LEARNING OUTFIT COMPATIBILITY WITH GRAPH ATTENTION NETWORK AND VISUAL-SEMANTIC EMBEDDING

Jianfeng Wang[1], Xiaochun Cheng[2,*], Ruomei Wang[1], and Shaohui Liu[3,4]

[1]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
[2] Department of Computer Science, Middlesex University, London, UK
[3] School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
[4] Peng Cheng Laboratory, Shenzhen China

2021 IEEE International Conference on Multimedia and Expo (ICME) | 978-1-6654-3864-3/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICME51207.2021.9428401

ABSTRACT

Fashion recommendation is an essential component of user shopping that it is capable of selecting and presenting fascinating items to customers. The fact that humans exhibit inconsistencies for fashion items in their choice is known to all due to the visual aesthetic features and fine-grained differences of fashion items. Previous research on fashion recommendations mainly focuses on sequential models, most of them only consider complex similarity relationships in fashion compatibility while neglecting the real-world compatible information often desired in practical applications. To learn the fashion compatibility and generate for the outfit, we propose an approach that jointly learns latent fashion concepts in visual-semantic space to measure compatibility between items. The fashion concepts are shaped by design elements such as color, material, and silhouette. Accordingly, we model a unified representation to learn different notions of similarity by mapping text descriptors and images into latent space to learn high-level representations. Experimental results reveal that our method effectively reaches the aimed results on the fill-in-the-blank and outfit compatibility tasks.

Index Terms— Outfits style, Fashion recommendation, Visual compatibility, Visual-semantic space

1. INTRODUCTION

Fashion has great commercial value due to its capacity for displaying personality and enhancing one's beauty. Beyond the significant commercial value, fashion recommendations have also implicitly modulated people's daily actions and choices. For example, the online fashion market shown by e-commerce platforms (e.g., Amazon, Polyvore) can influence people's buying and shopping decisions. Apparently, fashion recommendation aims to stimulate the desire for shopping by actively exploring to curate exciting experiences for their shoppers. Therefore, the research of fashion compatibility is the main research content in fashion recommendation. Fashion researchers and industry experts have achieved remarkable progress in fashion related research. Despite effectiveness, most of the existing methods model the entire image of fashion outfits to the sequence relationship, and therefore, part features of fashion image are considered. However, the fashion outfits are different from the other items due to the unique pattern, material, and different clothing may have different visually compatible and share similar style [1]. Besides, the compatibility between the outfits is not only affected by the visual aesthetic features of these items but also requires taking rich semantic information into account.

In this work, we propose a unified representation model based on heterogeneous information to jointly learn fashion compatibility relationships with their dependencies in the entire outfit. Specifically, we consider a set of items to represent the outfit and convert groups of compatibility items into a graph, which shares the latent visual information between items in outfits to make compatibility predictions. We view an outfit as a set of fashion items and consider the visual style for the compatibility and the diversity of outfit style (see Fig. 1). We can see the style of an element (a) is not suitable for outfit 2, so element(b) is recommended to outfit 2. Element (c) lacks in outfit 1, so it can be recommended for outfit 1. Our approach is evaluated on Polyvore[1], a popular social commerce web for fashion. The main contributions are as follows.

- We propose a novel approach to explore how visual information and side information can be jointly leveraged to realize more effective fashion recom-

* Corresponding Author: Xiaochun Cheng (X.Cheng AT mdx.ac.uk).

This research is supported by the Science and Technology Development in Guangdong Province (2020B010165001), Guangdong Basic and Applied Basic Research Foundation No.2019A1515011953 and in part by National Key Research and Development Program of China 2020YFB1406902, 2018YFC0832105 and 2018YFC0806802.

Fig. 1. Examples of outfits created by Polyvore[1] users. Different users have different fashion tastes, the recommendation goal is to meet the properties of the latent compositions of visual-semantics that define styles according to learn a style of coherent representation and discover the latent factors.

mendations. We make full use of visual information based on graph networks and text information to construct a new learning framework that jointly learns the embedding of visual and textual information in latent space.

- We present a model that learns the relationship between visual information and fashion styles, and research fashion items from different views. For the problem of data imbalance between classes, we adopt the focus loss factor function to increase the difference between different styles.
- The experimental results demonstrate that the fashion compatibility based on the graph networks and visual-semantic features achieves the expected effect. In addition, the results show the excellent performance that each item of outfits excellent retrieves fashion-sensitive collocations for the remaining items.

