Multimodal Compatibility Modeling via Exploring the Consistent and Complementary Correlations

Weili Guan
Monash University
Clayton, Australia
honeyguan@gmail.com

Haokun Wen
Shandong University
Shandong, China
whenhaokun@gmail.com

Xuemeng Song*
Shandong University
Shandong, China
sxmustc@gmail.com

Chung-Hsing Yeh
Monash University
Clayton, Australia
chunghsing.yeh@monash.edu

Xiaojun Chang*
RMIT University
Melbourne, Australia
cxj273@gmail.com

Liqiang Nie
Shandong University
Shandong, China
nieliqiang@gmail.com

ABSTRACT

Existing methods towards outfit compatibility modeling seldom explicitly consider multimodal correlations. In this work, we explore the consistent and complementary correlations for better compatibility modeling. This is, however, non-trivial due to the following challenges: 1) how to separate and model these two kinds of correlations; 2) how to leverage the derived complementary cues to strengthen the text and vision-oriented representations of the given item; and 3) how to reinforce the compatibility modeling with text and vision-oriented representations. To address these challenges, we present a comprehensive multimodal outfit compatibility modeling scheme. It first nonlinearly projects each modality into separable consistent and complementary spaces via multi-layer perceptron, and then models the consistent and complementary correlations between two modalities by parallel and orthogonal regularizations. Thereafter, we strengthen the visual and textual representation of items with complementary information, and further induct both the text-oriented and vision-oriented outfit compatibility modeling. We ultimately employ the mutual learning strategy to reinforce the final performance of compatibility modeling. Extensive experiments demonstrate the superiority of our scheme.

CCS CONCEPTS

• **Information systems** → **Multimedia and multimodal retrieval**.

KEYWORDS

Compatibility Modeling, Multimodal Correlations, Consistency and Complementarity

*Xuemeng Song (sxmustc@gmail.com) and Xiaojun Chang (cxj273@gmail.com) are corresponding authors.

MM '21, October 20–24, 2021, Chengdu, Sichuan Province, China.

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475392

ACM Reference Format:
Weili Guan, Haokun Wen, Xuemeng Song, Chung-Hsing Yeh, Xiaojun Chang, and Liqiang Nie. 2021. Multimodal Compatibility Modeling via Exploring the Consistent and Complementary Correlations. In *Proceedings of the 29th ACM Int'l Conference on Multimedia (MM '21), Oct. 20–24, 2021, Chengdu, Sichuan Province, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3475392

1 INTRODUCTION

In modern society, clothing plays an increasingly important role in people's social life, as a compatible outfit can largely improve one's appearance. Nevertheless, not all people grow keen sense of aesthetics, and hence often find it difficult to make compatible outfits. To this end, outfit compatibility modeling, aiming to automatically evaluate the compatibility of a given outfit, has become an emerging research topic.

Existing methods on outfit compatibility modeling can be roughly classified into three groups: pair-wise, list-wise, and graph-wise modeling. The pair-wise modeling [9, 22] mainly justifies the compatibility between two items. It is, however, suboptimal when justifying outfits with more than two items, since it lacks a global view of the outfit. As to the list-wise one, it deems the outfit as a list of items in a predefined order and evaluates the outfit compatibility with neural networks, like Bi-directional Long Short-Term Memory (Bi-LSTM) [5, 10]. Notably, the underlying assumption is somehow inappropriate by treating a set of unstructured items as an ordered sequence. Moving one step forward, recent studies organize the outfit as an item graph and employ graph neural networks to fulfil the compatibility modeling task. Despite the significance of existing methods, they mainly focus on exploring the visual modality of the fashion items, and seldom investigate the item's textual aspect, i.e., the textual description. In fact, textual descriptions of fashion items usually contain the key features, which benefit the item representation learning. Although some studies have attempted to incorporate the textual modality, they simply adopt the early/late fusion or consistency regularization to boost the performance. Nevertheless, the correlations among multimodalities are complex and sophisticated, which are not clearly separated and explicitly modeled yet.

In this work, we work towards outfit compatibility modeling via exploiting the multimodal correlations. It is, however, non-trivial considering the following facts. 1) In a sense, the visual and textual modalities characterize the same item, and thus should share certain

Figure 1: Illustration of the consistent and complementary correlations between the visual and textual modalities. In (a), both the text and image reflect the color (Dark Blue) and category (Shorts) of the item. In (b), the text reveals the item's material (Leather) and brand (New Ace) that is hardly derived visually, but fails to describe the pattern (Stripe) position.

consistency. As shown in Figure 1 (a), both the visual and textual modalities deliver the item's features of "color" and "category". Meanwhile, the user-generated text may provide complementary information to the visual image, like the item brand "New Ace" and material "Leather" in Figure 1 (b), yet certain features are hard to be described by textual sentences but easy to be visualized by the image, like the stripe position of the item in Figure 1 (b). Consistent and complementary contents are often mixed in each modality and may be nonlinearly separable. Therefore, how to explicitly separate and model them poses one challenge. 2) How to leverage the correlation modeling results to strengthen the text and vision-oriented representation of the given item forms another challenge. And 3) outfit compatibility modeling can be derived separately from vision or text-oriented representations, which indeed characterizes the item from different angles. We argue that these two kinds of modeling share certain common knowledge on the outfit compatibility evaluation and are capable of reinforcing each other. How to get the two kinds of modeling mutually enhanced and thus boost the final compatibility modeling result constitutes the last challenge.

To address the aforementioned challenges, we devise a comprehensive **M**ulti**M**odal **O**utfit **C**ompatibility **M**odeling scheme, MM-OCM for short. As shown in Figure 2, MM-OCM consists of four components: a) multimodal feature extraction, b) multimodal correlation modeling, c) compatibility modeling, and d) mutual learning. The first component extracts the textual and visual features of the given item via two separate Convolution Neural Networks (CNNs) [34] and Long Short-Term Memory (LSTM) networks [13], respectively. The reason of introducing two separate feature extractors is to facilitate the later mutual learning. As to the second component, it aims to separate and model the consistent and complementary correlations. Considering the fact that these two kinds of correlations may be not separable in the original visual and textual feature spaces, we therefore employ the multi-layer perceptrons to nonlinearly project the image/text feature into the consistent and complementary space, where the multimodal consistency and complementarity can be captured, respectively. In the third component, we incorporate the disengaged complementary content in the textual (visual) modality to complement the visual (textual) feature embedding and obtain the text (vision)-oriented representation. Thereafter, we build two independent graph convolutional networks to model outfit compatibility, namely text-oriented compatibility modeling (T-OCM) and vision-oriented compatibility modeling (V-OCM). Ultimately, the fourth component targets at mutually transferring knowledge from one compatibility modeling to guide the other one. Once MM-OCM converges, we average the compatibility scores predicted by T-OCM and V-OCM as the final result. Extensive experiments on the real-world dataset demonstrate the superiority of our MM-OCM scheme as compared to several state-of-the-art baselines. As a byproduct, we released the codes to benefit other researchers[1].

