Collocation and Try-on Network: Whether an Outfit is Compatible

Na Zheng[1], Xuemeng Song[1*], Qingying Niu[1], Xue Dong[1], Yibing Zhan[2], Liqiang Nie[1*]
[1]Shandong University, Shandong, China, [2]JD Explore Academy, Beijing, China
{zhengnagrape,sxmustc,qingyingn,dongxue.sdu,nieliqiang}@gmail.com,zhanyibing@jd.com

ABSTRACT

Whether an outfit is compatible? Using machine learning methods to assess an outfit's compatibility, namely, fashion compatibility modeling (FCM), has recently become a popular yet challenging topic. However, current FCM studies still perform far from satisfactory, because they only consider the collocation compatibility modeling, while neglecting the natural human habits that people generally evaluate outfit compatibility from both the collocation (discrete assess) and the try-on (unified assess) perspectives. In light of the above analysis, we propose a Collocation and Try-On Network (CTO-Net) for FCM, combining both the collocation and try-on compatibilities. In particular, for the collocation perspective, we devise a disentangled graph learning scheme, where the collocation compatibility is disentangled into multiple fine-grained compatibilities between items; regarding the try-on perspective, we propose an integrated distillation learning scheme to unify all item information in the whole outfit to evaluate the compatibility based on the latent try-on representation. To further enhance the collocation and try-on compatibilities, we exploit the mutual learning strategy to obtain a more comprehensive judgment. Extensive experiments on the real-world dataset demonstrate that our CTO-Net significantly outperforms the state-of-the-art methods. In particular, compared with the competitive counterparts, our proposed CTO-Net significantly improves AUC accuracy from 83.2% to 87.8% and MRR from 15.4% to 21.8%. We have released our source codes and trained models to benefit other researchers[1].

CCS CONCEPTS

• **Information systems** → **Retrieval tasks and goals**; *World Wide Web*.

KEYWORDS

Fashion Compatibility Modeling, Try-on Knowledge Distillation, Disentangled Graph Neural Network.

[1]https://compatibilitymodel.wixsite.com/cto-net.

* Xuemeng Song (sxmustc@gmail.com) and Liqiang Nie (nieliqiang@gmail.com) are corresponding authors.

MM '21, October 20–24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475691

Figure 1: Illustration of the compatibility modeling from both collocation and try-on perspectives.

ACM Reference Format:
Na Zheng[1], Xuemeng Song[1*], Qingying Niu[1], Xue Dong[1], Yibing Zhan[2], Liqiang Nie[1*]. 2021. Collocation and Try-on Network: Whether an Outfit is Compatible . In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3475691

1 INTRODUCTION

With the blossom of e-commerce, an increasing number of people have turned to online shopping for fashion garments. Enjoying the convenience provided by the online fashion market, people also tend to be overwhelmed by the numerous online clothing items. Specifically, they are frequently encountered with problems such as: "does this T-shirt match my jeans" or "which shirt is better for the skirt". Towards this end, fashion compatibility modeling (FCM) that aims to automatically assess the compatibility (*i.e.*, matching score) of items in an outfit has gained growing research attention.

There have been numerous researches of FCM proposed in the literature [2, 3, 8, 16, 18, 27, 34, 37]. Despite the impressive progress, current FCM methods still perform far from satisfactory. They only focus on modeling the compatibility relationship among discrete items in an outfit but overlook the human habits on fashion compatibility evaluation. In fact, people usually evaluate the matching degree of a given outfit from not only the collocation angle (*i.e.*, in a discrete manner) but also the try-on angle (*i.e.*, in a unified manner). In light of this, we aim to devise a comprehensive

FCM scheme that evaluates the outfit compatibility from both the discrete collocation and unified try-on angles, as shown in Figure 1.

However, combining both the collocation and try-on angles is non-trivial due to the following challenges. 1) The visual compatibility among discrete items in an outfit can be affected by multiple latent factors (*e.g.*, color, material and style), which indicates that the overall compatibility among items can be decoupled into multiple latent fine-grained visual compatibility. We argue that exploring the latent fine-grained compatibility would make the task more tractable, thus improving model performance. Therefore, capturing the latent fine-grained compatibility among discrete items to enhance FCM is the first challenge. 2) A simple solution towards try-on compatibility modeling is to evaluate the compatibility through an outfit's real try-on appearance image. However, the real try-on appearance is usually unavailable in practice. Therefore, how to utilize the limited training try-on appearance images of outfits to model the try-on compatibility for outfits without the real try-on images constitutes another challenge. And 3) the compatibility of the same outfit evaluated from the collocation and try-on angles should be intrinsically consistent. Therefore, utilizing the latent consistency to integrate the collocation and try-on compatibility modeling seamlessly, in order to boost the model performance, poses the third challenge.

