



Creating an AI fashioner through deep learning and computer vision

Caner Balim¹ · Kemal Ozkan²

Received: 18 December 2022 / Accepted: 27 March 2023 / Published online: 8 April 2023
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2023

Abstract

Fashion is a multibillion-dollar industry that concerns many people both socially and culturally. Thanks to social networks, there is a lot of data about the fashion industry on the internet. This has led researchers to shift their attention to this area, especially recently. This paper proposes an end-to-end framework to build an AI fashioner that can diagnose clothing compatibility and generate recommendations to improve compatibility. First, fashion compatibility reviews are analyzed, and incompatible clothing items are identified for each outfit combination. Next, the items of clothing that make up the outfit combination are separated using Mask R-CNN. Then, the incompatible clothing items were removed from the outfit combination, and the most similar outfit combinations were identified among the compatible clothing items. In addition, an attribute detection network was developed to extract the attributes of compatible outfits with the same category in the detected compatible outfits. Finally, recommendation sentences are generated using the detected attributes, and encoder-decoder models are used to train a deep network that generates recommendations from clothing images. Extensive experiments based on existing datasets demonstrate the effectiveness of the proposed method.

Keywords Outfit recommendation · Encoder-decoder networks · Outfit segmentation · Image captioning · Image processing

1 Introduction

Clothing and styling have a great impact on the popularity of the fashion field since they play a critical role in people's lives. Outfit compatibility helps people to cover their weaknesses and show their status. That is why people prefer to wear matching clothes. Add to this the ease of access to data with digitalization, and researchers' interest in this field has increased. In the last few years, there have been studies regarding clothing detection (Liu et al. 2016; Sidnev et al. 2021), clothes segmentation (Zhang et al. 2020), clothes image retrieval (Kang et al. 2020a, b; Ji et al. 2020) and outfit compatibility (Han et al. 2017; Sun et al. 2020a; Li

et al. 2016; Song et al. 2017; Kavitha et al. 2020; Vasileva et al. 2018; Wang et al. 2019).

Outfit compatibility is based on measuring the compatibility of several clothing items. In the literature, the approaches for outfit compatibility can be divided into two main types: outfit compatibility prediction and outfit compatibility recommendation. The task of outfit compatibility prediction is a binary classification problem that predicts the compatibility score of an outfit (Han et al. 2017; Li et al. 2016; Vasileva et al. 2018; McAuley et al. 2015). The Outfit Compatibility Recommendation task first estimates the compatibility level for a given outfit and then, if the level is bad, identifies incompatible clothes and determines how to create a compatible outfit to improve compatibility (Wang et al. 2019; Yang et al. 2021; Chen et al. 2019a; Han 2022). Most of the previous work on outfit compatibility prediction has attempted to measure the distance between visual features of clothing using Euclidean distance, Mahalanobis distance and Siamese networks (Vasileva et al. 2018; McAuley et al. 2015; He et al. 2016; Veit et al. 2015). Unlike metric learning-based methods, there are also studies that model the fashion compatibility problem as a sequence (Li et al. 2016). To diagnose outfit compatibility, there are systems that use Bayesian Personalized Ranking, multilayer convolutional

✉ Caner Balim
cbalim@aku.edu.tr

Kemal Ozkan
kozkan@ogu.edu.tr

¹ Department of Computer Computer Programming, Sandikli Vocational School of Higher Education, Afyon Kocatepe University, Afyonkarahisar, Turkey

² Department of Computer Engineering, Faculty of Engineering and Architecture, Eskisehir Osmangazi University, Eskisehir, Turkey

networks and contextual metadata (Wang et al. 2019; Han 2022; Lin et al. 2020).

On the other hand, image captioning is one of the popular fields that has gained a place in the literature in recent years. Image captioning-like problems are called sequence-to-sequence problems in the literature. In order to solve such problems, it is necessary to find the dependencies and connections between sequential elements in the input. Although theoretically possible, Recurrent Neural Networks (RNN) suffer from the problem of long-term dependencies due to the problem of vanishing gradients. Different RNN models have been proposed with minor modifications to overcome this problem. Long Short-Term Memory (LSTM) has the ability to remember or forget sequence elements using input, output and forget gates. Cho et al. (2014) and Sutskever et al. (2014) proposed the use of an encoder-decoder model for the machine translation problem that is entirely based on RNN and LSTM. However, the mechanism of encoding the entire input in a single vector in RNNs leads to the same problems. To solve this problem, attention mechanism has

been developed by researchers. In this way, instead of encoding the entire input in only one vector, the RNNs utilize the generated attention vector in the decoding phase. Although LSTM + attention mechanism achieves better results than solutions using only LSTM, it is not possible to process the inputs in parallel due to the sequential operation of the RNNs. For this reason, the training time of RNN models is very high. To solve the parallelization problem, Transformers has been proposed which tries to solve sequential-to-sequential problems with only the attention mechanism (Vaswani et al. 2017).

In this study, an AI fashioner is developed to solve the outfit compatibility problems with deep learning techniques. The proposed model first predicts the outfit compatibility and if there are incompatible items among the clothing items that make up the outfit, it helps to create more compatible outfits by suggesting compatible clothing items instead of these incompatible items. A summary of the whole study is shown in Fig. 1. Initially, the dataset is analyzed and the incompatible clothing items are identified. In the next stage,

