Partially Supervised Compatibility Modeling

Weili Guan, *Member, IEEE*, Haokun Wen, Xuemeng Song, *Senior Member, IEEE*, Chun Wang, Chung-Hsing Yeh, *Senior Member, IEEE*, Xiaojun Chang, *Senior Member, IEEE*, and Liqiang Nie, *Senior Member, IEEE*

***Abstract*— Fashion Compatibility Modeling (FCM), which aims to automatically evaluate whether a given set of fashion items makes a compatible outfit, has attracted increasing research attention. Recent studies have demonstrated the benefits of conducting the item representation disentanglement towards FCM. Although these efforts have achieved prominent progress, they still perform unsatisfactorily, as they mainly investigate the visual content of fashion items, while overlooking the semantic attributes of items (e.g., color and pattern), which could largely boost the model performance and interpretability. To address this issue, we propose to comprehensively explore the visual content and attributes of fashion items towards FCM. This problem is non-trivial considering the following challenges: a) how to utilize the irregular attribute labels of items to partially supervise the attribute-level representation learning of fashion items; b) how to ensure the intact disentanglement of attribute-level representations; and c) how to effectively sew the multiple granulairites (i.e, coarse-grained item-level and fine-grained attribute-level) information to enable performance improvement and interpretability. To address these challenges, in this work, we present a partially supervised outfit compatibility modeling scheme (PS-OCM). In particular, we first devise a partially supervised attribute-level embedding learning component to disentangle the fine-grained attribute embeddings from the entire visual feature of each item. We then introduce a disentangled completeness regularizer to prevent the information loss during disentanglement. Thereafter, we design a hierarchical graph convolutional network, which seamlessly integrates the attribute- and item-level compatibility modeling, and enables the explainable compatibility reasoning. Extensive experiments on the real-world dataset demonstrate that our PS-OCM significantly outperforms the state-of-the-art baselines. We have released our source codes and well-trained models to benefit other researchers (https://site2750.wixsite.com/ps-ocm).**

***Index Terms*— Partial supervision, disentangled representation, fashion compatibility estimation, graph convolutional network.**

Manuscript received 19 October 2021; revised 17 May 2022; accepted 21 June 2022. Date of publication 6 July 2022; date of current version 14 July 2022. This work was supported in part by the National Natural Science Foundation of China under Grant U1936203 and in part by the Australian Research Council (ARC) Discovery Early Career Researcher Award (DECRA) under Grant DE190100626. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Soma Biswas. *(Corresponding authors: Xuemeng Song; Xiaojun Chang.)*

Weili Guan and Chung-Hsing Yeh are with the Faculty of Information Technology, Monash University, Clayton, VIC 3800, Australia (e-mail: honeyguan@gmail.com; chunghsing.yeh@monash.edu).

Haokun Wen, Xuemeng Song, and Chun Wang are with the School of Computer Science and Technology, Shandong University, Qingdao 266000, China (e-mail: whenhaokun@gmail.com; sxmustc@gmail.com; wclegends1998@gmail.com).

Xiaojun Chang is with the Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, NSW 2007, Australia (e-mail: cxj273@gmail.com).

Liqiang Nie is with the School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen 518055, China (e-mail: nieliqiang@gmail.com).

Digital Object Identifier 10.1109/TIP.2022.3187290

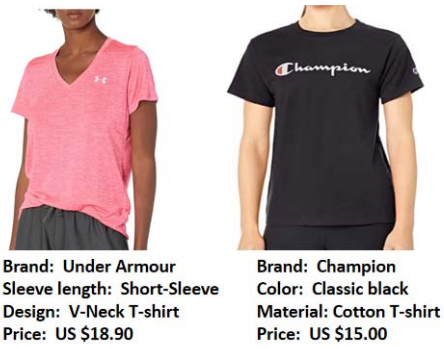


Fig. 1. Illustration of fashion items and their attribute labels.

I. Introduction

ALONG with the economic development, numerous fashion products have sprung up in the virtual and physical shops, such as bags, scarves, shoes, and skirts. Undoubtedly, they do beautify our lives. Nevertheless, they also bring in many troubles, especially for those who lack the sense of beauty. Moreover, people are easily overwhelmed by the abundant fashion items, thus it is difficult to find the desired fashion piece to make a compatible outfit with their wardrobe [1], [2]. Consequently, fashion compatibility modeling, which justifies whether the given set of fashion items makes a compatible outfit, is highly desired.

Owing to this practical value, fashion compatibility modeling has attracted increasing research attention. Over the past few years, many deep learning approaches have been explored [3]–[8]. Due to that people tend to habitually justify the compatibility of fashion items by considering a series of attribute-oriented questions, like whether the color/material of all composing items are compatible, a few recent studies have resorted to investigating the attribute-level representations by representation disentanglement [9], [10]. Despite their promising performance, they still have a key limitation: they mainly focus on the visual content of fashion items, while overlooking the item's semantic attributes. In fact, the attribute labels of items usually contain rich information that characterizes the key parts of the items [11]–[15], which can be adopted to supervise the attribute-level representation learning, and hence promoting the model's performance as well as interpretability. In light of this, we aim to jointly explore the visual content and semantic attributes of fashion items.

However, fulfilling this goal is non-trivial due to the following challenges. 1) The attribute labels of fashion items are not unified or aligned. In other words, each item may have different attribute labels. For instance, as shown in

Figure 1, one T-shirt is labeled with attributes of price, sleeve length, design, and brand; while the other has color, material, brand, and price. Thereby, how to fully take advantage of these irregular attribute labels to partially supervise the attribute-level representation learning of fashion items poses a big challenge. 2) When disentangling the entire visual embedding into multiple attribute-level representations, how to ensure information intactness during the disentanglement is another challenge. 3) To comprehensively capture the compatibility among fashion items, we should incorporate both the coarse-grained item-level and fine-grained attribute-level information into the compatibility modeling. Accordingly, how to seamlessly sew multiple granularities to strengthen the learning performance constitutes another tough challenge.

To address these challenges, we present a partially supervised compatibility modeling scheme (PS-OCM). As shown in Figure 2, it consists of three key components: 1) partially supervised attribute-level embedding learning, 2) disentangled completeness regularization, and 3) hierarchical outfit compatibility modeling. To be more specific, the first component extracts visual features from each composing item of the given outfit via a pretrained model. It then turns to disentangle the visual feature vector into a set of fine-grained attribute embeddings, which is partially supervised by the irregular attribute labels of each fashion item. As to the second component, it works towards an intact disentanglement. This is accomplished by adopting two strategies: orthogonal residual embedding and visual representation reconstruction. An orthogonal residual embedding is introduced to compensate the information loss, and regularize the orthogonal relationship between the residual embedding and each attribute-level embedding. Meanwhile, it uses the deconvolution neural network to ensure that the original image can be reconstructed from the disentangled attribute-level and residual embeddings. As to the last component, it contains a hierarchical graph convolutional network, which models the outfit compatibility by jointly integrating the fine-grained attribute-level and coarse-grained item-level information. Ultimately, it fuses the attribute-level compatibility scores and the item-level one via a multi-layer perceptron (MLP) to derive the final compatibility score of the given outfit.

Our main contributions can be summarized in threefold:

- We disentangle the visual representation of each item into a set of attribute-level embeddings, and present a partially supervised disentangled learning method to strengthen the learning performance via taking advantage of the irregular attribute labels.
- To prevent information loss during the disentanglement, we devise a novel disentangled completeness regularizer, which is accomplished by jointly introducing an orthogonal residual embedding and visual representation reconstruction.
- We propose a hierarchical graph convolutional network, which is able to seamlessly integrate the attribute- and item-level compatibility modeling. Most importantly, based upon the convolutional results, we are able to intuitively explain the attribute-level compatibility.

The rest of the paper is organized as follows. Section 2 briefly reviews the related work. In Section 3, we detail the proposed PS-OCM scheme. Experimental results and comprehensive analyses are presented in Section 4, followed by the conclusion and future work in Section 5.

II. Related Work

Our work is closely related to the studies on fashion compatibility modeling, disentangled representation learning, and graph convolutional networks. We will elaborate the literature of these research lines, respectively.

A. Fashion Compatibility Modeling

Recent years have witnessed the growing research interest in fashion compatibility modeling due to its huge commercial value. According to the way of the outfit structure, existing efforts can be broadly summarized into three categories: pair-wise methods [16], [17], sequence-wise methods [4], [18], and graph-wise methods [3], [19].

The pair-wise methods mainly focus on the compatibility between two given items. For example, McAuley *et al.* [20] first proposed a general framework to model the human visual preference for a given pair of items based on the Amazon real-world co-purchase dataset. Following that, Song *et al.* [16] proposed a multimodal pair-wise compatibility modeling scheme, whereby the deep neural networks are used to model the compatibility between fashion items via the Bayesian Personalized Ranking (BPR) optimization [21]. Later, Han *et al.* [22] presented a prototype-guide interpretable compatibility modeling, which seamlessly integrates the latent compatible/incompatible prototype learning and compatibility modeling with the BPR. Moreover, Yang *et al.* [23] utilized category complementary relations to model category-respected compatibility between fashion items in a translation-based embedding space. Thereafter, Liu *et al.* [6] introduced an auxiliary complementary template generation network equipped with the pixel-wise consistency and compatible template regularization to improve the compatibility modeling performance. The limitation of the pair-wise methods is that it is cumbersome and time-consuming to directly apply them to analyze the real-world outfit that usually comprises more than two items. Moreover, it maybe inappropriate to capture the complex compatibility relation among multiple items by separating the outfit into a set of independent item pairs.