2. RELATED WORK

Early methods [2] [3] [4] [5] mainly focused on learning the metric distances, which map the items into a style space and estimate the distance between style vectors of items. Lately, some works [6] [7] model the compatibility of outfits directly by mapping items into several unified style spaces. Veit et al. [3] proposed a new framework to learn a feature transformation from images of items into a latent space that expresses compatibility. However, these works pairwise compatibility relationships rather than an outfit as a whole. On this basis, Wu et al. [1] define an outfit as a sequence that regards each item as a time step to sequentially predict the next item [1] [8]. These methods are independent make each pair of items to compare the distance between items by using Siamese or Triplet network to learn features between a similar pair and a similar-dissimilar triplet respectively, and only explore the first-order relationships between items. Besides, these works mainly focus on the pairwise relationships of compatibility.

Moreover, previous methods [9] [10] [11] are able to learn the distance between a pair of fashion items using metric learning, but they fail to handle the outfit generation task. Vasileva et al. [9] have been trying to jointly optimize item embeddings and category-specific complementary relationship in a unified space, and the methods [12] mainly performed pairwise comparisons among items based on the item information. However, outfits are treated as a whole that needs to consider contextual compatibility.

Although the previous works [13] [9] [14] take advantage of outfit representations to achieve prominent performance, the semantic correlations among items in an outfit are ignored and the compatibility of style is based on visual similarity. Apart from vision information, the semantic information is considered within the fashion recommendation. Due to the semantic gap between vision information and descriptive information, the feature representation is inconsistent. Inspired by attribute features, we focus on fashion attributes to research the style compatibility of outfits. The fundamental task is to realize the unified representation and comprehensive utilization of multi-modal information.

3. PROPOSED METHOD

We take fashion compatibility as an probability prediction problem and aim at modeling outfit compatibility from multiple modalities upon the conventional fashion graph. Fig. 2 gives an overview of our proposed framework that is formulated as a combination of graph models and visual-semantic models. In the framework, first of all, the image features are extracted by the efficientnet, and the visual features are enhanced by the graph attention network. The side information is extracted features through word2vec, and then word features and image features are projected to the visual-semantic space that measures the distance of the items.

[1]www.polyvore.com.