The contributions of this work are threefold:

- To the best of our knowledge, we are the first to fulfill the outfit compatibility modeling via clearly separating and explicitly modeling the consistent and complementary correlations among multiple modalities.
- We propose to strengthen the text (vision)-oriented representation of the given item by incorporating the complementary information embedded in the visual (textual) modality. Based upon these two kinds of representations, we build two parallel networks to model the outfit compatibility.
- We introduce the mutual learning strategy in the context of compatibility modeling, which reinforces each other via knowledge sharing.

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

2 RELATED WORK

Our work is related to fashion compatibility modeling and deep mutual learning.

2.1 Fashion Compatibility Modeling

Existing methods on fashion compatibility modeling can be roughly grouped into three categories: pair-wise methods [9, 21, 25, 26, 33, 38, 39], list-wise methods [10], and graph-wise methods [2, 3]. The first category focuses on studying the compatibility between two items. For example, McAuley et al. [26] used the linear transformation to map items into a latent space, where the compatibility relation between items can be measured. Following that, Song et al. [33] proposed a multimodal compatibility modeling scheme, where neural networks are used to model the compatibility between fashion items with the Bayesian Personalized Ranking (BPR) [32] optimization. Later, Vasileva et al. [36] studies the compatibility for the outfit with multiple fashion items based on the pairwise modeling, where the item category information is additionally considered.

One key limitation of this category is that it lacks a global view of the outfit and can hardly generate the optimal solution. As to the second category, it regards the outfit as a sequence of items in a fixed pre-predefined order. For example, Han et al. [10] employed the Bi-LSTM network to uncover the outfit compatibility. It is worth noting

[1] https://site2750.wixsite.com/mmocm.

that the underlying assumption used by the list-wise methods, i.e., the outfit can be represented as a sequence of ordered items, is questionable. Approaches in the third category model each outfit as an item graph and turn to graph neural networks [7, 19, 30] to fulfil the outfit compatibility modeling task. For example, Cui et al. [3] proposed Node-wise graph Neural Networks (NGNN) to promote the item representation learning. In addition, Cucurull et al. [2] addressed the problem using a graph neural network that learns to generate product embeddings conditioned on their context.

Although these studies have achieved significant success, they mainly focus on either simply exploring the visual modality of the outfit, or considering both the visual and textual modalities, while overlooking the sophisticated multimodal correlations.

2.2 Deep Mutual Learning

The idea of deep mutual learning is developed from the knowledge distillation, which was first introduced by Hinton et al. [12] for transferring the knowledge from a large cumbersome model to a small one, so as to improve the model portability. In particular, Hu et al. [16] designed an iterative teacher-student knowledge distillation approach, where the teacher network grasps certain knowledge, while the student one iteratively mimics the teacher's solution to a certain problem in order to improve its own performance. After that, the teacher-student knowledge distillation scheme attracts lots of attention [8, 43]. However, in many cases, it might be too difficult to obtain a teacher network with the clear domain knowledge. Accordingly, Zhang et al. [44] proposed a deep mutual learning method for the classification task, where there is no explicit static teacher but an ensemble of student learning collaboratively throughout the training process. Thereafter, many researchers have investigated the deep mutual learning in various domains, such as person re-identification [6, 40], image retrieval [37], and deep metric learning [31]. Despite the value of mutual learning in these fields, its potential in outfit compatibility modeling has been largely unexplored, which is the major concern of this work.

3 METHODOLOGY

In this section, we first formulate the research problem and then detail the proposed MM-OCM scheme.

3.1 Problem Formulation

We deem the outfit compatibility modeling task as a binary classification problem. Suppose that we have a training set Ω composed of N outfits, i.e, $\Omega = \{(O_i, y_i)\,|i = 1, \cdots, N\}$, where O_i is the i-th outfit, and y_i denotes the ground truth label. We set $y_i = 1$ if the outfit O_i is compatible, and $y_i = 0$ otherwise. Given an arbitrary outfit O, it can be represented as a set of fashion items, i.e., $O = \{o_1, o_2, \cdots, o_m\}$, where o_i is the i-th item, associated with a visual image v_i and a textual description t_i. The symbol m is a variable for different outfits, considering that the number of items in outfits is not fixed. Based on these training samples, we target at learning an outfit compatibility model $\mathcal{F}$ that is able to judge whether the given outfit O is compatible or not,

$$s = \mathcal{F}\left(\{(v_i, t_i)\}_{i=1}^{m} | \mathbf{\Theta}\right), \tag{1}$$

where $\mathbf{\Theta}$ is a set of to-be-learned parameters of our model, and s denotes the probability the given outfit is compatible.

3.2 MM-OCM

Based upon the research problem and notations, we present the comprehensive **M**ulti**M**odal **O**utfit **C**ompatibility **M**odeling scheme, MM-OCM. As shown in Figure 2, it consists of four key components: (a) multimodal feature extraction, (b) multimodal correlation modeling, (c) compatibility modeling, and (d) mutual learning.

3.2.1 Multimodal Feature Extraction. We first introduce the visual and textual feature extraction.

Visual Feature Extraction. To extract visual features, we utilize the CNNs, which have shown compelling success in many computer vision tasks [11, 14, 15, 23, 24, 28, 42]. As to facilitate the mutual enhancement between the T-OCM and the V-OCM, which are alternatively optimized, we employ two separate CNNs to extract the visual features. Specifically, given the outfit O, the visual feature of the i-th item in the outfit can be obtained as follows,

$$\begin{cases} \hat{\mathbf{v}}_i = \mathrm{CNN}_1\left(v_i\right), \\ \tilde{\mathbf{v}}_i = \mathrm{CNN}_2\left(v_i\right), \end{cases} \tag{2}$$

where $\hat{\mathbf{v}}_i \in \mathbb{R}^{d_v}$ and $\tilde{\mathbf{v}}_i \in \mathbb{R}^{d_v}$ refer to the visual features to be processed by the following T-OCM and V-OCM, respectively. The symbol d_v is the dimension of the extracted visual feature embedding. CNN_1 and CNN_2 denotes the corresponding CNNs for the T-OCM and V-OCM, respectively.

