the rate ≥ 0.1. The number of selected labels is $c_p = 15$, and the selected label distribution is shown in Fig. 4, from which we can see that the label distribution is smoother than that in Fig. 3, but multi-label classification on the selected data is also challenging as the label distribution is still imbalanced.

Table 2
Performance comparison based on Resnet18 backbone.

Method	mAP	CP	CR	CF1	OP	OR	OF1
Resnet18	37.28	22.76	19.42	20.96	65.64	38.69	48.68
C-Tran	32.15	20.13	33.33	25.10	59.86	63.70	61.72
MGTN	35.98	31.75	37.89	34.55	55.01	59.78	57.29
CSRA	41.41	39.79	33.46	36.35	65.26	60.15	62.60
MCAR	39.86	40.41	30.34	34.66	64.15	51.42	57.09
M3TR	35.99	20.25	33.28	25.18	60.70	63.61	62.12
ASL	37.97	20.34	33.20	25.20	61.06	63.46	62.24
TDRG	38.17	40.16	33.98	36.81	60.14	54.26	57.05
Ours	**47.99**	**42.39**	**40.42**	**41.38**	**66.54**	**66.48**	**66.51**

Table 3
Performance comparison based on Resnet50 backbone.

Method	mAP	CP	CR	CF1	OP	OR	OF1
Resnet50	35.12	23.59	14.36	17.85	64.92	29.10	40.18
C-Tran	39.36	28.92	29.66	29.29	61.78	51.31	56.06
MGTN	37.20	34.22	28.76	31.25	56.35	43.15	48.87
CSRA	42.87	38.57	34.01	36.15	66.02	60.76	63.28
MCAR	40.78	43.09	32.88	37.30	63.74	55.27	59.20
M3TR	34.93	20.22	33.33	25.17	60.59	**63.70**	62.10
ASL	39.74	32.19	32.07	29.62	64.0	55.29	59.33
TDRG	38.58	20.65	32.40	25.22	61.96	62.07	62.01
Ours	**47.65**	**43.47**	**40.85**	**42.14**	**67.72**	63.02	**65.28**

Table 4
Performance comparison based on Resnet101 backbone.

Method	mAP	CP	CR	CF1	OP	OR	OF1
Resnet101	35.01	21.79	22.66	22.22	64.67	44.72	52.88
C-Tran	36.74	30.33	30.13	30.23	62.85	56.24	59.36
MGTN	36.79	27.73	27.84	27.78	61.64	46.26	52.85
CSRA	42.65	39.64	34.75	37.04	65.71	61.28	63.42
MCAR	39.31	33.48	30.76	32.06	65.45	61.15	30.76
M3TR	34.61	20.22	33.23	25.14	60.57	63.53	62.02
ASL	39.34	37.10	30.89	29.59	62.90	53.74	57.96
TDRG	40.40	36.09	30.98	33.34	65.04	54.66	59.40
Ours	**47.17**	**43.31**	**42.02**	**42.66**	**66.93**	**63.82**	**65.34**

4.2. Personalized compatibility evaluation (RQ1)

To evaluate the performance of the proposed method and the comparison method in terms of personalized compatibility prediction, we conducted extensive experiments on the O4U dataset. The experimental results of all methods based on the backbones of Resnet18, Resnet50, and Resnet101 on the O4U dataset are shown in Tables 2–4, respectively. From the results, we can make the following observations:

(1) The proposed method can obtain better performance with a large margin for almost all metrics than the comparison methods, demonstrating the superiority of the proposed framework. The potential reason is that the two transformer encoders of the proposed method take random amounts of known labels for training and it can discover the potential correlations among different feature and label combinations and generalize to the complicate prediction during the testing stage. Additionally, different from C-Tran which also uses transformer mechanism, the proposed method utilizes the second transformer to learn the independence among the initial physical attribute embeddings and the visual feature and attribute embeddings that obtained from the first transformer encoder, to improve the prediction performance.

(2) Compared to the Resnet baseline, all the other methods can achieve better performance, especially on the CF1 and OF1, which are the most important metrics (Chen, Wei, et al., 2019). This indicates the effectiveness of the additional components in these methods. For

Table 5
Fashion attribute prediction results of the proposed method based on different backbones.