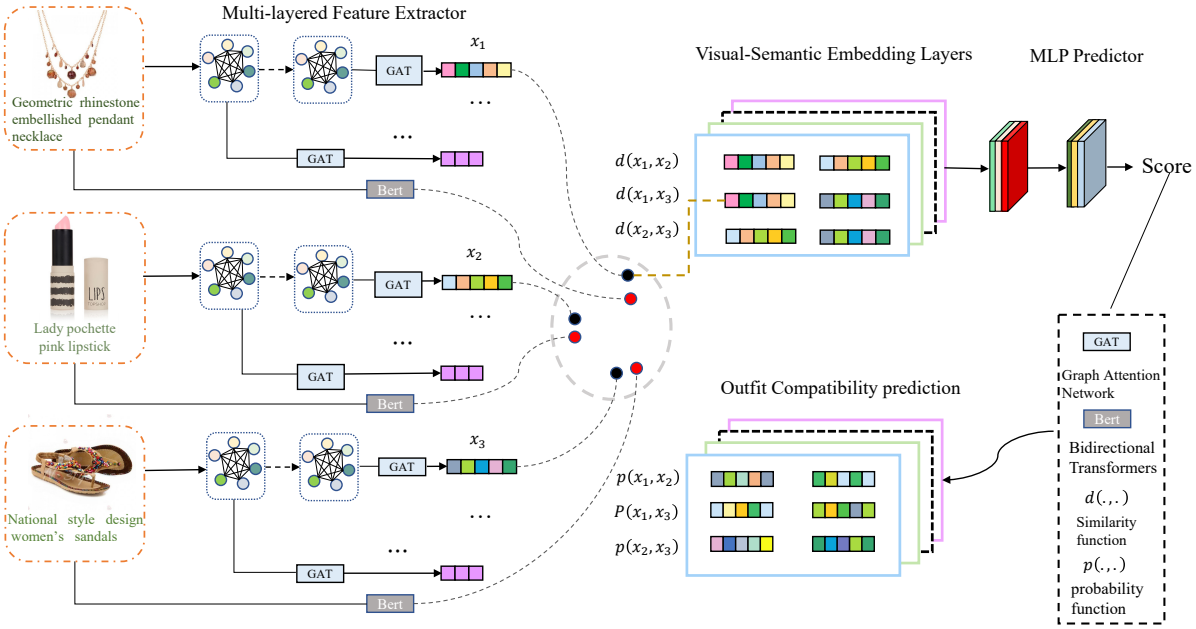

Fig. 2. Overview of the proposed method. The graph is constructed that consists of feature extraction layer, visual semantic embedding layer and outfit compatibility prediction layer. The compatibility prediction layers are used to calculate the compatibility score.

3.1. Multi-layered Feature Extractor Layer

In this work, we define a set of outfits as $S = \{s_1, s_2, \cdots, s_i, \cdots, s_n\}$ consisting of n items, which each item has rich information including content image, item descriptions, etc. We project the item descriptions, whole images of outfits into a latent space, and aim to predict the compatibility score of the outfit.

The images are represented from the visual space to the style space that models the relationship between items by adopting high level layer in the network. Based on the relationship between items in outfits, we construct a fashion outfit graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{E}$ denotes the edges and $\mathcal{E} \in \mathcal{V} \times \mathcal{V}$. For each item, $s_i \in \mathcal{V}$ in the outfit as an individual node [1], and employ the graph attention network to model the visual relationship of outfits. $\boldsymbol{r}_i$ is the feature representation derived from the efficientnet for the i-th fashion item in the outfits [1].

Considering the difference among the categories, we have a linear mapping matrix and obtain the representation of item s_i in the latent space as follows,

$$\boldsymbol{g}_i = \tanh\left(\boldsymbol{W}_i^r \boldsymbol{r}_i\right), \tag{1}$$

where the node takes features $\boldsymbol{g}_i$ as inputs to enrich its own features, which can be updated by aggregating the information of neighbor node s_j [15]. We can obtain the representation of item s_i by the graph attention in the latent space as follows,

$$\boldsymbol{f}_i = Relu\left(\frac{1}{h}\sum_{h=1}^{h}\sum_{j\in N_i} a_{ij}^h \boldsymbol{g}_j \boldsymbol{w}_g^h\right), \tag{2}$$

where $\boldsymbol{f}_i$ denotes the visual feature of the s_i and h is the number of heads. $\boldsymbol{w}_g^h$ is the weight matrix that was used to compute the importance between the source graph nodes and target graph nodes, and it can be calculated as:

$$\alpha_{ij}^h = \frac{\exp\left(\sigma\left(\mathbf{a}^\top[\boldsymbol{f}_i \| \boldsymbol{f}_j]\right)\right)}{\sum_{j\in\mathcal{N}(i)}\exp\left(\sigma\left(\mathbf{a}^\top[\boldsymbol{f}_i \| \boldsymbol{f}_j]\right)\right)}. \tag{3}$$

Where, σ is nonlinear LeakyRelu function, and $\mathbf{a}^\top$ is denoted as the normalized adjacency matrix. The residual connection was applied to further enhance visual representation with the original feature maps. Therefore, the residual connection was applied to the $\boldsymbol{f}_{i-1}$ and we obtain $\boldsymbol{f}_i$ as follows:

$$\boldsymbol{f}_i = \boldsymbol{f}_{i-1} + LayerNorm(F((\boldsymbol{f}_{i-1}))), \tag{4}$$

where F is the feed forward neural network layer. During the training, we use the sigmoid function to get the output score as the class probability as follow,

$$q_i = \text{softmax}(\text{sigmoid}\left(\boldsymbol{w}_i \boldsymbol{f}_i + \boldsymbol{b}_i\right)), \tag{5}$$

where $\boldsymbol{w}_i \in \mathbb{R}^f$ and $\boldsymbol{b}_i \in \mathbb{R}^f$ are the weight matrix and bias vector, respectively.