Textual Feature Extraction. Due to its prominent performance in textual representation learning [1, 17, 27, 29, 41], we adopt LSTM to extract the textual feature of the given item[2]. Similar to the visual feature extraction, we also use two separate LSTMs, i.e., LSTM_1 and LSTM_2, to obtain the textual features for T-OCM and V-OCM, respectively. Formally, we have

$$\begin{cases} \hat{\mathbf{t}}_i = \mathrm{LSTM}_1\left(t_i\right), \\ \tilde{\mathbf{t}}_i = \mathrm{LSTM}_2\left(t_i\right), \end{cases} \tag{3}$$

where $\hat{\mathbf{t}}_i \in \mathbb{R}^{d_t}$ and $\tilde{\mathbf{t}}_i \in \mathbb{R}^{d_t}$ refer to the text features for the following T-OCM and V-OCM, respectively. d_t is the feature dimension. To facilitate the multimodal fusion, we set $d_t = d_v = d$ in this work.

3.2.2 Multimodal Correlation Modeling. As illustrated in Figure 1, we argue that the visual image and textual description may possess certain consistency and complementarity information. Inspired by this, instead of unreasonably fusing the general multimodal features, we propose to clearly separate and explicitly model the consistent and complementary contents of each modality, whereby we expect the consistent content of a modality is able to capture the alignment information between two modalities, and the complementary one of a modality is able to encode the supplement information to the other modality.

In particular, we first introduce two MLPs to separate the consistent and complementary parts of each modality, respectively. Mathematically, we have

$$\begin{cases} \hat{\mathbf{v}}_i^s = \mathrm{MLP}_v^s\left(\hat{\mathbf{v}}_i\right), \hat{\mathbf{t}}_i^s = \mathrm{MLP}_t^s\left(\hat{\mathbf{t}}_i\right), \\ \hat{\mathbf{v}}_i^p = \mathrm{MLP}_v^p\left(\hat{\mathbf{v}}_i\right), \hat{\mathbf{t}}_i^p = \mathrm{MLP}_t^p\left(\hat{\mathbf{t}}_i\right), \end{cases} \tag{4}$$

[2]Before fed into the LSTM, the text is first tokenized into standard vocabularies.

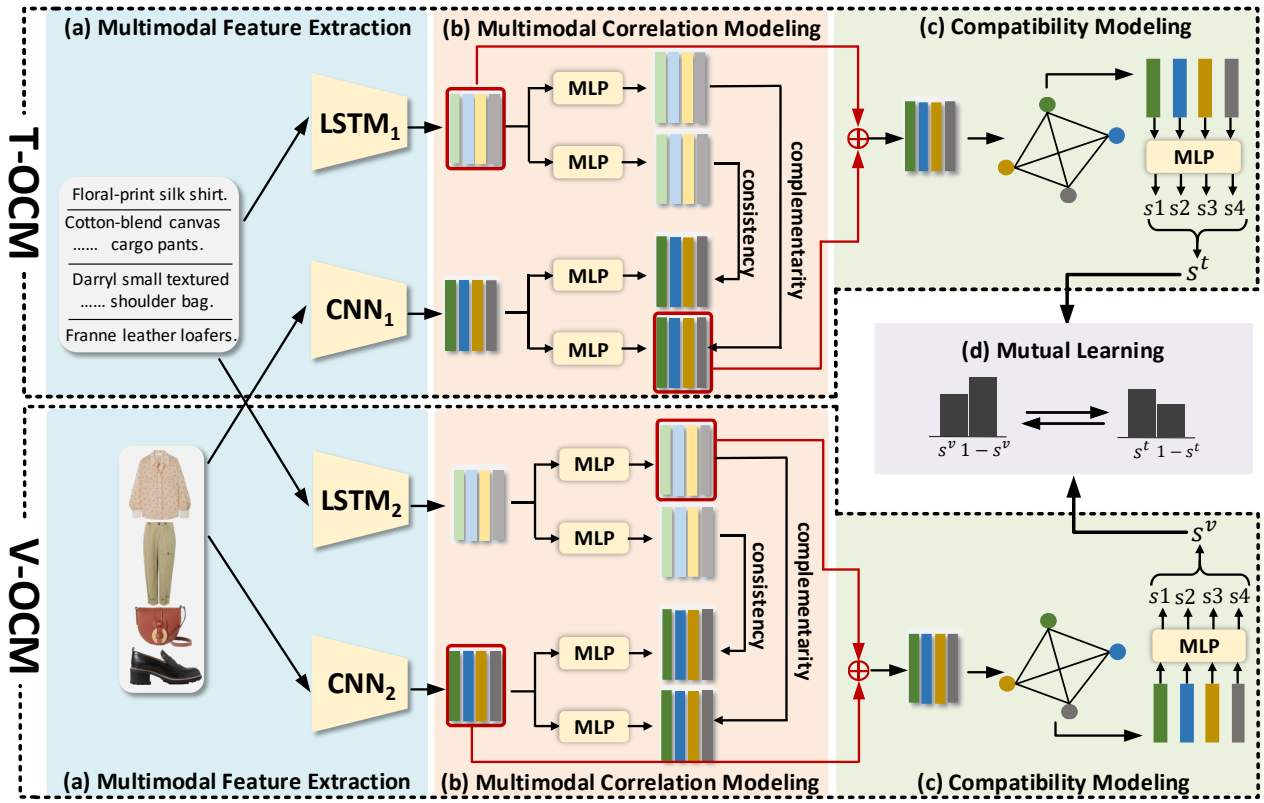


Figure 2: Illustration of the proposed MM-OCM scheme. It consists of four key components: (a) multimodal feature extraction, (b) multimodal correlation modeling, (c) compatibility modeling, and (d) mutual learning.

where $\hat{\mathbf{v}}_i^s$ and $\hat{\mathbf{v}}_i^p$ respectively denote the consistent and complementary representation of the visual modality, and $\hat{\mathbf{t}}_i^s$ and $\hat{\mathbf{t}}_i^p$ denote that of the textual modality. It is noteworthy that the consistent and complementary parts are probably inseparable within the original low dimensional space. After non-liner mapping via MLPs, we are capable of projecting them into a high dimensional space, whereby the consistent and complementary parts are distinguishable.

