Modality-Oriented Graph Learning Toward Outfit Compatibility Modeling

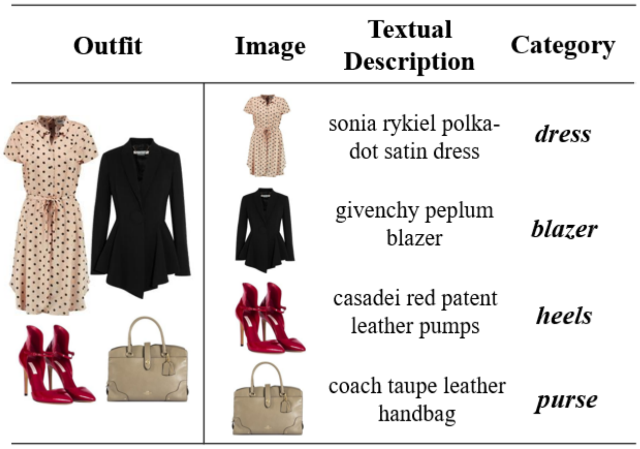
Xuemeng Song, *Member, IEEE,* Shi-Ting Fang, Xiaolin Chen, Yinwei Wei, *Member, IEEE,* Zhongzhou Zhao, and Liqiang Nie, *Senior Member, IEEE*

***Abstract*—Outfit compatibility modeling, which aims to automatically evaluate the matching degree of an outfit, has drawn great research attention. Regarding the comprehensive evaluation, several previous studies have attempted to solve the task of outfit compatibility modeling by integrating the multi-modal information of fashion items. However, these methods primarily focus on fusing the visual and textual modalities, but seldom consider the category modality as an essential modality. In addition, they mainly focus on the exploration of the intra-modal compatibility relation among fashion items in an outfit but ignore the importance of the inter-modal compatibility relation, *i.e.,* the compatibility across different modalities between fashion items. Since each modality of the item could deliver the same characteristics of the item as other modalities, as well as certain exclusive features of the item, overlooking the inter-modal compatibility could yield sub-optimal performance. To address these issues, a multi-modal outfit compatibility modeling scheme with modality-oriented graph learning is proposed, dubbed as MOCM-MGL, which takes both the visual, textual, and category modalities as input and jointly propagates the intra-modal and inter-modal compatibilities among fashion items. Experimental results on the real-world Polyvore Outfits-ND and Polyvore Outfits-D datasets have demonstrated the superiority of our proposed model over existing methods.**

***Index Terms*—Outfit Compatibility Modeling, Multi-modal Recommendation, Graph Convolutional Network**

I. Introduction

RECENT years have witnessed the flourish of the online fashion industry, which reveals people's increasing desires for fashion clothing. However, not everyone is sensitive to the appreciation of beauty and good at making compatible outfits. Thanks to the prosperity of online fashion-oriented websites (*e.g.,* SSENSE[1] and CHICTOPIA[2]), a large number of compatible outfits composed by fashion experts are publicly available, which opens the door for investigating solutions for automatically rating the matching degree of an outfit, *i.e.,* outfit compatibility modeling. As shown in Figure 1, each outfit usually consists of multiple fashion items, each of which is characterized by an image, a textual description, and category information. Therefore, to fully utilize the cues delivered by different modalities of fashion items and comprehensively model the compatibility of outfits, many research efforts [1], [2] have attempted to tackle the problem of the outfit compatibility modeling with the multi-modal information of fashion items.



Outfit	Image	Textual Description	Category
		sonia rykiel polka-dot satin dress	***dress***
		givenchy peplum blazer	***blazer***
		casadei red patent leather pumps	***heels***
		coach taupe leather handbag	***purse***

Fig. 1. Example of an outfit composition.

Despite their remarkable performance, they mainly suffer from the following two key limitations. 1) Prior studies mainly focus on visual and textual modalities, and few of them utilize the category information of fashion items. Additionally, these few studies [3], [4] mainly focus on using items' categories to supervise the model learning, but fail to regard the category information as one essential input modality, *i.e.,* comparable to the visual and textual modalities. And 2) previous efforts mainly focus on the intra-modal compatibility, *i.e.,* the compatibility relation between the same modalities of fashion items in an outfit, but overlook the inter-modal compatibility, *i.e.,* the compatibility relation between different modalities of fashion items, thereby probably causing sub-optimal performance. The underlying philosophy is twofold: a) since different modalities

Manuscript received September 21, 2021; revised November 22, 2021; accepted November 26, 2021. This work was supported in part by the Shandong Provincial Natural Science Foundation under Grant ZR2019JQ23, in part by the Key Research and Development Program of Shandong (Major scientific and technological innovation projects) under Grant 2020CXGC010111, in part by the National Natural Science Foundation of China under Grant U1936203, and in part by the Young Creative Team in Universities of Shandong Province under Grant 2020KJN012. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Liangliang Cao. *(Corresponding author: Liqiang Nie.)*

Xuemeng Song, Shi-Ting Fang, and Liqiang Nie are with the School of Computer Science and Technology, Shandong University, Qingdao, Shandong, China (e-mail: sxmustc@gmail.com; fangshiting@mail.sdu.edu.cn; nieliqiang@gmail.com).

Xiaolin Chen is with the School of Software, Shandong University, Jinan, Shandong, China (e-mail: cxlicd@gmail.com).

Yinwei Wei is with the School of Computing, National University of Singapore, Singapore (e-mail: weiyinwei@hotmail.com).

Zhongzhou Zhao is with the DAMO Academy, Alibaba Group, China (e-mail: zhongzhou.zhaozz@alibaba-inc.com).

[1] https://www.ssense.com/.

[2] https://www.chictopia.com/.

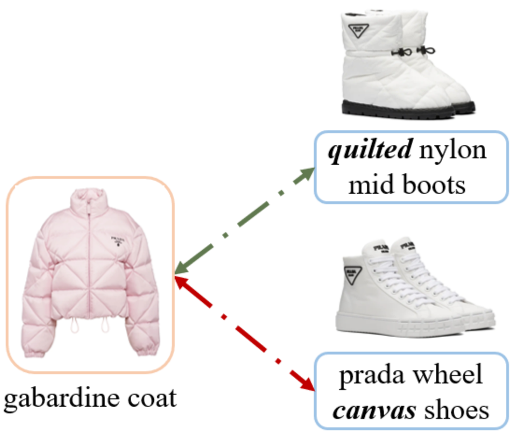


Fig. 2. Examples of the inter-modal compatibility relation. The green and red arrows represent the compatible and incompatible relation, respectively.

of an item tend to reflect the same characteristics of the fashion item [5], [6], incorporating the inter-modal compatibility (*e.g.,* visual-textual compatibility) can supplement the intra-modal compatibility (*e.g.,* visual-visual compatibility) and strengthen the overall compatibility estimation from an auxiliary perspective. b) Meanwhile, different modalities of the same fashion item can also emphasize different aspects of the same item. For example, the visual modality is more likely to reveal the color and pattern of the item, while the textual modality tends to deliver its material and brand. As seen from Figure 2, the given coat is visually compatible with both pairs of shoes. However, if the inter-modal compatibilities between the image of the coat and the textual descriptions of the two pairs of shoes are investigated, it would be easy to determine that the given top is more suitable to go with the quilted boots rather than the canvas shoes.

Towards the outlined limitations, we propose to incorporate items' category information with their content information (*i.e.,* visual and textual modalities) and jointly model their intra-modal and inter-modal compatibilities to optimize the outfit compatibility modeling. However, this is non-trivial due to the following challenges. 1) Undoubtedly, the visual modality plays a pivotal role in outfit compatibility modeling. It usually delivers not only the low-level visual features (*e.g.,* color, shape) but also the high-level visual features (*e.g.,* style) of fashion items. Therefore, how to thoroughly explore the low-level and high-level visual features and thus benefit the compatibility modeling poses a key challenge for us. 2) Since each outfit always comprises various fashion items, among which there is no clear order, and the matching degree between each pair of items affects the outfit compatibility, we model the outfit as an item graph. Moreover, similar to existing studies [7], [8], we resort to Graph Convolutional Networks (GCNs) [9] to fulfill the outfit compatibility modeling. Accordingly, how to effectively propagate both the intra-modal and inter-modal compatibilities among the fashion graph to derive the outfit compatibility also constitutes an essential challenge. And 3) essentially, one key step of the outfit compatibility modeling is to learn an accurate latent representation of the outfit that can capture the compatibility of the outfit. Therefore, how to seamlessly unify the multi-modal information of fashion items to derive the latent outfit representation is another crucial challenge.

To address the aforementioned challenges, a multi-modal outfit compatibility modeling scheme with modality-oriented graph learning is presented, dubbed as MOCM-MGL. As shown in Figure 3, MOCM-MGL consists of three modules: *Multi-modal Embedding*, *Modality-oriented Graph Learning*, and *Outfit Compatibility Estimation*. The multi-modal embedding module comprises three encoders to extract the visual, textual, and category features of fashion items. In particular, multiple intermediate convolutional layers of the CNN are adopted to derive both the low-level and high-level visual features. In addition, the TextCNN [10] is utilized to embed the textual modality, and directly assign the to-be-learned embedding vector to each category. The modality-oriented graph learning module introduces a multi-modal item graph for each outfit and propagates both the intra-modal and inter-modal compatibility relation among fashion items to refine the fashion item representations. Notably, instead of simply using the 1-D co-occurrence frequency of categories, the edge between two item nodes is defined by a multi-dimensional embedding to encode the complex compatibility relation between two items. Ultimately, the outfit compatibility estimation module derives the latent outfit representation by aggregating all the composing items' representations, and based on that, estimates the outfit compatibility with the Multi-Layer Perceptron (MLP) [11].