To address the aforementioned challenges, we propose the **C**ollocation and **T**ry-**O**n **Net**work (CTO-Net) for FCM. As shown in Figure 2, CTO-Net consists of two core parts: *Collocation Compatibility Modeling (CCM)* and *Try-on Compatibility Modeling (TCM)*. CCM implicitly fulfills the fine-grained compatibility modeling. Specifically, to uncover the latent factors influencing the compatibility, CCM injects disentangled item representations into a graph convolutional network (GCN) and devises a new disentangled compatibility propagation module to adaptively propagate the fine-grained compatibility relationships among items. TCM fulfills the try-on representation learning by exploiting the teacher-student knowledge distillation scheme. In particular, the teacher network is first trained using unsupervised self-encoding. Then, the student network imitates the output of the teacher network and hence derives the accurate try-on representation directly from the outfit's discrete composing items. To strengthen the try-on representation learning, we incorporate the category information of each item as the context, which remains untapped by previous studies. In addition, we employ the mutual learning strategy [43] to encourage both the CCM and TCM to transfer knowledge from each other and further boost the final compatibility modeling performance. Experimental results on the real-world dataset demonstrate the superiority of our CTO-Net over the state-of-the-art methods.

The main contributions of this paper are summarized as follows:

1) Inspired by the human habits towards the fashion compatibility evaluation, we present a novel framework, *i.e.*, CTO-Net, to comprehensively analyze the fashion compatibility from both the discrete collocation and unified try-on angles. In particular, the two compatibility modeling schemes get mutually enhanced by absorbing knowledge from each other.

2) We propose a novel disentangled graph learning scheme, which is capable of analyzing the fine-grained collocation compatibility through propagating the compatibility between discrete items based on their disentangled representations.

3) We propose an integrated distillation learning scheme for try-on compatibility modeling, which utilizes the limited try-on appearance images for guiding the network to learn a reliable try-on representation based on the discrete fashion items in the outfit.

2 RELATED WORK

This work is related to fashion compatibility modeling, graph convolutional network, and knowledge distillation.

Fashion Compatibility Modeling. Existing studies on FCM can be summarized into three groups, *i.e.*, pair-wise [20, 23, 27, 28, 30], list-wise [4, 5, 8], and set-wise [3, 16, 37]. The pair-wise methods focus on compatibility between a pair of items and derive the compatibility by measuring all the pairs of items in an outfit. Apparently, the pair-wise methods do not treat the outfit as a whole and hence suffer from the sub-optimal performance. As to directly evaluate the compatibility for outfits with multiple items, increasing efforts have been dedicated to the list-wise and set-wise manners. Specifically, the list-wise methods assume the outfit as an ordered sequence of items, and employ the bi-directional LSTM [8] or GRU [2], to uncover the sequential compatibilities among items. Beyond that, the set-wise methods behave in a more flexible manner by treating an outfit as a set of unordered items and employing either GCN or self-attention mechanism to evaluate an outfit's matching degree [3, 37]. Although huge success, existing methods cannot achieve comprehensive compatibility modeling, as they overlook that humans usually evaluate the outfit compatibility from both the discrete collocation and unified try-on angles.

Graph Convolutional Network. Mathias et al. [25] proposed the graph convolutional network (GCN), which generalizes convolution operation to the graph domain [13]. Due to its remarkable capability of representation learning, GCN has been widely explored in various tasks, including recommendation [31], information retrieval [6], and visual comprehension [36, 39]. Recently, in the fashion domain, as each outfit can be abstracted as an item graph, several GCN-based methods, like NGNN [3] and HFGN [37], have been proposed for FCM. The key of these methods is to update the item embedding with its context (the other fashion items) in the outfit. Different from these methods that only propagate the general item embedding, in this work, we conducted the fine-grained item-item relationship propagation among items with the disentangled representation learning.

Knowledge Distillation. Due to the remarkable performance in various tasks [11, 14, 32, 40], knowledge distillation has attracted growing research attention recently. Knowledge distillation adopts the teacher-student network and aims to enhance the student network by transferring the knowledge from a powerful teacher network to the student network [9]. Recently, several studies have been dedicated to exploring the potency of the knowledge distillation to enhance the performance in the fashion domain [1, 27]. For example, Song et al. [27] proposed to incorporate fashion domain knowledge to facilitate fashion compatibility evaluating. Beyond the existing efforts, we worked on transferring the try-on knowledge gained from the teacher network with the real try-on image, to improve the try-on representation directly based on the outfit's discrete composing items.