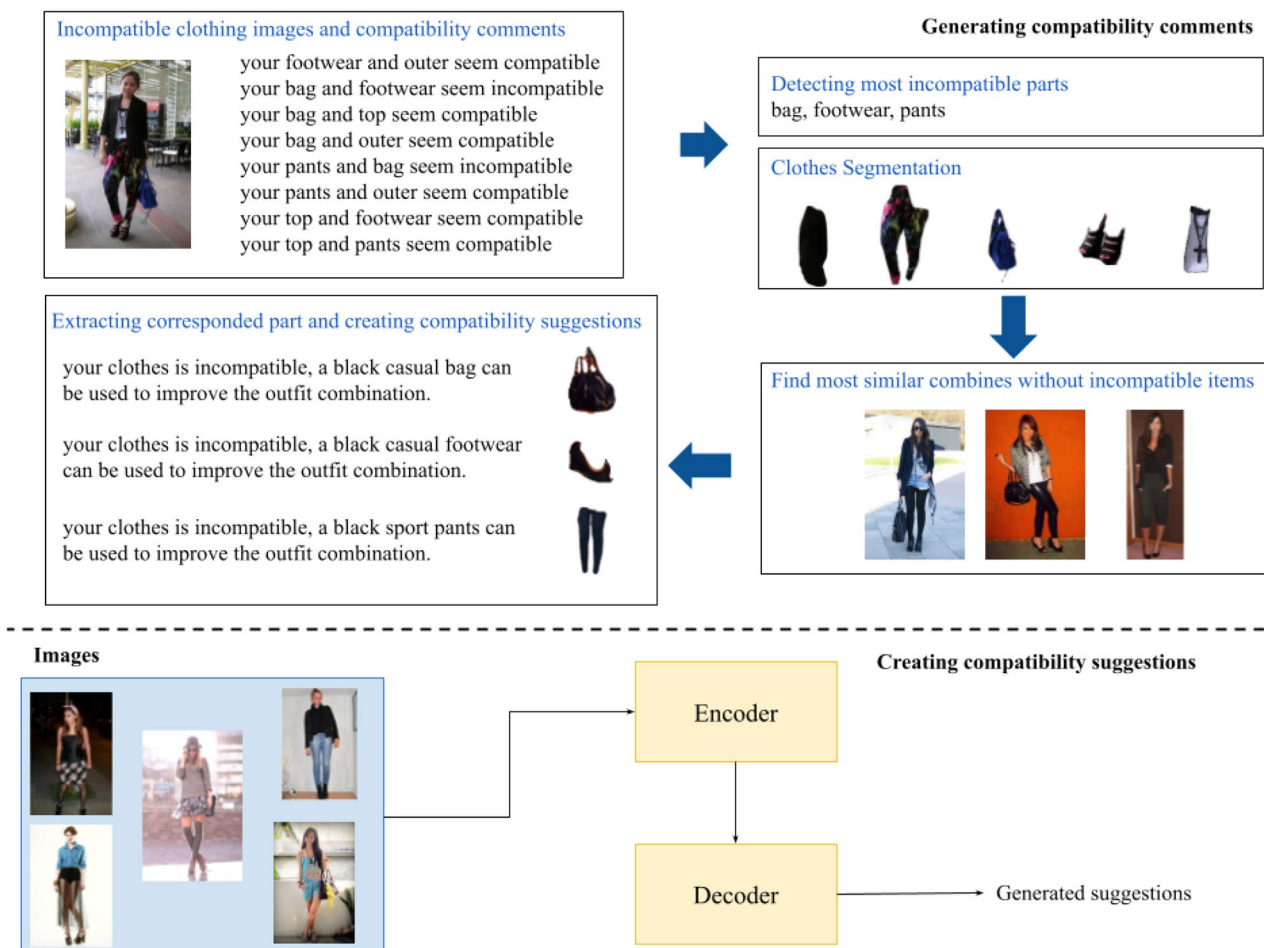


Fig. 1 Overview. This paper presents an AI fashioner model that learns the relationships between compatible clothes and gives suggestions for reorganizing incompatible outfits

the outfit images are segmented using Mask R-CNN. The mismatched clothing items are respectively removed from the combination image set and the images of the remaining items are compared with the compatible outfit's image sets with the same clothing categories in the dataset. In the comparison phase, low and high-level visual features are extracted and combined to identify matching combinations with the most similar correlations across compatible clothing items. The items that will replace the incompatible parts from the detected matched combinations are given to the attribute detection network. According to the results obtained from this network, different compatibility improvement suggestions are presented for each incompatible combination by utilizing Convolutional Neural Network (CNN) and Transformer.

The main contributions of the proposed work are as follows:

1. We develop an end-to-end framework problem to make the outfit compatibility task more user-friendly. This framework not only predict and diagnose outfit compatibility but also gives suggestions to improve compatibility.
2. We combine high-level and low-level features to generate reviews of outfit combinations and propose a joint attribute detection module to understand real-world clothing images.
3. Extensive simulations are conducted on two different datasets that contain real-world images. Rather than the classic top–bottom compatibility, reviews and suggestions are created over twelve different clothing categories.

The structure of this work is as follows. In Sect. 2, we briefly review relevant works on fashion compatibility, fashion attribute detection, and image captioning. In Sect. 3, we provide a detailed description of the proposed work. In Sect. 4, we present the results of our simulations, showing the effectiveness of the proposed approach. Finally, we present our conclusions in Sect. 5.

2 Related work

Many previous studies assume fashion compatibility as a metric learning problem. Veit et al. (2015) train a Siamese CNN to learn compatibility across co-purchasing items. McAuley et al. (2015) utilize parameterized distance metrics to learn relationships between co-purchased item pairings. They use CNNs for feature extraction. Chen and He (2018) propose a deep mixed-category metric learning framework that is based on triplet-loss to recommend complementary fashion items.

There have also been alternative approaches to metric learning regarding the sequential processes. Han et al. (2017) use Bi-directional-LSTM to model outfit generation. Li et al. (2016) use RNNs to predict fashion outfit compatibility. There are also methods that use transformation vectors from clothing feature representations to estimate clothing compatibility (Li et al. 2019; Lu et al. 2021).

Image captioning is a challenging area of work involving both image processing and natural language processing (NLP). Image Captioning studies often employ NLP techniques and deep learning models such as CNNs and RNNs to analyze the visual content of the images and generate descriptive captions. Besides these techniques, state-of-the-art approaches mainly use attention techniques to generating captions for images (Anderson et al. 2017; Herdade et al. 2019; Xu et al. 2015). Recent studies comparing image captioning methods have shown that Transformer models using only the attention mechanism yield the most successful results. Transformer was introduced in 2017 in an article called "Attention Is All You Need" as an encoder-decoder architecture based on layers of attention (Vaswani et al. 2017).

There are also studies that address the problem of fashion and image captioning. This involves using machine learning techniques to automatically generate textual descriptions of fashion images, which can be helpful for various applications such as automatic title creation, image search, and style recommendation. Chen et al. (2019b) proposed a Personalized Outfit Generation (POG) model by connecting both user preferences and individual items with transformer architecture. Chen et al., proposed a method for fashion recommendation that combines attention-weighted visual features and GRU-based weak supervised learning. Their approach uses image region-level features and user review information to make recommendations (Chen et al. 2019a). Park et al. (2022) proposed an improved Transformer model for a conversational system that recommends a desired fashion item based on the conversation between the user and the system and fashion image information. Li et al., proposed a framework for clothes image captioning that combines attribute detection, visual attention, and LSTM. The framework uses attention mechanisms to focus on relevant image regions, allowing for more accurate caption generation (Li et al. 2021). Balim and Ozkan (2021) proposed a system for the automatic generation of product titles of fashion images using CNN, LSTM, and Global Vectors. Goenka et al., proposed a new pre-trained transformer model called FashionVLP for fashion image retrieval. The model uses prior knowledge from large image-text corpora to improve performance on the task of fashion image retrieval (FashionVLP 2022). Yang et al. (2020) proposed a set of strategies to improve the performance of captioning for clothing and

created the FACAD dataset that can be used in fashion captioning studies.