Resnet18					Resnet50				Resnet101			
Parts	mAP	WP	WR	WF1	mAP	WP	WR	WF1	mAP	WP	WR	WF1
U	67.07	58.31	62.22	60.2	71.70	65.14	65.70	65.42	73.34	67.16	67.74	67.45
B	77.07	71.02	72.59	71.8	83.53	75.49	76.27	75.88	85.32	76.88	78.20	77.53
S	82.21	77.43	77.04	77.23	85.49	80.8	79.70	80.24	87.41	83.03	81.71	82.36
G	73.45	69.30	70.27	69.78	76.16	73.62	69.64	71.57	79.41	73.94	75.1	74.51

Table 6
The performance of the proposed method with different inputs fed to the transformer encoder.

Input	mAP	CP	CR	CF1	OP	OR	OF1
Ours_cs1	41.1	27.8	26.3	27	62.6	46.7	53.5
Ours_cs2	44.9	38.4	36.4	37.4	66.1	60.1	63
Ours_cs3	33	15.9	26.7	19.9	59	50.4	54.3
Ours	47.17	43.31	42.02	42.66	66.93	63.82	65.34

the proposed method, the potential reason for the high performance is that the correlations among the visual features, the fashion attributes, and the physical attributes are explored by the transformer encoder for learning more effective representations for multi-label classification.

(3) Although the performance of all methods is hard to satisfy the real application of online personalized fashion compatibility evaluation, the proposed work in this paper is still necessary and valuable as it provides a new viewpoint to consider the personalized compatibility towards individual physical information, outfit visual features, and the corresponding fashion item attributes, which is consistent with human perception of fashion aesthetics.

4.3. Fashion attribute prediction (RQ2)

To evaluate the effectiveness of the proposed method on fashion attribute prediction, we use attribute label embeddings of the top, bottom, shoes, and bag that learned from the first transformer encoder in the proposed framework for multi-fashion-attribute prediction on the corresponding item images. The experimental results of the proposed method based on Resnet18, Resnet50, and Resnet101 are shown in Table 5, where the mAP, WP, WR, and WF1 are used as they can calculate the weighted average result to account for the attribute label imbalance problem. From Table 5, we can know that the proposed method is effective for fashion attribute classification, and tends to obtain better performance when the neural network gets deeper.

4.4. Outfit-attribute attention explanation (RQ3-4)

RQ3-4 is too abstract to answer with quantitative metrics, and we explore the potential effect of the attention mechanism in the proposed method on attribute prediction and compatibility evaluation via attention visualization. In Figs. 5, 11(a), 11(b), 12(a) and 12(b) (the figures except Fig. 5 are shown in the appendix due to page limitation), the results on the left show the predicted fashion attributes of top, bottom, shoes, and bag and the corresponding visual attention regions over the images, while the results on the right show the predicted incompatible physical labels towards the outfit with corresponding attention regions. From the results, we can draw the following interesting points:

(1) According to the predicted fashion attributes and the corresponding highlighted attention regions, we can know that the attention mechanism tends to locate the key regions of the images with distinct patterns, colors, or shadings. For example, the pants can be recognized by highlighting the crotch of the pants in the second row in Figs. 5 and 12(a) while skirts in the second row in Figs. 11(a) and 11(b) have no such information.

(2) However, locating the aforementioned key features may mislead the model into generating fake predictions, especially when two fashion attributes have a similar property. For example, the silhouette of the pants in Fig. 5 is “A” while the model predicts it as “H” due to the H-like attention shape. In addition, the model tends to predict most fashion attributes as the attributes that have high occurrence frequency. For example, the neckline of the top item in Fig. 12(b) is predicted as “R,” but it is “Turtle”. Although the model has correctly highlighted the neckline region of the top item, it provides a fake prediction as the “R” attribute occurs more frequently than the “Turtle” attribute.

(3) The figures show that the physical attribute prediction tends to be more related to body shape than other aspects. This is reasonable as the aspect of the body figure is of high occurrence frequency on the dataset, as shown in Fig. 4. The highlighted attention regions of the body figure in Figs. 5 and 12(a) indicate that the attention regions can perceive the shape or edge of the items while the highlighted attention regions of skin color tend to be related to the color or textures. The observation is consistent with human aesthetics as we usually consider silhouette characteristics for a given body shape and perceive the visual color matching for individual skin color.