In the existing data set, the distribution of data is unbalanced among different fashion items. For example, for style, fashion styles favored by common sense usually contain more samples; From the item level, common types of items usually appear more frequently than unusual items. However, graph structure can solve the imbalance problem, since the graph network models the relationship between items. In order to deal with the style imbalance problem of the optional fashion style classifier, we adopt a focus loss factor for cross-entropy as the loss function,

$$\mathcal{L}_v = \sum_{i=1}^{z} y_i \left(1 - q_i\right)^{\gamma} \log q_i, \tag{6}$$

where z is the number of style classes, q_i is the predicted probability for class i, y_i is the ground truth style indicator (0 or 1), and γ is a hyper-parameter.

3.2. Visual-Semantic Embedding

For a pair of fashion items, there is image information as well as text information, which is used as the description of the items. To take advantage of the features of text descriptions accompanying items, we adopt the bert model to learn the vector representation of sentences. For the traditional sentence vector generation method, the word embedding method is used to take the weighted average, but this method fails to understand the semantics of the context. Since the same word may have different meanings in different contexts, it will be represented as the same word embedding by the traditional continuous bag-of-words method. The advantage of bert method is that it can handle the semantic relationship well to generate sentence vectors. For a pair of text features ($\boldsymbol{t}_i, \boldsymbol{t}_j$) corresponding to items pair ($\boldsymbol{s}_i, \boldsymbol{s}_j$), $\boldsymbol{t}_i$ is the semantic vector in the description derived from the corresponding full sentence vector of item s_i [1]. The visual feature and semantic feature are fused to form a unified feature representation in the visual semantic space, the visual-semantic embedding function is defined as follow,

$$\boldsymbol{x}_i = \sigma(w_{tf}\left[\boldsymbol{t}_i; \boldsymbol{f}_i\right]), \tag{7}$$

where w_{tf} is a weight matrix to learn weights, and σ is the sigmoid function. Given a training sample i, j from a set of fashion items, we can predict the compatibility probability between item s_i and item s_j as follow,

$$p(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{\exp\left(\boldsymbol{x}_i \boldsymbol{x}_j^T\right)}{\sum_{j \in \mathcal{N}(i)} \exp\left(\boldsymbol{x}_i \boldsymbol{x}_j^T\right)}. \tag{8}$$

Compatibility is measured with the function $d\left(.,.\right)$, the compatibility function is defined as follow,

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{w}^{(i,j)}) = \|\boldsymbol{x}_i \odot \boldsymbol{w}^{(i,j)} - \boldsymbol{x}_j \odot \boldsymbol{w}^{(i,j)}\|_2^2. \tag{9}$$

Here, $\boldsymbol{w}^{(i,j)}$ is weight for pair $\{\boldsymbol{x}_i, \boldsymbol{x}_j\}$, $\odot$ represents the component-wise multiplication, $d\left(.,.\right)$ is the similarity function between the two embeddings, $\|\cdot\|_2^2$ denotes the euclidean norm. The pair $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is compatible, meaning that the two items appear together in an outfit, while $\boldsymbol{x}_k$ is a randomly sampled item of the same type as $\boldsymbol{x}_j$ that has not been seen in an outfit with $\boldsymbol{x}_i$ [9]. Finally, the unified features are embedded in the joint space by minimizing the following triplet loss,

$$L_{vse} = \max\left(0, m - d\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) + d\left(\boldsymbol{x}_i, \boldsymbol{x}_k\right)\right). \tag{10}$$

The distance is forced to be smaller than the distance from unmatched items $\boldsymbol{x}_k$ by some margin m.

For each sample pair, we not only need to consider the self-similarity of the sample pair, but also its relative similarity with other sample pairs. We take advantage of the adaptive weight mechanism. (s^+, s^-) are set to represent the training samples containing compatible items (s^+) and incompatible items (s^-), the loss is formulated as,

$$\begin{aligned}\mathcal{L}_m = \frac{1}{n}\sum_{i=1}^{n}\Bigg\{ &\frac{1}{\alpha}\log\left[1 + \sum_{j \in s^+} e^{-\alpha(d(x_i, x_j) - \lambda)}\right] \\ &+ \frac{1}{\beta}\log\left[1 + \sum_{k \in s^-} e^{\beta(d(x_i, x_k) - \lambda)}\right]\Bigg\},\end{aligned} \tag{11}$$

where positive examples (s^+) contain two compatible items, whereas negative examples (s^-) contain incompatible items, α, β, λ are hyper-parameters as in loss. The total loss function is as follows,

$$\mathcal{L}\left(\boldsymbol{x}, \lambda_1, \lambda_2, \lambda_3, \theta\right) = \lambda_1 \mathcal{L}_v + \lambda_2 \mathcal{L}_{vse} + \lambda_3 \mathcal{L}_m, \tag{12}$$

where λ_{1-3} are the scalar parameters. θ is the parameters of model.