There are studies to explain compatibility (Han 2022; Lin et al. 2020; Sun et al. 2020b; Tangseng and Okatani 2019; VisQu 2022; Kaicheng et al. 2021). Lin et al. (2020) introduced the ExpFashion dataset, which contains contextual metadata for top and bottom fashion items. They also proposed a neural network-based approach for generating outfit recommendations with abstractive comments. The proposed approach uses the metadata in the ExpFashion dataset to generate personalized recommendations for individual users. Han et al. propose a Bayesian Personalized Ranking (BPR) framework named PAICM, giving also suggestions on how to modify those incompatible outfits to make them appealing to the user (Han 2022). Wang et al., studied the problem of diagnosing outfit compatibility and proposed a multi-layered comparison network for predicting the compatibility of different fashion items. The network uses gradient information to make its predictions, allowing it to take into account the relationships between different items in an outfit (Wang et al. 2019). Mo et al. (2022) presented a model for fashion compatibility assessment utilizing low and high-level features based on multilayer convolutional networks and Transformer for explainable evaluation and recommendation. Yang et al. (2021) uses the attribute information of fashion items to explain their compatibility. Balim and Ozkan (2023) used image processing techniques and transformers to perform the task of diagnosing fashion compatibility and generated explanations over body images.

While previous studies in the literature have mostly focused on the fit of tops and bottoms, this study proposes a system that checks the compatibility of twelve different clothing categories. In addition, while existing studies

Table 1 Categorical information from the ModaNet dataset

Clothing categories	Fine grained categories
Footwear	Footwear
Boots	Boots
Top/blouse/t-shirt/shirt	Top
Pants/jeans/leggings/ long socks	Pants
Dress	Dress
Coat/jacket/suit/blazers/cardigan/sweater/Jumpsuits/ Rompers/vest	Outerwear
Skirt	Skirt
Bag	Bag
Shorts	Shorts
Belt	Belt
Sunglasses	Sunglasses
Hat/ headwear/ headband	Headwear
Scarf/tie	Scarf&tie

focused on a single recommendation, in this study, a correlation matrix was created using low and high-level features and three recommendations were generated for each mismatched outfit combination.

3 The proposed method

In this section, we present the proposed model, mainly based on three stages: data preparation, fashion comment generation and fashion captioning. The overall structure of the proposed system is shown in Algorithm 1.

Algorithm 1:	
	Input: image X , encoder network e , decoder network e^{-1} , CrossEntropyLoss L , comments C
1	foreach batch of samples $\{x_k\}_{k=1}^N$ do
2	foreach $k \in \{1, \dots, N\}$ do
	// Take C_k as an input and find the most incompatible parts:
3	$iC_k = \text{detecting_incompatible_parts}(C_k)$
	// Segment clothes from image:
4	$S_k = \text{outfit_segmentation}(X_k)$
	// Find best compatible matches from the similarity detection module:
5	$R_k = \text{find_compatible_outfit_items}(S_k, iC_k)$
	// Detect characteristics from the attribute detection module:
6	$ch_k = \text{detect_item_characteristics}(R_k)$
	// Generate compatibility comments:
7	$c_k = \text{create_fashioner_comments}(C_k, ch_k)$
	// Generate encoded image:
8	$E_k = e(S_k, c_k)$
	// Generate comments from image encodings:
9	$c_k^1 = e^{-1}(S_k)$
	// Update the networks e and e^{-1} to minimize the loss
10	$loss = L(c_k, c_k^1)$

3.1 Data preparation

3.1.1 Detection of incompatible items

In this study, we use the ModAI dataset which is produced for use in fashion studies (Balim and Özkan 2023). The ModAI consists of real-world outfit images and comments about the compatibility of outfit. In the dataset, comments are about the compatibility or incompatibility of paired clothing categories such as footwear-top: compatible, top-scarf: incompatible, etc. The ModAI dataset contains 11,010 outfit images and 25,916 compatibility comments about clothing pairs for each outfit. It contains twelve clothing fine-grained clothing categories inspired by the ModaNet dataset (Mo et al. 2022) which is used with Mask R-CNN for the segmentation stage. These categories are shown in Table 1. In the ModaNet dataset, footwear and boot are categorized separately, while in the ModAI dataset, these categories are grouped under a single category as footwear.

In the ModAI dataset, the compatibility comments of each outfit contain compatible or incompatible keywords. We use these keywords for detecting incompatible parts. If all the comments on an outfit do not mention incompatible parts, we consider that outfit as compatible. Otherwise, we make a list of clothes mentioned as incompatible and identify clothes that are more mentioned as incompatible. If the number of incompatible clothing categories is less than three, we add more than one item from the same category to the incompatible clothing list. Our goal is to generate three recommendation comments for each incompatible outfit. The whole process is presented in Algorithm 2.

3.1.2 Outfit segmentation

In real-life images taken for outfit compatibility problems, the clothes are located on the human body. The proposed model uses a segmentation technique called Mask R-CNN to extract relevant clothing items from the human body. Mask R-CNN is basically a built on Faster Region Based Convolutional Neural Networks (Faster R-CNN) (Ren et al. 2016; He et al. 2018). Faster R-CNN is a popular object detection algorithm that uses a CNN to perform both object classification and bounding box regression. Given an input image, Faster R-CNN first generates a set of proposals for potential object locations, and then processes these proposals using the CNN to predict the class label and bounding box coordinates for each object in the image. Faster R-CNN basically performs the learning process in 4 steps:

- A CNN architecture is used to extract vectors of features from images. (ResNet-101 architecture is preferred for this study.)
- These feature vectors are sent to the Region Proposal Network to generate candidate frames.
- The Region of Interest pooling layer reduces candidate sizes to the same size.
- After the calculations, the features extracted from the proposed frames are determined, the class of the object is determined and the bounding box coordinates are determined.

Like Faster R-CNN, Mask R-CNN uses a CNN to perform object classification and bounding box regression. However, it also includes an additional branch in the network that is used to predict the mask for each object. This branch takes the features extracted by the CNN and

Algorithm 2: *detecting_incompatible_parts*

```

Input: comments for a sample image  $I$ , comments count for one image  $T$ 
Output: generated fashioner comments for a sample image  $I$ 
1 let  $IC$  be a list of clothing that consist of incompatible outfits
2 let  $mIC$  be a list of three most incompatible clothes
3 if  $T$  not contains "incompatible" keyword:
4   return "your combination is compatible"
5 else
6   for  $i:=0$  to  $T$  do
7     append ( $C_i[0]$ ,  $C_i[1]$ ) to  $IC$ 
8   end for
9   end if
10   $mIC \leftarrow$  get_first_three_item (sort ( $IC$ )) sort clothes in ascending order and get first three clothes

```

processes them using additional layers to generate a binary mask that indicates the pixels belonging to the object. The mask is then combined with the bounding box prediction to generate the final object detection result, which includes the class label, bounding box coordinates, and mask for each detected object in the image.