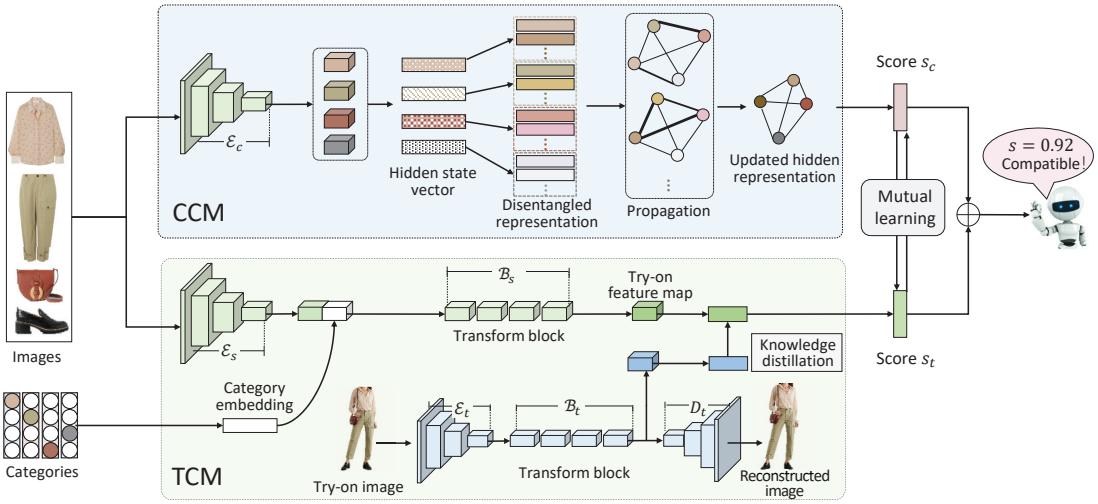

Figure 2: Illustration of the proposed CTO-Net, which consists of two core parts: CCM and TCM. In particular, CCM and TCM are responsible for modeling the fine-grained collocation compatibility and the unified try-on compatibility, respectively. Finally, the mutual learning is employed to encourage the knowledge sharing between CCM and TCM.

3 METHODOLOGY

3.1 Problem Formulation

Suppose we have the training set $\Omega = \{(O^i, y^i)|i = 1, ..., N\}$ composed of N outfits, where O^i is the i-th outfit, and y^i denotes the ground truth label. Specifically, $y^i = 1$ indicates that the outfit O^i is compatible and $y^i = 0$ otherwise. Each outfit O^i consists of M_i complementary fashion items $O^i = \{o^i_1, o^i_2, \cdots, o^i_{M_i}\}$, where o^i_j is the j-th item, associated with an image pixel array $\mathbf{o}^i_j$ and a category embedding $\mathbf{c}^i_j$. In particular, $\mathbf{c}^i_j \in \mathbb{R}^C$ is a one-hot vector and C is the total number of fashion item categories in the dataset. Besides, only the positive/compatible outfits, have their corresponding try-on appearance images. Accordingly, the whole training set Ω can be split into two sets: one set of positive outfits $\Omega_+ = \{(O^i, \mathbf{P}^i, y^i)|y^i = 1\}$ and one set of negative outfits $\Omega_- = \{(O^i, y^i)|y^i = 0\}$. $\mathbf{P}^i$ denotes the i-th outfit's try-on image pixel array. Based on these data, we aim to devise a FCM scheme $\mathcal{F}$ to evaluate the compatibility score s^i of a given outfit $O^i = \{o^i_1, o^i_2, \cdots, o^i_{M_i}\}$ as follows:

$$s^i = \mathcal{F}(\{o^i_j\}^{M_i}_{j=1}|\Theta), \tag{1}$$

where Θ is a set of to-be-learned parameters. Notably, we omit the superscript i in the rest of the paper for brevity.

3.2 Disentangled Graph Learning for CCM

Existing methods [37] utilize general representations of composing items to capture the underlying collocation compatibility. However, we argue that the compatibility relationship among discrete items can be influenced by multiple latent factors, like color, texture, and style. Hence, there simultaneously exist multiple latent fine-grained compatibility relationships among discrete items. In light of this, we propose the disentangled graph learning scheme for CCM, which consists of three key parts: graph initialization, disentangled item representation, and disentangled compatibility propagation.

3.2.1 Graph Initialization. For each outfit O, we first construct an undirected graph $\mathcal{G} = (\mathcal{H}, \mathcal{E})$, where $\mathcal{H} = \{h_j\}^M_{j=1}$ is the set of nodes corresponding to the M composing fashion items in the outfit, respectively, while $\mathcal{E} = \{(h_j, h_k)|j, k \in [1, \cdots, M], j \neq k\}$ is the set of edges indicating the relation among the composing items of the outfit. Each node h_j is associated with a hidden state vector $\mathbf{h}_j$ utilized for the compatibility propagation. Since the visual cue plays an important role in the compatibility modeling, we initialize the hidden state vector $\mathbf{h}_j$ with the visual feature of the corresponding item o_j. Specifically, we introduce a visual encoder $\mathcal{E}_c$, consisting of several convolutional layers to obtain the visual feature map $\mathbf{F}_j$ of each fashion item o_j. Then we obtain the visual representation $\mathbf{f}_j$ of the j-th item by reshaping the visual feature map $\mathbf{F}_j$ into a vector. Ultimately, we initialize the hidden state vector $\mathbf{h}_j$ with a linear transformation over the visual representation $\mathbf{f}_j$ as follows:

$$\mathbf{h}_j = \frac{\mathbf{W}_h\mathbf{f}_j + \mathbf{b}_h}{||\mathbf{W}_h\mathbf{f}_j + \mathbf{b}_h||_2}, \tag{2}$$

where $\mathbf{W}_h$ and $\mathbf{b}_h$ are the weight matrix and bias vector to be learned, respectively. The linear transformation aims to project the visual representation to a low-dimensional space and the normalization is used to ensure the numerical stability.

3.2.2 Disentangled Item Representation. We argue that utilizing a single overall hidden vector to capture the multi-facet fine-grained compatibility relationship among fashion items can be insufficient. Beyond existing studies, inspired by the recent advance of disentangled representation learning in various recommendation tasks [10, 33], we propose to disentangle the multi-facet compatibility relationship among discrete items and capture their fine-grained compatibility on the basis of GCN.