We then argue that the consistent representations of the two modalities are parallel, and the complementary representations are orthogonal. Accordingly, to regulate the consistent and complementary representations, we use the following objective functions:

$$\begin{cases} \mathcal{L}_s = \sum_{i=1}^{m} \{\cos(\hat{\mathbf{v}}_i^s, \hat{\mathbf{t}}_i^s)^2 + \cos(\tilde{\mathbf{v}}_i^s, \tilde{\mathbf{t}}_i^s)^2\}, \\ \mathcal{L}_p = \sum_{i=1}^{m} \{[\cos(\hat{\mathbf{v}}_i^p, \hat{\mathbf{t}}_i^p) - 1]^2 + [\cos(\tilde{\mathbf{v}}_i^p, \tilde{\mathbf{t}}_i^p) - 1]^2\}. \end{cases} \tag{5}$$

where $\mathcal{L}_s$ and $\mathcal{L}_p$ refer to the consistent and complementary regularizations, respectively.

3.2.3 Compatibility Modeling. We here first introduce the text/vision-oriented representation learning for each item, and we then present the text/vision-oriented compatibility modeling.

Text/Vision-oriented Representation Learning. Based upon the component of multimodal correlation modeling, we are able to derive the complementary cues of the textual (visual) modality from the visual (textual) one. Distinguished from the consistent parts that are shared between modalities, complementarity means exclusive and supplement information. Inspired by this, to learn comprehensive item representations and hence boost the outfit compatibility modeling performance, we introduce two multimodal fusion strategies: text-oriented multimodal fusion and vision-oriented multimodal fusion. As to the first one, we take the textual feature extracted by LSTM as the basis and additionally incorporate the complementary representation of the visual modality. By contrast, in the latter fusion strategy, we strengthen the visual feature extracted by CNN with the complementary representation of the textual modality. Specifically, based upon the consistent and complementary representation of each modality, we can derive the final item representations from different fusion schemes, which are achieved as follows,

$$\begin{cases} \hat{\mathbf{o}}_i = \hat{\mathbf{t}}_i + \hat{\mathbf{v}}_i^p, \\ \tilde{\mathbf{o}}_i = \tilde{\mathbf{v}}_i + \tilde{\mathbf{t}}_i^p, \end{cases} \tag{6}$$

where $\hat{\mathbf{o}}_i$ and $\tilde{\mathbf{o}}_i$ denote the final item representation based on the text-oriented multimodal fusion and vision-oriented multimodal fusion, respectively.

Text/Vision-oriented Compatibility Modeling. Similar to previous studies, we employ Graph Convolutional Network (GCN) to flexibly model the compatibility of the outfit with variable number of items. In particular, we adopt two GCNs, one for the T-OCM, while the other for the V-OCM. Regarding the limited space, we take the

T-OCM as an example, since the V-OCM can be derived in the same way. In particular, for each outfit O composed of m fashion items, we first construct an indirected graph $G = (\mathcal{E}, \mathcal{R})$. $\mathcal{E} = \{o_i\}_{i=1}^{m}$ is the set of nodes, corresponding to the items of the given outfit O. Meanwhile, $\mathcal{R} = \{(o_i, o_j) | i, j \in [1, \cdots, m]\}$ stands for the set of edges. In this work, for each pair of items o_i and o_j in the outfit, we introduce an edge. During learning, each node o_i is associated with a hidden state vector $\mathbf{h}_i$, which keeps dynamically updated to fulfil the information propagation over the graph. For T-OCM, we initialize the hidden state vector for the i-th node based on the text-oriented representation of the i-th item, namely, $\mathbf{h}_i = \hat{\mathbf{o}}_i$.

The information propagation from the item o_j to item o_i is defined as follows:

$$\mathbf{m}_{j \to i} = \phi[\mathbf{W}_{pp}(\mathbf{h}_i \odot \mathbf{h}_j) + \boldsymbol{b}_{pp}], \tag{7}$$

where $\mathbf{W}_{pp} \in \mathbb{R}^{d \times d}$ and $\mathbf{b}_{pp} \in \mathbb{R}^d$ denote the weight matrix and bias vector to be learned; $\phi(\cdot)$ is a nonlinear activation function, which is set as LeakyReLU; $\mathbf{h}_i \odot \mathbf{h}_j$ accounts for the interaction between the fashion item o_i and o_j; $\odot$ is the element-wise product operation. By summarizing the information propagated from all neighbours, the hidden state vector corresponding to the item o_i can be updated as follows,

$$\mathbf{h}_i^* = \phi\left(\mathbf{W}_0 \mathbf{h}_i + \mathbf{b}_0\right) + \sum_{o_j \in \mathcal{N}_i} \mathbf{m}_{j \to i}, \tag{8}$$

where $\mathbf{W}_0 \in \mathbb{R}^{d \times d}$ and $\mathbf{b}_0 \in \mathbb{R}^d$ denote the weight matrix and bias vector to be learned; $\mathcal{N}_i$ stands for the set of neighbour nodes of the node o_i and $\mathbf{h}_i^* \in \mathbb{R}^d$ is the updated hidden representation of the item o_i.

We ultimately feed the updated item representation to a MLP, consisting of two fully-connected layers, to derive its probability of being a compatible outfit as follows,

$$\begin{cases} s_t^i = \mathbf{W}_2\left[\psi\left(\mathbf{W}_1 \mathbf{h}_i^* + \mathbf{b}_1\right)\right] + \mathbf{b}_2, \\ s_t = \sigma\Big(\frac{1}{m}\sum_{i=1}^{m} s_t^i\Big), \end{cases} \tag{9}$$

where $\mathbf{W}_1$, $\mathbf{b}_1$, $\mathbf{W}_2$, and $\mathbf{b}_2$ are the to be learned layer parameters. $\psi(\cdot)$ refers to the Relu active function, and $\sigma(\cdot)$ denotes the Sigmoid function to ensure the compatibility probability falling in the range of $[0, 1]$. Notably, in the same way, we can derive the compatible probability of the outfit by V-OCM, which is termed as s_v.

3.2.4 Mutual Learning. In a sense, no matter the text-oriented item representation or the vision-oriented one, i.e., $\hat{\mathbf{o}}_i$ and $\tilde{\mathbf{o}}_i$, both of them fuse the multimodal data of an item. Therefore, the information encoded by these two representations should be largely aligned, and hence the corresponding outfit compatibility modeling should yield similar outputs. Meanwhile, since they emphasize the different aspects of the item and therefore may complement each other from a global view. Therefore, the knowledge learned by one compatibility modeling could be able to guide the other one. Inspired by this, we turn to the deep mutual learning knowledge distillation scheme to regularize these two compatibility modeling results, making them mutually reinforced.

Unlike the traditional teacher-student knowledge distillation network, mutual learning replaces the one way knowledge transferring from the static pre-trained teacher to the student with the mutual knowledge distillation. In particular, an ensemble of student networks are employed to learn collaboratively. In our context, the T-OCM and the V-OTM can be treated as two student networks, and optimized alternatively. Namely, in each iteration, we only train one student network, while keeping the other fixed, which temporarily is acted as the teacher.