In this paper, we use the ModaNet dataset to train the Mask R-CNN. This dataset includes 55,176 annotated images with pixel-level segments, polygons, and bounding boxes, covering thirteen fine-grained categories introduced for use in clothing segmentation and feature estimation research (Zheng et al. 2018). After the training phase, we learn the pixel information for clothing items in the fine-grained categories given in Table 1. In order to use the ModAI and the ModaNet datasets jointly, we update the garments segmented as boots to footwear.

In this work, we use the segmentation model in three different stages. Firstly, we use this model in the similarity detection stage to separate clothes from outfit images. Then, we also use this model in the attribute detection module to extract the correct clothing items from e-commerce images. Lastly, we use this model to generate fashion compatibility comments for element-wise feature extraction.

3.2 Fashion comment generation

Fashion comment generation aims to create fashioner-like reviews that address incompatible clothes and make suggestions to improve compatibility for each incompatible outfits. This phase has two components: The similarity detection module and fashion attribute detection module. More details of each component are described in the following sections.

3.2.1 The similarity detection module

In the similarity detection module, we aim to find clothes that can be recommended instead of mismatched items in incompatible outfits. For this purpose, we first segment all the clothes from outfit images using the Mask R-CNN segmentation technique described in Sect. 3.2. Then, we remove the mismatched items from the incompatible outfits and perform a similarity measurement stage with all the compatible outfits in the dataset. Our goal is to identify compatible outfits that are most similar to incompatible outfits without mismatched items. In this way, it will be possible to recommend compatible clothing items in the same category instead of mismatched items.

In this module, correlation matrices are created for each outfit, using low-level features and high-level features together to detect similar combinations of clothes using

latent structures between outfits. First, using ResNet-101 which is trained on ImageNet, high-level features are extracted for all segmented clothing items in the dataset. The high-level features mostly reflect features such as fashion style and overall compatibility (Wang et al. 2019). Color information, which is very important for compatibility is used as low-level features are then added to these features to create clothing representation vectors. Although RGB is the most widely-used color space in the digital environment, it can be seen that pixels with the same color value have different color values depending on the light brightness, especially in images taken from the real world. Since real-world images contain different images taken in various poses and places with light tones, it is difficult to extract the color features of clothes correctly. In this study, HSV (Hue, saturation, value) color space is preferred over RGB color space in order to obtain a system that is less affected by the amount of light brightness in the environment. The HSV space consists of hue (H), saturation (S), and brightness (V) components. In this study, images are converted from RGB space to HSV space, and H and S values are used, while the V value is not included in order to minimize the degree of brightness.

Given an outfit with clothing items, we denote each different category with l and we denote the outfit as $X = [x_1, x_2, \dots, x_l] \in R^{l \times d}$, where is the d dimensionality of the fashion images. The image embeddings from each outfit are respectively calculated as F , hue histogram as h and saturation histogram as s for a clothing item. By concatenating F , h and s , we obtain v :

$$v_i = c(h_i, s_i, F_i) \quad (1)$$

where h_i and s_i are i th color features, F_i represents feature vector of i th fashion item in an outfit. As a result, we have a representation of a clothing item as $v_i \in R^d$.

The Pearson correlation coefficient (PCC) is a statistical measure of the linear correlation between two variables. It is commonly used in research to identify positive or negative relationships between variables and is calculated using the mean and standard deviation of the two variables. The PCC ranges from -1 to 1 , where -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation. In the present study, the PCC is used to detect the similarity between each item in same outfit. To calculate the degree of correlation between the representation vectors v_i and v_j of two clothes; the PCC can be calculated as follows:

$$\rho_{v_i, v_j} = \frac{\text{Cov}(v_i, v_j)}{\sigma_{v_i} \sigma_{v_j}} = \frac{\sum |(v_i - \mu v_i)(v_j - \mu v_j)|}{\sigma_{v_i} \sigma_{v_j}} \quad (2)$$

Using this method, the PCC values between the compatible clothes in each outfit is calculated and a correlation

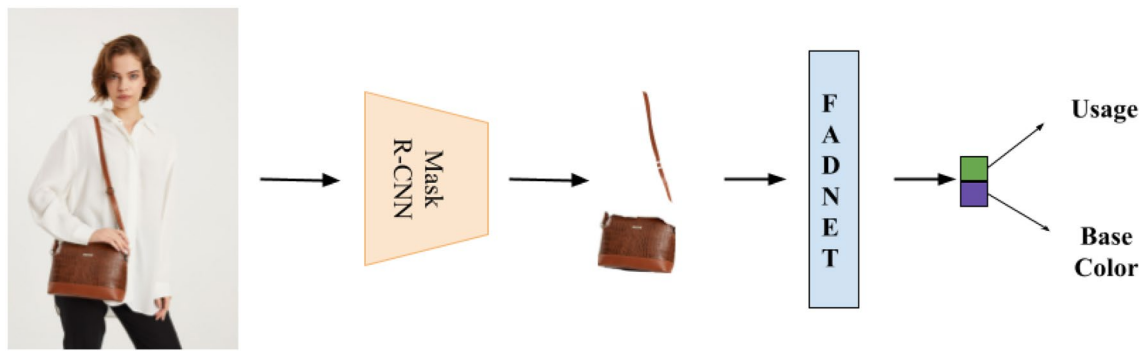


Fig. 2 The creation of the attribute detection module

Table 2 Outputs of the attribute detection network

Attributes	Categories
Color	Multi color, Black, Blue, Pink, Cream, Red, Brown, White, Yellow, Grey, Green, Purple, Orange
Usage	Casual, Formal, Sport

matrix is created for representing compatible clothes. Finally, by examining the similarities of the correlation matrices, it is determined which clothes could be recommended instead of incompatible clothes.

3.2.2 The fashion attribute detection module (FAD-NET)

An AI fashioner model needs to make recommendations for replacing mismatched pieces, as real-life fashionistas do. To generate a fashion recommendation, the model needs to identify the characteristics of the clothes to be recommended. At this stage, a neural network is developed that takes clothing images (the identified clothes in the previous section) and produces basic clothing characteristics which are important for compatibility, as output. The whole process is shown in Fig. 2.