In particular, we suppose that there are L latent factors $\{f_1, f_2, \cdots, f_L\}$ influencing the compatibility relationship among items. For each latent factor f_l, we employ a condition mask $\mathbf{C}_l \in \mathbb{R}^d$

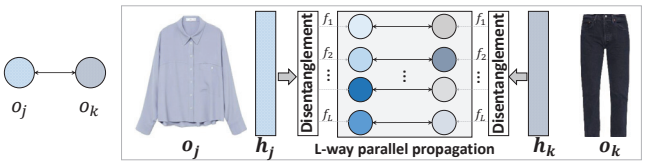

Figure 3: Illustration of L-way parallel propagation between two nodes (items).

Figure 4: Examples of different factors that dominate the compatibility between different items.

to derive the disentangled item representation $\mathbf{u}_j^l$ pertaining to the l-th latent factor as follows:

$$\mathbf{u}_j^l = \mathbf{h}_j \odot \mathbf{C}_l, \tag{3}$$

where $\odot$ denotes the element-wise product.

Intuitively, to promote the fine-grained compatibility modeling for CCM, we expect that each disentangled representation can focus on only one latent factor, namely, the disentangled representations for different latent factors should be as independent as possible. Therefore, we utilize L1 regularization on the condition masks to encourage the sparsity and disentanglement [29, 34] as follows:

$$\mathcal{L}_{mask} = \frac{1}{L}\sum_{l=1}^{L} ||\mathbf{C}_l||_1. \tag{4}$$

3.2.3 Disentangled Compatibility Propagation. After obtained the disentangled representation for each node (item), we proceed to the disentangled compatibility propagation over the graph. In particular, we employ L-way parallel propagation to deliver the fine-grained compatibility between items, as shown in Figure 3. One naive way to fulfill the information propagation between nodes o_j and o_k is to equally aggregate all the parallel propagations. However, the factor that dominates the compatibility between different item pairs may be different. For example, as shown in Figure 4, the factor that influences the compatibility for the left pair of items is most likely to be the pattern, while that for the right one is the style. In light of this, we incorporate the factor importance in the disentangled compatibility relationship propagation with the attention mechanism [19, 21, 22, 24, 35]. Specifically, given items o_j and o_k, we assess the attention score for their collocation compatibility on the l-th latent factor as follows:

$$\begin{cases} \mathbf{e}_{j,k} = \mathbf{W}_a^2(\delta(\mathbf{W}_a^1(\mathbf{h}_j||\mathbf{h}_k) + \mathbf{b}_a^1)) + \mathbf{b}_a^2, \\ a_{j,k}^l = \dfrac{\exp(e_{j,k}^l)}{\sum_{p=1}^{L}\exp(e_{j,k}^p)}, l \in \{1, 2, \cdots, L\}, \end{cases} \tag{5}$$

where $||$ denotes the concatenation operator; $\mathbf{W}_a^1$, $\mathbf{W}_a^2$, $\mathbf{b}_a^1$, and $\mathbf{b}_a^2$ refer to the attention network parameters. $\delta(\cdot)$ denotes the Tanh activate function. $\mathbf{e}_{j,k} = [e_{j,k}^1, e_{j,k}^2, \cdots, e_{j,k}^L] \in \mathbb{R}^L$ is the vector of intermediate attention scores for L factors.

Based on these attention scores, we define the message passing from the item o_k to the item o_j as follows:

$$\mathbf{m}_{k\to j} = \phi\big(\mathbf{W}_p \sum_{l=1}^{L} a_{j,k}^l(\mathbf{u}_j^l \odot \mathbf{u}_k^l) + \mathbf{b}_p\big), \tag{6}$$

where $\mathbf{W}_p$ and $\mathbf{b}_p$ denote the weight and bias to be learned, respectively. $\phi(\cdot)$ denotes the LeakyReLU activate function. Notably, instead of directly propagating the embedding of the k-th item to the j-th item, we focus more on the interaction between them, which is supposed to be the underlying compatibility between them. In particular, $\mathbf{u}_j^l \odot \mathbf{u}_k^l$ accounts for the compatibility relationship between items o_j and o_k in terms of the l-th latent factor.

Then by summarizing the passing message from all neighbors, the hidden state vector of item o_j can be updated as follows:

$$\mathbf{h}_j^* = \phi(\mathbf{W}_e\mathbf{h}_j + \mathbf{b}_e) + \sum_{\mathbf{o}_k \in \mathcal{N}_j} \frac{1}{|\mathcal{N}_j|}\mathbf{m}_{k\to j}, \tag{7}$$

where $\mathbf{W}_e$ and $\mathbf{b}_e$ denote the weight matrix and bias to be learned, respectively. $\mathcal{N}_j$ stands for the set of neighbor nodes of the node o_j, *i.e.*, all the complementary items for item o_j in outfit O. $\mathbf{h}_j^* \in \mathbb{R}^d$ is the updated hidden representation of item o_j, which absorbs information from both the neighbors and the node itself.