4.5. Potential relationship towards fashion attributes and physical attributes (RQ5)

The compatibility of fashion attributes and physical attributes can be intuitive or inexplicit under different situations. To explore the potential relationship between fashion attributes and physical attributes, Figs. 6 and 7 show the heatmaps with fashion attributes marked on the horizontal axis and the physical attributes marked on the vertical axis. The heatmap is extracted from the proposed framework’s attention matrix from the last transformer. Although defining explicit matching rules for the physical attributes and different fashion attributes is challenging, we still can observe some interesting points as the following:

(1) In Fig. 6, the body shapes of the spoon and bottom hourglass have similar sensitivity to the fashion attributes of the top category. This is reasonable as they have similar characteristics that the upper body looks slimmer than the lower body. Interestingly, the body shapes of triangles and inverted triangles are completely opposite and look intuitively different. A user with a triangle body shape is recommended to try a T-shirt with a slim silhouette, while the user with an inverted triangle body shape should avoid such kinds of items. In Fig. 6, the highlighted regions on the right show that the inverted triangle body shape is sensitive to the slim silhouette attribute while the triangle body shape is not so relevant.

(2) Additionally, the body shapes of triangle, spoon, bottom hourglass, and inverted triangle present various sensitivity in the bottom silhouette, especially when the silhouette is slim. As highlighted in Fig. 7, the slim silhouette has a strong effect on the compatibility of the triangle, spoon, and bottom hourglass, while the inverted triangle tends to fit the attribute more easily. This is consistent with our human perception of body shape characteristics and silhouette design.

(3) The highlighted row in Fig. 6 indicates that a tall user tends to easily fit an item with any fashion attributes while the short user may face difficulty doing so. Both Figs. 6 and 7 indicate that skin color and hair color are more sensitive to different fashion attributes than the height. It is because color matching usually plays an important role in outfit compatibility learning. The variations in skin colors or hair colors may affect the overall visual compatibility, and thus they are more sensitive to different fashion attributes.

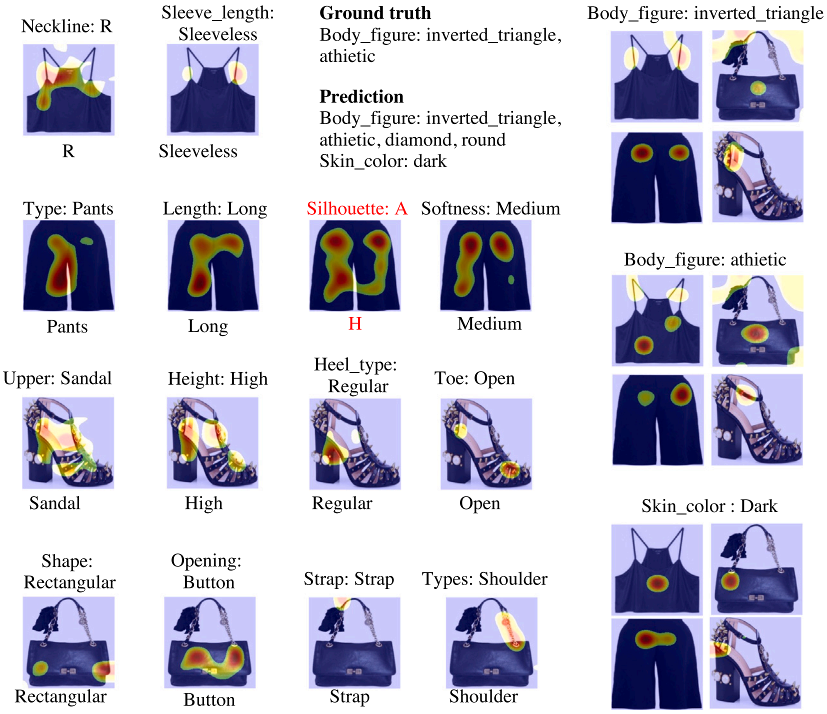

Fig. 5. Example: attention visualization on fashion attribute and personalized outfit compatibility.