With growing e-commerce systems, the amount of tagged data related to fashion has increased. These data are used in many studies about the fashion area. For generating an attribute detection module, e-commerce website data is used to extract specific garment features that are frequently used by fashioners. Approximately, 50 k clothing images and different clothing characteristics are downloaded from e-commerce sites using web scraping techniques. After the data cleaning steps, there are 40 k clothing images and characteristics for using in the attribute detection module. When the

identified clothing images are analyzed in the training step, the images mostly include people, fashion products in different clothing categories or different backgrounds. To overcome this problem and identify the correct garment image, the image segmentation technique described in Sect. 3.1 and a number of pre-processing techniques are used.

After these steps, the characteristic detection phase is started. At this stage, a fully connected network is developed that produces color and usage information as output using the weights of networks trained on ImageNet, a popular approach in the literature. This network takes the clothing image as input and produces color and usage type as output. The details of the outputs of the network are shown in Table 2.

In their study on outfit compatibility evaluation, Wang et al. assumed that the first layers in the deep neural network tend to learn low-level features like color from fashion images (Wang et al. 2019). Mo et al. cited that the last layers show abstract characteristics such as style and usage (Kaicheng et al. 2021). These inferences are consistent with the general principle that early layers of CNNs tend to learn low-level features, while later layers learn higher-level abstractions. Based on these studies, a feed-forward neural network is developed to detect the base color and usage which are the characteristics that are often used from fashioners. The network is made up of a multi-output CNN fusion network that concatenates different convolutional layers and outputs the attribute tags. The feature learning layers consist of a pre-trained ResNet-101 network, which is pre-trained on ImageNet, where the fully connected layer takes different convolutional layers. The final outputs of the network are rectified by the sigmoid function, which maps the outputs to the range [0,1], allowing them to be interpreted as probabilities. This allows the network to generate a set of predicted attribute tags for each input image. The whole attribute detection process is shown in Fig. 3.

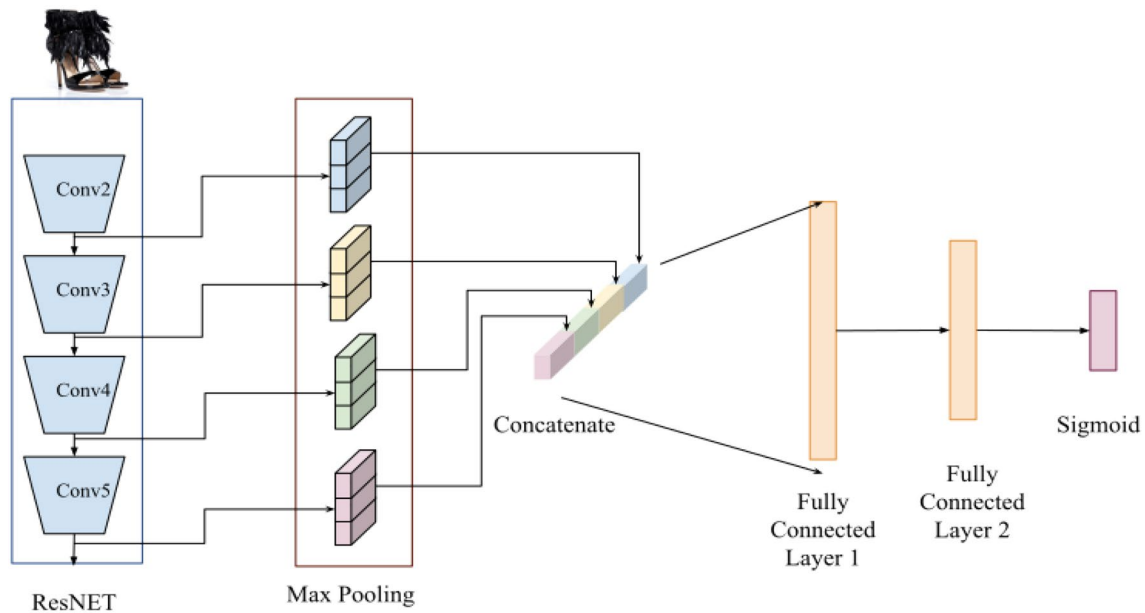


Fig. 3 FADNET attribute detection process. We concatenate the last four block of the ResNet-101 network as feature vector. Next, we use two fully connected layers for extracting clothing attributes

3.3 Fashion captioning

Fashion captioning is the part where fashioner's recommendations are created over the images. In this study, transformer neural networks are used to generate fashion recommendations. The transformer architecture was originally proposed for NLP tasks, such as machine translation, but it has since been successfully applied to a wide range of other tasks. One of the key advantages of the transformer is its flexibility, which allows it to be easily adapted to different problem domains. In the case of image captioning, the encoder part of the transformer can be used to process the visual input, while the decoder part can be used to generate a natural language description of the image. Transformers consist of N x number of repeating modules between encoders and decoders. Most of the elements of these modules are composed of the Multi Head Attention and position-wise feed forward network. The core component of the Multi Head Attention mechanism is the scaled dot-product attention which consists of queries $Q = \{q_1, q_2, q_3, \dots, q_{1_q}\}, q_i \in \mathbb{R}^{d_q}$, keys $K = \{k_1, k_2, k_3, \dots, k_{1_k}\}, k_i \in \mathbb{R}^{d_k}$ and values $V = \{v_1, v_2, v_3, \dots, v_{1_v}\}, v_i \in \mathbb{R}^{d_v}$, is formulated as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3)$$

In Eq. (3), Q (Query) is a matrix containing the query (vector of words selected and compared with other words), K (Key) is a matrix containing the keys (vector representations of all words except the selected word) and V (Value) is a matrix containing the values of the words (whole sentence) multiplied by the calculated weight. In the above equation, d_k is a scaling factor that helps to improve the learning process. The multi-head attention mechanism is shown in Eqs. (4) and (5).

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O \quad (4)$$

$$\text{head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \quad (5)$$

The multi-head attention mechanism uses multiple heads with the same architecture, each of which applies attention independently to the input data. These heads can be repeated h times. By applying self-attention multiple times, with each head attending to different parts of the input sequence, the transformer is able to capture a wide range of relationships between the input tokens. Because each head has its own set of weight matrices, it can learn different aspects of the input data, such as grammar or semantics. This allows the transformer to represent the input data in a more rich and diverse way, improving its ability to perform a variety of NLP tasks.