Finally, we feed the updated item representation that embodies its compatibility towards all the other items in the outfit to a multi-layer perceptron (MLP) to acquire its compatibility score towards the whole outfit. Thereafter, by summing the compatibility scores of all items in the outfit, we can derive the outfit's collocation compatibility. Formally, we have:

$$s_c = \frac{1}{M}\sum_{j=1}^{M} \sigma(\mathbf{W}_c^2(\phi(\mathbf{W}_c^1\mathbf{h}_j^*))), \tag{8}$$

where $\sigma(\cdot)$ denotes the Sigmoid function to normalize the compatibility score. $\mathbf{W}_c^1$ and $\mathbf{W}_c^2$ are the layer parameters.

3.3 Integrated Distillation Learning for TCM

The real try-on appearance image of an outfit is usually unavailable in practice. Although existing method [5] directly generates the outfit try-on appearance image based on the discrete items with a template generator, it suffers from the large input-output misalignment of the template generator. Instead, we resort to the teacher-student knowledge distillation scheme [32]. In particular, the teacher network is to learn the reliable try-on representation by unsupervised self-encoding with the target try-on appearance image, while the student network is to learn the global try-on representation with the guidance of the teacher network.

3.3.1 Teacher Network. To learn a valid representation of the try-on appearance and to guide the student network, we employ the auto-encoder network [41] as the teacher network, which has proven to be effective in the unsupervised visual encoding [42]. In particular, we adopt the ResNet-like architecture that has shown remarkable performance in various generation tasks [44, 45] as the teacher network $\mathcal{T}$. To be more specific, the teacher

network comprises an encoder $\mathcal{E}_t$ for compressing the real try-on appearance image $\mathbf{P}$ of the outfit O into the visual feature map as $\mathbf{z}_t = \mathcal{E}_t(\mathbf{P})$, a transform block $\mathcal{B}_t$ for converting the visual feature map into the more expressive one as $\mathbf{b}_t = \mathcal{B}_t(\mathbf{z}_t)$, and a decoder $\mathcal{D}_t$ for reconstructing the original try-on appearance image based on $\mathbf{b}_t$ as $\hat{\mathbf{P}} = \mathcal{D}_t(\mathbf{b}_t)$. For optimization, we introduce L1 regularization to minimize the discrepancy between the reconstructed try-on appearance $\hat{\mathbf{P}}$ and the ground truth one $\mathbf{P}$ as follows:

$$\mathcal{L}_t = \sum_{\mathbf{P} \in \mathcal{P}_+} ||\hat{\mathbf{P}} - \mathbf{P}||_1. \tag{9}$$

3.3.2 Student Network. Beyond existing studies that only focus on the visual cues of fashion items [5], we also take into account the category metadata in the try-on appearance learning. The major concern is that the category information of fashion items plays a pivotal role in the spatial arrangement of fashion items in an outfit try-on appearance, *e.g.*, the sweater or shirt is always on top of the trousers or jeans. Specifically, we devise the student network $\mathcal{S}$ with a visual encoder $\mathcal{E}_s$ for merging the visual cues of the composing items into the global visual embedding $\mathbf{z}_s$, a multi-modal fusion block $\mathcal{M}_s$ for fusing $\mathbf{z}_s$ and the category embedding into the latent feature map $\mathbf{m}_s$, and a transform block $\mathcal{B}_s$ for converting $\mathbf{m}_s$ into the try-on feature map $\mathbf{b}_s$ as follows:

$$\begin{cases} \mathbf{z}_s = \mathcal{E}_s(\mathbf{O}), \\ \mathbf{m}_s = \mathcal{M}_s(\mathbf{z}_s, \text{Rep}(\mathbf{W}_s^c \mathbf{C})), \\ \mathbf{b}_s = \mathcal{B}_s(\mathbf{m}_s), \end{cases} \tag{10}$$

where $\mathbf{O} = [\mathbf{o}_1, \mathbf{o}_2, \cdots, \mathbf{o}_M]$ and $\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_M]$ are the visual and category cues of items in the outfit. $\mathbf{W}_s^c \in \mathbb{R}^{d \times (C \times M)}$ is the weight matrix of the linear transformation to learn the global category embedding of the outfit, while $\text{Rep}(\cdot)$ refers to replicating the global category embedding to form the category feature map with the same shape of visual feature map $\mathbf{z}_s$.

3.3.3 Knowledge Distillation. Regarding the knowledge distillation from teacher network to student network, it is natural to regulate the latent representation of the try-on appearance learned by both teacher and student networks to be similar. In particular, we first resort to the global average pooling (GAP) [17], which has shown remarkable performance in discriminative visual property extraction [38], to summarize the learned try-on representations of the teacher and student networks as $\mathbf{f}_s = \text{GAP}(\mathbf{b}_s)$ and $\mathbf{f}_t = \text{GAP}(\mathbf{b}_t)$, respectively. Then, using L1 regularization, we have:

$$\mathcal{L}_s = ||\mathbf{f}_s - \mathbf{f}_t||_1. \tag{11}$$

Similar with CCM, we employ a fully-connected layer to obtain the try-on compatibility score s_t as follows:

$$s_t = \sigma(\mathbf{W}_s \mathbf{f}_s + \mathbf{b}_s), \tag{12}$$

where $\mathbf{W}_s$ and $\mathbf{b}_s$ are layer parameters to be learned.