By contrast, the sequence-wise methods regard the outfit as an ordered list of items and utilize sequential neural networks to uncover the complex compatibility relationship among them. For instance, Han *et al.* [4] employed a Bi-LSTM network to sequentially model the compatibility relationships among the fashion items in a given outfit. Later, Dong *et al.* [18] presented a multi-modal try-on-guided compatibility modeling framework to jointly characterize the discrete interaction and try-on appearance of the outfit, where Bi-LSTM is used for discrete interaction modeling. One key limitation of the sequence-based methods is that there is no explicit and fixed order of items in an outfit. Moreover,

the sequential neighborhood dependency in a given outfit is less stronger as compared to the tokens in a sentence. Taking this case as an intuitive example, in the sequence of $< top, bottom, shoes >$, the top may be tightly correlated with the shoes instead of the bottom.

Moving one step forward, the graph-wise methods treat each outfit as an item graph, whereby each node represents an item and each edge bridges two items. Based upon the constructed graph, graph neural networks and their variants are designed to calculate the outfit compatibility. For example, Cui *et al.* [3] proposed the node-wise graph neural network (NGNN) towards fashion compatibility modeling. This method constructs a category-oriented fashion graph, where each node represents a category, and accordingly, each outfit can be abstracted as a subgraph consisted with the corresponding category nodes of its composing items. The outfit compatibility score is calculated based on the learned item representations with the attention mechanism. In addition, Cucurull *et al.* [24] utilized a graph neural network to learn the items' embeddings conditioned on their context, and cast the task of FCM as an edge prediction problem. Moreover, Li *et al.* [25] developed a hierarchical fashion graph network (HFGN) for personalized outfit recommendation, which models the relationship among users, items, and outfits simultaneously. Although these methods have achieved great success and outperformed the pair-wise and list-wise methods, they overlook the valuable attribute labels associated with fashion items and therefore neglect the potential of explicitly representing the items from the attribute perspective. Beyond these studies, in this work, we jointly explore the visual and attribute information of fashion items, and model the fashion compatibility by integrating multi-granularities, i.e., attribute- and item-levels.

B. Disentangled Representation Learning

Disentangled representation learning [26] targets at learning multiple factorized representations to capture the latent explanatory factors reside in the observed data, which has drawn increasing research attention from various domains, such as the recommendation domain [27], [28] and computer vision domain [11]–[13], [29]–[31]. For example, in the recommendation domain, Hu *et al.* [32] proposed a graph neural news recommendation model with unsupervised preference disentanglement, where a neighborhood routing mechanism is introduced to dynamically identify the latent preference factors affecting the user's click on a piece of news. In addition, Wang *et al.* [33] presented a disentangled graph collaborative filtering model to mine the fine-grained user-item relationships.

As the compatibility relationship among fashion items can be influenced by multiple latent factors, like color, texture, and style, some researchers also incorporated the disentangled representation to address the task of fashion compatibility modeling. For example, Zheng *et al.* [9] devised a disentangled graph learning scheme, where the collocation compatibility is disentangled into multiple fine-grained compatibilities among fashion items. Similarly, Guan *et al.* [34] presented a comprehensive multimodal outfit compatibility modeling scheme, which not only explores the fine-grained outfit compatibility with disentangled item representations, but also explicitly models the consistent and complementary correlations between the visual and textual modalities of items. Despite of their significant value, the existing efforts mostly overlook the potential of the semantic labels in supervising the disentangled representation learning. Therefore, in this work, we propose to utilize the irregular attributes as the partial supervision to guide the disentangled representation learning of items and introduce the completeness regularizer to prevent the information loss during disentanglement.

C. Graph Convolutional Network

Graph Neural Network (GNN) is devised to learn effective graph representations by updating the node embedding via information aggregation from the node's neighbors. Initially, Gori *et al.* [35] utilized the graph neural network to model the relationship among a set of items. To remedy the long-term message propagation problem, Li *et al.* [36] introduced the Gate Recurrent Units (GRU) in the propagation process. Although GNNs can be applied to most types of graphs, it is hard to train for a fixed point. Inspired by this, Kipf *et al.* [37] introduced the Graph Convolutional Network (GCN), which applies the convolutional operations directly on graphs by updating each node's representation via the information aggregation from its neighbor nodes. In order to improve the model generalization ability, Hamilton *et al.* [38] presented a general inductive framework to learn a function that generates embeddings by sampling and aggregating features from a node's local neighborhood. Thus far, GCNs have been widely explored in various tasks, including but not limited to the tasks of visual comprehension [39], [40], natural language processing [41], recommendation [42], [43], and image recognition [44]. By virtue of its powerful modeling capabilities for unstructured data, we elaborate a hierarchical GCN-based outfit compatibility modeling scheme, where the attribute-level and item-level compatibility modeling is jointly investigated.

III. Methodology

In this section, we first formulate the research problem, and then detail the proposed partially supervised outfit compatibility modeling scheme (PS-OCM for short).

A. Problem Formulation

In this work, we cast the outfit compatibility modeling task as a binary classification problem, i.e., *whether the given outfit is compatible*. Suppose we have a set of N outfits, denoted as $\Omega = \{(O_i, y_i)\}_{i=1}^{N}$, where O_i is the i-th outfit, and y_i denotes its corresponding compatibility label. Specifically, $y_i = 1$ if the outfit O_i is compatible, and $y_i = 0$, otherwise. In addition, we have a set of fashion items $\mathcal{I}$ distributed over T categories. For simplicity, we temporally omit the subscript i of each outfit. An outfit O comprises K fashion items, i.e., $\{I_1, I_2, \ldots, I_K\}$, where $I_i \in \mathcal{I}$ is the i-th composing item of the outfit. Considering that the number of items in

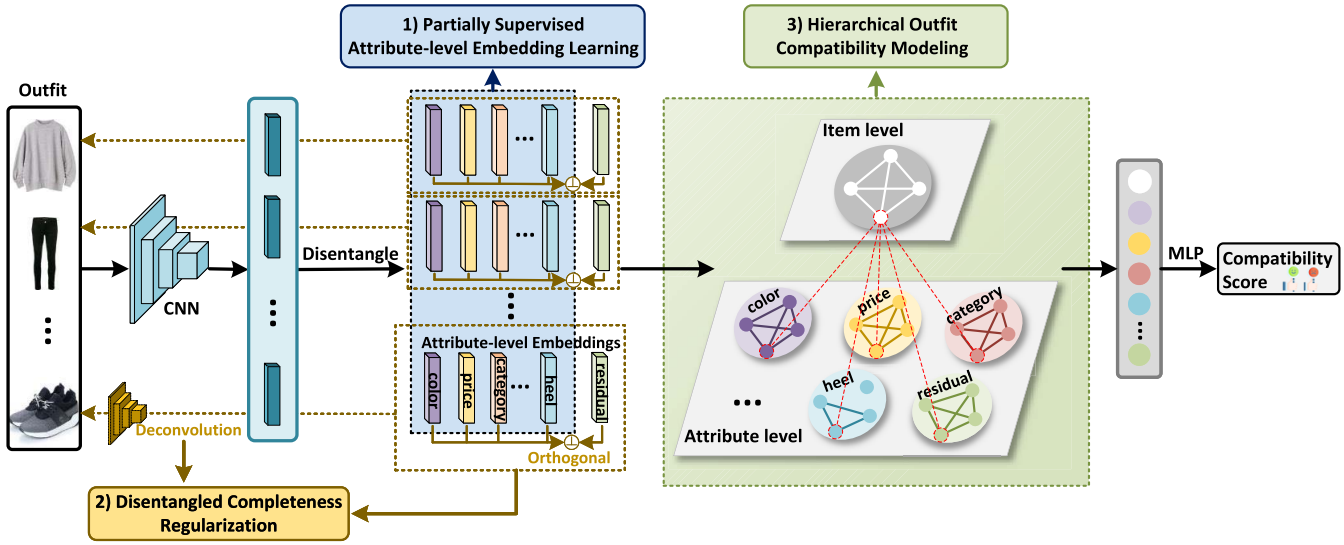


Fig. 2. Illustration of our proposed PS-OCM scheme. It consists of three components: partially supervised attribute-level embedding learning, disentangled completeness regularization, and hierarchical outfit compatibility modeling.

an outfit is not fixed, K is hence a variable. Each item I_i is associated with a visual image V_i and a set of attribute labels $\mathcal{L}_i$. We heuristically pre-defined a set of attributes (*e.g.*, the *color* and *material*) $\mathcal{A} = \{a_m\}_{m=1}^{M}$ that can be applied to characterize all the fashion items, where a_m is the m-th attribute, and M is the total number of attributes. Moreover, each attribute has a set of corresponding attribute values, *e.g.*, *red* and *black* are two possible values for the attribute *color*. We then formally use $\mathcal{V}_m = \{v_m^n\}_{n=1}^{N_m}$ to denote all the possible values for the attribute a_m, and N_m is the corresponding total number of values. Therefore, the set of attribute labels of the i-th item can be written as $\mathcal{L}_i = \{l_i^1, l_i^2, \ldots, l_i^M\}$, where l_i^m denotes the i-th item's m-th attribute label, $l_i^m \in \mathcal{V}_m$ if the item I_i has m-th attribute, otherwise $l_i^m = none$. Usually there are two possible reasons leading to $l_i^m = none$: one is the intrinsic flaws of the dataset due to loose user-generated annotation, and the other is that items of certain categories essentially cannot present certain attributes (*e.g.*, the *trousers* does not have the attribute of *sleeve length*).