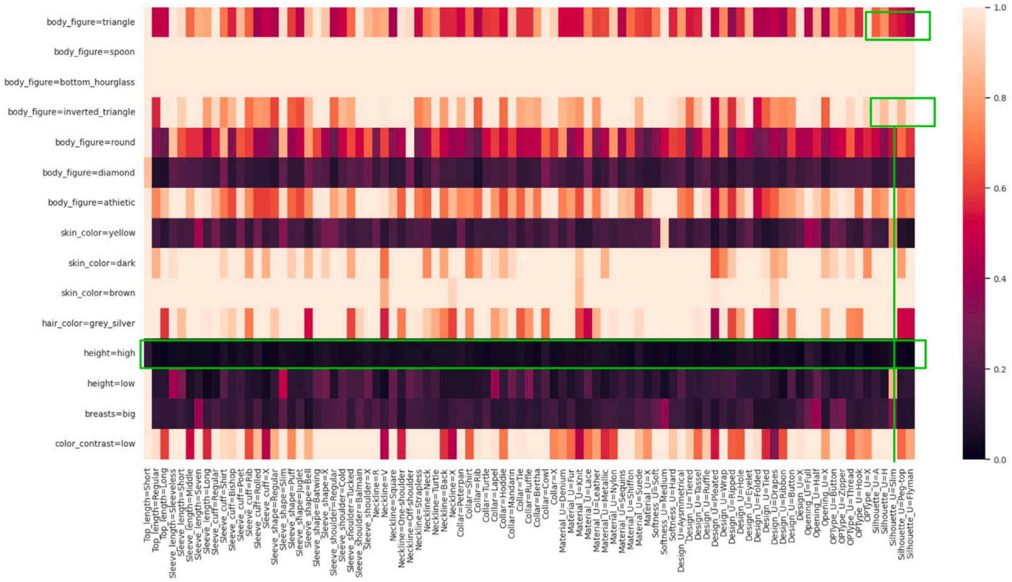

Fig. 6. The heatmap visualization of the physical attributes towards fashion attributes of top category.

4.6. Ablation study (RQ6)

In this section, the performance of the proposed method is examined under different cases for personalized outfit compatibility prediction.

(1) Fig. 8 presents the CF1 of the proposed method with various values of the parameter λ. The curve in Fig. 8 keeps improving with increasing values at the beginning and it reaches the peak when the value of λ is 1.0. For simplicity, we set $\lambda = 1.0$ for the proposed method in all experiments.

(2) For the objective function in Eq. (25), the overall loss composes of attribute loss, personalized loss, and multi-label smoothing loss. To explore the effect of each component, we denote the following variations of the proposed method:

Ours_var1: There has no attribute loss and smoothing loss in the objective loss function.

Ours_var2: There has smoothing loss but no attribute loss in the objective loss function.

Ours_var3: There has to attribute loss but no smoothing loss in the objective loss function.

The performance of the variations is shown in Fig. 9 and the result indicates the effectiveness of each component of the proposed method.

(3) The transformer encoder plays an important role in learning the correlation between the visual feature embeddings and the label embeddings. To explore how different inputs of the transformer affect the performance of physical label prediction, we use different combinations of the embeddings to feed to the transformer.

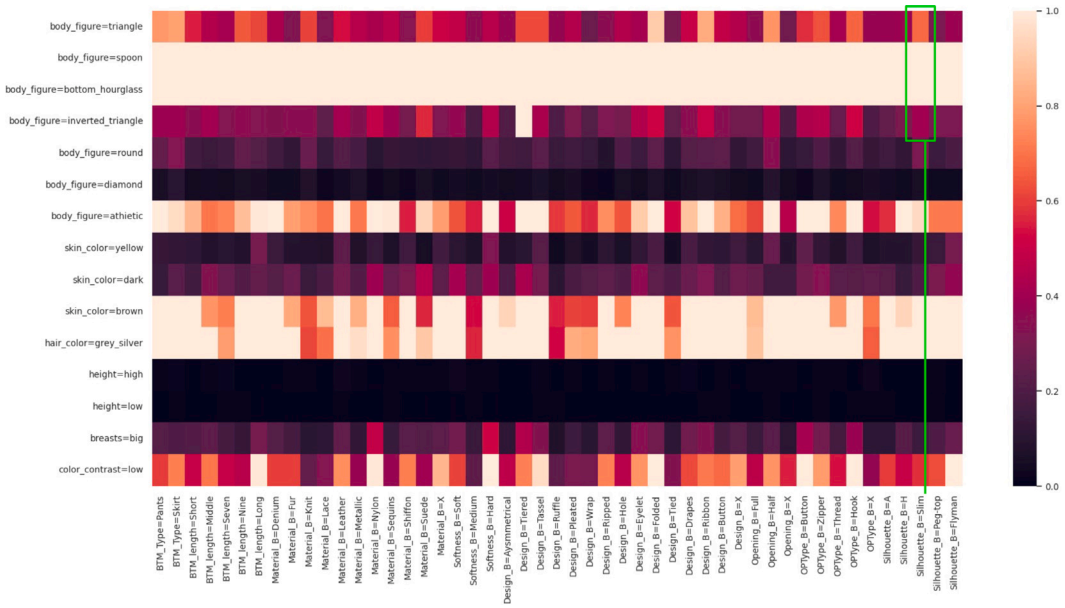

Fig. 7. The heatmap visualization of the physical attributes towards fashion attributes of bottom category.