3.4 Mutual Learning based Joint Optimization

Similar to [5, 15], we cast the compatibility modeling as a binary classification task. For each outfit O, we adopt the following cross-entropy losses for CCM and TCM:

$$\begin{cases} \mathcal{L}_{ce}^c = -y\log(s_c) - (1-y)\log(1-s_c), \\ \mathcal{L}_{ce}^t = -y\log(s_t) - (1-y)\log(1-s_t), \end{cases} \tag{13}$$

Table 1: Data split provided by FOTOS. In FOTOS, each outfit contains 4 clothing items at most.

#item	training	validating	testing	total
2	4,674	47	486	5, 207
3	4, 797	50	469	5,316
4	418	3	44	465
total	9, 889	100	999	10,988

where y is the ground truth label of the outfit. $\mathcal{L}_{ce}^c$ and $\mathcal{L}_{ce}^t$ denote the classification losses for CCM and TCM, respectively.

Moreover, although CCM and TCM model the compatibility from different angles, for the same outfit, their evaluation should still be consistent. In other words, the knowledge of CCM can be used for guiding the TCM and vice versa. Therefore, we incorporate the mutual learning strategy [43] to encourage their knowledge sharing with each other. In particular, we adopt the most popular Kullback-Leibler divergence regularization as follows:

$$\begin{cases} \mathcal{L}_{kl}^{t \to c} = KL(\mathbf{p}_c || \mathbf{p}_t), \\ \mathcal{L}_{kl}^{c \to t} = KL(\mathbf{p}_t || \mathbf{p}_c), \end{cases} \tag{14}$$

where $\mathbf{p}_c = [s_c, 1-s_c]^{\mathrm{T}}$ and $\mathbf{p}_t = [s_t, 1-s_t]^{\mathrm{T}}$. $\mathcal{L}_{kl}^{t \to c}$ and $\mathcal{L}_{kl}^{c \to t}$ refer to the regularization for CCM and TCM, respectively.

Taking all the training samples into account, our final objective function can be formulated as follows:

$$\begin{cases} \mathcal{L}^c = \sum_{\Omega} (\mathcal{L}_{ce}^c + \mathcal{L}_{kl}^{t \to c} + \lambda_m \mathcal{L}_{mask}), \\ \mathcal{L}^t = \sum_{\Omega} (\mathcal{L}_{ce}^t + \mathcal{L}_{kl}^{c \to t}) + \sum_{\Omega_+} \mathcal{L}_s, \end{cases} \tag{15}$$

where λ_m refers to the trade-off hyper-parameter. Notably, $\mathcal{L}_s$ is optimized by the set of positive outfits Ω_+, since the set of negative ones is unavailable. Overall, we alternatively optimize the CCM and TCM modules. In particular, for each module optimization, only the corresponding parameters need to be optimized. Once the whole network gets well-optimized, we estimate the overall compatibility score s for a given outfit O as follows:

$$s = \frac{1}{2}(s_c + s_t). \tag{16}$$

4 EXPERIMENTS

4.1 Experimental Settings

Dataset. For evaluation, we used the public dataset FOTOS [5], which consists of 10, 988 compatible outfits composed by 20, 318 fashion items. Each fashion item is associated with a visual image and the category metadata. Distinguished from other datasets, apart from the discrete items information, each outfit in FOTOS also has a corresponding try-on image, which enables the optimization of our try-on compatibility learning scheme. For the fair comparison, we adopted the public training/validating/testing data split provided by FOTOS. The detailed statistics are shown in Table 1.

Evaluation Task and Metric. Following TryOn-CM [5], we evaluated the performance of our model with two tasks: the top-n recommendation [26] and the positive/negative outfit classification, where we followed the same data settings in [5]. For evaluation

Table 2: Performance comparison among different methods. * denotes the statistical significance for $p < 0.01$, compared with the strongest baselines highlighted with the underline.

Method	AUC	MRR	HR			
			@1	@10	@100	@200
BRR-DAE	0.742	0.087	0.046	0.165	0.552	0.741
PAICM	0.692	0.057	0.024	0.110	0.468	0.662
CSA-Net	0.701	0.061	0.040	0.107	0.486	0.689
NCR	0.646	0.034	0.012	0.064	0.376	0.616
LSTM-VSE	0.794	0.118	0.065	0.226	0.642	0.809
TryOn-CM	0.832	0.134	0.061	0.290	0.721	0.852
NGNN	0.698	0.055	0.022	0.102	0.478	0.687
HFGN	0.816	0.154	0.080	0.312	0.704	0.834
CANN	0.820	0.146	0.081	0.267	0.702	0.837
CTO-Net	**0.878***	**0.218***	**0.134***	**0.395***	**0.800***	**0.899***

metrics, we utilized the Mean Reciprocal Ranking (MRR) and the Hit Rate (HR) @1, 10, 100, and 200 for the former task, during the Area Under Curve (AUC) for the latter one.