In this work, we target at learning an outfit compatibility model $\mathcal{F}$ to judge whether a given outfit O is compatible or not. It is formulated as follows,

$$s = \mathcal{F}\left(\{(V_i, \mathcal{L}_i)\}_{i=1}^{K} | \mathbf{\Theta}\right), \tag{1}$$

where $\mathbf{\Theta}$ refers to the to-be-learned parameters of our model, and s denotes the compatible probability of the given outfit. Table I summarizes the main notations used in this work.

B. PS-OCM

As shown in Figure 2, PS-OCM consists of three components: 1) partially supervised attribute-level embedding learning, 2) disentangled completeness regularization, and 3) hierarchical outfit compatibility modeling. We elaborate them as follows.

TABLE I
SUMMARY OF THE MAIN NOTATIONS

Notation	Explanation
O_i	The i-th outfit.
y_i	The i-th outfit compatibility label.
I_i	The i-th item of an outfit.
a_m	The m-th attribute.
v_m^n	The n-th attribute value of the m-th attribute.
l_i^m	The i-th item's m-th attribute label.
$\mathbf{v}_i$	The extracted visual feature of the i-th item.
$\mathbf{e}_i^j$	The j-th disentangled attribute-level embedding of the i-th item.
p_i^m	The binarized mask indicating whether the i-th item has the m-th attribute label.
q_t^m	The binarized mask denoting whether the m-th attribute is meaningful to items of the t-th category.

1) Partially Supervised Attribute-Level Embedding Learning: This component aims to derive the fine-grained attribute-level representation of the fashion item, which is the basis for the following hierarchical outfit compatibility modeling. Given an outfit, we first extract the visual feature of each composing item via the Convolutional Neural Networks (CNN), which have obtained remarkable success in many computer vision tasks [45], [46]. Specifically, we obtain the overall visual feature embedding of the i-th item in the outfit O as follows,

$$\mathbf{v}_i = \text{CNN}\left(\mathbf{V}_i\right), \tag{2}$$

where $\mathbf{V}_i$ refers to the i-th item image in its raw RGB pixels, $\mathbf{v}_i \in \mathbb{R}^{D_v}$ denotes the extracted visual feature of the i-th item, and D_v is the dimension of the visual feature. In this work, the function CNN refers to the ResNet18 [47] pretrianed on ImageNet.

As aforementioned, we have pre-defined a set of M attributes to characterize all the items. Accordingly, we disentangle the visual feature of each item I_i, i.e., $\mathbf{v}_i$, into M

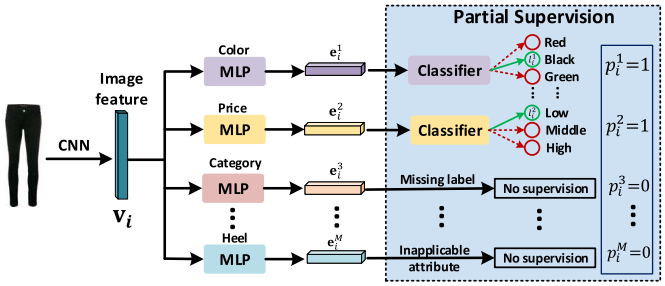


Fig. 3. Partially supervised attribute-level embedding learning.

attribute-level embeddings. We argue that the attributes are not linearly separable, and hence accomplish this task by the non-linear mapping of the MLP. Mathematically, we have

$$\begin{cases} \mathbf{e}_i^1 = \mathrm{MLP}_1(\mathbf{v}_i), \\ \mathbf{e}_i^2 = \mathrm{MLP}_2(\mathbf{v}_i), \\ \quad \vdots \\ \mathbf{e}_i^M = \mathrm{MLP}_M(\mathbf{v}_i), \end{cases} \tag{3}$$

where $\mathbf{e}_i^j \in \mathbb{R}^{D_e}$ $(j = 1, \ldots, M)$ denotes the j-th disentangled attribute-level embedding of the i-th item, and D_e is the dimension.

Different from existing studies that focus on the unsupervised disentangled representation learning, we argue that even the irregular attribute labels of fashion items contain rich cues. Therefore, they can be used to supervise the attribute-level embedding learning and hence strengthen the final compatibility modeling performance. Thereby, we further utilize M MLPs as the attribute classifiers to explore the attribute labels. As aforementioned, the attribute labels of fashion items are irregular. We thus introduce a binary mask $\mathbf{p}_i$ for each item I_i in the outfit to select the available attribute labels of the i-th item. In particular, we define the mask as $\mathbf{p}_i = [p_i^1, p_i^2, \ldots, p_i^M]$, where $p_i^m = \phi(l_i^m)$, and $\phi(\cdot)$ is an indicator function defined as follows,

$$\phi(x) = \begin{cases} 0 & x \text{ is } none, \\ 1 & else. \end{cases} \tag{4}$$

By utilizing the binarized mask, if and only if the item has the corresponding attribute label, we enforce the supervision over the embedding for that attribute. In particular, we adopt the cross-entropy loss to achieve the partial supervision. Formally, for a given outfit O consisting of K items, the partial supervision loss function is formulated as follows,

$$\mathcal{L}_{ps} = \sum_{i=1}^{K} \sum_{m=1}^{M} -\log\left(p\left(l_i^m \mid \mathcal{C}^m\left(\mathbf{e}_i^m\right)\right)\right) p_i^m, \tag{5}$$

where $\mathcal{C}^m(\cdot)$ is the label classifier for the m-th attribute, $\mathbf{e}_i^m$ is the disentangled embedding of the m-th attribute, and l_i^m is the ground truth attribute label. We illustrate the procedure of partially supervised disentangled attribute-level embedding in Figure 3.

2) Disentangled Completeness Regularization: To prevent information loss during the disentangling process which may degrade the model performance, we devise a disentangled completeness regularizer, as illustrated in Figure 2. In particular, we rely on two strategies to regulate the disentangling process: orthogonal residual embedding and visual representation reconstruction.

a) Orthogonal residual embedding: There may be some implicit visual properties of the item that cannot be represented by the pre-defined set of attributes. We thus introduce another special attribute *residual* to compensate the information loss during the disentangled representation learning. Specifically, similar to the M attribute-level embeddings, we adopt another MLP to derive the residual attribute embedding via,

$$\mathbf{e}_i^{M+1} = \mathrm{MLP}_{M+1}(\mathbf{v}_i), \tag{6}$$

where $\mathbf{e}_i^{M+1} \in \mathbb{R}^{D_e}$ denotes the residual attribute embedding.

Since the residual attribute embedding acts as a compensation for fully representing the item, we argue that it should be complementary to other M attribute-level embeddings that have clear semantics. In other words, the residual embedding should be orthogonal to each other attribute-level embedding. It is worth noting that although we disentangle the visual feature of each fashion item into M attribute-level embeddings, certain embeddings of the given item maybe meaningless, since some attributes are not universal and cannot be applied to certain items. For example, we can discuss the attribute *sleeve length* for a *T-shirt* but not *trousers*, and the attribute *heel* for a pair of *shoes* rather than a *T-shirt*. In the light of this, for each item category, we should define a set of meaningful attributes to guarantee the effective orthogonal regularization. Towards this end, we first build the *category-attribute* associations. For the t-th category, we take the union set of attributes used to label items in the t-th category as the whole set of applicable attributes, denoted as $\mathcal{T}_t$. We then introduce a mask $\mathbf{q}_t = [q_t^1, q_t^2, \ldots, q_t^M]$ to select the meaningful attributes for the t-th item category, where $q_t^m = 1$ if the pre-defined m-th attribute belongs to the applicable attribute set $\mathcal{T}_t$, otherwise $q_t^m = 0$. It is worth noting that in the aforementioned partial supervision module, only the attribute-level embeddings that have corresponding labels are triggered. Whereas in the this orthogonal regularization, we further utilize the attribute-level embedding that even has no corresponding label, as long as it can be possibly presented by this item.