The main contributions can be summarized threefold:

- We present a novel multi-modal outfit compatibility modeling scheme with modality-oriented graph learning, MOCM-MGL. To the best of our knowledge, this is the first attempt to fully exploit the visual, textual, and category modalities with GCN for outfit compatibility modeling.
- The intra-modal and inter-modal compatibility relation between two fashion items is clearly defined and unified to thoroughly model the outfit compatibility.
- Extensive experiments conducted on the Polyvore Outfits-ND and Polyvore Outfits-D datasets demonstrate the superiority of our proposed method over the state-of-the-art methods. As a byproduct, we have released the codes and involved parameters to benefit other researchers[3].

The remainder of this paper is organized as follows. Section II briefly reviews the related work. Section III details the proposed MOCM-MGL. The experimental results and analyses are presented in Section IV, followed by the conclusion and future work in Section V.

II. RELATED WORK

A. *Outfit Compatibility Modeling*

The recent flourish of the fashion industry has promoted researchers to pay attention to many fashion analysis tasks, such as clothing retrieval [12], compatibility modeling [13],

[3] https://outfitcompatibility.wixsite.com/mocm-mgl.

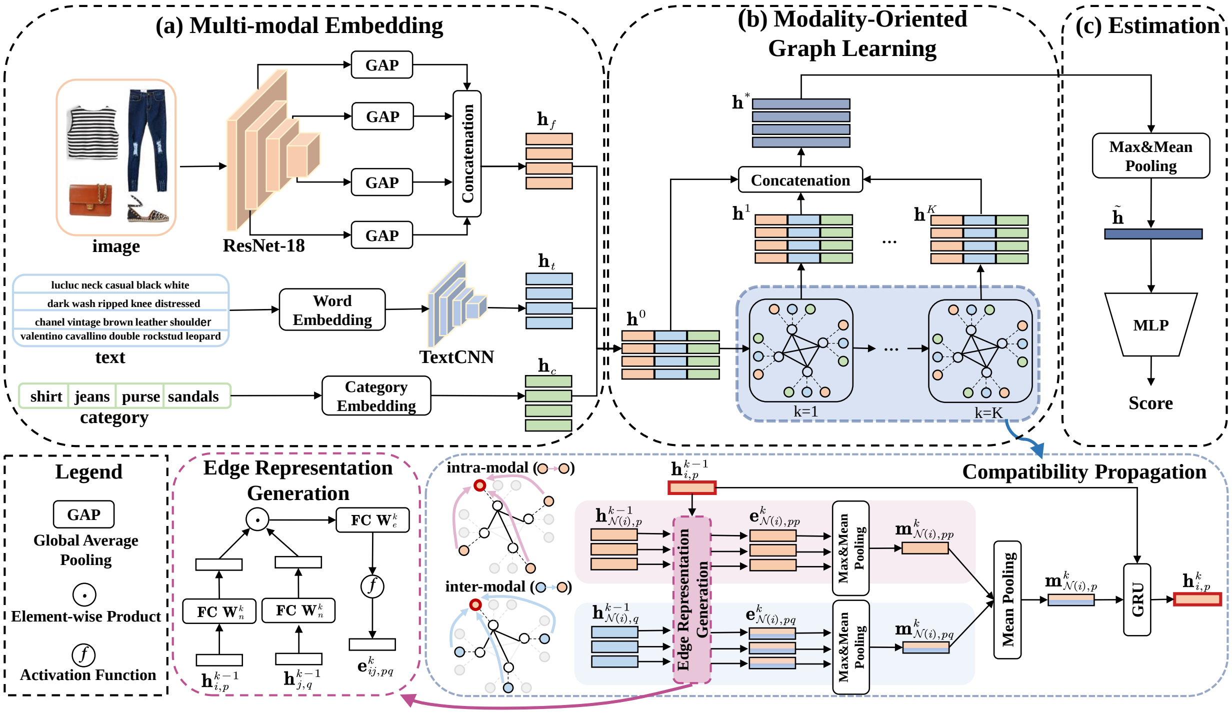


Fig. 3. Illustration of the proposed scheme, which consists of three modules: *Multi-modal Embedding*, *Modality-oriented Graph Learning*, and *Outfit Compatibility Estimation*. The multi-modal embedding module extracts the multi-modal features of fashion items, and the modality-oriented graph learning module refines the representation of each fashion item by absorbing its intra-modal and inter-modal compatibility relation with the other items. Ultimately, the outfit compatibility estimation module first aggregates the composing items' representations and then uses the MLP to estimate the outfit compatibility score.

[14], fashion trend prediction [15], and clothing recommendation [16], [17]. In particular, as the key to many fashion-oriented applications, such as complementary item retrieval [18] and personal capsule wardrobe creation [19], outfit compatibility modeling has drawn great research attention. According to the input information of fashion items, existing outfit compatibility modeling studies can be broadly grouped into two categories: single-modal methods and multi-modal methods.

Single-modal methods only utilize the visual or textual modality of fashion items. Apparently, the visual modality plays a significant role in outfit compatibility modeling, as many characteristics of items, such as color and shape, are mainly encoded by visual information. Therefore, existing efforts mainly exploit the visual information of fashion items. For example, Tangseng *et al.* [20] defined an outfit as a few ordered slots, corresponding to the common item categories (range from outerwear to accessory), and concatenated the visual representations of all the composing items in the outfit as the outfit representation. In addition, Cucurull *et al.* [21] built a graph with all fashion items in the dataset, where each node is initialized by the corresponding visual feature and receives the message from its neighborhood to learn the contextual item embedding. Apart from the visual modality, Chaidaroon *et al.* [22] investigated the potential of the textual modality of fashion items in the outfit compatibility modeling, where a text-based neural compatibility ranking model is proposed. Although great progress has been made by these works, they utilize only one modality of fashion items and overlook the potential to combine the multi-modal information of fashion items.

Multi-modal methods involve more than one modality of fashion items. For example, Han *et al.* [23] proposed a bidirectional LSTM method to sequentially model the outfit compatibility by predicting the next item conditioned on previous items, where the visual semantic embedding (VSE) [24] is used to capture the inter-modal consistency of visual and textual modalities. Apparently, this method only considers the consistency between two modalities of fashion items and neglects the complementarity between them. Towards this end, several researchers have been trying to use the fusion strategy (*i.e.,* early fusion and late fusion) to integrate the multi-modal information. 1) Early fusion based approaches typically fuse the input features extracted from each modality into a single representation before compatibility modeling [25]. For example, Tan *et al.* [26] fused the visual and textual features of fashion items by the element-wise product operation, while Yang *et al.* [27] and Sun *et al.* [28] directly combined the visual and textual features of each item by the concatenation operation before feeding the item feature into the compatibility modeling module. In addition, Laenen *et al.* [29] used the attention mechanism to fuse the visual and textual features, and projected the multi-modal representations to the type-specific compatibility spaces. 2) Late fusion based methods [7], [30] first perform the compatibility modeling directly over each modality feature, and then linearly combine the estimated outfit compatibility scores from different modalities. For example, Cui *et al.* [7] introduced the Node-wise Graph Neural

Network (NGNN) for the outfit compatibility modeling from each modality. The overall outfit compatibility score is derived by a weighted summation of the scores obtained by the visual and textual modalities. In a sense, both early fusion based and late fusion based methods overlook the importance of the inter-modal compatibility relation between fashion items in the outfit compatibility modeling, which is the major concern of this work.

It is worth mentioning that some multi-modal methods have incorporated the category information as an indicator to guide the outfit compatibility modeling. For example, Vasileva *et al.* [3] presented a pair-wise outfit compatibility modeling scheme, where a category-specific embedding space is introduced for each pair of categories. Besides, Wang *et al.* [4] learned the overall compatibility from all category-specified pairwise similarities between fashion items, and used the backpropagation gradients to diagnose incompatible factors. Differently, as a major novelty, we take the category modality as one essential input modality, *i.e.,* comparable to the visual and textual modalities, to enhance the outfit compatibility modeling performance with GCN.

B. Graph Convolutional Network

Recently, Graph Neural Networks (GNNs) have attracted increasing research attention due to the great expressive power of graphs [31]. The concept of GNNs is first proposed in [32], which extends the neural networks for processing the data represented in graph domains. The target of GNNs is to learn a state embedding for each node, which absorbs the information of one's neighborhood. To remedy the limitation that CNNs cannot be operated on non-Euclidean data, some researchers redefined the notion of convolution for graph data and proposed GCNs [33].

GCNs have been widely used in many tasks and domains, such as computer vision [34], [35], natural language processing [36], [37], and recommender systems [38]–[41]. Recently, in the fashion domain, as each outfit can be abstracted as an item graph, several GCN-based methods, such as NGNN [7], HFGN [8], and Neural Graph Filtering [42], have been proposed for fashion compatibility modeling. The rationale of these methods is to update the item embedding with its context in the outfit. Different from the methods that only propagate the general item embedding, in this work, we conducted a modality-oriented GCN, which jointly propagates the intra-modal and inter-modal compatibility relation among fashion items in an outfit.

III. METHODOLOGY

In this section, we first present the notations and problem formulation and then detail the proposed multi-modal outfit compatibility modeling scheme.