4. EXPERIMENTAL RESULTS

4.1. Experimental Settings and training details

Given an outfit consisting of garments in polyvore, we construct a graph with visual information in latent space and infer the outfit's compatibility. To stabilize the predictions, our method is implemented by extending the pytorch framework. The efficientnet-b5 network was used to extract the features of fashion images which the embedding size is set to 2048. For each dataset, we train all models at the settings for 6000 epochs for good convergence. For optimization, we adopt the adam optimizer and the propagation steps are set to be 3, hidden size is 12 and learning rate is set to be 0.001. All networks are trained on the TITAN 1080. Due to the GPU memory limitation, the batch size is set to be 16.

4.2. Comparison with State-of-the-art Methods

We use the polyvore dataset released by a popular fashion website polyvore.com. After discarding outfits containing a single item, as well as items that miss type information, the final dataset contains 17,316 for training samples, 3,076 for testing samples, and 4,000 for validation. In the compatibility experiments, some data sets contained less than 3 items. We discarded them and selected four samples in outfits for testing. We replace each item in a ground truth outfit by randomly selecting another item of the same category.

We set up two tasks to evaluate the performance of our model. The two tasks are the fashion compatibility task and fill-in-the-blank that generate the recommends garments. We score the candidate outfit to determine the compatibility of items within outfits. Another task is to select the most compatible items from the outfit to form overall outfits and evaluate the performance by accuracy on the answered questions. Thus, we evaluate performance simultaneously by the FIFB Acc and the Compat AUC. To evaluate our proposed method, we make a detailed comparison with existing methods: SiameseCNNs [9], LMT[2], Bi-LSTM [1], NGNN [6], GCN [13] and MCAN [16]. The comparison results are presented in Table 1. In Table 1, our method obtains a 1%-2% improvement in the accuracy of the fill-in-the-blank. In addition, our methods can learn styles-inherent style and improve the outfit compatibility auc from 0.9600 to 0.9809.

Table 1. The performance of our proposed method with different methods

Method	FIFB Acc.(%)	Compat. AUC (%)
LMT [2]	50.91	67.82
SiameseNet [9]	48.09	70.87
Bi-LSTM [1]	67.01	84.27
GCN [13]	76.82	91.08
NGNN [6]	78.13	97.22
MCAN [16]	86.50	96.00
Ours	87.75	98.09

4.3. Ablation Study

In order to analyze the effectiveness of each part in the framework, we perform an ablation study. The experiments are shown in Table 2, in which VE is a visual feature module, TE is a text feature module, and VSE is a visual semantic embedded module. It can be observed from the table that the performance of the graph neural network is not as good as that of the graph attention module, and the model containing visual information shows excellent compatibility performance. Modules with text information and visual information can better express the compatibility of the model.

Table 2. Ablation study for each part in our framework including the visual embedding module (VE), text embedding module(TE) and visual semantic embedding (VSE) .

Method	FIFB Acc. (%)	Compat. AUC(%)
GCN	76.82	91.08
NGNN	78.13	97.22
GAT	86.91	90.40
VE	85.04	89.52
TE	83.28	83.16
VSE	87.75	98.09

4.4. Case Study

Fig. 3. Example of automatically revising the incompatible outfits.

For each compatible outfit in Fig. 3, we select compatible outfits by randomly swapping one piece for another piece, and we replace one garment from the whole set and predict the compatibility of each garment in the candidate items.

Fig. 4 shows the lady-like looks outfits generated by our algorithm. In the figure, there are two examples of items with the queried item on the left. The generated outfit corresponding to items is outfit 1 and outfit 2 on the right. We can see that outfit 1 and outfit 2 generated according to the shirt is a coarse-grained style outfit with leather material and dark colors with denim and pastel colors. The compatibility score of outfit 1 is 0.9051, and the compatibility score of outfit 2 is 0.9125. The compatibility scores show that outfit1 and outfit2 are matched

Fig. 4. Personalized capsules tailored for user preference.

to the knit shirt. In contrast, outfit 3 and outfit 4 generated according to the knit shirt is a fine-grained style outfit. The compatibility score of outfit 3 is 0.8763, and the compatibility score of outfit 4 is 0.9079. The outfit 3 and outfit 4 have the gradely compatible scores for the shirt. The generated outfit have fine compatibility between items.