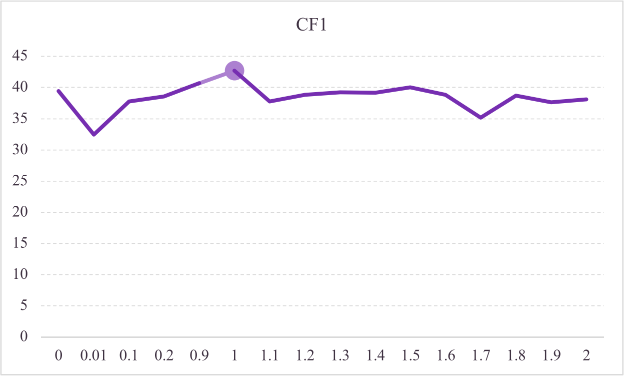

Fig. 8. The performance of the proposed method with different values of the parameter λ.

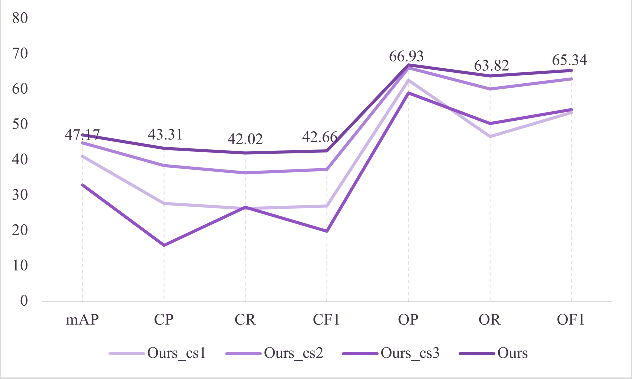

Fig. 10. Performance of the proposed method with different inputs fed to the transformer encoder.

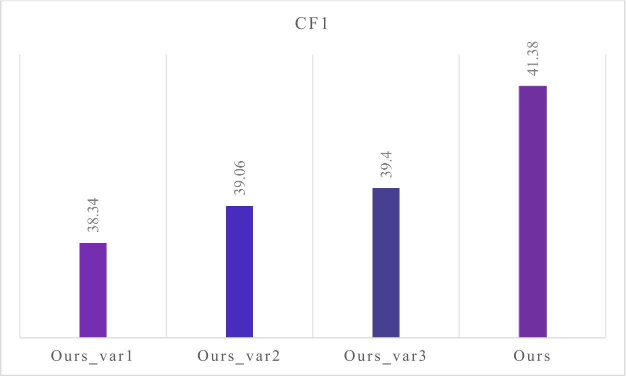

Fig. 9. The performance of different variations of the proposed method.

Table 6 shows the detailed metric values of the proposed method with different inputs fed to the transformer, where Ours_cs1, Ours_cs2, Ours_cs3 and Ours corresponds to the cases when the input is $[l'_{att}, l_p]$, $[h, l_p]$, $[z, l'_{att}, l_p]$, and $[h, l'_{att}, l_p]$ for the last transformer encoder, l_p denotes the initial physical label embedding, l'_{att}, h denote attribute label embedding and feature embedding after the first transformer encoder, z is the feature embedding after the Resnet feature extractor. Fig. 10 demonstrates the results for intuitive comparison of different cases. From Table 6 and Fig. 10, we can see that the performance of the proposed method is affected by different combinations of the embeddings. Comparing with Ours_cs1 and Ours_cs2, the proposed method presents high superiority with $[h, l'_{att}, l_p]$ fed to the last transformer, which indicates that exploring the correlation among the learned embeddings of visual features, fashion attributes and physical labels with the seconder transformer encoder in Fig. 2 is effective for personalized outfit compatibility prediction.