A. Notations and Problem Formulation

Since different modalities (*e.g.,* the visual image, text description, and category) can deliver different aspects of fashion items, we propose to explore all the modalities of fashion items to comprehensively measure the compatibility score of outfits. Suppose that we have a set of Q fashion items $\mathcal{I} = \{x_i\}_{i=1}^{Q}$, coming from N_c categories. Each fashion item $x_i \in \mathcal{I}$ is attached with a visual image, a textual description, and a category, termed as f_i, t_i, and c_i, respectively. Based on these items, we can derive a set of N training outfit samples $\Omega = \{(\mathcal{O}^j, y^j)|j = 1, \cdots, N\}$, where $\mathcal{O}^j$ is the j-th outfit, and y^j is the ground truth label that indicates whether the outfit is compatible or not. Specifically, $y^j = 1$ denotes that the j-th outfit $\mathcal{O}^j$ is compatible, and $y^j = 0$ otherwise. Each outfit can be regarded a set of fashion items, *i.e.,* $\mathcal{O}^j = \left\{x_1^j, x_2^j, \cdots, x_{S_j}^j\right\}$, where $x_i^j \in \mathcal{I}$ denotes the i-th composing item of the outfit $\mathcal{O}^j$, and S_j represents the total number of fashion items in the outfit $\mathcal{O}^j$. Notably, since each outfit can be composed of various fashion items, S_j is variable. Based on these data, we aim to devise a comprehensive multi-modal outfit compatibility modeling scheme $\mathcal{F}$, which is capable of integrating the multi-modal information of its composing fashion items toward the accurate outfit compatibility estimation. Mathematically, we have:

$$\hat{y}^j = \mathcal{F}(\{f_i^j, t_i^j, c_i^j\}_{i=1}^{S_j}|\boldsymbol{\Theta}), \tag{1}$$

where f_i^j, t_i^j, and c_i^j represent the visual image, textual description, and category of the i-th item of the j-th outfit, respectively. $\boldsymbol{\Theta}$ is the set of to-be-learned parameters, and $\hat{y}^j$ denotes the estimated compatibility score of the outfit $\mathcal{O}^j$. For brevity, we omit the superscript j of the j-th outfit $\mathcal{O}^j$ in the rest of this paper.

B. Multi-modal Embedding

First, we resort to the following encoders to learn the visual, textual, and category representation of each fashion item, respectively.

Image Encoder. Regarding the visual image of each item, we utilize the CNN to extract its visual features. It is well known that the CNN comprises multiple convolutional layers, where the shallow layers can capture the low-level visual features, such as the color of the item, while the deep layers can capture the high-level features, such as the style of the item [43]. Since both the low-level and high-level visual features would affect the compatibility among fashion items, similar to the work [4], we take both the shallow and deep layers' outputs into consideration to learn the visual representation for each item instead of only using the final layer's output. In particular, we resort to the Global Average Pooling operation (GAP) [44], which has shown remarkable performance in the discriminative visual property extraction [45], to summarize the learned visual representations. Formally, given the image f_i of the fashion item x_i, we can obtain its visual feature as follows,

$$\mathbf{f}_i = \left[\text{GAP}\left(Conv^1\left(f_i\right)\right), \cdots, \text{GAP}\left(Conv^L\left(f_i\right)\right)\right], \tag{2}$$

where $\mathbf{f}_i \in \mathbb{R}^{d_f}$ is the visual feature of the item x_i, d_f is the dimension of the visual feature, and $[\ ,\ ]$ denotes the concatenation operation. In addition, $Conv^l$ represents the l-th convolutional layer used for visual encoding of CNN, and L is the total number of convolutional layers.

Text Encoder. To embed the textual description of each fashion item, we adopt the TextCNN, which has achieved astonishing success in various natural language processing tasks [46], [47]. In particular, we first represent the textual description (*i.e.*, a sequence of words) as a matrix, each column of which refers to a word embedding learned by the pre-trained word2vec [48]. We then employ the CNN architecture to extract the semantic information of the text description of the given fashion item. Specifically, given the textual description t_i of the fashion item x_i, we obtain its textual feature as follows,

$$\mathbf{t}_i = TextCNN\left(t_i\right), \tag{3}$$

where $\mathbf{t}_i \in \mathbb{R}^{d_t}$ denotes the extracted textual feature, and d_t is its dimension.

Category Encoder. In addition to the visual and textual information, the category information of the composing items also plays an important role in the outfit compatibility estimation. Different from previous studies that only incorporate the category information to guide the outfit compatibility modeling, we propose to regard the category as a unique input modality. To represent the discrete categories, we introduce a category embedding matrix $\mathbf{C} \in \mathbb{R}^{N_c \times d_c} = \{\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_{N_c}\}$, where N_c is the total number of categories in the dataset, d_c is the dimension of the category feature, and $\mathbf{c}_k$ denotes the embedding for the k-th category. Therefore, for each fashion item x_i, its category feature $\mathbf{c}_i$ can be obtained according to its category information c_i.

Ultimately, based on the above three encoders, for each fashion item x_i, we can obtain its visual feature $\mathbf{f}_i$, textual feature $\mathbf{t}_i$, and category feature $\mathbf{c}_i$.

C. Modality-oriented Graph Learning

Since each outfit comprises a set of fashion items with no clear order, we treat each outfit as an item graph and hence resort to the GCN to explore its outfit compatibility. In particular, we construct an item graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ for each outfit $\mathcal{O}$, where $\mathcal{V}$ denotes the set of nodes, each of which represents a composing item, and $\mathcal{E}$ stands for the set of edges representing the compatibility relation among items. We assume that the compatibility between each pair of items should be considered in the outfit compatibility modeling, and thus make the fashion graph a complete graph. Namely, there is an edge between each pair of nodes.

1) Node Initialization: Different from the conventional methods that assign each node with a single hidden state vector, we attribute each node with three modality-oriented hidden state vectors, corresponding to the three modalities. Concretely, for each fashion item x_i, we employ linear transformations to map its multi-modal features into a common space to derive the modality-oriented hidden state vectors as follows,

$$\begin{cases} \mathbf{h}_{i,1}^{0} = \mathbf{W}_f \mathbf{f}_i + \mathbf{b}_f, \\ \mathbf{h}_{i,2}^{0} = \mathbf{W}_t \mathbf{t}_i + \mathbf{b}_t, \\ \mathbf{h}_{i,3}^{0} = \mathbf{W}_c \mathbf{c}_i + \mathbf{b}_c, \end{cases} \tag{4}$$

where $\mathbf{h}_{i,1}^{0} \in \mathbb{R}^{d}$, $\mathbf{h}_{i,2}^{0} \in \mathbb{R}^{d}$, and $\mathbf{h}_{i,3}^{0} \in \mathbb{R}^{d}$ are the initial hidden representations of the i-th item's visual, textual and category modalities, respectively. For ease of the following presentation, without losing the generality, we arrange the visual, textual and category modalities as the first, second, and third modalities of fashion items, respectively. $\mathbf{W}_f \in \mathbb{R}^{d_f \times d}$, $\mathbf{W}_t \in \mathbb{R}^{d_t \times d}$, and $\mathbf{W}_c \in \mathbb{R}^{d_c \times d}$ are the linear mapping matrices, $\mathbf{b}_f \in \mathbb{R}^{d}$, $\mathbf{b}_t \in \mathbb{R}^{d}$ and $\mathbf{b}_c \in \mathbb{R}^{d}$ are the biases, where d is the dimension of the hidden state representation.

2) Edge Representation Generation: Previous GCN-based studies [7], [8] on outfit compatibility modeling mainly utilize edges to indicate the graph topological information and assign each edge with a scalar. Beyond that, we model the edge between two items with a multi-dimensional feature rather than a one-dimensional weight, which is capable of encoding the complex compatibility relation between items.

As mentioned above, apart from the intra-modal compatibility, the interaction of different modalities between fashion items can also deliver the compatibility relation between fashion items. Towards this end, we introduce the fine-grained edge representation $\mathbf{e}_{ij,pq}$ to capture the compatibility between the p-th modality of node v_i and the q-th modality of node v_j, where $p, q = 1, 2, 3$. It is worth noting that 1) when $p = q$, $\mathbf{e}_{ij,pq}$ represents the intra-modal compatibility relation, and 2) when $p \neq q$, $\mathbf{e}_{ij,pq}$ represents the inter-modal compatibility relation. Regarding the fine-grained edge representation generation, it is worth noting that the order of items in each item pair does not influence the underlying compatibility relation Accordingly, in this work, we employ the symmetric element-wise product function to generate the edge representation. Specifically, we produce the fine-grained edge representation $\mathbf{e}_{ij,pq}$ for the k-th propagation step as follows,

$$\mathbf{e}_{ij,pq}^{k} = \alpha\left(\mathbf{W}_e^{k}\left(\mathbf{W}_n^{k}\mathbf{h}_{i,p}^{k-1} \odot \mathbf{W}_n^{k}\mathbf{h}_{j,q}^{k-1}\right) + \mathbf{b}_e^{k}\right), \tag{5}$$

where $\mathbf{h}_{i,p}^{k-1}$ is the hidden representation of the p-th modality of the node v_i, and $\mathbf{h}_{j,q}^{k-1}$ is the hidden representation of the q-th modality of the node v_j. $\mathbf{W}_n^{k} \in \mathbb{R}^{d \times d}$, $\mathbf{W}_e^{k} \in \mathbb{R}^{d \times d_e}$, and $\mathbf{b}_e^{k} \in \mathbb{R}^{d_e}$ are the parameters for the edge representation generation in the k-th propagation step. $\mathbf{W}_n^{k}$ is the weight matrix of the linear transformation to project the node embedding to latent compatibility space, while $\mathbf{W}_e^{k}$ is the weight matrix of the linear transformation to further compress the latent compatibility relation into a lower-dimensional space, where the compatibility relation with all the other nodes is aggregated. In particular, to facilitate the following mean pooling and max pooling based compatibility aggregation, we make $d_e = d/2$. $\alpha\left(\cdot\right)$ is the ReLU activation function, and $\odot$ denotes the element-wise product operation.