5. CONCLUSIONS

This paper presents a novel model that learns the compatibility relationships in visual-semantic space to lead a more natural understanding of fashion style. Our work explores the outfit generation framework that captures latent fashion style from visual information jointly in language learning of visual compatibility. Further, we reseach a collection of fashion items, and adopt a loss factor to solve the problem of data imbalance between classes. Furthermore, our evaluation shows that our method effectively achieves the expect in performance by checking whether two items from different outfits fit together with compatibility.

6. REFERENCES

[1] Xintong Han, Zuxuan Wu, Yu-Gang Jiang, and Larry S. Davis, "Learning fashion compatibility with bidirectional lstms," in MM, 2017, pp. 1078–1086.

[2] Julian J. McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel, "Image-based recommendations on styles and substitutes," in SIGIR, 2015, pp. 43–52.

[3] Andreas Veit, Balazs Kovacs, Sean Bell, Julian J. McAuley, Kavita Bala, and Serge J. Belongie, "Learning visual clothing style with heterogeneous dyadic co-occurrences," in ICCV, 2015, pp. 4642–4650.

[4] Ruiping Yin, Kan Li, Jie Lu, and Guangquan Zhang, "Enhancing fashion recommendation with visual compatibility relationship," in WWW, 2019, pp. 3434–3440.

[5] Guang-Lu Sun, Jun-Yan He, Xiao Wu, Bo Zhao, and Qiang Peng, "Learning fashion compatibility across categories with deep multimodal neural networks," Neurocomputing, vol. 395, pp. 237–246, 2020.

[6] Zeyu Cui, Zekun Li, Shu Wu, Xiaoyu Zhang, and Liang Wang, "Dressing as a whole: Outfit compatibility learning based on node-wise graph neural networks," in WWW, 2019, pp. 307–317.

[7] Xue Dong, Jianlong Wu, Xuemeng Song, Hongjun Dai, and Liqiang Nie, "Fashion compatibility modeling through a multi-modal try-on-guided scheme," in SIGIR, 2020, pp. 771–780.

[8] Suthee Chaidaroon, Yi Fang, Min Xie, and Alessandro Magnani, "Neural compatibility ranking for text-based fashion matching," in SIGIR. 2019, pp. 1229–1232, ACM.

[9] Mariya I. Vasileva, Bryan A. Plummer, Krishna Dusad, Shreya Rajpal, Ranjitha Kumar, and David A. Forsyth, "Learning type-aware embeddings for fashion compatibility," in ECCV, 2018, pp. 405–421.

[10] Rishabh Misra, Mengting Wan, and Julian J. McAuley, "Decomposing fit semantics for product size recommendation in metric spaces," in RecSys. 2018, pp. 422–426, ACM.

[11] Yen-Liang Lin, Son Tran, and Larry S. Davis, "Fashion outfit complementary item retrieval," in CVPR. 2020, pp. 3308–3316, IEEE.

[12] Xun Yang, Xiangnan He, Xiang Wang, Yunshan Ma, Fuli Feng, Meng Wang, and Tat-Seng Chua, "Interpretable fashion matching with rich attributes," in SIGIR, 2019, pp. 775–784.

[13] Guillem Cucurull, Perouz Taslakian, and David Vázquez, "Context-aware visual compatibility prediction," in CVPR, 2019, pp. 12617–12626.

[14] Xue Dong, Xuemeng Song, Fuli Feng, Peiguang Jing, Xin-Shun Xu, and Liqiang Nie, "Personalized capsule wardrobe creation with garment and user modeling," in MM, 2019, pp. 302–310.

[15] Xiang Wang, Xiangnan He, Yixin Cao, Meng Liu, and Tat-Seng Chua, "KGAT: knowledge graph attention network for recommendation," in SIGKDD. ACM, 2019, pp. 950–958.

[16] Xuewen Yang, Dongliang Xie, Xin Wang, Jiangbo Yuan, Wanying Ding, and Pengyun Yan, "Learning tuple compatibility for conditional outfit recommendation," in MM, 2020, pp. 2636–2644.