Ultimately, we have the following orthogonal regularization,

$$\begin{aligned} \mathcal{L}_{or} &= \sum_{i=1}^{K} \sum_{m=1}^{M} \left[\cos\left(\hat{\mathbf{e}}_i^m, \mathbf{e}_i^{M+1}\right)\right]^2 \\ &= \sum_{i=1}^{K} \sum_{m=1}^{M} \left[\cos\left(q_{t_i^*}^m \mathbf{e}_i^m, \mathbf{e}_i^{M+1}\right)\right]^2, \end{aligned} \tag{7}$$

where $\cos(\cdot,\cdot)$ is the cosine similarity function, and $t_i^* \in \{1, 2, \ldots, T\}$ refers to the category of the i-th item. It is worth mentioning that once the m-th attribute cannot be applied to the item I_i, i.e., $q_{t_i^*}^m = 0$, we will ignore the orthogonal regularization between that attribute-level embedding and the residual one.

b) Visual representation reconstruction: To avoid information loss during disentangled representation learning, we regulate the disentangled embeddings to be able to reconstruct the original item visual representation. In light of this, we feed the concatenation of the meaningful disentangled attribute-level embeddings of the item I_i and the residual one into the deconvolutional neural network [48]. It is formulated as,

$$\hat{\mathbf{V}}_i = \mathcal{D}\left(\left[q_{t_i^*}^1 \mathbf{e}_i^1 \| q_{t_i^*}^2 \mathbf{e}_i^2 \| \ldots, q_{t_i^*}^M \mathbf{e}_i^M \| \mathbf{e}_i^{M+1}\right]\right), \tag{8}$$

where the binary masks $\mathbf{q}_{t_i^*}^m$'s are used to select the meaningful attribute embeddings of the item I_i, $[\cdot\|\cdot]$ refers to the concatenation operation, $\mathcal{D}(\cdot)$ denotes the deconvolutional neural network, and $\hat{\mathbf{V}}_i$ stands for the reconstructed visual representation of the i-th item. We hereafter utilize l_2 loss to regulate the distance between the reconstructed visual representation and the origin one via,

$$\mathcal{L}_{rec} = \sum_{i=1}^{K} \left\| \hat{\mathbf{V}}_i - \mathbf{V}_i \right\|_F^2. \tag{9}$$

Combining the losses of both the orthogonal residual embedding and the visual representation reconstruction constraints, we reach the final loss for regularizing the disentangled completeness as follows,

$$\mathcal{L}_{dc} = \mathcal{L}_{or} + \mathcal{L}_{rec}. \tag{10}$$

3) Hierarchical Outfit Compatibility Modeling: Inspired by previous studies [3], [24], we leverage GCNs to model the outfit compatibility. Beyond existing work, we design a novel hierarchical graph convolutional network, which is capable of modeling the complex compatibility relations among items in an outfit from both attribute and item levels. In particular, the attribute-level compatibility modeling aims to investigate the fine-grained compatibility among fashion items, while the item-level one targets at summarizing the the coarse-grained outfit compatibility from the item level.

a) Attribute-level compatibility modeling: Regarding the attribute-level compatibility modeling, given an outfit, we first construct $M+1$ parallel compatibility modeling graphs $\mathcal{G}_a^m = \left(\mathcal{N}_a^m, \mathcal{E}_a^m\right), (m = 1, 2, \ldots, M+1)$, with each devised to model the outfit compatibility from an attribute aspect.[1] In particular, $\mathcal{N}_a^m$ and $\mathcal{E}_a^m$ refer to the set of nodes and edges of the graph $\mathcal{G}_a^m$, respectively. In the graph $\mathcal{G}_a^m$, each node refers to a composing item of the outfit that has the corresponding attribute, i.e., a_m. Notably, as aforementioned, not every attribute can be applied to all the items, e.g., the attribute *sleeve length* cannot be used to characterize a pair of *trousers*. Therefore, for different attributes, different number of items are applicable for the attribute-level compatibility modeling. In other words, graphs corresponding to different attributes may have different numbers of nodes. Towards this end, for the ease of presentation, we still deploy K item nodes for all these graphs, i.e., $\mathcal{N}_a^m = \left\{\hat{n}_i^m\right\}_{i=1}^K$, where $\hat{n}_i^m$ is the i-th node in the graph $\mathcal{G}_a^m$. However, some nodes in these graphs will be defined as the virtual isolated ones and inactive during the attribute-level compatibility propagation.

During the learning process, each node $\hat{n}_i^m$ is associated with a hidden state vector $\mathbf{h}_i^m$, which will be updated to fulfil the compatibility information propagation over the graph. We initialize the hidden vector of the node $\hat{n}_i^m$ by,

$$\mathbf{h}_i^m = \begin{cases} q_{t_i^*}^m \mathbf{e}_i^m, & m \in \{1, 2, \ldots, M\}, \\ \mathbf{e}_i^{M+1}, & m = M+1. \end{cases} \tag{11}$$

In this way, if the m-th attribute can be applied to the item of the i-th node, we will initialize the node with the item's corresponding attribute feature. Otherwise the node will be initialized with an all-zero vector, making it an isolated node in the graph, and it will not join the subsequent compatibility information propagation. Regarding the edge construction for each graph, we introduce an edge between each pair of non-isolated nodes, i.e., each pair of meaningful items in the corresponding attribute-level compatibility modeling.

To simplify the notation, considering that the parallel attribute-level compatibility modeling for different attributes follow the same learning process, we temporally remove all the superscripts m from the above notations and present the general attribute-level compatibility modeling scheme as an example. Inspired by Graph Attention Networks (GAT) [49], we employ the attention mechanism to make each node adaptively absorb compatibility information from the neighbors. Formally, we have

$$\alpha_{ij} = \frac{\exp\left(\mathbf{W}_a \left[\mathbf{h}_i \| \mathbf{h}_j\right]\right)}{\sum_{n_k \in \mathcal{N}_i} \exp\left(\mathbf{W}_a \left[\mathbf{h}_i \| \mathbf{h}_k\right]\right)}, \tag{12}$$

where α_{ij} indicates the importance of the node n_j's hidden state to the node n_i, $\mathbf{W}_a$ is a weight matrix to perform the linear transformation, $[\cdot\|\cdot]$ refers to the concatenation operation, and $\mathcal{N}_i$ denotes the neighborhood of node n_i. Once the attention weights α_{ij}'s are obtained, they are then used to propagate information from the neighbors of node n_i to the node itself by,

$$\mathbf{h}_i' = \omega\left\{\mathbf{W}_u\left[\sum_{n_j \in \mathcal{N}_i} \alpha_{ij}\left(\mathbf{h}_i \odot \mathbf{h}_j\right)\right] + \mathbf{b}_u\right\}, \tag{13}$$

where $\odot$ denotes the element-wise multiplication, $\mathbf{W}_u$ and $\mathbf{b}_u$ are the parameters of the fully-connected layer, and ω refers to the nonlinear activation function LeakyReLU. In a sense, the element-wise multiplication $\mathbf{h}_i \odot \mathbf{h}_j$ indicates the compatibility information between the items I_i and I_j. More generally, instead of propagating the features of node n_i's neighbors, we propagate the compatibility information between node n_i and its neighbors, which has proven to be effective in tackling the outfit compatibility modeling task [34].

Based upon the above inference and computation, the updated hidden representation of node n_i is written as,

$$\tilde{\mathbf{h}}_i = \omega\left(\mathbf{W}_o \mathbf{h}_i + \mathbf{b}_o\right) + \mathbf{h}_i', \tag{14}$$

where $\mathbf{W}_o$ and $\mathbf{b}_o$ denote the weight matrix and bias to be learned, respectively. The symbol ω denotes the LeakyReLU function. We ultimately feed the updated hidden

[1] As aforementioned, the residual attribute is also incorporated as a special implicit attribute.

node embeddings into a MLP to derive the attribute-specific compatibility score of the given outfit via,

$$\begin{cases} c_i = \mathbf{W}_2 \left[\psi \left(\mathbf{W}_1 \tilde{\mathbf{h}}_i + \mathbf{b}_1 \right) \right] + \mathbf{b}_2, \\ c = \frac{1}{K} \sum_{i=1}^{K} c_i, \end{cases} \tag{15}$$

where $\mathbf{W}_1$, $\mathbf{W}_2$, $\mathbf{b}_1$, and $\mathbf{b}_2$ are the parameters of the MLPs, the symbol ψ denotes the ReLu active function, and c is the compatibility score. Following the above general scheme, we can obtain all the attribute-level compatibility scores, denoted as $\mathbf{c}_a = \left[c^1, c^2, \ldots, c^M, c^{M+1}\right]$, as well as the updated hidden attribute-level embeddings of each node/item, i.e., $\tilde{\mathbf{h}}_i = [\tilde{\mathbf{h}}_i^1, \tilde{\mathbf{h}}_i^2, \ldots, \tilde{\mathbf{h}}_i^M, \tilde{\mathbf{h}}_i^{M+1}]$.

b) Item-level compatibility modeling: Similar to attribute-level compatibility modeling, we also construct a compatibility modeling graph $\mathcal{G}_o = (\mathcal{N}_o, \mathcal{E}_o)$ at the overview item level, where $\mathcal{N}_o$ and $\mathcal{E}_o$ refer to the node set and the edge set, respectively. The difference is that we initialize the hidden vector of the i-th node in the graph $\mathcal{G}_o$ from two aspects: the item's original visual feature $\mathbf{v}_i$, and the updated attribute-level item embedding $\tilde{\mathbf{h}}_i^m$'s from the attribute-level compatibility modeling scheme. In this way, a more comprehensive overview representation of the item is derived. Specifically, for the i-th node in the graph $\mathcal{G}_o$, we initialize its hidden vector as follows,

$$\mathbf{g}_i = \left[\mathbf{v}_i \| \mathbf{W}_h \left(\left[\tilde{\mathbf{h}}_i^1 \| \tilde{\mathbf{h}}_i^2 \| \cdots \| \tilde{\mathbf{h}}_i^{M+1} \right] \right) \right], \tag{16}$$

where $[\cdot\|\cdot]$ denotes the concatenation operation, and $\mathbf{W}_h \in \mathbb{R}^{D_v \times D_e(M+1)}$ is the to-be-learned weight matrix, which projects the attribute-level embeddings to the same space of the entire visual one. Following the same information propagation scheme as the attribute-level compatibility modeling, we can obtain the item-level compatibility score c_o.