3) Intra-modal and Inter-modal Compatibility Propagation: During the intra-modal and inter-modal compatibility propagation, we make each modality of each node absorb the fine-grained compatibility information from its connected edges to update its hidden state vector. Without losing the generality, as an example, we present the compatibility aggregation process toward the p-th modality of the node v_i as follows,

$$\begin{cases} \mathbf{m}_{\mathcal{N}(i),pq}^{k} = \mathrm{AGG}\left(\left\{\mathbf{e}_{ij,pq}^{k}, \forall j \in \mathcal{N}\left(i\right)\right\}\right), \\ \mathbf{m}_{\mathcal{N}(i),p}^{k} = \dfrac{1}{M}\displaystyle\sum_{q=1}^{M} \mathbf{m}_{\mathcal{N}(i),pq}^{k}, \end{cases} \tag{6}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3134164, IEEE Transactions on Multimedia

where $\mathcal{N}(i)$ is the neighbors of node v_i, *i.e.,* the nodes connected to the node v_i in the graph. $\mathbf{m}_{\mathcal{N}(i),pq}^k \in \mathbb{R}^d$ denotes the aggregated compatibility information from the q-th modality of the node's neighbors toward the p-th modality of the node v_i in the k-th propagation step, while $\mathbf{m}_{\mathcal{N}(i),p}^k \in \mathbb{R}^d$ represents the aggregated compatibility information from all the modalities of the node's neighbors toward the p-th modality of the node v_i in the k-th propagation step. M is the total number of modalities, which is 3 in our context. $\mathrm{AGG}(\cdot)$ is the aggregation function, which is implemented with both the mean and max pooling operations. Specifically, we have $\mathbf{m}_{\mathcal{N}(i),pq}^k =$

$$\left[\gamma_{mean}\left(\left\{\mathbf{e}_{ij,pq}^k, \forall j \in \mathcal{N}(i)\right\}\right), \gamma_{max}\left(\left\{\mathbf{e}_{ij,pq}^k, \forall j \in \mathcal{N}(i)\right\}\right)\right], \tag{7}$$

where $\gamma_{mean}(\cdot)$ and $\gamma_{max}(\cdot)$ are the mean and max pooling operations, respectively. In a sense, the mean and max pooling operations are used for extracting the average and most prominent information from the connected edges, respectively.

Then, we adopt the Gated Recurrent Unit (GRU) [49] to selectively absorb the compatibility information from the node's neighbors and the original hidden information of the node. Specifically, we define the modality representation update function for each node as follows,

$$\begin{cases} \mathbf{z}_{i,p}^k = \sigma\left(\mathbf{W}_z^k\left[\mathbf{m}_{\mathcal{N}(i),p}^k, \mathbf{h}_{i,p}^{k-1}\right] + \mathbf{b}_z^k\right), \\ \mathbf{r}_{i,p}^k = \sigma\left(\mathbf{W}_r^k\left[\mathbf{m}_{\mathcal{N}(i),p}^k, \mathbf{h}_{i,p}^{k-1}\right] + \mathbf{b}_r^k\right), \\ \tilde{\mathbf{h}}_{i,p}^k = \tanh\left(\mathbf{W}_h^k\left[\mathbf{m}_{\mathcal{N}(i),p}^k, \mathbf{r}_{i,p}^k \odot \mathbf{h}_{i,p}^{k-1}\right] + \mathbf{b}_h^k\right), \\ \mathbf{h}_{i,p}^k = \left(1 - \mathbf{z}_{i,p}^k\right) \odot \mathbf{h}_{i,p}^{k-1} + \mathbf{z}_{i,p}^k \odot \tilde{\mathbf{h}}_{i,p}^k, \end{cases} \tag{8}$$

where $\mathbf{W}_z^k \in \mathbb{R}^{2d\times d}$, $\mathbf{W}_r^k \in \mathbb{R}^{2d\times d}$, and $\mathbf{W}_h^k \in \mathbb{R}^{2d\times d}$ are weight matrices of the update function, while $\mathbf{b}_z^k \in \mathbb{R}^d$, $\mathbf{b}_r^k \in \mathbb{R}^d$, and $\mathbf{b}_h^k \in \mathbb{R}^d$ are biases. $\mathbf{z}_{i,p}^k$ and $\mathbf{r}_{i,p}^k$ are update gate vector and reset gate vector, respectively. $\sigma(\cdot)$ is the sigmoid activation function, and $tanh(\cdot)$ is the tanh activation function. $\mathbf{h}_{i,p}^k$ denotes the hidden representation of the p-th modality of the item x_i in the k-th propagation step. As can be seen, the node update function (*i.e.,* GRU) takes both the hidden modality representation of node v_i and the aggregated compatibility information $\mathbf{m}_{\mathcal{N}(i),p}^k$ as the input. In this manner, the updated representation of the node v_i comprises not only the item intrinsic characteristics but also the compatibility relation with connected items.

D. Outfit Compatibility Estimation

After K propagation steps, we obtain a series of multimodal hidden representations of fashion item x_i, namely $\left\{\mathbf{h}_i^0, \cdots, \mathbf{h}_i^K\right\}$, where $\mathbf{h}_i^k = \left[\mathbf{h}_{i,1}^k, \mathbf{h}_{i,2}^k, \mathbf{h}_{i,3}^k\right]$, and $k = 1, \cdots, K$. Since the representations obtained at different propagation layers absorb the neighbor compatibility information at different levels, toward the comprehensive representation, we concatenate them to constitute the final representation of each fashion item x_i as follows,

$$\mathbf{h}_i^* = \left[\mathbf{h}_i^0, \cdots, \mathbf{h}_i^K\right], i = 1, \cdots, S_j. \tag{9}$$

Thereafter, we define the final representation of the outfit based on these composing items' representations. Notably, instead of using the concatenation of all composing items' representations, we further apply a pooling layer that includes both the max pooling and the mean pooling operations to derive the whole outfit representation. We expect that the max pooling and mean pooling operations can capture the most prominent and overall features of all composing items' hidden states, respectively. Specifically, we obtain the final embedding for each outfit $\mathcal{O}$ as follows,

$$\tilde{\mathbf{h}} = \left[\gamma_{mean}\left(\left\{\mathbf{h}_i^*, \forall v_i \in \mathcal{V}\right\}\right), \gamma_{\max}\left(\left\{\mathbf{h}_i^*, \forall v_i \in \mathcal{V}\right\}\right)\right]. \tag{10}$$

Ultimately, an MLP with two layers is empirically chosen as the final compatibility estimation, in which the outfit embedding is fed into to compute the final compatibility score of the outfit $\mathcal{O}$ as follows,

$$\hat{y} = \sigma\left(\mathbf{W}_2\left(\alpha\left(\mathbf{W}_1\tilde{\mathbf{h}} + \mathbf{b}_1\right)\right) + \mathbf{b}_2\right), \tag{11}$$

where $\mathbf{W}_1 \in \mathbb{R}^{d_o \times d'}$, $\mathbf{W}_2 \in \mathbb{R}^{d' \times 1}$, $\mathbf{b}_1 \in \mathbb{R}^{d'}$ and $\mathbf{b}_2 \in \mathbb{R}^1$ are the parameters of the MLP, where d_o is the dimension of $\tilde{\mathbf{h}}$, and d' is the number of hidden units of the MLP. σ is the sigmoid active function, used for projecting the estimated compatibility score into the range of $[0, 1]$, and making the estimated compatibility score can be regarded as the probability that the outfit is compatible.

Optimization To optimize the proposed model, we adopt the binary cross-entropy loss, which shows the great superiority in the classification task [50], [51], formally,

$$\mathcal{L}_{clf} = -y\log(\hat{y}) - (1 - y)\log(1 - \hat{y}), \tag{12}$$

where $\hat{y}$ and y denote the estimated score and the ground truth label, respectively. Inspired by [4], to encourage the CNN to encode normalized representations in the latent space, we add additional loss to penalize the training process as follows,

$$\mathcal{L}_{emb} = \sum_{i=1}^{S} \|\mathbf{f}_i\|_2, \tag{13}$$

where S represents the number of fashion items in an outfit sample, and $\|\cdot\|_2$ denotes the Euclidean norm of a vector. Ultimately, the final objective function can be formulated as follows,

$$\mathcal{L}_{total} = \sum_{\Omega}\left(\mathcal{L}_{clf} + \lambda_1\mathcal{L}_{emb}\right) + \lambda_2\|\mathbf{\Theta}\|_F^2, \tag{14}$$

where λ_1 and λ_2 are the trade-off hyper-parameters, controlling the weights for the normalization loss and overfitting regularization loss, respectively. As aforementioned, Ω is the training set, and $\mathbf{\Theta}$ refers to the set of to-be-learned parameters. $\|\cdot\|_F$ denotes the Frobenius norm of a matrix.

IV. EXPERIMENT

To evaluate the proposed method, we conducted extensive experiments on the real-world dataset by answering the following research questions:

- RQ1: Does MOCM-MGL achieve better performance than state-of-the-art methods?
- RQ2: How does each component affect the MOCM-MGL?

- RQ3: How does each modality influence the performance?
- RQ4: How about the sensitivity of MOCM-MGL for certain vital hyper-parameters?