The position-wise feed forward network is another important component of the transformer architecture. It is typically applied after the self-attention or multi-head attention layers, and it is used to process the representations generated

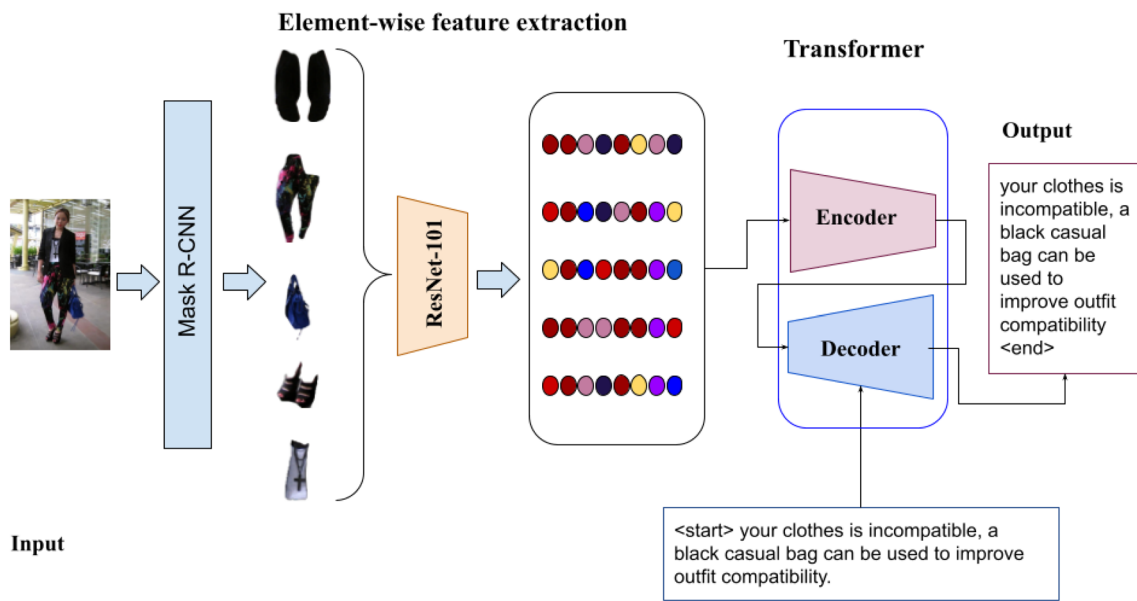


Fig. 4 Fashioner comment generation process

Table 3 The statistics of the ModAI and the Polyvore-T datasets

	Incompatible	Compatible	Item	Outfit
ModAI	4,258	6,752	–	11,010
Polyvore-T	–	–	142,480	21,889

by these layers further, generating a more abstract and condensed representation of the input data. The definition is as follows:

$$FFN(x) = W_2 \max(0, W_1 x + b_1) + b_2 \tag{6}$$

In this equation, x is the input to the position-wise feed forward network, and W_1 and b_1 are the weight matrix and bias vector of the first linear transformation, respectively. The non-linear activation function is typically a rectified linear unit, which applies the function $\max(0, x)$ element-wise to the input. The output of the activation function is then transformed using a second linear transformation, with weight matrix W_2 and bias vector b_2 , to produce the final output of the position-wise feed forward network. This layer helps to capture higher-level features in the input data, improving the performance of the transformer on a variety of tasks.

One of the key differences between the transformer architecture and other popular deep learning models, such as RNNs, is that the transformer does not explicitly incorporate information about the order of the input sequence. To address this limitation, the transformer architecture includes a mechanism for incorporating positional information into the input data. This is typically done by adding a set of fixed

"spatial coding" vectors to the input sequence, which encode the relative positions of the input tokens.

The Original Transformer consists of 6 blocks with both encoder and decoder (Vaswani et al. 2017). In this study, the features of the segmented clothing images are extracted using different ResNet-101 architecture and fashioner comments are generated using Transformer. The whole process is shown in Fig. 4.

4 Experiments

4.1 Dataset

The performance of the proposed method was evaluated on two datasets. The ModAI dataset is mentioned in Sect. 3. To evaluate the proposed work in another domain different from the ModAI, the Polyvore-T dataset is used, which is widely used in fashion compatibility studies (Wang et al. 2019). In their study, Wang et al., generated incompatible examples by randomly replacing some items in the Polyvore-T dataset. In this study, we follow the same path to evaluate our work in the Polyvore-T dataset and randomly replace some combinations with different clothes while preserving the clothing categories. The negative examples are generated by randomly replacing an item in positive outfits with another item of the same type from different outfits. The attributes of the ModAI and Polyvore datasets are listed in Table 3. The datasets are divided into training, validation, and testing sets with a ratio of 7:1:2 for use in the experiments.

4.2 Training details

All experiments are conducted on a NVIDIA Quadro RTX 5000 graphics card. The ResNet-101 pre-trained on ImageNet is used as the backbone. Different hyper-parameters such as different number of heads and number of layers were used for the transformer. The batch size for the experiments on the ModAI is 32 while that on the Polyvore-T dataset is 64. Adaptive learning rate optimization algorithm (Adam) is used during the training process with learning rate 0.0001. Dropout is a regularization technique that is commonly used in neural networks to prevent overfitting. In the context of the methods described in the previous statement, it is likely that dropout is applied to the input sequences or to the intermediate representations computed by the encoder and decoder blocks. During training, the models are trained for a maximum of 50 epochs, and the training process is stopped early if the performance on a validation set starts to deteriorate. Finally, during inference, the models use a beam search strategy with a beam size of 3 to generate output sequences. This means that the models consider a set of 3 most likely sequences at each step and select the best one to extend based on the predicted probabilities of the next tokens. The results are reported on standard machine translation and image captioning evaluation metrics including CIDEr (Vedantam et al. 2014), BLEU (Papineni et al. 2001), ROUGE-L (Lin 2004) and METEOR (Banerjee and Lavie 2005).

4.3 Performance analysis

4.3.1 Quantitative results

The results of the proposed method with different hyper-parameters are shown in Table 4. The best results are highlighted in bold. Compatibility comments are generated based

Table 4 Results of different hyperparameters

Fine Tuning	Head Size	Layer Size	BLEU4	METEOR	ROUGE-L	CIDEr
+	1	1	0.5019	0.3873	0.7124	0.6331
+	2	2	0.5156	0.3944	0.7338	0.6504
+	3	3	0.4938	0.3841	0.7045	0.6697
–	1	1	0.4545	0.3684	0.7520	0.5692
–	2	2	0.4580	0.3689	0.7462	0.5259
–	3	3	0.4712	0.3745	0.7584	0.5738

Table 5 Results according to different methods and datasets

	Similarity technique	Dataset	BLEU4	METEOR	ROUGE-L	CIDEr
CNN	Correlation matrix similarity	ModAI	0.4821	0.3665	0.6971	0.5988
CNN + Color Features	Euclidean distance	ModAI	0.4902	0.3734	0.7124	0.6184
CNN + Color Features	Correlation matrix similarity	ModAI	0.5156	0.3944	0.7338	0.6504
CNN + Color Features	Euclidean distance	Polyvore-T	0.4496	0.3683	0.6857	0.4809
CNN + Color Features	Correlation matrix similarity	Polyvore-T	0.4705	0.3753	0.6988	0.5103

on correlation-based similarity and experiments were conducted using different fine-tuning, head, and number of layers. It can be observed that there is a clear difference between the variations with and without fine-tuning. It is also observed that different numbers of heads and layers increase the system performance up to a certain number.




