Algorithm 1 The Training Procedure of Our MM-OCM.

Input: Training set Ω, hyper-parameters λ, η, and μ.
Output: Parameters $\mathbf{\Theta_1}$ in the T-OCM, and parameters $\mathbf{\Theta_2}$ in the V-OCM.

```
Initialize neural network parameters Θ1 and Θ2.
repeat
    Sample minibatch from Ω.
    Update the parameters Θ1 according to L_t in Eqn.(12).
    Update the parameters Θ2 according to L_v in Eqn.(12).
until Convergence
```

We cast the compatibility modeling as a binary classification task, and adopt the widely-used cross-entropy loss for both T-OCM and V-OCM. Accordingly, we have the objective functions,

$$\begin{cases} \mathcal{L}_{ce}^t = -ylog(s_t) - (1-y)log(1-s_t), \\ \mathcal{L}_{ce}^v = -ylog(s_v) - (1-y)log(1-s_v), \end{cases} \tag{10}$$

where y refers to the ground truth label of the outfit O. $\mathcal{L}_{ce}^t$ and $\mathcal{L}_{ce}^v$ are the objective functions for the T-OCM and V-OCM, respectively.

To encourage the two student networks to learn from each other, we adopt the Kullback Leibler (KL) divergence loss function to penalize the distance between the evaluation results of the T-OCM and V-OCM as follows,

$$\begin{cases} \mathcal{L}^{v->t} = s_v log\frac{s_v}{s_t} + (1-s_v)log\frac{(1-s_v)}{(1-s_t)}, \\ \mathcal{L}^{t->v} = s_t log\frac{s_t}{s_v} + (1-s_t)log\frac{(1-s_t)}{(1-s_v)}. \end{cases} \tag{11}$$

Notably, we use $\mathcal{L}_{v->t}$ for training T-OCM, and $\mathcal{L}_{t->v}$ for training V-OCM. Finally, we have

$$\begin{cases} \mathcal{L}_t = \mathcal{L}_{ce}^t + \lambda\mathcal{L}^{v->t} + \eta\mathcal{L}_s + \mu\mathcal{L}_p, \\ \mathcal{L}_v = \mathcal{L}_{ce}^v + \lambda\mathcal{L}^{t->v} + \eta\mathcal{L}_s + \mu\mathcal{L}_p, \end{cases} \tag{12}$$

where λ, η, and μ are trade-off hyper-parameters. $\mathcal{L}_t$ and $\mathcal{L}_v$ are the final loss functions for the T-OCM and V-OCM, respectively. In a sense, each compatibility modeling component (i.e., T-COM or V-OCM) not only learns to correctly predict the true label of the training instances, but also learns to mimic the output of the other compatibility modeling component, where the consistent and complementary regularizations are also jointly satisfied. Notably, although both $\mathcal{L}_t$ and $\mathcal{L}_v$ have the consistent and complementary regularizations, i.e., $\mathcal{L}_s$ and $\mathcal{L}_p$, the parameters to be optimized for them are distinguished, where the regularizations in $\mathcal{L}_t$ target at optimizing the T-OCM, while that in $\mathcal{L}_s$ aim to learn parameters of V-OCM. Algorithm 1 summarizes the alternative optimization procedure of our MM-OCM. Once our MM-OCM is well-trained, we will take the average of the predicted compatibility probabilities of the V-OCM and T-OCM as the final compatibility probability of the outfit.

Table 1: Performance comparison between our proposed MM-OCM scheme and other baselines over two datasets. The baselines were re-trained by their released codes. The best results are in boldface, and the second best are underlined.

Method	Polyvore Outfits		Polyvore Outfits-D	
	Compat. AUC	FITB Accuracy	Compat. AUC	FITB Accuracy
Bi-LSTM (Han et al. 2017) [10]	0.68	42.20%	0.65	40.10%
Type-aware (Vasileva et al. 2018) [36]	<u>0.87</u>	<u>56.60%</u>	0.78	47.30%
SCE-NET (Tan et al. 2019) [35]	0.83	52.80%	<u>0.82</u>	<u>52.10%</u>
NGNN (Cui et al. 2019) [3]	0.75	53.02%	0.68	42.49%
Context-aware (Cucurull et al. 2019) [2]	0.81	55.63%	0.77	50.34%
HFGN (Li et al. 2020) [20]	0.84	49.90%	0.70	39.03%
MM-OCM	**0.93**	**63.40%**	**0.88**	**58.02%**

4 EXPERIMENT

In this section, we conducted experiments over two real-world datasets by answering the following research questions.

- **RQ1**: Does MM-OCM outperform state-of-the-art baselines?
- **RQ2**: How does each module affect MM-OCM?
- **RQ3**: How is the qualitative performance of MM-OCM?

4.1 Experimental Settings

4.1.1 Datasets. In this work, we chose the Polyvore dataset constructed by Vasileva et al. in [36], which contains more unique items and outfits than other public datasets. Meanwhile, it also provides textual descriptions of items, which enables our multi-modal compatibility modeling. According to the dataset split protocal, this dataset provides the following two versions. 1) Polyvore Outfits: It consists of 53,306 outfits for training, 5,000 outfits for validation, and 10,000 outfits for testing. In this dataset, an item may simultaneously appear in both the training, validation and testing phases. 2) Polyvore Outfits-D: This dataset contains 16,995 outfits, 15,145 outfits, and 15,145 outfits for training, validation, and testing, respectively. Compared with Polyvore Outfits, Polyvore Outfits-D is a more challenging version, since there is no item appears in more than one split.

4.1.2 Evaluation Tasks. Similar to [2, 3, 10, 35, 36], we justified our proposed MM-OCM scheme with two specific tasks: Outfit Compatibility Estimation and Fill-in-the-blank (FITB). The former task is to classify whether a given outfit is compatible, where a threshold of 0.5 is introduced to derive the class label for each outfit sample. We adopted the area under a receiver operating characteristic curve (AUC) as the corresponding evaluation metric. The latter task is to choose one item from a set of candidates with one positive item and three negative items, for a given incomplete outfit. For this task, we applied accuracy as the evaluation metric. Notably, Polyvore dataset provides the corresponding data split for these two tasks, which is directly applied in this paper.