5. Conclusion

This paper proposes a personalized outfit compatibility prediction method, which can be regarded as a multi-label classification problem given a set of fashion images from different categories. The method considers the personalized outfit compatibility problem from a new viewpoint, while the existing methods usually consider the personalized outfit learning from the fashion item with specific patterns or attributes or the users' visual preference to achieve private stylist. Instead, we propose connecting outfit compatibility with individual physical attributes, such as body figure, hairstyle, skin color, etc., for complete

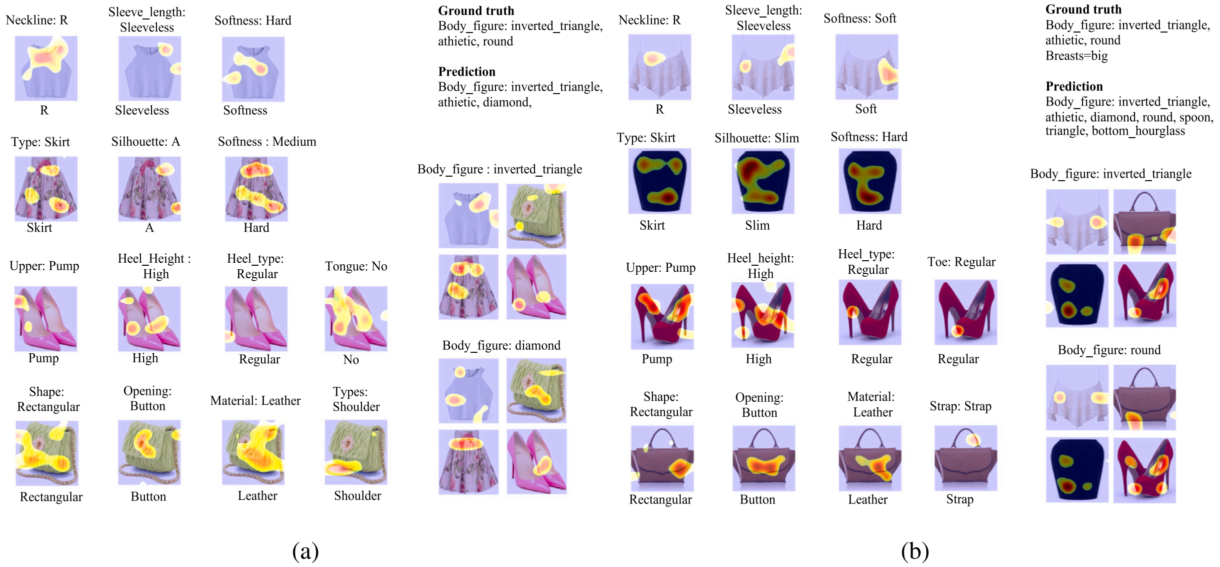

Fig. 11. Examples: attention visualization on fashion attribute and personalized outfit compatibility.

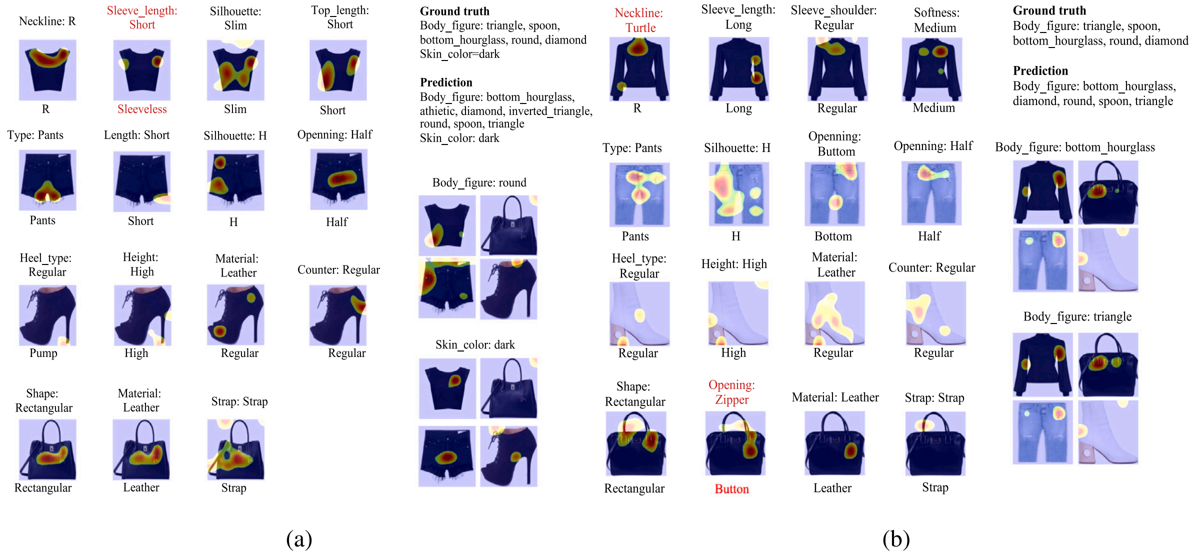

Fig. 12. Examples: attention visualization on fashion attribute and personalized outfit compatibility.