Implementation Details. As for CCM, we devised the visual encoder $\mathcal{E}_c$ with one 1-strided convolutional layer and four 2-strided convolutional layers, where the numbers of filters are 32, 64, 128, 256, and 512, respectively. The shape of the visual feature map generated by $\mathcal{E}_c$ is $512 \times 8 \times 8$, and the dimension of the hidden state vector is set to 512. The number of latent factors is set to 5.

Regarding TCM, we implemented the teacher network $\mathcal{T}$ with an encoder $\mathcal{E}_t$ sharing the same network architecture with $\mathcal{E}_c$, followed by the transform block $\mathcal{B}_t$ composed of 6 residual blocks, as well as the decoder $\mathcal{D}_t$ with four 2-strided deconvolutional layers and one 1-strided convolutional layer. The numbers of filters are set to 32, 64, 128, 256, 512, 512, 512, 512, 512, 512, 512, 256, 128, 64, 32 and 3, respectively. All convolutional layers above are followed by the Instance Normalization and ReLU function, except for the last layer, which takes the Tanh function. In addition, we implemented the student network $\mathcal{S}$ with an encoder $\mathcal{E}_s$ sharing the same network architecture with $\mathcal{E}_t$, a multi-modal fusion block $\mathcal{M}_s$ by a 1-strided convolutional layer, and a transform block $\mathcal{B}_s$ by four residual blocks. The numbers of filters are set to 32, 64, 128, 256, 512, 512, 512, 512, 512 and 512, respectively. Ultimately, we obtained $\mathbf{f}_t \in \mathbb{R}^{512}$ and $\mathbf{f}_s \in \mathbb{R}^{512}$ with the GAP layer. Similar to the study [32], we first pre-trained the teacher network and then fixed the teacher network for guiding the student network learning.

For optimization, we adopted the Adam [12] optimizer with $\beta_1 = 0.5$, $\beta_2 = 0.999$, a fixed learning rate of 0.0002, and the batch size of 32 for all experiments. As for outfits with less than 4 items, we used zero-paddings. Ultimately, we empirically found that the proposed method achieves the optimal performance with the hyper-parameter λ_m in Eqn.(15) as 0.0005. In particular, we reported the average results of eight dependent experiments of our method.

4.2 On Model Comparison

To verify the effectiveness of our CTO-Net, we adopted the following state-of-the-art methods for comparison.

BRR-DAE [28] aims to model the coherent relation among multi-modalities of items based on a dual auto-encoder network.

PAICM [7], as a pair-wise method, utilizes matrix factorization to learn several compatible/incompatible prototypes to promote the explainable fashion compatibility modeling.

CSA-Net [29] introduces a category-based subspace attention network to flexibly learn latent representations for pair-wise fashion items based on the category labels.

NCR [2] explores the textual information in terms of the semantic and lexical aspects towards outfit compatibility modeling, where the outfit compatibility is learned in a list-wise manner.

LSTM-VSE [8] exploits the latent discrete item interaction by a bi-directional LSTM and visual-semantic consistency to facilitate the outfit compatibility modeling.

TryOn-CM [5] is the first to leverage the try-on appearance image of an outfit to boost the performance of fashion compatibility.

NGNN [3] is the first attempt to employ a graph to uncover the complex relationships among multiple complementary items.

HFGN [37] develops a hierarchical fashion graph network to unify the fashion compatibility modeling and personalized outfit recommendation. In our context, we only employed the item graph module for compatibility estimation.

CANN [16] learns the computational visual coherence by fully exploring the attention mechanism in a set-wise manner.

Table 2 shows the performance comparison among different methods, where the statistically significant test is performed between CTO-Net with the strongest baselines (highlighted with the underline). It is worth noting that since BPR-DAE, PAICM, NCR, LSTM-VSE, and TryOn-CM have been also adopted by the work [5], we directly referred to their performance in [5]. Note that we employed the same data settings with [5] for all experiments. From Table 2, we have the following observations: 1) Our CTO-Net consistently outperforms all the baselines, including both pair-wise, list-wise, and set-wise methods, by a large margin with respect to all metrics, which demonstrates the superiority of our proposed framework. 2) CTO-Net shows superiority over TryOn-CM, which also considers the try-on appearance to boost the compatibility modeling performance. This indicates the robustness of our integrated distillation learning scheme in obtaining the reliable try-on representation. The detailed comparison between CTO-Net and TryOn-CM will be given in Section 4.4. 3) On average, pair-wise methods (*i.e.*, BRP-DAE, PAICM, and CSA-Net) perform worse than list-wise and set-wise methods (*i.e.*, LSTM-VSE, TryOn-CM, HFGN, CANN, and ours). The philosophy behind may be that separately modeling the compatibility between item pairs in the outfit cannot accurately discover the complicated relationships among them and hence yields sub-optimal performance. And 4) it is unexpected that the set-wise method NGNN performs worse than the pair-wise methods. The possible reason is that NGNN focuses on propagating category-oriented fashion compatibility. Thus, it performs unsatisfactorily in our context, where the negative outfit shares the same item category as the positive one.