A. Experimental Settings

1) Dataset: To evaluate the proposed method, we adopted the Polyvore Outfits dataset [3], which is widely utilized by several works on fashion analysis [26], [52]. This dataset is collected from the Polyvore fashion website. In the light of whether fashion items overlap in the training, validation, and testing dataset, the Polyvore Outfits dataset provides two dataset versions: the nondisjoint and disjoint versions, termed as Polyvore Outfits-ND and Polyvore Outfits-D, respectively. There are a total of $68,306$ outfits in the Polyvore Outfits-ND, divided into three sets: training set ($53,306$ outfits), validation set ($5,000$ outfits), and testing set ($10,000$ outfits). As for the disjoint version, Polyvore Outfits-D, there are a total of $32,140$ outfits, where $16,995$ outfits for training, $3,000$ outfits for validation, and $15,145$ outfits for testing. Each outfit in the Polyvore Outfits-ND has at least 2 items and up to 19 items, while that in the Polyvore Outfits-D has at least 2 items and up to 16 items. Each fashion item in these two datasets contains multiple modalities, *e.g.,* the visual image, textual description, popularity score, and category information. Regarding the category information of fashion items, there are 11 coarse-grained categories and 154 fine-grained categories in the Polyvore Outfits dataset. In particular, we utilized the visual images, textual descriptions, and category information of fashion items in this work.

2) Evaluation tasks: To evaluate the proposed model, we conducted experiments on two tasks: compatibility estimation and fill-in-the-blank (FITB) fashion recommendation.

Compatibility estimation: This task is to estimate a compatibility score for a given outfit. Different from the previous study [23] that generates negative outfits randomly without any restriction, we replaced each item in the positive compatible outfit with another randomly selected item in the same category, which makes the task more challenging and practical. The ratio of positive and negative samples is set to $1:1$. The positive samples are labeled as 1, while the negative samples are labeled as 0. To evaluate the performance, similar to previous studies [23], [52], we selected the area under the receiver operating characteristic curve (AUC) as the evaluation metric.

FITB fashion recommendation: Given an incomplete outfit and a target item annotated with the question mark, this task aims to select the most compatible fashion item from a candidate item set to fill in the blank and transform the given incomplete outfit into a compatible and complete one. This task is rather practical since people always need to buy a garment to match the garments they already have. Concretely, we constructed the FITB question by randomly selecting an item from a positive/compatible outfit as the target item and replacing it with a blank. We then randomly selected 3 items in the same category along with the target item to form the candidate set. The performance on this task is evaluated by the accuracy (ACC) of choosing the correct item from the candidate items.

3) Implementation Details: For the image encoder, we selected the ResNet-18 [53] as the backbone network and used the output of its final 4 convolutional layers (*i.e.,* conv2_x, conv3_x, conv4_x, and conv5_x) to derive the multi-layer visual representation according to Eqn. (2). In this case, $L = 4$ in Eqn. (2). Regarding the text encoder, we first employed the pre-trained word2vec tool to obtain the 300-D vector for each word, and then fed the concatenation of all word vectors into the TextCNN. In particular, the TextCNN is equipped with $100 * 5$ filters in 3 distinct sizes $[2, 3, 4]$. Ultimately, we captured a 300-D textual representation for each item. As for the category encoder, we empirically used the fine-grained category information and set the dimension of the category vector as 256. Accordingly, the number of category embeddings N_c is 154.

For the optimization, we employed the adaptive moment (Adam) [54] estimation method. We adopted the grid search strategy to determine the optimal values for the hyper-parameters (*i.e.,* λ_1 and λ_2) among the values $\{5e^r \mid r\epsilon - 5, \cdots, -1\}$. In addition, the learning rate, batch size, the number of propagation steps K, and the dimension of the hidden state d for all methods are searched in $\left[1e^{-3}, 5e^{-4}, 1e^{-4}, 5e^{-5}, 1e^{-5}\right]$, $[24, 32, 64, 128, 256]$, $[1, 2, 3, 4, 5]$, and $[16, 32, 64, 128, 256]$, respectively. The proposed model is fine-tuned based on training set and validation set for 15 epochs, and the performance on testing set is reported. We experimentally found that the model achieves the optimal performance with the initial learning rate is $5e^{-5}$ and decays by a factor of 0.5 every 10 epochs, the batch size of 32, the number of propagation steps $K = 4$, and the dimension of the hidden state $d = 64$, respectively. The hyper-parameters λ_1 and λ_2 in the loss function are $5e^{-3}$ and $5e^{-5}$, respectively. All experiments are implemented by PyTorch.

B. On Model Comparison (RQ1)

To validate the effectiveness of our MOCM-MGL, we chose the following state-of-the-art methods as baselines.

- **Bi-LSTM** [23]: By viewing an outfit as a sequence, this method exploits the latent item interaction by the bidirectional LSTM and utilizes the VSE to capture the inter-modal consistency.
- **Concatenation-Visual** [20]: This method concatenates the visual features of all fashion items into a vector, and then uses an MLP as the binary classifier to compute the outfit compatibility score.
- **Concatenation-All**: For a fair comparison, this method concatenates the visual, textual, and category features of all fashion items into a vector, and then uses an MLP to estimate the outfit compatibility. The encoders are the same as our proposed model.
- **Pooling** [1]: This is an early fusion based method that first concatenates the visual, textual, and category features of each fashion item, and then applies the average pooling operation to aggregate fashion items.

TABLE I
PERFORMANCE COMPARISON ON POLYVORE OUTFITS-ND AND POLYVORE OUTFITS-D. † INDICATES THE RESULTS ARE CITED FROM [29].

Method	Polyvore Outfits-ND		Polyvore Outfits-D	
	AUC(%)	ACC(%)	AUC(%)	ACC(%)
Bi-LSTM	66.24	38.11	62.72	37.43
Concatenation-Visual	85.21	49.93	78.62	43.05
Concatenation-All	87.61	51.35	80.23	45.14
Pooling	89.09	56.58	83.99	51.37
Type-Aware	87.23	57.78	84.49	55.85
SCE-Net	87.09	57.80	84.22	55.44
NGNN	87.12	51.79	83.61	48.37
ABF†	89.99	61.90	87.48	60.78
MOCM-MGL	**93.26**	**63.26**	**90.79**	**61.05**

- **Type-Aware** [3]: This method maps the item pairs into the category-specific embedding spaces, and estimates the outfit compatibility by averaging all distances of the item pairs in the spaces.
- **SCE-Net** [26]: Different from Type-Aware, this method learns condition-aware embeddings by item's own characteristics without explicit category supervision. In particular, this method also uses the early fusion strategy, which integrates the visual and textual features of fashion items by the element-wise product operation.
- **NGNN** [7]: This method constructs a subgraph for each outfit, where each node represents a category and edges represent interactions among nodes. In this way, the item representation can be enhanced by that of the items in the same outfit. The outfit compatibility is jointly modeled from two channels of NGNN, whose inputs are the visual and textual modalities.
- **ABF** [29]: This is an attention-based fusion method that utilizes the attention mechanism to fuse the visual and textual features of fashion items. Since the experiment setting in [29] is consistent with ours, we directly cited the results.

Notably, all methods use the ResNet-18 as the backbone network for a fair comparison. Table I shows the performance comparison among different approaches on the Polyvore Outfits-ND and Polyvore Outfits-D datasets under different tasks. From this table, the following observations can be made:

1) MOCM-MGL surpasses all the baselines by a large margin with respect to all metrics, which demonstrates the superiority of our proposed framework.
2) Concatenation-All outperforms Concatenation-Visual, which verifies the effectiveness of integrating the multi-modal information of fashion items.
3) The performance of SCE-Net is similar to that of Type-Aware, demonstrating the great potential of learning condition-aware embeddings by item's own characteristics instead of explicit category indicator.
4) ABF shows superiority over all multi-modal baselines, which reflects the superiority of utilizing the attention mechanism to fuse the multi-modal information.
5) It is unexpected that the graph-wise method NGNN performs worse than the pair-wise methods (*i.e.,* Type-Aware and SCE-Net). The possible reason is that NGNN focuses on propagating category-oriented fashion compatibility. However, in the context of this study, the negative outfit shares the same item category as the positive one, which is hardly handled by NGNN.

C. On Ablation Study (RQ2)

To explore the contribution of each component of the proposed model, we introduced the following derivatives from the model.

- **w/o-MGL**: To explore the effect of the proposed modality-oriented graph learning scheme, we disabled the module by directly concatenating the visual, textual, and category features of each fashion item obtained by the multi-modal embedding module, and then fed it into the outfit compatibility estimation.
- **w/o-Inter**: To validate the necessity of exploring the inter-modal compatibility among fashion items, we redefined the edge representation between two item nodes as $\mathbf{e}_{ij}^{k} = \alpha\left(\mathbf{W}_e^k\left(\mathbf{W}_n^k\mathbf{h}_i^{k-1} \odot \mathbf{W}_n^k\mathbf{h}_j^{k-1}\right) + \mathbf{b}_e^k\right)$, where $\mathbf{h}_i^k = \left[\mathbf{h}_{i,1}^k, \mathbf{h}_{i,2}^k, \mathbf{h}_{i,3}^k\right]$. In this way, only the intra-modal compatibility is considered.
- **w/o-Edge**: To investigate the importance of edge-based compatibility relation modeling, we removed the edge representation generation unit and directly aggregated information from the hidden states of neighbors, *i.e.,* we changed Eqn. (6) to $\mathbf{m}_{\mathcal{N}(i),pq}^k = \text{AGG}\left(\left\{\mathbf{h}_{j,q}^k, \forall j \in \mathcal{N}\left(i\right)\right\}\right)$.
- **w/o-GRU**: To verify whether it is necessary to retain the original hidden information of the node when updating the node representation, we removed the GRU unit and only utilized the aggregation information, *i.e.,* we changed Eqn. (8) to $\mathbf{h}_{i,p}^k = \mathbf{m}_{\mathcal{N}(i),p}^k$.
- **w/o-MultiLayer**: To explore the importance of integrating representations obtained at different propagation layers, we treated the representation obtained at the final K-th propagation layer as the updated item representation, *i.e.,* we made $\mathbf{h}_i^* = \mathbf{h}_i^K$ in Eqn. (9).
- **w/o-MeanPool**: To validate the function of the mean pooling operation in the outfit compatibility estimation module, we only employed the max pooling operation to generate the final outfit embedding, *i.e.,* we rewrote Eqn. (10) as $\tilde{\mathbf{h}} = \gamma_{max}\left(\{\mathbf{h}_i^*, \forall v_i \in \mathcal{V}\}\right)$.
- **w/o-MaxPool**: Similarly, we removed the max pooling operation in the outfit compatibility estimation module to learn its effect by making $\tilde{\mathbf{h}} = \gamma_{mean}\left(\{\mathbf{h}_i^*, \forall v_i \in \mathcal{V}\}\right)$ in Eqn. (10).