compatibility learning. It is the first attempt to solve the personalized outfit compatibility evaluation problem via interacting with physical information and fashion attributes. To evaluate the performance of the proposed method on incompatible physical label prediction over an outfit, we conduct extensive experiments on the O4U dataset. The quantitative and qualitative results on the dataset verified the superiority of the proposed method compared with state-of-the-art methods.

However, since the attribute distribution of fashion items and the physical label distribution of outfits are imbalanced, the personalized multi-label classification task is challenging and the performance of the proposed method as well as the comparison methods are not satisfying. In the future, on the one hand, we need to collect more outfit data with rich fashion attribute and physical attribute annotations by supervised or unsupervised techniques to release the label imbalance problem. On the other hand, we need to develop an enhanced model which can discover the fashion aesthetics relationship among visual fashion elements and personalization information for effective personalized fashion compatibility learning.

CRediT authorship contribution statement

Dongmei Mo: Conceptualization, Methodology, Formal analysis, Writing – original draft. **Xingxing Zou:** Conceptualization, Writing – review & editing. **Kaicheng Pang:** Methodology, Writing – review & editing. **Wai Keung 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.

Data availability

Data will be made available on request.

Acknowledgments

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

Chen, X., Chen, H., Xu, H., Zhang, Y., Cao, Y., Qin, Z., & Zha, H. (2019). Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval* (pp. 765–774).

Chen, W., Huang, P., Xu, J., Guo, X., Guo, C., Sun, F., Li, C., Pfadler, A., Zhao, H., & Zhao, B. (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, Z.-M., Wei, X.-S., Wang, P., & Guo, Y. (2019). Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5177–5186).

Cui, Z., Li, Z., Wu, S., Zhang, X.-Y., & Wang, L. (2019). Dressing as a whole: Outfit compatibility learning based on node-wise graph neural networks. In *The world wide web conference* (pp. 307–317).

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255). Ieee.

Gao, B.-B., & Zhou, H.-Y. (2021). Learning to discover multi-class attentional regions for multi-label image recognition. *IEEE Transactions on Image Processing, 30*, 5920–5932.

Ge, W., Yang, S., & Yu, Y. (2018). Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1277–1286).

Guan, W., Jiao, F., Song, X., Wen, H., Yeh, C.-H., & Chang, X. (2022). Personalized fashion compatibility modeling via metapath-guided heterogeneous graph learning. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval* (pp. 482–491).

Guan, W., Wen, H., Song, X., Yeh, C.-H., Chang, X., & Nie, L. (2021). Multimodal compatibility modeling via exploring the consistent and complementary correlations. In *Proceedings of the 29th ACM international conference on multimedia* (pp. 2299–2307).

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).

Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., & Wang, Y. (2021). Transformer in transformer. *Advances in Neural Information Processing Systems*, *34*, 15908–15919.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Hidayati, S. C., Goh, T. W., Chan, J.-S. G., Hsu, C.-C., See, J., Wong, L.-K., Hua, K.-L., Tsao, Y., & Cheng, W.-H. (2020). Dress with style: Learning style from joint deep embedding of clothing styles and body shapes. *IEEE Transactions on Multimedia*, *23*, 365–377.

Hidayati, S. C., Hsu, C.-C., Chang, Y.-T., Hua, K.-L., Fu, J., & Cheng, W.-H. (2018). What dress fits me best? Fashion recommendation on the clothing style for personal body shape. In *Proceedings of the 26th ACM international conference on multimedia* (pp. 438–446).

Hu, Y., Yi, X., & Davis, L. S. (2015). Collaborative fashion recommendation: A functional tensor factorization approach. In *Proceedings of the 23rd ACM international conference on multimedia* (pp. 129–138).

Kaicheng Pang, W. W. (2022). Dress well via fashion cognitive learning. arXiv preprint arXiv:2208.00639.

Lanchantin, J., Wang, T., Ordonez, V., & Qi, Y. (2021). General multi-label image classification with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16478–16488).