The results of the different techniques used to generate similarity-based recommendations and their comparison with the dataset are shown in Table 5. The best results are highlighted in bold. It can be seen that correlation-based similarity detection is more successful than distance-only techniques. It is also observed that more successful results are obtained with the ModAI dataset.

It can also be observed from Table 5 that using only high-level attributes or only low-level attributes in the fashion recommendation generation phase produces less successful results than using both together.

4.3.2 Qualitative results

The qualitative results of the proposed method are shown in Table 6. The results show that the proposed model's performance is quite adequate. If the system finds the outfit combination compatible, it directly generates the result "your combination is compatible". If there are elements that make the combination incompatible, the proposed model suggests new elements that can be used instead of incompatible elements to make the combination compatible. In the event that there is only one item in the outfit that breaks the outfit compatibility, the proposed model suggests different pieces in the same category that will harmonize the outfit. It is noteworthy that the system sometimes confuses items with similar coordinates in the images. For example, the system suggests a skirt when it should

Table 6 Example results from the proposed AI fashioner

Original image	Segmented clothes	Compatibility label	Suggestions		
			1	2	3
		Compatible			
		Compatible			
		Incompatible	 A black casual bag can be used to improve the outfit combination.	 A pink casual top can be used to improve the outfit combination.	 A gray sport short can be used to improve the outfit combination.
		Incompatible	 A brown casual bag can be used to improve the outfit combination.	 A white casual footwear can be used to improve the outfit combination.	 A colorful casual dress can be used to improve the outfit combination.
		Incompatible	 A black casual footwear can be used to improve the outfit combination.	 A colorful casual top can be used to improve the outfit combination.	 A blue sport short can be used to improve the outfit combination.

suggest an item in the shorts category. Again, the system may suggest a jacket instead of suggesting a top. These errors can be reduced by increasing the success of segmentation techniques.

5 Conclusion

This paper aims to create an artificial fashioner using different levels of visual features, transformers and relationships between clothes. The proposed work not only

detects compatibility but also identifies the elements that break the compatibility and defines the characteristics of the elements that can replace the incompatible elements to harmonize the outfit. In the next step, it provides multiple suggestions to correct and increase the outfit compatibility like real fashion designers. Experiments on two different real-world datasets demonstrate the success of the proposed system. For future work, we are planning to explore different aspects of fashion to explain how to better diagnose the compatibility of clothes. We also want to

address the issue of creating a personal artificial fashioner, as fashion is very subjective.

Author contributions CB: data curation, visualization, investigation, methodology, software, validation, writing—original draft. KÖ: supervision, conceptualization, investigation, writing- reviewing and editing.

Data availability Datasets are available in public repositories: The ModAI dataset can be accessed by anyone at <https://doi.org/10.1016/j.eswa.2022.119305>. The Polyvore-T are openly available at <https://doi.org/10.1145/3343031.3350909>.

Declarations

Conflict of interests 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.

References

- Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, Zhang L (2017) Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp 6077–6086
- Balim C, Özkan K (2021) Ürün görsellerini kullanarak e-ticaret sistemleri için ürün başlığı oluşturulması. *Int J 3D Rint Technol Dig Ind* 5:614–624. <https://doi.org/10.46519/ij3dptdi.991789>
- Balim C, Özkan K (2023) Diagnosing fashion outfit compatibility with deep learning techniques. *Expert Syst Appl* 215:119305. <https://doi.org/10.1016/j.eswa.2022.119305>
- Banerjee S, Lavie A (2005) METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, pp 65–72
- Chen L, He Y (2018) Dress fashionably: learn fashion collocation with deep mixed-category metric learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 32, no. 1
- Chen X, Chen H, Xu H, Zhang Y, Cao Y, Qin Z, Zha H (2019a) 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. Association for Computing Machinery, New York. <https://doi.org/10.1145/3331184.3331254>
- Chen W, Huang P, Xu J, Guo X, Guo C, Sun F, Li C, Pfadler A, Zhao H, Zhao B (2019b) POG: personalized outfit generation for fashion recommendation at alibaba iFashion. In: Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining, pp 2662–2670
- Cho K, van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. [arXiv:1406.1078](https://arxiv.org/abs/1406.1078) [cs, stat]
- FashionVLP (2022) Vision language transformer for fashion retrieval with feedback. <https://www.amazon.science/publications/fashionvlp-vision-language-transformer-for-fashion-retrieval-with-feedback>. Accessed 8 Aug 2022
- Han X, Wu Z, Jiang Y-G, Davis LS (2017) Learning fashion compatibility with bidirectional LSTMs. In: MM 2017—proceedings of the 2017 ACM multimedia conference, pp 1078–1086. Doi: <https://doi.org/10.1145/3123266.3123394>
- Han X (2022) Prototype-guided Attribute-wise Interpretable Scheme for Clothing Matching. In: Proceedings of the 42nd International ACM SIGIR conference on research and development in information retrieval. <https://doi.org/10.1145/3331184.3331245>. Accessed 7 Aug 2022
- He R, Packer C, McAuley J (2016) Learning compatibility across categories for heterogeneous item recommendation. In: Proceedings—IEEE international conference on data mining, ICDM, pp 937–942
- He K, Gkioxari G, Dollár P, Girshick R (2017) Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp 2961–2969
- Herdade S, Kappeler A, Boakye K, Soares J (2019) Image captioning: transforming objects into words. *Advances in neural information processing systems* 32
- Ji Y-H, Jun H, Kim I, Kim J, Kim Y, Ko B, Kook H-K, Lee J, Lee S, Park S (2020) An effective pipeline for a real-world clothes retrieval system. [arXiv:2005.12739](https://arxiv.org/abs/2005.12739) [cs]
- Kaicheng P, Xingxing Z, Wong WK (2021) modeling fashion compatibility with explanation by using bidirectional LSTM. In: 2021 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW), pp 3889–3893. <https://doi.org/10.1109/CVPRW53098.2021.00432>
- Kang Z, Pan H, Hoi SCH, Xu Z (2020a) Robust graph learning from noisy data. *IEEE Trans Cybern* 50:1833–1843. <https://doi.org/10.1109/TCYB.2018.2887094>
- Kang Z, Lu X, Liang J, Bai K, Xu Z (2020b) Relation-guided representation learning. [arXiv:2007.05742](https://arxiv.org/abs/2007.05742) [cs, stat]
- Kavitha K, Kumar SL, Pravalika P, Sruthi K, Lalitha RVS, Rao NVK (2020) Fashion compatibility using convolutional neural networks. *Mater Today: Proc.* <https://doi.org/10.1016/j.matpr.2020.09.365>
- Li Y, Cao L, Zhu J, Luo J (2016) Mining fashion outfit composition using an end-to-end deep learning approach on set data. *IEEE Trans Multimedia* 19:1946–1955. <https://doi.org/10.1109/TMM.2017.2690144>
- Li X, Ye Z, Zhang Z, Zhao M (2021) Clothes image caption generation with attribute detection and visual attention model. *Pattern Recogn Lett* 141:68–74. <https://doi.org/10.1016/j.patrec.2020.12.001>
- Li K, Liu C, Kumar R, Forsyth D (2019) Using discriminative methods to learn fashion compatibility across datasets. *J Environ Sci (China)* (English Ed)
- Lin CY (2004) Rouge: a package for automatic evaluation of summaries. In: Text summarization branches out, pp 74–81
- Lin Y, Ren P, Chen Z, Ren Z, Ma J, de Rijke M (2020) Explainable outfit recommendation with joint outfit matching and comment generation. *IEEE Trans Knowl Data Eng* 32:1502–1516. <https://doi.org/10.1109/TKDE.2019.2906190>
- Liu Z, Luo P, Qiu S, Wang X, Tang X (2016) DeepFashion: powering robust clothes recognition and retrieval with rich annotations. In: Proceedings of IEEE conference on computer vision and pattern recognition (CVPR)
- Lu S, Zhu X, Wu Y, Wan X, Gao F (2021) Outfit compatibility prediction with multi-layered feature fusion network. *Pattern Recogn Lett* 147:150–156. <https://doi.org/10.1016/j.patrec.2021.04.009>
- McAuley J, Targett C, Shi Q, Hengel A (2015) van den: image-based recommendations on styles and substitutes. In: SIGIR 2015—Proceedings of the 38th International ACM SIGIR conference on research and development in information retrieval, pp 43–52
- Mo D, Zou X, Wong W (2022) Neural stylist: towards online styling service. *Expert Syst Appl* 203:117333. <https://doi.org/10.1016/j.eswa.2022.117333>