Taking both the attribute- and item-level compatibility modeling results into account, we feed the concatenation of the attribute- and item-level compatibility scores, i.e., $\mathbf{c} = [c_a \| \mathbf{c}_o]$, into the MLP to get the final compatibility probability score as follows,

$$s = \sigma \left\{ \mathbf{W}_4 \left[\psi \left(\mathbf{W}_3 \mathbf{c} + \mathbf{b}_3 \right) \right] + \mathbf{b}_4 \right\}, \tag{17}$$

where $\mathbf{W}_3$, $\mathbf{W}_4$, $\mathbf{b}_3$, and $\mathbf{b}_4$ are the parameters of the MLP, the symbol ψ denotes the ReLU active function, and σ refers to the Sigmoid active function. We finally adopt the cross-entropy loss to optimize our proposed PS-OCM, and reach the following formulation,

$$\mathcal{L}_{hc} = -y log(s) - (1 - y) log(1 - s), \tag{18}$$

where y is the ground truth compatibility label for the outfit O. Accordingly, the total loss for our PS-OCM can be written as follows,

$$\mathcal{L} = \mathcal{L}_{hc} + \lambda \mathcal{L}_{ps} + \mu \mathcal{L}_{dc}, \tag{19}$$

where λ and μ are trade-off hyper-parameters.

Interpretability. In a sense, the semantic attributes have explicit meaning and can be used naturally to interpret the compatibility evaluation result. In particular, we can identify the prominent attributes that contributing to the final compatibility evaluation most, according to the absolute values of these attribute-specific compatibility scores, i.e., c^m's.

TABLE II
ATTRIBUTES AND THE POSSIBLE VALUE

Attribute	Possible Value	Total Number
Color	Grey, Black, Green, $\cdots$	12
Price	Low, Middle, High.	3
Brand	ABISTE, FURLA, BEIGE, $\cdots$	5,180
Category	Trousers, Belt, Handbag, $\cdots$	61
Variety	Coat, Bag, Cosmetics, $\cdots$	20
Material	Fur, Leather, Denim, $\cdots$	37
Pattern	Stripe, Embroidery, Animal, $\cdots$	15
Design	Turtleneck, Frill, Ribbons, $\cdots$	23
Heel	Chunky, Pin, High, $\cdots$	6
Dress Length	Short, Middle, Long.	3
Sleeve Length	Sleeveless, Long, Short, $\cdots$	4

IV. EXPERIMENT

In this section, we first introduce the experimental settings, and then detail the experiments that we conducted on a real-world dataset by answering the following research questions:

- **RQ1**: How does the hyperparameters affect our model?
- **RQ2**: Does PS-OCM outperform existing methods?
- **RQ3**: How does each component affect PS-OCM?
- **RQ4**: What is the intuitive evaluation result of PS-OCM?

A. Experimental Settings

1) Datasets: To justify our model, we resorted to the public dataset IQON3000 [50], due to the fact that each item in IQON3000 has not only the visual image, but also several semantic attributes, such as the color and category. In particular, IQON3000 consists of 308,747 outfits, composed by 672,335 items. In total, there are 11 attributes provided by this dataset. Table II shows the possible value examples and the corresponding number for each attribute. To ensure the quality of the dataset, we empirically sampled 20,000 compatible outfits, each of which consists of at least 2 but no more than 10 items. Since the dataset only provides the compatible outfits, it is needed to compose the incompatible ones for training. Specifically, for each compatible outfit, we replaced each of its composing items with a randomly sampled item from the same category to construct the incompatible outfit. In this manner, we end up with a set of 40,000 compatible/incompatible outfits. We then divided it into the training set, validation set, and test set according to the ratio of 8 : 1 : 1.

2) Evaluation Tasks and Metrics: Similar to previous studies [3], [4], [17], [24], [34], we justified our proposed PS-OCM scheme with two specific tasks: outfit compatibility estimation and fill-in-the-blank (FITB). The former task is to evaluate the compatibility score of a given outfit, where we adopted the AUC (Area Under the ROC curve) [51] as the corresponding evaluation metric. The latter task is to choose one item from a set of candidates (i.e., one positive item and three negative items), for a given incomplete outfit (with an item missing). For this task, we composed each candidate item with the given items as a complete outfit, and used the well-trained model to compute its compatibility score. We then chose the item with the highest score as the answer. For this task, we applied the accuracy as the evaluation metric.

TABLE III

PERFORMANCE COMPARISON BETWEEN OUR PROPOSED PS-OCM AND OTHER BASELINE METHODS ON TWO TASKS OVER THE IQON3000 DATASET. NOTABLY, THE BASELINE METHODS WERE RE-TRAINED BY THE RELEASED CODES. THE BEST RESULTS ARE IN BOLD, WHILE THE SECOND BEST RESULTS ARE UNDERLINED

Method	Compatibility AUC	FITB Accuracy
Type-aware [17]	0.6688	0.3901
SCE-NET [55]	0.6792	0.3783
Bi-LSTM [4]	0.7739	0.3813
NGNN [3]	0.7591	0.4002
HFGN [19]	0.8243	0.4511
MM-OCM [34]	0.8444	0.4661
OCM-CF [10]	0.8402	0.4825
MOCM-MGL [56]	<u>0.8929</u>	<u>0.5160</u>
PS-OCM	0.9009	0.5412
PS-OCM-ResNet50	0.9029	0.5746
PS-OCM-SwinTransformer	**0.9295**	**0.5853**

3) Implementation Details: For the image encoder, we employed the ResNet18 [47] pre-trained on ImageNet [52] as the backbone, and modified the last layer to make the output feature dimension as 256. Pertaining to the MLPs that obtain the disentangled attribute-level embeddings, we set the output dimension to 64. For each label classifier, we implemented it by the two-layer MLP with the LeakyReLU activation, whose output dimension is set to the number of corresponding attribute values. As for the deconvolutional neural network, we stacked five transposed convolution layers, and the first four layers are followed by a Batch Normalization [53] and ReLU activation, while the last layer is followed by a Tanh activation to scale the output values. We selected Adam [54] as the training optimizer, with a fixed learning rate of 0.0001. We empirically set the batch size as 32, and both trade-off hyper-parameters, i.e., λ and μ in Eqn.(19), as 1. All the experiments are implemented by PyTorch over a server equipped with 4 GeForce RTX 2080 Ti GPUs, and the random seeds for model initialization are fixed for the reproducibility.

B. On Hyper-Parameters (RQ1)

In addition, we studied the influence of the key hyper-parameters, including the trade-off parameters λ and μ in Eqn.(19), the output feature dimension of the image encoder, and the depth of the GCNs in our hierarchical outfit compatibility modeling component.

The trade-off parameters λ and μ are searched among values of $[0.01, 0.05, 0.1, 0.5, 1]$. Figure 4(a) and Figure 4(b) show the performance of our model on the testing set with different hyperparameter values. As can be seen, our method performs best when $\lambda = 1$ and $\mu = 1$. This suggests that both the partially supervised attribute embedding learning component and disentangled completeness regularization component contribute to the model. As for the feature dimension, it is searched among values of [64, 128, 256, 512, 1024, 2048]. As can be seen from Figure 4(c), our model is not sensitive to this parameter in the FITB task when the dimension does not exceed 1024. For the sake of efficiency, the feature dimension is set to 256. Moreover, Figure 4(d) shows the performance of our model with the number of GCN layers ranging from 1 to 5. As can be seen, our model performs generally stable when the number of GCN layers is no more than 3. However, when the number of GCN layers keeps increasing, our model's performance significantly drops. This observation is similar to that reported in [57], and can be attributed to that the more layers may lead to the overfitting problem and hence hurt the model's performance. Overall, our model achieves the best performance with only one GCN layer.

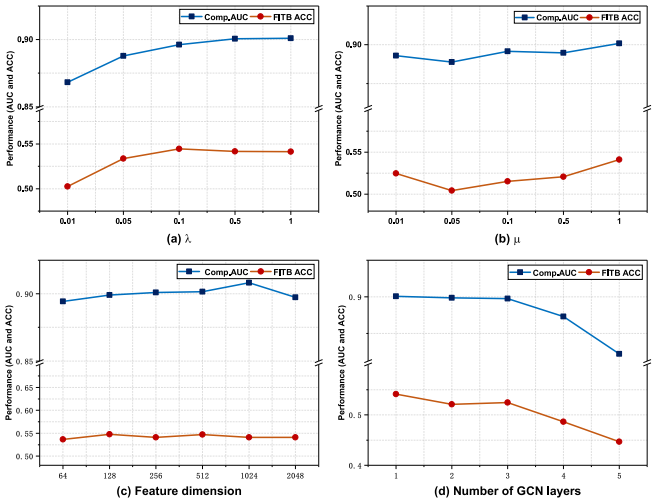


Fig. 4. Influence of trade-off parameters (a) λ and (b) μ, (c) feature dimension, and (d) number of GCN layers on two tasks.