Liu, S., Feng, J., Song, Z., Zhang, T., Lu, H., Xu, C., & Yan, S. (2012). Hi, magic closet, tell me what to wear!. In *Proceedings of the 20th ACM international conference on multimedia* (pp. 619–628).

Lu, Z., Hu, Y., Jiang, Y., Chen, Y., & Zeng, B. (2019). Learning binary code for personalized fashion recommendation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10562–10570).

Mo, D., Zou, X., & Wong, W. (2022). Neural stylist: Towards online styling service. *Expert Systems with Applications*, Article 117333.

Nguyen, H. D., Vu, X.-S., & Le, D.-T. (2021). Modular graph transformer networks for multi-label image classification. In *Proceedings of the AAAI conference on artificial intelligence (vol. 35)* (pp. 9092–9100).

Pardo-Sixtos, L. F., López-Monroy, A. P., Shafaei, M., & Solorio, T. (2022). Hierarchical attention and transformers for automatic movie rating. *Expert Systems with Applications*, Article 118164.

Ridnik, T., Ben-Baruch, E., Zamir, N., Noy, A., Friedman, I., Protter, M., & Zelnik-Manor, L. (2021). Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 82–91).

Song, X., Feng, F., Liu, J., Li, Z., Nie, L., & Ma, J. (2017). Neurostylist: Neural compatibility modeling for clothing matching. In *Proceedings of the 25th ACM international conference on multimedia* (pp. 753–761).

Song, X., Han, X., Li, Y., Chen, J., Xu, X.-S., & Nie, L. (2019). GP-BPR: Personalized compatibility modeling for clothing matching. In *Proceedings of the 27th ACM international conference on multimedia* (pp. 320–328).

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818–2826).

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).

Tzelepi, M., Passalis, N., & Tefas, A. (2021). Online subclass knowledge distillation. *Expert Systems with Applications*, *181*, Article 115132.

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* (pp. 390–405).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

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).

Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z.-H., Tay, F. E., Feng, J., & Yan, S. (2021). Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 558–567).

Zhan, H., & Lin, J. (2021). Pan: Personalized attention network for outfit recommendation. In *2021 IEEE international conference on image processing* (pp. 2663–2667). IEEE.

Zhang, C.-B., Jiang, P.-T., Hou, Q., Wei, Y., Han, Q., Li, Z., & Cheng, M.-M. (2021). Delving deep into label smoothing. *IEEE Transactions on Image Processing*, *30*, 5984–5996.

Zhao, J., Yan, K., Zhao, Y., Guo, X., Huang, F., & Li, J. (2021). Transformer-based dual relation graph for multi-label image recognition. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 163–172).

Zhao, J., Zhao, Y., & Li, J. (2021). M3tr: Multi-modal multi-label recognition with transformer. In *Proceedings of the 29th ACM international conference on multimedia* (pp. 469–477).

Zhu, F., Li, H., Ouyang, W., Yu, N., & Wang, X. (2017). Learning spatial regularization with image-level supervisions for multi-label image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5513–5522).

Zhu, K., & Wu, J. (2021). Residual attention: A simple but effective method for multi-label recognition. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 184–193).

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

Dongmei Mo received the Ph.D. degree from The Hong Kong Polytechnic University, Hong Kong. She is currently a postdoctoral fellow in The Laboratory for Artificial Intelligence in Design Limited (AiDLab). Her research interests are computer vision and AI in fashion.

Xingxing Zou received the Ph.D. degree from The Hong Kong Polytechnic University, Hong Kong. She is currently a technical project manager with the Laboratory for Artificial Intelligence in Design (AiDLab). Her major research interest is AI in fashion.

Kaicheng Pang: Kaicheng Pang is now pursuing the Ph.D. degree in School of Fashion and Textiles, The Hong Kong Polytechnic University. His current research interests are computer vision and AI in fashion.

Wai Keung Wong received the Ph.D. degree from The Hong Kong Polytechnic University, Hong Kong. He is currently a full Professor with The Hong Kong Polytechnic University and concurrently serving as the Centre Director of the Laboratory for Artificial Intelligence in Design. He has authored or co-authored more than 170 papers in refereed journals and conferences, including the IEEE Transactions on Neural Networks and Learning Systems (TNNLS), IEEE Transactions on Cybernetics, IEEE Transactions on Image Processing, Pattern Recognition, Information Science, Decision Support Systems, etc.. His main research focuses on integrating artificial intelligence (AI) with fashion and textiles, particularly in machine vision, pattern recognition and deep learning.