- Papineni K, Roukos S, Ward T, Zhu W-J (2001) BLEU: a method for automatic evaluation of machine translation. *ACL* 2011:311–318. <https://doi.org/10.3115/1073083.1073135>
- Park YJ, Jo BC, Lee KU, Kim KS (2022) Improved transformer model for multimodal fashion recommendation conversation system. *J Korea Contents Assoc* 22:138–147. <https://doi.org/10.5392/JKCA.2022.22.01.138>
- Qu W (2022) Visual and textual jointly enhanced interpretable fashion recommendation. *IEEE Journals & Magazines*. <https://ieeexplore.ieee.org/document/9046774>. Accessed 7 Aug 2022.
- Ren S, He K, Girshick R, Sun J (2016) Faster R-CNN: towards real-time object detection with region proposal networks. *arXiv:1506.01497 [cs]*
- Sidnev A, Krapivin A, Trushkov A, Krasikova E, Kazakov M, Viryasov M (2021) DeepMark++: real-time clothing detection at the edge. In: Presented at the proceedings of the IEEE/CVF winter conference on applications of computer vision
- Song X, Feng F, Liu J, Li Z, Nie L, Ma J (2017) NeuroStylist: neural compatibility modeling for clothing matching. In: Presented at the October 23. <https://doi.org/10.1145/3123266.3123314>
- Sun GL, He JY, Wu X, Zhao B, Peng Q (2020a) Learning fashion compatibility across categories with deep multimodal neural networks. *Neurocomputing* 395:237–246. <https://doi.org/10.1016/j.neucom.2018.06.098>
- Sun P, Wu L, Zhang K, Fu Y, Hong R, Wang M (2020b) Dual learning for explainable recommendation: towards unifying user preference prediction and review generation. In: Proceedings of the web conference 2020b, pp 837–847. Association for Computing Machinery, New York
- Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. *Advances in neural information processing systems* 27
- Tangseng P, Okatani T (2019) Toward explainable fashion recommendation. *Arxiv*. <https://doi.org/10.48550/arXiv.1901.04870>
- Vasileva MI, Plummer BA, Dusad K, Rajpal S, Kumar R, Forsyth D (2018) Learning type-aware embeddings for fashion compatibility. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), 11220 LNCS, pp 405–421
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: *Advances in neural information processing systems*, pp 5999–6009. Neural information processing systems foundation
- Vedantam R, Zitnick CL, Parikh D (2014) CIDEr: consensus-based image description evaluation. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp 4566–4575
- Veit A, Kovacs B, Bell S, McAuley J, Bala K, Belongie S (2015) Learning visual clothing style with heterogeneous dyadic co-occurrences. In: Proceedings of the IEEE international conference on computer vision, pp 4642–4650
- Wang X, Wu B, Ye Y, Zhong Y (2019) Outfit compatibility prediction and diagnosis with multi-layered comparison network. In: *MM 2019* —Proceedings of the 27th ACM international conference on multimedia, pp 329–337. <https://doi.org/10.1145/3343031.3350909>
- Xu K, Ba JL, Kiros R, Cho K, Courville A, Salakhutdinov R, Zemler RS, Bengio Y (2015) Show, attend and tell: Neural image caption generation with visual attention. In: 32nd international conference on machine learning, ICML, pp 2048–2057. International Machine Learning Society (IMLS)
- Yang X, Zhang H, Jin D, Liu Y, Wu C-H, Tan J, Xie D, Wang J, Wang X (2020) Fashion captioning: towards generating accurate descriptions with semantic rewards. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics). 12358 LNCS, pp 1–17
- Yang X, Song X, Feng F, Wen H, Duan L-Y, Nie L (2021) Attribute-wise Explainable Fashion Compatibility Modeling. *ACM Trans Multimedia Comput Commun Appl* 17:361–3621. <https://doi.org/10.1145/3425636>
- Zhang H, Sun Y, Liu L, Wang X, Li L, Liu W (2020) ClothingOut: a category-supervised GAN model for clothing segmentation and retrieval. *Neural Comput Appl* 32:4519–4530. <https://doi.org/10.1007/s00521-018-3691-y>
- Zheng S, Yang F, Kiapour M, Piramuthu R (2018) ModaNet: a large-scale street fashion dataset with polygon annotations. In: Presented at the October 15. <https://doi.org/10.1145/3240508.3240652>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.