C. On Model Comparison (RQ2)

To validate the effectiveness of our proposed scheme, we chose the following baselines for comparison, including the pair-wise, sequence-wise, and graph-wise models.

- **Type-aware** [17] devises the type-specific embedding spaces according to the item types, to facilitate the outfit compatibility measurement. The visual-semantic loss is utilized to incorporate the visual and textual information.
- **SCE-NET** [55] embeds the item visual features into multiple semantic subspaces by multiple condition masks, and uses the multimodal features to derive the importance weights for different subspace features to obtain the final item representations.
- **Bi-LSTM** [4] takes items in an outfit as a sequence, and exploits the latent item interaction by a bi-directional LSTM. Notably, the textual information is also adopted to regularize the outfit compatibility modeling by the visual-semantic consistency loss.
- **NGNN** [3] represents each outfit as a graph, and utilizes an attention mechanism to calculate the outfit compatibility score. For multimodal features, NGNN designs two graph channels, and derives the final compatibility score with the late fusion.
- **HFGN** [19] develops a hierarchical fashion graph network to jointly fulfill the fashion compatibility modeling and personalized outfit recommendation, where a category-oriented fashion graph is built for each outfit. It only uses the visual features.

- **MM-OCM** [34] explicitly models the consistent and complimentary relations between the visual and textual modalities of fashion items by the parallel and orthogonal regularizations. Moreover, MM-OCM jointly unifies the text-oriented and vision-oriented outfit compatibility modeling with the mutual learning strategy.
- **OCM-CF** [10] directly learns the context-aware global outfit representation by GCNs and the multi-head attention mechanism, and employs multiple network branches to explore the hidden complementary factors that affect the outfit compatibility.
- **MOCM-MGL** [56] proposes a multi-modal outfit compatibility modeling with modality-oriented graph learning. It takes both visual, textual, and category modalities as input and jointly propagates the intra-modal and inter-modal compatibilities among fashion items in the outfit.[2]
- **PS-OCM-Resnet50/PS-OCM-SwinTransformer**. To study the effect of utilizing different backbones to extract the image features, we replaced the Resnet18 backbone to Resnet50 and SwinTransformer [58], respectively.

Table III shows the performance of different methods on the outfit compatibility estimation task and fill-in-the-blank task. Notably, the baseline methods are re-trained by the released corresponding codes over the IQON3000 dataset. From this table, we had the following observations.

1) The pair-wise methods, i.e., Type-aware and SCE-NET, achieve the worst performance on both two tasks. This maybe due to the fact that the pair-wise methods mainly justify the local compatibility between two items, lacking the global view of the whole outfit.
2) The sequence-wise method, i.e., Bi-LSTM, performs better than the pair-wise methods, but worse than the graph-wise methods, i.e., HFGN and MM-OCM. On the one hand, this confirms the advantage of treating the outfit as a unified sequence rather than the item pairs. On the other hand, this implies that treating the outfit as an ordered sequence of fashion items is still suboptimal. This may be attributed to that the sequence-wise method can suffer from the cumulative error propagation problem, since it computes the outfit compatibility score by keeping predicting the next item with the previous ones.
3) Our methods consistently surpass all the baseline methods on both tasks. This confirms the advantage of our scheme that utilizes the irregular attribute labels to provide the partial supervision to strengthen the item representation learning and employs the hierarchical graph convolutional network to integrate the attribute-level and item-level outfit compatibility learning.
4) PS-OCM-SwinTransformer performs better than both PS-OCM and PS-OCM-ResNet50, indicating the superiority of swin transformer in image feature extraction and hence boost the final performance.

To gain deep insights about our proposed PS-OCM, we further checked the performance of our PS-OCM for outfits with different numbers of composing items on the two tasks. In particular, we reported the performance of our model for outfits with number of composing items ranging from 2 to 10. As can be seen from Figure 5, our PS-OCM is generally not sensitive to the composing numbers, which indicates that our model PS-OCM has the capacity of handling the compatibility modeling for outfits with various numbers of items.

[2]For fair comparison, the attribute information is utilized as the pure text in MOCM-MGL.

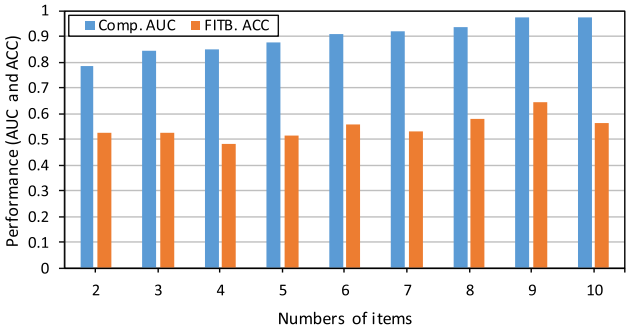


Fig. 5. Performance of PS-OCM for outfits with different numbers of items.

D. On Ablation Study (RQ3)

To justify each component in our model, we conducted ablation experiments on the following derivatives.

- **w/o Partial_Supervision**: To explore the effect of the partially supervised attribute embedding learning component, we removed the partial supervision loss by setting $\lambda = 0$ in Eqn.(19).
- **w/o Orthogonal**: To study the effect of the orthogonal regularization during the visual attributes disentanglement, we removed the orthogonal regularization $\mathcal{L}_{or}$ in Eqn.(10).
- **w/o Reconstruction**: To validate the necessity of visual representation reconstruction learning, we removed the visual representation reconstruction constraint $\mathcal{L}_{rec}$ in Eqn.(10).
- **w/o Hierarchical_Graph**: To validate the function of hierarchical graph compatibility modeling component, we removed this part by directly concatenating the attribute-level embeddings of each outfit to obtain the overall outfit representation and passing it to a MLP to get the outfit's compatibility score.
- **Attribute-level_Only**: To verify the importance of coarse-grained item-level information, this derivative only utilizes the fine-grained attribute-level compatibility modeling part in the hierarchical graph compatibility modeling component.
- **Item-level_Only**: Similarly, to justify the necessity of introducing the fine-grained attribute-level compatibility modeling, we removed it from the hierarchical outfit compatibility modeling network.

Based on ablation experimental illustrated in Table IV, we found that our model consistently outperforms all the above derivatives on both tasks, which demonstrates the effectiveness of each component in our proposed PS-OCM. Specifically, we have the following detailed observations.

TABLE IV

ABLATION STUDY OF OUR PROPOSED PS-OCM ON IQON3000 DATASET. THE BEST RESULTS ARE IN BOLD

Method	Compatibility AUC	FITB Accuracy
w/o Partial_Supervision	0.8433	0.4866
w/o Orthogonal	0.8938	0.5293
w/o Deconvolution	0.8909	0.5293
w/o Hierarchical_Graph	0.8197	0.4459
Attribute-level_Only	0.8848	0.5337
Item-level_Only	0.8720	0.5292
PS-OCM	**0.9009**	**0.5412**

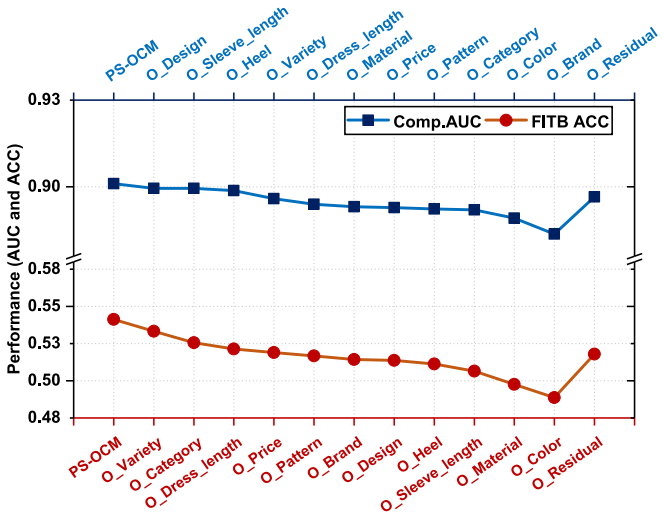


Fig. 6. Comparison of the effect of removing each single attribute from our PS-OCM on two tasks.

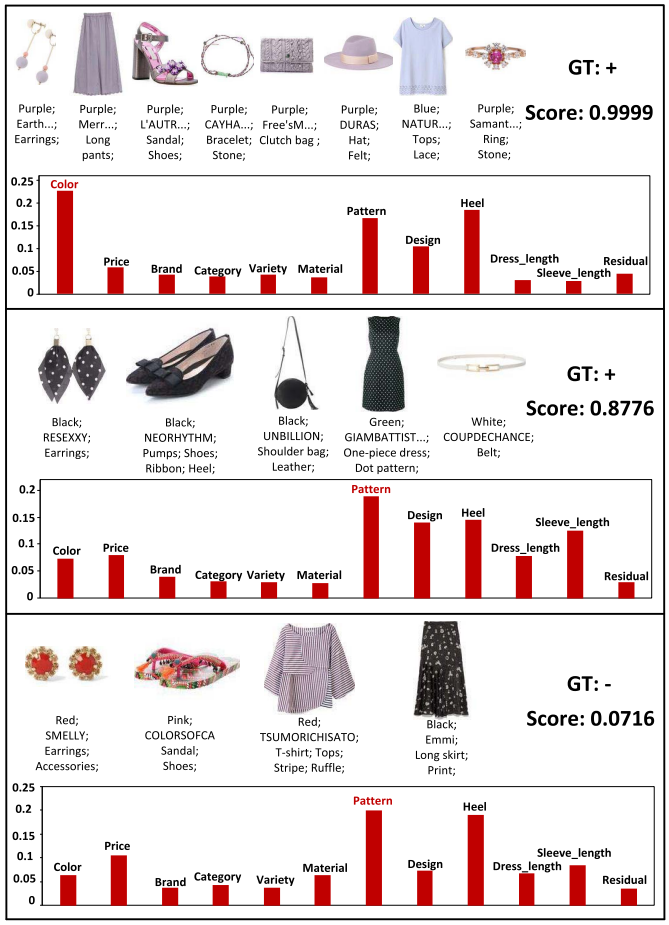


Fig. 7. Case study of PS-OCM on the outfit compatibility estimation task.