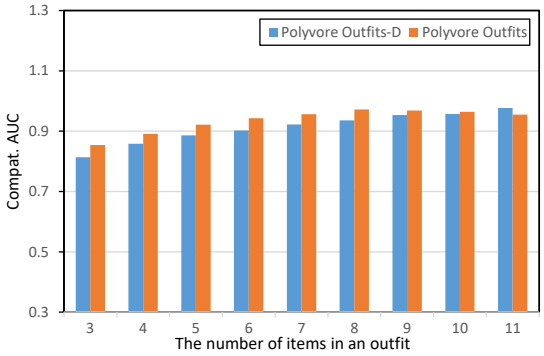
4.1.3 Implementation Details. For the image encoder, we selected the ImageNet [4] pre-trained ResNet18 [11] as the backbone, and modified the last layer to make the output feature dimension as 256. Regarding the text encoder, we set the word embedding size to 512, and the dimension of the hidden layer in LSTM to 256. We alternatively trained the T-OCM and V-OCM by the Adam optimizer [18] with a fixed learning rate of 0.0001, and the batch size of 16. The trade-off hyper-parameters in Eqn.(11) are set as $\lambda = \eta = \mu = 1$. In particular, we launched 10-fold cross validation for each experiment, and reported the average results. All the experiments are implemented by PyTorch over a server equipped with 4 NVIDIA TITAN Xp GPUs, and the random seeds are fixed for the reproducibility.

4.2 On Model Comparison (RQ1)

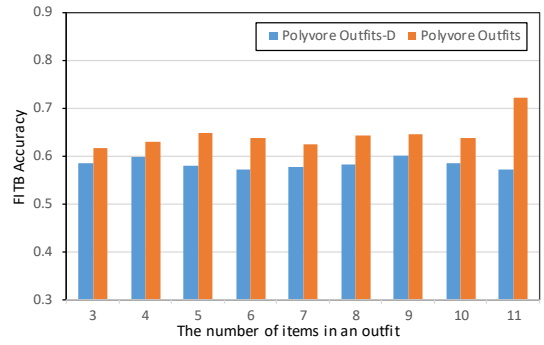
To validate the effectiveness of our proposed scheme, we chose the following baselines for comparison.

- **Bi-LSTM** [10] takes the items in an outfit as a sequence ordered by the item category and fulfils the fashion compatibility modeling with Bi-LSTM. For fair comparison, we only utilized the visual information.
- **Type-aware** [36] designs type-specific embedding spa-ces according to the item category. It also utilizes the textual information by the common-used visual-semantic loss.
- **SCE-NET** [35] is a pair-wise method, which utilizes multiple similarity condition masks to embed the item features into different semantic subspaces. This method also takes into account the textual information.
- **NGNN** [3] employs the GNN to tackle the compatibility modeling task, where the node is updated by a gate mechanism. For multimodal features, NGNN designs two graph channels, and the final compatibility score is derived in a weighted average manner.
- **Context-aware** [2] regards fashion compatibility modeling as an edge prediction problem, where a graph auto-encoder framework is introduced. Only visual features are employed.
- **HFGN** [20] shares the same spirits with NGNN, and builds a category-oriented graph. Additionally, it introduces a R-view attention map and a R-view score map to compute the compatibility score. This baseline only uses the visual features.

Table 1 shows the performance comparison among different methods on two datasets under two tasks. From this table, we had the following observations: 1) Among all the baselines, Bi-LSTM performs the worst, which suggests that modeling the outfit as an ordered list of items is not reasonable. 2) The methods that use multimodal features gain more promising results (e.g., Type-aware on Polyvore Outfits and SCE-NET on Polyvore Outfits-D) compared with those only utilize the visual ones (i.e., HFGN and Context-aware) . This implies that taking both visual and textual modalities into account is rewarding in the outfit compatibility



(a) Outfit Compatibility Estimation



(b) Fill-in-the-blank (FITB)

Figure 3: Performance of our MM-OCM in two tasks for outfits with different numbers of items.

modeling task. And 3) MM-OCM consistently surpasses all baseline methods on the two datasets under both tasks. This indicates the advantage of our scheme that utilizes the multimodal correlation modeling and mutual learning in the context of outfit compatibility modeling. Notably, we performed the ten-fold t-test between our proposed scheme and each of the baselines. We observed that all the p-values are much smaller than 0.05, and we hence concluded that the MM-OCM is significantly better than the baselines.

To gain deeper insights, we further checked the performance of our MM-OCM in two tasks for outfits with different numbers of items. In particular, we only reported the results for the number of items ranging from 3 to 11, where the others are too small in the test set to be displayed. As can be seen from Figure 3, our MM-OCM is not sensitive to the number of items in the outfit, which indicates that our method has the capacity of handling the compatibility modeling for outfits with variable items.

4.3 On Ablation Study (RQ2)

To verify the importance of each component in our model, we also compared MM-OCM with the following derivatives.

- **w/o Correlation**: To explore the effect of the multimodal correlation modeling, we removed this component by setting $\hat{\mathbf{o}}_i = \hat{\mathbf{t}}_i$ and $\tilde{\mathbf{o}}_i = \tilde{\mathbf{v}}_i$ in Eqn.(6).
- **w/o Mutual**: To study the effect of the mutual learning component, we removed the knowledge distillation between the T-OCM and V-OCM by setting $\lambda = 0$.

Table 2: Ablation study of our proposed MM-OCM scheme on two datasets. The best results are in boldface.

Method	Polyvore Outfits		Polyvore Outfits-D	
	Compat. AUC	FITB Accuracy	Compat. AUC	FITB Accuracy
w/o Correlation	0.91	52.91%	0.87	54.47%
w/o Mutual	0.92	60.80%	0.86	55.62%
Image_Only	0.90	57.80%	0.85	52.85%
Text_Only	0.79	42.28%	0.74	35.45%
Concat_Directly	0.91	58.67%	0.79	49.73%
Concat_Ensemble	0.91	59.31%	0.80	52.04%
MM-OCM	**0.93**	**63.40%**	**0.88**	**58.02%**

- **Image_Only** and **Text_Only**: The two derivatives are set to verify the importance of visual and textual information. Specifically, for the Image_Only, we removed T-OCM by setting $\tilde{\mathbf{o}}_i = \tilde{\mathbf{v}}_i$, while for the Text_Only, V-OCM was removed by seting $\hat{\mathbf{o}}_i = \hat{\mathbf{t}}_i$.
- **Concat_Directly**: To gain more insights into our manners of utilizing visual and textual information, we directly concatenated the visual and textual features of each item and fed them to a MLP to get $\hat{\mathbf{o}}_i$. Accordingly, the correlation modeling and mutual learning are simultaneously removed.
- **Concat_Ensemble**: To further investigate whether the improvement of MM-OCM is achieved by the ensemble of more models, we also derived Concat_Ensemble from Concat_Directly by introducing two LSTM and two CNN encoders.