Table II shows the performance comparison between MOCM-MGL and its derivatives. From this table, we obtained the following observations:

1) Our model consistently surpasses all derivations across all metrics, demonstrating the effectiveness of each component in the proposed MOCM-MGL.
2) MOCM-MGL demonstrates superiority over w/o-MGL, which implies that the modality-oriented GCN can propagate the intra-modal and inter-modal compatibility relation among fashion items, and therefore boost the expressiveness of item representations.

TABLE II
ABLATION STUDY ON POLYVORE OUTFITS-ND AND POLYVORE OUTFITS-D DATASETS.

Method	Polyvore Outfits-ND		Polyvore Outfits-D	
	AUC(%)	ACC(%)	AUC(%)	ACC(%)
w/o-MGL	92.31	61.84	88.93	58.44
w/o-Inter	92.95	62.87	89.18	59.27
w/o-Edge	92.92	62.75	89.76	59.35
w/o-GRU	92.45	62.41	88.97	58.77
w/o-MultiLayer	93.07	62.53	89.57	58.94
w/o-MeanPool	93.13	62.29	90.07	60.16
w/o-MaxPool	92.72	61.87	89.08	58.12
MOCM-MGL	**93.26**	**63.26**	**90.79**	**61.05**

TABLE III
THE PERFORMANCE OF OUR PROPOSED METHOD WITH DIFFERENT MODALITY COMBINATIONS.

Method	Polyvore Outfits-ND		Polyvore Outfits-D	
	AUC(%)	ACC(%)	AUC(%)	ACC(%)
Visual	90.77	57.45	85.95	52.93
Visual+Category (coarse)	90.85	59.60	85.98	54.64
Visual+Category (fine)	91.02	59.67	86.01	54.75
Textual	77.02	40.33	75.62	39.01
Textual+Category (coarse)	78.41	41.66	76.71	40.92
Textual+Category (fine)	79.75	42.11	76.95	41.15
Visual+Textual	92.55	61.05	89.17	57.55
All (coarse)	93.06	62.40	90.01	59.81
All (fine)	**93.26**	**63.26**	**90.79**	**61.05**

3) MOCM-MGL outperforms w/o-Inter, implying the necessity of investigating the inter-modal compatibility among fashion items, to fully explore the fine-grained compatibility relation among items.
4) MOCM-MGL achieves better performance than w/o-Edge. This confirms the benefit of the edge-based compatibility relation modeling and the compatibility relation propagation during the outfit compatibility modeling.
5) MOCM-MGL surpasses w/o-GRU, which implies that selectively absorbing the compatibility information from the nodes' neighbors and the original hidden information of the node can boost the model performance.
6) w/o-MultiLayer performs worse than our MOCM-MGL. This implies that different propagation layers indeed absorb the neighbor compatibility information at different levels, and contribute to the comprehensive outfit compatibility estimation.
7) MOCM-MGL shows superiority over w/o-MaxPool and w/o-MeanPool. This suggests that both the most prominent and the overall hidden states of fashion items are beneficial to the outfit compatibility modeling. Additionally, we observed that w/o-MeanPool outperforms w/o-MaxPool, which reflects that the max pooling operation is more effective than the mean pooling operation. This indicates that the most prominent feature of all composing items' hidden states, compared with the overall feature, has a greater impact on the outfit compatibility estimation.

D. On Modality Comparison (RQ3)

To investigate the influence of different modalities (*i.e.,* visual image, textual description, and category) on the performance, we compared the MOCM-MGL with different modality combinations. Notably, due to the concern that the negative outfit shares the same item categories as the positive outfit, we did not adopt the method that only utilizes category information for comparison. In addition, there are two kinds of item categories: coarse-grained categories and fine-grained categories. Therefore, there are nine modality combinations: Visual, Visual+Category (coarse), Visual+Category (fine), Textual, Textual+Category (coarse), Textual+Category (fine), Visual+Textual, All (coarse), and All (fine), where coarse, fine, and All indicate that coarse-grained categories, fine-grained categories, and all the three modalities are used, respectively. Table III shows the performance of our model with the nine different modality combinations. As can be seen from Table III, we observed that:

1) Visual outperforms Textual. This demonstrates that the visual modality is more effective than the textual feature for the outfit compatibility modeling.
2) Multi-modal Visual+Textual achieves better performance than single-modal Visual and Textual. This indicates that the visual and textual modalities of fashion items complement each other toward the outfit compatibility estimation.
3) Visual+Textual performs better than Visual+Category and Textual+Category. This may be attributed to the fact that the visual image and textual description deliver more content-related features of fashion items than the category information.
4) All surpasses Visual+Textual, indicating that incorporating the category information as one essential modality does improve the model performance.
5) The methods with fine-grained categories perform better than those with coarse-grained categories. This may be due to the fact that fine-grained categories provide more detailed information on fashion items, which facilitates outfit compatibility modeling.

To gain an intuitive understanding of the impact of the multi-modal integration, we showed several results obtained by MOCM-MGL with different modality combinations (*i.e.,* Visual, Textual, and All) on the FITB task in Figure 4. We found that only considering a single modality of fashion items may lead to incorrect choices. For instance, in the first example, Textual chooses the wrong answer d. This may be due to the fact that the textual description of the answer d and that of the given gloves shares the same color, *i.e.,* "black". Nevertheless, further incorporating the visual modality, the method All gives the correct answer c. Regarding the third example, Visual fails to give the correct answer, while All does. This makes sense, as the textual description of the ground truth answer c shares the same pattern with the given striped hoodie. These examples demonstrate the necessity of incorporating the complementary multi-modal information towards outfit compatibility modeling.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3134164, IEEE Transactions on Multimedia

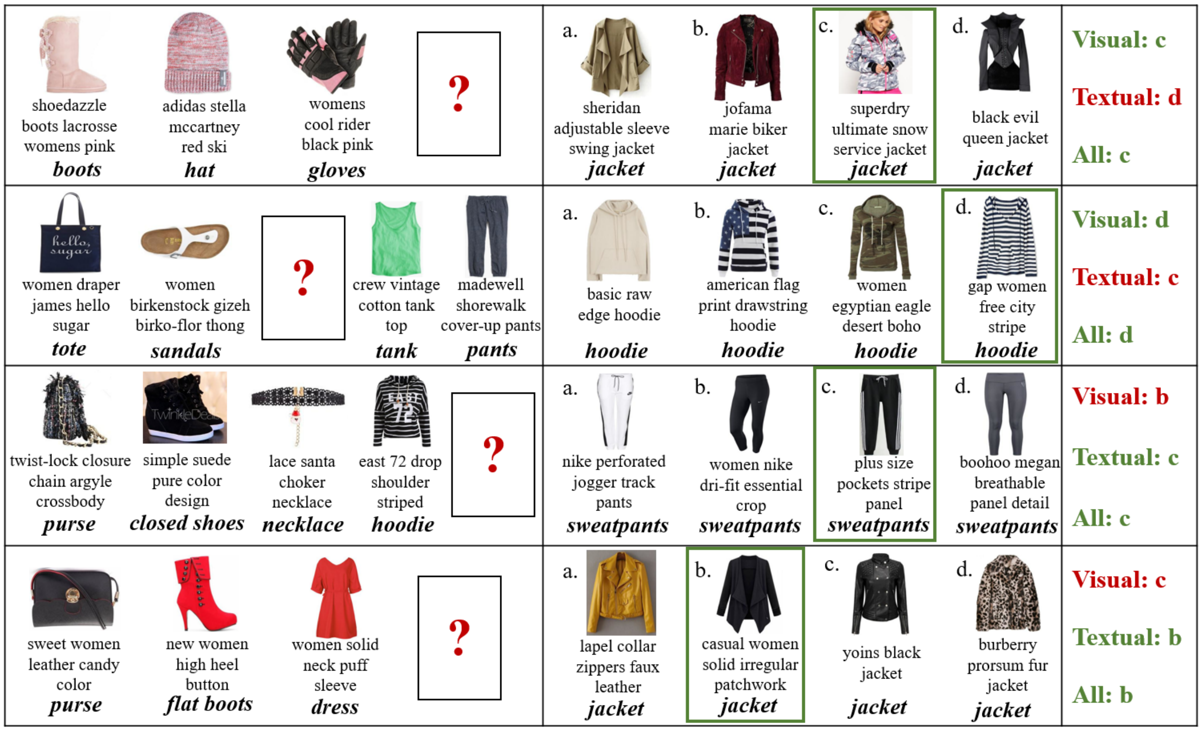

shoedazzle boots lacrosse womens pink
boots
adidas stella mccartney red ski
hat
womens cool rider black pink
gloves
?
a.
sheridan adjustable sleeve swing jacket
jacket
b.
jofama marie biker jacket
jacket
c.
superdry ultimate snow service jacket
jacket
d.
black evil queen jacket
jacket
Visual: c
Textual: d
All: c
women draper james hello sugar
tote
women birkenstock gizeh birko-flor thong
sandals
?
crew vintage cotton tank top
tank
madewell shorewalk cover-up pants
pants
a.
basic raw edge hoodie
hoodie
b.
american flag print drawstring hoodie
hoodie
c.
women egyptian eagle desert boho
hoodie
d.
gap women free city stripe
hoodie
Visual: d
Textual: c
All: d
twist-lock closure chain argyle crossbody
purse
simple suede pure color design
closed shoes
lace santa choker necklace
necklace
east 72 drop shoulder striped
hoodie
?
a.
nike perforated jogger track pants
sweatpants
b.
women nike dri-fit essential crop
sweatpants
c.
plus size pockets stripe panel
sweatpants
d.
boohoo megan breathable panel detail
sweatpants
Visual: b
Textual: c
All: c
sweet women leather candy color
purse
new women high heel button
flat boots
women solid neck puff sleeve
dress
?
a.
lapel collar zippers faux leather
jacket
b.
casual women solid irregular patchwork
jacket
c.
yoins black jacket
jacket
d.
burberry prorsum fur jacket
jacket
Visual: c
Textual: b
All: b