1) The performance of w/o Partial_Supervision significantly drops, as compared to PS-OCM, indicating that partially supervised attribute embedding learning component is indeed helpful to strengthen the visual representation learning performance. 2) Both w/o Orthogonal and w/o Reconstruction are inferior to PS-OCM, which suggests that it is essential to consider the orthogonal regularization and visual feature reconstruction to prevent the visual information loss during the visual feature disentanglement and guarantee the completeness for the disentanglement. And 3) w/o Hierarchical_Graph delivers the worst performance, reflecting the overall effectiveness of our proposed hierarchical outfit compatibility modeling component. Moreover, both Attribute-level_Only and Item-level_Only performs better then w/o Hierarchical_Graph, which confirms the necessity of jointly incorporating the attribute-level and item-level compatibility modeling modules. In a sense, this also reflects that the fine-grained attribute-level features and the overview item-level features complement each other to certain level toward the outfit compatibility modeling.

As the partially supervised attribute-level embedding learning contributes the key novelty of our work, we further studied the effect of removing each attribute embedding from the training phase of our PS-OCM. As aforementioned, we had 12 attributes, including 11 concrete attributes in the original dataset and one "residual" attribute we newly defined. Accordingly, we omitted each of the 12 attributes from our model, and hence obtained 12 derivatives of our model, with each named as O_{each_attribute}. Figure 6 shows the performance of our PS-OCM and its derivatives on the two tasks. As can be seen, removing any concrete attribute (e.g., the design or color) hurts our model's performance, which verifies that each concrete attribute contributes to the outfit compatibility modeling. In particular, we noticed that the color attribute greatly affects our model's performance on both tasks, which is reasonable, as the color attribute is the most straightforward influential factor on the outfit compatibility modeling. Meanwhile, we found that O_residual underperforms our PS_OCM. This reflects the importance of the residual attribute, and indicates its capability of compensating the information loss during the attribute representation disentanglement.

E. On Case Study (RQ4)

To get the intuitive understanding of our model, we also conducted the case study of our method in the two tasks: outfit compatibility estimation and fill-in-the-blank.

Figure 7 shows several testing examples of our model on the outfit compatibility estimation task, where the importance distribution of attributes, i.e., the normalization of the absolute values of the attribute-level compatibility scores, is also given

Fig. 8. Case study of PS-OCM and its several derivatives as well as the best baseline MOCM-MGL on the FITB task.

to intuitively demonstrate the interpretability of our model. As can be seen from the first example, our model yields the correct compatibility estimation, and captures the color attribute as the most important influential factor. This is reasonable as the color presented by the outfit is harmonious. In second example, our model also gives a high compatible probability score, and identifies that the pattern attribute is the most important factor. As we can see, the earrings and the dress in the given outfit do consistently present the dotted pattern. Accordingly, the result makes sense. As for the last incompatible example, our PS-OCM gives a low compatibility score, and the pattern attribute is also captured as the most important factor contributing to the incompatible estimation result. From this example we found that the striped pattern of the T-shirt, spotted pattern of the skirt, and floral pattern of the sandal indeed form no compatible look.

Figure 8 shows several testing results of our PS-OCM, compared with its several derivatives and MOCM-MGL which gains the best performance among baselines. In particular, the first column refers to the questions of the fill-in-the-blank task, and the second column lists the corresponding four options, where the ground truth item is denoted with a green box. The last column shows the choice yielded by each method and indicates whether the choice is true or not by a green tick and red cross. As can be seen from the first example in Figure 8, only w/o Partial_Supervision and MOCM-MGL fail to give the correct choice, i.e, the second item that has the same geometric pattern with the given earrings. This suggests the effectiveness of incorporating irregular attribute information as the partial supervision. In the second example, all the derivatives chose the false item, which further demonstrates the importance of each designed component in PS-OCM. Regarding the last example, although all our methods fail to give the correct answer, we noticed that their chosen items also go well with the given question items, especially from the color perspective. This also implies the effectiveness of our model.

V. Conclusion and Future Work

In this work, towards outfit compatibility modeling, we present a novel partially supervised compatibility modeling, named PS-OCM, which consists of three key components: 1) partially supervised attribute embedding learning; 2) disentangled completeness regularization; and 3) hierarchical outfit compatibility modeling. In particular, we first present a partially supervised disentangled learning method to disentangle the visual representation of each item into several attribute-level embeddings. In addition, we devise the disentangled completeness regularization to prevent the information loss during disentanglement. Finally, we design a hierarchical graph convolutional network that jointly performs the attribute- and item-level compatibility modeling. Extensive experiments have been conducted on a real-world dataset with two popular tasks: the outfit compatibility prediction and fill-in-the-blank. The encouraging experiment results validate the superiority of our proposed model and the importance of its each component. In addition, we found that our PS-OCM is not sensitive to the number of items in the outfit, and removing each attribute, including the introduced residual one, from the embedding disentanglement will hurt the model's performance. This shows that each attribute could affect the outfit compatibility modeling to some extent.

The limitation of our work is that currently we only evaluate the outfit compatibility from the general standard. In fact, there may be some subjective factors influencing the outfit compatibility evaluation, namely, for the same garment, different users may have different evaluations. Therefore, in future, we intend to study the personalized fashion compatibility modeling, where the user's preference would be explored.

REFERENCES

[1] W.-L. Hsiao and K. Grauman, "Creating capsule wardrobes from fashion images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7161–7170.

[2] W. Cheng, S. Song, C. Chen, S. C. Hidayati, and J. Liu, "Fashion meets computer vision: A survey," *ACM Comput. Surveys*, vol. 54, no. 4, pp. 72:1–72:41, 2021.

[3] 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.

[4] 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.

[5] X. Song, F. Feng, X. Han, X. Yang, W. Liu, and L. Nie, "Neural compatibility modeling with attentive knowledge distillation," in *Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jun. 2018, pp. 5–14.

[6] J. Liu, X. Song, Z. Ren, L. Nie, Z. Tu, and J. Ma, "Auxiliary template-enhanced generative compatibility modeling," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 3508–3514.

[7] A. Revanur, V. Kumar, and D. Sharma, "Semi-supervised visual representation learning for fashion compatibility," in *Proc. 15th ACM Conf. Recommender Syst.*, Sep. 2021, pp. 463–472.

[8] S. C. Hidayati *et al.*, "Dress with style: Learning style from joint deep embedding of clothing styles and body shapes," *IEEE Trans. Multimedia*, vol. 23, pp. 365–377, 2021.

[9] N. Zheng, X. Song, Q. Niu, X. Dong, Y. Zhan, and L. Nie, "Collocation and try-on network: Whether an outfit is compatible," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 309–317.

[10] T. Su, X. Song, N. Zheng, W. Guan, Y. Li, and L. Nie, "Complementary factorization towards outfit compatibility modeling," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 4073–4081.

[11] X. He, Y. Peng, and J. Zhao, "Which and how many regions to gaze: Focus discriminative regions for fine-grained visual categorization," *Int. J. Comput. Vis.*, vol. 127, no. 9, pp. 1235–1255, Sep. 2019.

[12] J. Zhao, Y. Peng, and X. He, "Attribute hierarchy based multi-task learning for fine-grained image classification," *Neurocomputing*, vol. 395, pp. 150–159, Jun. 2020.

[13] L. Zhang, S. Huang, and W. Liu, "Intra-class part swapping for fine-grained image classification," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 3208–3217.

[14] X. Yang, X. Song, F. Feng, H. Wen, L. Duan, and L. Nie, "Attribute-wise explainable fashion compatibility modeling," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 17, no. 1, pp. 36:1–36:21, 2021.

[15] H. Zhan, J. Lin, K. E. Ak, B. Shi, L. Duan, and A. C. Kot, "A^3-FKG: Attentive attribute-aware fashion knowledge graph for outfit preference prediction," *IEEE Trans. Multimedia*, vol. 24, pp. 819–831, 2022.

[16] 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.

[17] M. I. Vasileva, B. A. Plummer, K. Dusad, S. Rajpal, R. Kumar, and D. A. Forsyth, "Learning type-aware embeddings for fashion compatibility," in *Proc. Eur. Conf. Comput. Vis.*, vol. 11220. Cham, Switzerland: Springer, 2018, pp. 405–421.

[18] 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.

[19] 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.

[20] J. McAuley, C. Targett, Q. Shi, and A. van den Hengel, "Image-based recommendations on styles and substitutes," in *Proc. 38th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2015, pp. 43–52.