Table 2 shows the ablation results of our MM-OCM. From this table, we gained the following observations. 1) w/o Correlation performs worse than our MM-OCM, which proves the effectiveness of the proposed multimodal consistency and complementarity modeling. 2) MM-OCM surpasses w/o Mutual, indicating that the mutual learning component is indeed helpful for integrating the T-OCM and V-OCM by transferring knowledge between the two modules. 3) Both Image_Only and Text_Only are inferior to MM-OCM, which suggests that it is essential to consider both visual and textual information to gain better outfit compatibility modeling effects. In addition, Image_Only outperforms Text_Only remarkably, which reflects that the image contains more useful information than the text, which is in consensus with the saying that "A picture is worth a thousand words". 4) Compared to our MM-OCM, Concat_Directly also delivers worse performance, implying that simply fusing visual and textual features is insufficient to explore the intrinsic correlation of the two modalities. This further verifies the superiority of our strategy that models the multimodal correlation and devises two schemes of multimodal fusion. Furthermore, it can be observed that on the more challenging dataset Polyvore Outfits-D, the results of Concat_Directly are better than that of Text_Only, but worse than that of Image_Only. This phenomenon indicates that an inappropriate multimodal fusion method will be less effective than only utilizing the more informative modality. And 5) Concat_Ensemble has limited improvement over Concat_Directly, and performs worse than our MM-OCM, which demonstrates that the superiority of our model is not mainly induced by the ensemble of more models.

(a) Outfit Compatibility Estimation **(b) Fill-in-the-blank (FITB)**

Figure 4: Qualitative results of MM-OCM on (a) outfit compatibility estimation, and (b) fill-in-the-blank.

4.4 On Case Study (RQ3)

To gain a thorough understanding of our model, we also conducted qualitative evaluation of our method.

Figure 4 intuitively shows several testing examples on the outfit compatibility estimation and fill-in-the-blank tasks. From Figure 4 (a), we observed that, for the example in the first row, which contains items with the consistent black color and elegant style, our MM-OCM is able to assign it with a high compatible probability. For the example in the middle row, the four items share the compatible tone and material, and thus obtain the high compatibility score. As for the outfit in the last row with obvious incompatible colors, e.g., green does not go well with red, our MM-OCM gives a low compatibility score. From Figure 4 (b), we can see that our method has the ability to choose the most suitable item from the candidate set to form a compatible outfit. For the example in the first row, the outfit lacks a pair of shoes and our MM-OCM correctly selects the first item by attributing a high compatibility score. As can be seen, the selected item matches well with other items in the query. As to the example in the second row, although our method chooses the correct answer (item D), it also gives a high compatibility score to the item *B*, since these two items are both dark jackets of the same style. This reconfirms the compatibility modeling capabilities of our model.

5 CONCLUSION AND FUTURE WORK

In this work, we solve the outfit compatibility modeling problem by exploring the multimodal correlations. In particular, we clearly separate and explicitly model the consistent and complimentary relations between the visual and textual modalities. This is accomplished by nonlinearly projecting the consistent and complementary contents into the separable spaces, whereby they are respectively formulated by parallel and orthogonal regularizers. We then apply the complementary information to strengthen the vision- and text-oriented representations. Based upon these two kinds of representations, two compatibility modeling brunches are derived and reinforced by mutual learning via knowledge transferring. Extensive experiments over two benchmark datasets have verified the effectiveness of our proposed MM-OCM scheme, as compared with several state-of-the-art baselines. In future, we plan to apply our multimodal correlation modeling methods to enhance other research problems, such as multiple social network analysis.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China, No.:U1936203; X. Chang gratefully acknowledge the support of Australian Research Council (ARC) Discovery Early Career Researcher Award (DECRA) under grant no. DE190100626.

REFERENCES

[1] Yuxiao Chen, Jianbo Yuan, Quanzeng You, and Jiebo Luo. 2018. Twitter Sentiment Analysis via Bi-sense Emoji Embedding and Attention-based LSTM. In *ACM Multimedia Conference on Multimedia Conference*. ACM, 117–125.

[2] Guillem Cucurull, Perouz Taslakian, and David Vázquez. 2019. Context-Aware Visual Compatibility Prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 12617–12626.

[3] Zeyu Cui, Zekun Li, Shu Wu, Xiaoyu Zhang, and Liang Wang. 2019. Dressing as a Whole: Outfit Compatibility Learning Based on Node-wise Graph Neural Networks. In *Porceedings of the World Wide Web Conference*. ACM, 307–317.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. ImageNet: A Large-scale Hierarchical Image Database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 248–255.

[5] Xue Dong, Jianlong Wu, Xuemeng Song, Hongjun Dai, and Liqiang Nie. 2020. Fashion Compatibility Modeling through a Multi-modal Try-on-guided Scheme. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 771–780.

[6] Yixiao Ge, Dapeng Chen, and Hongsheng Li. 2020. Mutual Mean-Teaching: Pseudo Label Refinery for Unsupervised Domain Adaptation on Person Re-identification. In *International Conference on Learning Representations*. OpenReview.net, 1–15.

[7] William L. Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive Representation Learning on Large Graphs. In *Advances in Neural Information Processing Systems*. MIT Press, 1024–1034.

[8] Xianjing Han, Xuemeng Song, Yiyang Yao, Xin-Shun Xu, and Liqiang Nie. 2020. Neural Compatibility Modeling With Probabilistic Knowledge Distillation. *IEEE Transactions on Image Processing* 29 (2020), 871–882.

[9] Xianjing Han, Xuemeng Song, Jianhua Yin, Yinglong Wang, and Liqiang Nie. 2019. Prototype-guided Attribute-wise Interpretable Scheme for Clothing Matching. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 785–794.

[10] Xintong Han, Zuxuan Wu, Yu-Gang Jiang, and Larry S. Davis. 2017. Learning Fashion Compatibility with Bidirectional LSTMs. In *Proceedings of the ACM on Multimedia Conference*. ACM, 1078–1086.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 770–778.

[12] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the Knowledge in a Neural Network. *CoRR* abs/1503.02531 (2015).

[13] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (1997), 1735–1780.

[14] Yupeng Hu, Meng Liu, Xiaobin Su, Zan Gao, and Liqiang Nie. 2021. Video Moment Localization via Deep Cross-modal Hashing. *IEEE Transactions on Image Processing* 30 (2021), 4667–4677.