Fig. 4. Comparison of Visual, Textual, and All on the FITB task. The black, green and red bold fonts represent the category information of fashion items, true chosen, and false chosen, respectively. The items highlighted in the green boxes are the ground truth.

E. Hyper-parameter Discussion (RQ4)

In this section, we examined how the number of propagation steps K, the number of convolutional layers of Resnet-18 used for visual encoding (*i.e.,* L in Eqn. (2)), and the number of composing items affect the performance of our method.

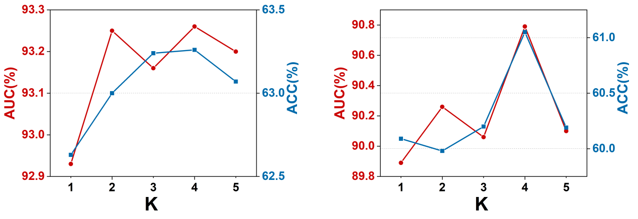
To explore the impact of the number of propagation steps, we evaluated our model's performance on two tasks with two datasets by changing K from 1 to 5 with a step of 1. As shown in Figure 5, our model achieves the optimal performance when K is 4. This suggests that it is necessary to propagate several runs so that the fashion items can absorb the neighbor compatibility information thoroughly at different levels. Moreover, when K is higher than 4, the performance drops. One possible reason is that superfluous information propagation might introduce more noise into the node representations, therefore leading to a negative effect.

We then studied the influence of the number of convolutional layers of Resnet-18 used for visual encoding, *i.e.,* L in Eqn. (2), on the model performance. In particular, we varied the number of convolutional layers used for visual encoding from 1 to 4. Specifically, $L = 1$ indicates that we only used the output of the final layer conv5_x of Resnet-18 for visual encoding, while $L = 2$ refers to that we used the output of the final two layers (*i.e.,* conv4_x, and conv5_x) of Resnet-18. The cases of $L = 3$ and $L = 4$ can be similarly derived. In a sense, the larger the L, the shallower layers' output would be incorporated. Figure 6 shows the performance of our model on the two tasks with the two datasets. As can be seen from Figure 6, the model's performance grows with integrating more convolutional layers' output, which indicates that each convolutional layer has its contribution to boosting the visual encoding. The possible reason is that the shallow layers can capture the low-level visual features of the item, while the deep ones can capture the high-level features, both of which benefit the visual encoding of fashion items and hence boost the outfit compatibility estimation performance.

To gain deeper insights, we examined the performance of our proposed model regarding outfits with different numbers of composing items. In particular, the testing set is divided according to the number of fashion items, ranging from 2 to 8. Figure 7 shows the performance of our proposed method with different testing configurations. As can be seen, our method performs well in all settings, verifying the effectiveness of our method to handle outfit compatibility modeling with different composing item numbers. In addition, our method performs better for outfits with multiple (*i.e.,* more than 2) fashion items compared to those with two fashion items. This may be due to that the benefit of modeling the comparability between two items with a graph is limited.

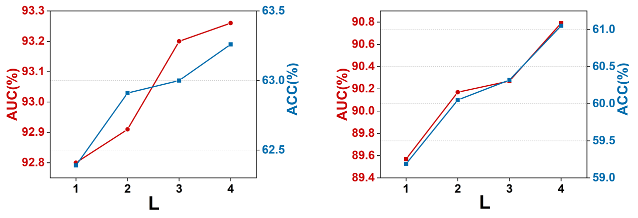
V. Conclusion and Future Work

In this work, we present a multi-modal outfit compatibility modeling scheme with modality-oriented graph learning, named MOCM-MGL, which fully exploits the visual, textual, and category modalities with GCN. Different from previous work, we treat the category information of fashion items as a unique and comparable modality to the visual and textual modalities. In addition, the proposed MOCM-MGL jointly unifies the intra-modal and inter-modal compatibility relation among fashion items. Extensive experiments have



(a) Polyvore Outfits-ND. (b) Polyvore Outfits-D.

Fig. 5. Effect of the number of propagation steps, *i.e.,* K, on Polyvore Outfits-ND and Polyvore Outfits-D datasets.



(a) Polyvore Outfits-ND. (b) Polyvore Outfits-D.

Fig. 6. Effect of the number of convolutional layers of ResNet-18 *i.e.,* L, on Polyvore Outfits-ND and Polyvore Outfits-D datasets.

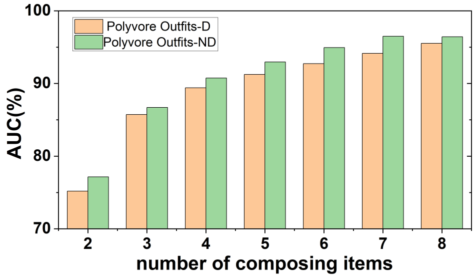


Fig. 7. Performance of our proposed method regarding outfits with different number of composing items on Polyvore Outfits-ND and Polyvore Outfits-D datasets.

been conducted on the Polyvore Outfits-ND and Polyvore Outfits-D datasets. The experimental results demonstrate the superiority of MOCM-MGL, suggesting that employing a modality-oriented GCN to propagate the intra-modal and inter-modal compatibility relation among fashion items is helpful to boost the model performance. In addition, integrating the multi-modal information of fashion items greatly improves the outfit compatibility estimation performance. One limitation of our work is that we currently ignore the attribute context of fashion items in the outfit compatibility modeling. In the future, we plan to take the attribute context of fashion items into account to boost the model performance.

References

[1] Y. Li, L. Cao, J. Zhu, and J. Luo, "Mining fashion outfit composition using an end-to-end deep learning approach on set data," *IEEE Trans. Multimedia*, vol. 19, no. 8, pp. 1946–1955, Aug. 2017.

[2] X. Dong, J. Wu, X. Song, H. Dai, and L. Nie, "Fashion compatibility modeling through a multi-modal try-on-guided scheme," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr*, Jul. 2020, pp. 771–780.

[3] M. I. Vasileva, B. A. Plummer, K. Dusad, S. Rajpal, R. Kumar, and D. Forsyth, "Learning type-aware embeddings for fashion compatibility," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 405–421.

[4] X. Wang, B. Wu, and Y. Zhong, "Outfit compatibility prediction and diagnosis with multi-layered comparison network," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 329–337.

[5] X. Lu, L. Zhu, Z. Cheng, L. Nie, and H. Zhang, "Online multi-modal hashing with dynamic query-adaption," in *Proc. 42nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr*, Jul. 2019, pp. 715–724.

[6] L. Zhu, X. Lu, Z. Cheng, J. Li, and H. Zhang, "Deep collaborative multi-view hashing for large-scale image search," *IEEE Trans. Image Process.*, vol. 29, pp. 4643–4655, 2020.

[7] Z. Cui, Z. Li, S. Wu, X.-Y. Zhang, and L. Wang, "Dressing as a whole: Outfit compatibility learning based on node-wise graph neural networks," in *Proc. World Wide Web Conf.*, May 2019, pp. 307–317.

[8] X. Li, X. Wang, X. He, L. Chen, J. Xiao, and T.-S. Chua, "Hierarchical fashion graph network for personalized outfit recommendation," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr*, Jul. 2020, pp. 159–168.

[9] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Proc. Adv. Neural Inf. Process. Syst*, Dec. 2017, pp. 1024–1034.

[10] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Oct. 2014, pp. 1746–1751.

[11] M. W. Gardner and S. Dorling, "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences," *Atmos. Environ.*, vol. 32, no. 14-15, pp. 2627–2636, Aug. 1998.

[12] X. Liang, L. Lin, W. Yang, P. Luo, J. Huang, and S. Yan, "Clothes co-parsing via joint image segmentation and labeling with application to clothing retrieval," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1175–1186, Jun. 2016.

[13] X. Song, F. Feng, J. Liu, Z. Li, L. Nie, and J. Ma, "Neurostylist: Neural compatibility modeling for clothing matching," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 753–761.

[14] J. Liu, X. Song, Z. Chen, and J. Ma, "Neural fashion experts: I know how to make the complementary clothing matching," *Neurocomputing*, vol. 359, no. 24, pp. 249–263, 2019.

[15] X. Gu, Y. Wong, P. Peng, L. Shou, G. Chen, and M. S. Kankanhalli, "Understanding fashion trends from street photos via neighbor-constrained embedding learning," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 190–198.