[21] H. Zhao and X. Wang, "Bi-group Bayesian personalized ranking from implicit feedback," in *Proc. 2nd Int. Conf. Comput. Sci. Softw. Eng.*, 2019, pp. 452–461.

[22] X. Han, X. Song, J. Yin, Y. Wang, and L. Nie, "Prototype-guided attribute-wise interpretable scheme for clothing matching," in *Proc. 42nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2019, pp. 785–794.

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

[24] G. Cucurull, P. Taslakian, and D. Vázquez, "Context-aware visual compatibility prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12617–12626.

[25] 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.

[26] J. Ma, P. Cui, K. Kuang, X. Wang, and W. Zhu, "Disentangled graph convolutional networks," in *Proc. Int. Conf. Mach. Learn.*, vol. 97, 2019, pp. 4212–4221.

[27] J. Ma, C. Zhou, P. Cui, H. Yang, and W. Zhu, "Learning disentangled representations for recommendation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 5712–5723.

[28] Y. Zhang, G. Lai, M. Zhang, Y. Zhang, Y. Liu, and S. Ma, "Explicit factor models for explainable recommendation based on phrase-level sentiment analysis," in *Proc. 37th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2014, pp. 83–92.

[29] L. Ma, Q. Sun, S. Georgoulis, L. Van Gool, B. Schiele, and M. Fritz, "Disentangled person image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 99–108.

[30] H. Chen, Y. Deng, Y. Li, T.-Y. Hung, and G. Lin, "RGBD salient object detection via disentangled cross-modal fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 8407–8416, 2020.

[31] F. Yang, J. Chang, C. Tsai, and Y. F. Wang, "A multi-domain and multi-modal representation disentangler for cross-domain image manipulation and classification," *IEEE Trans. Image Process.*, vol. 29, pp. 2795–2807, 2020.

[32] L. Hu *et al.*, "Graph neural news recommendation with unsupervised preference disentanglement," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 4255–4264.

[33] X. Wang, H. Jin, A. Zhang, X. He, T. Xu, and T.-S. Chua, "Disentangled graph collaborative filtering," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2020, pp. 1001–1010.

[34] W. Guan, H. Wen, X. Song, C.-H. Yeh, X. Chang, and L. Nie, "Multimodal compatibility modeling via exploring the consistent and complementary correlations," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 2299–2307.

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

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

[37] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–14.

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

[39] Y. Cai *et al.*, "Exploiting spatial-temporal relationships for 3D pose estimation via graph convolutional networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2272–2281.

[40] X. Zhang, C. Xu, and D. Tao, "Context aware graph convolution for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14321–14330.

[41] L. Zhong, J. Cao, Q. Sheng, J. Guo, and Z. Wang, "Integrating semantic and structural information with graph convolutional network for controversy detection," in *Proc. ACL*, 2020, pp. 515–526.

[42] X. Wang, X. He, M. Wang, F. Feng, and T.-S. Chua, "Neural graph collaborative filtering," in *Proc. 42nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2019, pp. 165–174.

[43] Y. Wei, X. Wang, L. Nie, X. He, R. Hong, and T.-S. Chua, "MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1437–1445.

[44] X. Wang, Y. Ye, and A. Gupta, "Zero-shot recognition via semantic embeddings and knowledge graphs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6857–6866.

[45] T. Sun, Y. Wang, J. Yang, and X. Hu, "Convolution neural networks with two pathways for image style recognition," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4102–4113, Sep. 2017.

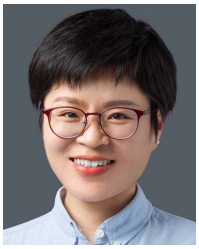
[46] Y. Hu, M. Liu, X. Su, Z. Gao, and L. Nie, "Video moment localization via deep cross-modal hashing," *IEEE Trans. Image Process.*, vol. 30, pp. 4667–4677, 2021.

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

[48] J. Li, J. Jia, and D. Xu, "Unsupervised representation learning of image-based plant disease with deep convolutional generative adversarial networks," in *Proc. 37th Chin. Control Conf. (CCC)*, Jul. 2018, pp. 9159–9163.
[49] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–12.
[50] 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.
[51] H. Zhang, Z.-J. Zha, Y. Yang, S. Yan, Y. Gao, and T.-S. Chua, "Attribute-augmented semantic hierarchy: Towards bridging semantic gap and intention gap in image retrieval," in *Proc. 21st ACM Int. Conf. Multimedia*, Oct. 2013, pp. 33–42.
[52] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[53] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.
[54] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–15.
[55] R. Tan, M. Vasileva, K. Saenko, and B. Plummer, "Learning similarity conditions without explicit supervision," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10372–10381.
[56] X. Song, S.-T. Fang, X. Chen, Y. Wei, Z. Zhao, and L. Nie, "Modality-oriented graph learning toward outfit compatibility modeling," *IEEE Trans. Multimedia*, early access, Dec. 9, 2021, doi: 10.1109/TMM.2021.3134164.
[57] X. Chen, X. Song, R. Ren, L. Zhu, Z. Cheng, and L. Nie, "Fine-grained privacy detection with graph-regularized hierarchical attentive representation learning," *ACM Trans. Inf. Syst.*, vol. 38, no. 4, pp. 37:1–37:26, 2020.
[58] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted Windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

Weili Guan (Member, IEEE) received the master's degree from the National University of Singapore. She is currently pursuing the Ph.D. degree with the Faculty of Information Technology, Monash University (Clayton Campus), Australia. After that, she joined Hewlett Packard Enterprise in Singapore as a Software Engineer and worked there for around five years. She has published many papers at the first-tier conferences and journals, like ACM MM, SIGIR, and IEEE TRANSACTIONS ON IMAGE PROCESSING. Her research interests are multimedia computing and information retrieval.

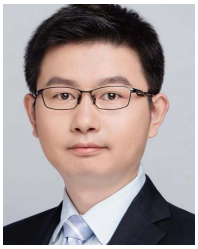
Haokun Wen received the B.E. degree from the Ocean University of China in 2019. He is currently pursuing the Graduate degree with the School of Computer Science and Technology, Shandong University. He has published several work in top conferences and journals, such as ACM SIGIR, ACM MM, and *ACM TOMM*. His research interests include multimedia computing and information retrieval.



Xuemeng Song (Senior Member, IEEE) received the B.E. degree from the University of Science and Technology of China in 2012 and the Ph.D. degree from the School of Computing, National University of Singapore in 2016. She is currently an Associate Professor with Shandong University, China. She has published several papers in the top venues, such as ACM SIGIR, MM, and *ACM TOIS*. She has served as a reviewer for many top conferences and journals. Her research interests include information retrieval and multimedia analysis.

Chun Wang received the B.E. degree from the School of Computer Science and Technology, Shandong University, Shandong, in 2019, where he is currently pursuing the Graduate degree with the Department of Computer Science and Technology. His research interests include natural language generation, information retrieval, and data mining.

Chung-Hsing Yeh (Senior Member, IEEE) received the B.S. degree in engineering and the M.S. degree in management science from the National Cheng Kung University, Taiwan, and the Ph.D. degree in information systems from Monash University, Australia. He is currently an Associate Professor with the Faculty of Information Technology, Monash University, and a Visiting Professor with the National Cheng Kung University. His current research interests include applied artificial intelligence, optimization modeling, image processing for design and decision analysis, and multicriteria decision analysis.



Xiaojun Chang (Senior Member, IEEE) is currently a Professor with the Faculty of Engineering and Information Technology, Australian Artificial Intelligence Institute, University of Technology Sydney. He is also the Director of the ReLER Laboratory. He is also an Honorary Professor with the School of Computing Technologies, RMIT University, Australia, where he was an Associate Professor with the School of Computing Technologies, before joining UTS. After graduation, he subsequently worked as a Postdoctoral Research Fellow with the School of Computer Science, Carnegie Mellon University, a Lecturer, and a Senior Lecturer with the Faculty of Information Technology, Monash University, Australia. He has focused his research on exploring multiple signals (visual, acoustic, textual) for automatic content analysis in unconstrained or surveillance videos. His team has won multiple prizes from international grand challenges which hosted competitive teams from MIT, University of Maryland, Facebook AI Research (FAIR) and Baidu VIS, and aim to advance visual understanding using deep learning. For example, he won the first place in the TrecVID 2019—Activity Extended Video (ActEV) challenge, which was held by National Institute of Standards and Technology, USA.

Liqiang Nie (Senior Member, IEEE) received the B.Eng. degree from Xi'an Jiaotong University and the Ph.D. degree from the National University of Singapore (NUS). He is currently the Dean of the Department of Computer Science and Technology, Harbin Institute of Technology (Shenzhen). He has coauthored/authored more than 100 papers and four books, received more than 15,000 Google Scholar citations. His research interests lie primarily in multimedia computing and information retrieval. He is an Associate Editor of IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, *ACM ToMM*, and *Information Science*. Meanwhile, he is the Regular Area Chair of ACM MM, NeurIPS, IJCAI, and AAAI. He is a member of ICME Steering Committee. He has received many awards, like ACM MM and SIGIR Best Paper Honorable Mention in 2019, SIGMM Rising Star in 2020, TR35 China 2020, DAMO Academy Young Fellow in 2020, and SIGIR Best Student Paper in 2021.