[15] Yupeng Hu, Liqiang Nie, Meng Liu, Kun Wang, Yinglong Wanga, and Xiansheng Hua. 2021. Coarse-to-Fine Semantic Alignment for Cross-modal Moment Localization. *IEEE Transactions on Image Processing* 30 (2021), 5933–5943.

[16] Zhiting Hu, Xuezhe Ma, Zhengzhong Liu, Eduard H. Hovy, and Eric P. Xing. 2016. Harnessing Deep Neural Networks with Logic Rules. In *Proceedings of the the Annual Meeting of the Association for Computational Linguistics*. The Association for Computer Linguistics, 2410–2420.

[17] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs. In *Proceedings of the ACM on Multimedia Conference*. ACM, 795–816.

[18] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the International Conference on Learning Representations*. OpenReview.net, 1–15.

[19] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *International Conference on Learning Representations*. OpenReview.net, 1–14.

[20] Xingchen Li, Xiang Wang, Xiangnan He, Long Chen, Jun Xiao, and Tat-Seng Chua. 2020. Hierarchical Fashion Graph Network for Personalized Outfit Recommendation. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 159–168.

[21] Yujie Lin, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Jun Ma, and Maarten de Rijke. 2020. Explainable Outfit Recommendation with Joint Outfit Matching and Comment Generation. *Transactions on Knowledge and Data Engineering* 32, 8 (2020), 1502–1516.

[22] Jinhuan Liu, Xuemeng Song, Liqiang Nie, Tian Gan, and Jun Ma. 2020. An End-to-End Attention-Based Neural Model for Complementary Clothing Matching. *ACM Transactions on Multimedia Computing, Communications and Applications*, 15, 4 (2020), 114:1–114:16.

[23] Meng Liu, Liqiang Nie, Xiang Wang, Qi Tian, and Baoquan Chen. 2019. Online Data Organizer: Micro-Video Categorization by Structure-Guided Multimodal Dictionary Learning. *IEEE Transactions on Image Processing* 28, 3 (2019), 1235–1247.

[24] Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, and Tat-Seng Chua. 2018. Cross-Modal Moment Localization in Videos. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 843–851.

[25] Si Liu, Jiashi Feng, Zheng Song, Tianzhu Zhang, Hanqing Lu, Changsheng Xu, and Shuicheng Yan. 2012. Hi, Magic Closet, Tell Me What to Wear!. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 619–628.

[26] Julian J. McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. Image-Based Recommendations on Styles and Substitutes. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 43–52.

[27] Liqiang Nie, Yongqi Li, Fuli Feng, Xuemeng Song, Meng Wang, and Yinglong Wang. 2020. Large-Scale Question Tagging via Joint Question-Topic Embedding Learning. *ACM Transactions on Information Systems* 38, 2 (2020), 20:1–20:23.

[28] Liqiang Nie, Xiang Wang, Jianglong Zhang, Xiangnan He, Hanwang Zhang, Richang Hong, and Qi Tian. 2017. Enhancing Micro-video Understanding by Harnessing External Sounds. In *Proceedings of the International ACM Conference on Multimedia*. ACM, 1192–1200.

[29] Liqiang Nie, Xiaochi Wei, Dongxiang Zhang, Xiang Wang, Zhipeng Gao, and Yi Yang. 2017. Data-Driven Answer Selection in Community QA Systems. *IEEE Transactions on Knowledge and Data Engineering* 29, 6 (2017), 1186–1198.

[30] Mathias Niepert, Mohamed Ahmed, and Konstantin Kutzkov. 2016. Learning Convolutional Neural Networks for Graphs. In *Proceedings of the International Conference on Machine Learning*. JMLR.org, 2014–2023.

[31] Wonpyo Park, Wonjae Kim, Kihyun You, and Minsu Cho. 2020. Diversified Mutual Learning for Deep Metric Learning. In *Proceedings of the European Conference on Computer Vision*. Springer, 709–725.

[32] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 452–461.

[33] Xuemeng Song, Fuli Feng, Jinhuan Liu, Zekun Li, Liqiang Nie, and Jun Ma. 2017. NeuroStylist: Neural Compatibility Modeling for Clothing Matching. In *Proceedings of the ACM on Multimedia Conference*. ACM, 753–761.

[34] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2818–2826.

[35] Reuben Tan, Mariya I. Vasileva, Kate Saenko, and Bryan A. Plummer. 2019. Learning Similarity Conditions Without Explicit Supervision. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 10372–10381.

[36] Mariya I. Vasileva, Bryan A. Plummer, Krishna Dusad, Shreya Rajpal, Ranjitha Kumar, and David A. Forsyth. 2018. Learning Type-Aware Embeddings for Fashion Compatibility. In *Proceedings of the European Conference on Computer Vision*. Springer, 405–421.

[37] Haokun Wen, Xuemeng Song, Xin Yang, Yibing Zhan, and Liqiang Nie. 2021. Comprehensive Linguistic-Visual Composition Network for Image Retrieval. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1369––1378.

[38] Xun Yang, Yunshan Ma, Lizi Liao, Meng Wang, and Tat-Seng Chua. 2019. TransNFCM: Translation-Based Neural Fashion Compatibility Modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press, 403–410.

[39] Xin Yang, Xuemeng Song, Fuli Feng, Haokun Wen, Ling-Yu Duan, and Liqiang Nie. 2021. Attribute-wise Explainable Fashion Compatibility Modeling. *ACM Transactions on Multimedia Computing* 17, 1 (2021), 36:1–36:21.

[40] Yunpeng Zhai, Qixiang Ye, Shijian Lu, Mengxi Jia, Rongrong Ji, and Yonghong Tian. 2020. Multiple Expert Brainstorming for Domain Adaptive Person Re-Identification. In *Proceedings of the European Conference on Computer Vision*. Springer, 594–611.

[41] Yibing Zhan, Jun Yu, Ting Yu, and Dacheng Tao. 2019. On Exploring Undetermined Relationships for Visual Relationship Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 5128–5137.

[42] Yibing Zhan, Jun Yu, Ting Yu, and Dacheng Tao. 2020. Multi-task Compositional Network for Visual Relationship Detection. *International Journal of Computer Vision* 128, 8 (2020), 2146–2165.

[43] Peng Zhang, Li Su, Liang Li, BingKun Bao, Pamela C. Cosman, Guorong Li, and Qingming Huang. 2019. Training Efficient Saliency Prediction Models with Knowledge Distillation. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 512–520.

[44] Ying Zhang, Tao Xiang, Timothy M. Hospedales, and Huchuan Lu. 2018. Deep Mutual Learning. In *IEEE Conference on Computer Vision and Pattern Recognition,*. IEEE, 4320–4328.