[16] X. Song, X. Han, Y. Li, J. Chen, X.-S. Xu, and L. Nie, "Gp-bpr: Personalized compatibility modeling for clothing matching," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 320–328.

[17] W. Chen, P. Huang, J. Xu, X. Guo, C. Guo, F. Sun, C. Li, A. Pfadler, H. Zhao, and B. Zhao, "Pog: personalized outfit generation for fashion recommendation at alibaba ifashion," in *Proc. 25th ACM SIGKDD Int. conf. Knowl. Discovery Data Mining*, Aug. 2019, pp. 2662–2670.

[18] H. Wen, X. Song, X. Yang, Y. Zhan, and L. Nie, "Comprehensive linguistic-visual composition network for image retrieval," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr*, Jul. 2021, pp. 1369–1378.

[19] X. Dong, X. Song, F. Feng, P. Jing, X.-S. Xu, and L. Nie, "Personalized capsule wardrobe creation with garment and user modeling," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 302–310.

[20] P. Tangseng, K. Yamaguchi, and T. Okatani, "Recommending outfits from personal closet," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2017, pp. 268–277.

[21] G. Cucurull, P. Taslakian, and D. Vazquez, "Context-aware visual compatibility prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12 617–12 626.

[22] S. Chaidaroon, Y. Fang, M. Xie, and A. Magnani, "Neural compatibility ranking for text-based fashion matching," in *Proc. 42nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr*, Jul. 2019, pp. 1229–1232.

[23] X. Han, Z. Wu, Y.-G. Jiang, and L. S. Davis, "Learning fashion compatibility with bidirectional lstms," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 1078–1086.

[24] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *arXiv:1411.2539*, Nov. 2014. [Online]. Available: https://arxiv.org/abs/1411.2539

[25] Y. Wei, X. Wang, W. Guan, L. Nie, Z. Lin, and B. Chen, "Neural multimodal cooperative learning toward micro-video understanding," *IEEE Trans. Image Process.*, vol. 29, pp. 1–14, 2020.

[26] R. Tan, M. I. Vasileva, K. Saenko, and B. A. Plummer, "Learning similarity conditions without explicit supervision," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10 372–10 381.

[27] X. Yang, Y. Ma, L. Liao, M. Wang, and T.-S. Chua, "Transnfcm: Translation-based neural fashion compatibility modeling," in *Proc. 33rd AAAI Conf. Artif. Intell.*, Jan. 2019, pp. 403–410.

[28] G.-L. Sun, J.-Y. He, X. Wu, B. Zhao, and Q. Peng, "Learning fashion compatibility across categories with deep multimodal neural networks," *Neurocomputing*, vol. 395, pp. 237–246, Jun. 2020.

[29] K. Laenen and M.-F. Moens, "A comparative study of outfit recommendation methods with a focus on attention-based fusion," *Inf. Process. Manage.*, vol. 57, no. 6, p. 102316, Nov. 2020.

[30] Z. Lu, Y. Hu, Y. Jiang, Y. Chen, and B. Zeng, "Learning binary code for personalized fashion recommendation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10 562–10 570.

[31] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph neural networks: A review of methods and applications," *AI Open*, vol. 1, pp. 57–81, 2020.

[32] M. Gori, G. Monfardini, and F. Scarselli, "A new model for learning in graph domains," in *Proc. Proc. IEEE Int. Joint Conf. Neural Netw.*, vol. 2, Jul. 2005, pp. 729–734.

[33] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip, "A comprehensive survey on graph neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 4–24, Jan. 2021.

[34] J. Wu, S.-H. Zhong, and Y. Liu, "Mvsgcn: A novel graph convolutional network for multi-video summarization," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 827–835.

[35] F. Ye, S. Pu, Q. Zhong, C. Li, D. Xie, and H. Tang, "Dynamic gcn: Context-enriched topology learning for skeleton-based action recognition," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 55–63.

[36] J. Liu, Z.-J. Zha, R. Hong, M. Wang, and Y. Zhang, "Deep adversarial graph attention convolution network for text-based person search," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 665–673.

[37] Z. Zhang, D. Xu, W. Ouyang, and L. Zhou, "Dense video captioning using graph-based sentence summarization," *IEEE Trans. Multimedia*, vol. 23, pp. 1799–1810, 2021.

[38] J. Zhang, Y. Yang, Q. Tian, L. Zhuo, and X. Liu, "Personalized social image recommendation method based on user-image-tag model," *IEEE Trans. Multimedia*, vol. 19, no. 11, pp. 2439–2449, 2017.

[39] Y. Wei, X. Wang, X. He, L. Nie, Y. Rui, and T.-S. Chua, "Hierarchical user intent graph network for multimedia recommendation," *IEEE Trans. Multimedia*, pp. 1–12, 2021.

[40] X. Lu, L. Zhu, L. Liu, L. Nie, and H. Zhang, "Graph convolutional multi-modal hashing for flexible multimedia retrieval," in *Proc. 21st ACM Int. Conf. Multimedia*, Oct. 2021, pp. 1414–1422.

[41] F. Liu, Z. Cheng, L. Zhu, Z. Gao, and L. Nie, "Interest-aware message-passing GCN for recommendation," in *Proc. World Wide Web Conf.*, Apr. 2021, pp. 1296–1305.

[42] X. Liu, Y. Sun, Z. Liu, and D. Lin, "Learning diverse fashion collocation by neural graph filtering," *IEEE Trans. Multimedia*, vol. 23, pp. 2894–2901, Aug. 2021.

[43] D. Kollias and S. P. Zafeiriou, "Exploiting multi-cnn features in cnn-rnn based dimensional emotion recognition on the omg in-the-wild dataset," *IEEE Trans. Affective Comput.*, vol. 12, no. 3, pp. 1–12, Jul. 2021.

[44] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv:1312.4400*, Dec. 2013. [Online]. Available: https://arxiv.org/abs/1312.4400

[45] X. Yang, X. Song, X. Han, H. Wen, J. Nie, and L. Nie, "Generative attribute manipulation scheme for flexible fashion search," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr*, Jul. 2020, pp. 941–950.

[46] A. Severyn and A. Moschitti, "Twitter sentiment analysis with deep convolutional neural networks," in *Proc. 38th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr*, Aug. 2015, pp. 959–962.

[47] B. Guo, C. Zhang, J. Liu, and X. Ma, "Improving text classification with weighted word embeddings via a multi-channel textcnn model," *Neurocomputing*, vol. 363, no. 21, pp. 366–374, Oct. 2019.

[48] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. 1st Int. Conf. Learn. Represent*, May 2013, pp. 1–12.

[49] Y. Li, D. Tarlow, M. Brockschmidt, and R. Zemel, "Gated graph sequence neural networks," in *Proc. 4th Int. Conf. Learn. Represent*, May 2016, pp. 1–20.

[50] S. Guo, W. Huang, X. Zhang, P. Srikhanta, Y. Cui, Y. Li, H. Adam, M. R. Scott, and S. Belongie, "The imaterialist fashion attribute dataset," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2019, pp. 3113–3116.

[51] N. Inoue, E. Simo-Serra, T. Yamasaki, and H. Ishikawa, "Multi-label fashion image classification with minimal human supervision," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 2261–2267.

[52] Y.-L. Lin, S. Tran, and L. S. Davis, "Fashion outfit complementary item retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3308–3316.

[53] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[54] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent*, May 2015.

Xuemeng Song received a B.E. from the University of Science and Technology of China in 2012, and a Ph.D. from the School of Computing, National University of Singapore in 2016. She is currently an associate professor of Shandong University, Jinan, China. Her research interests include the information retrieval and social network analysis. She has published several papers in the top venues, such as ACM SIGIR, MM, TIP, and TOIS. In addition, she has served as a reviewer for many top conferences and journals.

Shi-Ting Fang is currently pursing the M.S. degree with the School of Computer Science and Technology, Shandong University. Her current research interests include deep learning and multimedia computing.

Xiaolin Chen received the M.S. degree from Shandong University in 2021. She is currently pursuing the Ph.D. degree with the School of Software, Shandong University. Her research primarily focuses on information retrieval and multimedia computing. She has published papers in the top venues, including ACM SIGIR, TOIS and IEEE TMM. Moreover, she has served as reviewers for conferences and journals, such as ACM MM and Neurocomputing.

Yinwei Wei received his MS degree from Tianjin University and Ph.D. degree from Shandong University, respectively. Currently, he is a research fellow with National University of Singapore. His research interests include multimedia computing and information retrieval. He has published some academic papers in top conferences and journals, such as ACM MM, TIP, and TMM. Dr. Wei has served as the PC member for several conferences, such as ACM MM, AAAI, and IJCAI, and the reviewer for TPAMI, TIP, and TMM.

Liqiang Nie is currently a professor with the School of Computer Science and Technology, Shandong University. Meanwhile, he is the adjunct dean with the Shandong AI institute. He received his B.Eng. and Ph.D. degree from Xi'an Jiaotong University in July 2009 and National University of Singapore (NUS) in 2013, respectively. After PhD, Dr. Nie continued his research in NUS as a research follow for more than three years. His research interests lie primarily in multimedia computing and information retrieval. Dr. Nie has co-/authored more than 160 papers, received more than 5300 Google Scholar citations as of Oct. 2019. He is an AE of Information Science, an area chair of ACM MM 2018, a special session chair of PCM 2018, a PC chair of ICIMCS 2017. Meanwhile, he is supported by the program of "Thousand Youth Talents Plan 2016", "Qilu Scholar 2016", and "The Shandong Province Science Fund for Distinguished Young Scholars 2018". In 2017, he co-founded "Qilu Intelligent Media Forum".