To gain deeper insights, we looked into the performance of our model regarding outfits with different number (*i.e.*, M) of composing items. Figure 5 shows the performance comparison between our CTO-Net and the five strongest baselines, *i.e.*, BPR-DAE, LSTM-VSE, TryOn-CM, HFGN, and CANN, with different testing configurations. As can be seen, our method surpasses all baseline methods in all settings, verifying the effectiveness of our method to handle the

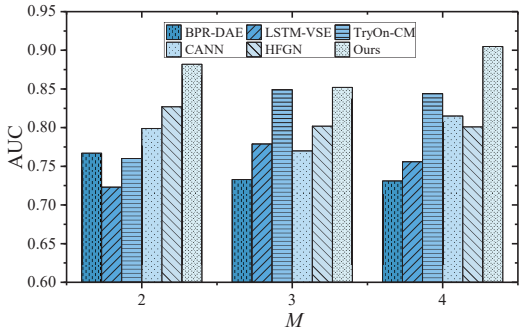

Figure 5: Performance comparison on compatibility evaluation in terms of different item numbers.

outfit compatibility with different composing item numbers. We also found that TryOn-CM outperforms other baselines, confirming that the try-on appearance modeling indeed benefits the FCM.

On Time Consumption. To study the efficiency of our CTO-Net, we compared the time consumption of TryOn-CM, HFGN, CANN and CTO-Net in Table 3. Notably, the input of all methods are images of an outfit's composing items. As can be seen, our CTO-Net shows the acceptable complexity during the training phase, while outperforms the baselines in terms of the testing phase. This demonstrates that our CTO-Net is efficient and practical for the real-world application scenarios.

4.3 On Ablation Study

To get a thorough understanding of our model, we conducted ablation experiments on the following derivatives.

CCM: This is a variant of our model that only uses the collocation compatibility modeling with the disentangled graph learning.

TCM: This is an implementation that only employs the integrated distillation learning-based TCM module.

TCM-w/o-SC and **w/o-Sc**: We removed the category input from the student network of both TCM and CTO-Net to learn its importance to try-on representation learning.

CCM-w/o-Dis and **w/o-Dis**: To validate the necessity of the fine-grained compatibility learning, we disabled the disentangled representation by setting $L = 1$ for CCM and CTO-Net, respectively.

CCM-w/o-InterP and **w/o-InterP**: To validate the function of the item-item compatibility propagation, we replaced $\mathbf{u}_j^l \odot \mathbf{u}_k^l$ with $\mathbf{u}_k^l$ in Eqn.(6) from both CCM and CTO-Net.

CCM-w/o-Att and **w/o-Att**: To study the effect of attention mechanism in the adaptive importance attribution for different latent factors, we equally aggregated all the parallel propagations regarding different latent factors in both CCM and CTO-Net.

w/o-Mut: We removed both $\mathcal{L}_{kl}^{c\to t}$ and $\mathcal{L}_{kl}^{t\to c}$ from Eqn.(15) to explore the importance of mutual learning strategy.

CCM-w/-Mut and **TCM-w/-Mut**: To learn the effect of mutual learning for CCM and TCM, we provided their corresponding results with the enhancement of the mutual learning, respectively.

Table 3: The comparison of time consumption. Training(s) and Testing(s) denote the time cost for training per epoch and testing per outfit, respectively.

Method	Training(s)	Testing(s)
TryOn-CM	430	0.0036
HFGN	**139**	0.0068
CANN	175	0.0079
CTO-Net	164	**0.0020**

Table 4: The ablation experiments of our proposed method.

Method	AUC	MRR	HR			
			@1	@10	@100	@200
CCM-w/o-Dis	0.825	0.125	0.063	0.241	0.719	0.847
CCM-w/o-InterP	0.786	0.076	0.031	0.154	0.638	0.805
CCM-w/o-Att	0.835	0.133	0.064	0.270	0.726	0.862
CCM	**0.848**	**0.151**	**0.079**	**0.309**	**0.751**	**0.871**
TCM-w/o-Sc	0.813	0.122	0.065	0.222	0.689	0.840
TCM	**0.835**	**0.134**	**0.072**	**0.255**	**0.723**	**0.845**
w/o-Sc	0.862	0.181	0.103	0.346	0.776	0.885
w/o-Dis	0.867	0.196	0.115	0.365	0.782	0.884
w/o-InterP	0.822	0.150	0.080	0.278	0.693	0.832
w/o-Att	0.862	0.186	0.110	0.339	0.771	0.889
CCM-w/-Mut	0.872	0.177	0.096	0.345	0.791	0.899
TCM-w/-Mut	0.854	0.175	0.099	0.319	0.764	0.873
w/o-Mut	0.867	0.190	0.112	0.346	0.784	0.893
CTO-Net	**0.878**	**0.218**	**0.134**	**0.395**	**0.800**	**0.899**

Table 4 shows the ablation experimental results. Based on Table 4, we have the following observations. 1) Our model consistently surpasses all derivations across all metrics, demonstrating the effectiveness of each component in our proposed CTO-Net. 2) Both CTO-Net and w/o-Mut consistently surpass CCM and TCM, which implies that only modeling the compatibility from one angle (either the collocation and try-on) cannot comprehensively capture the complex compatibility relationship among multiple items. 3) CTO-Net (TCM) shows superiority over w/o-Sc (TCM-w/o-Sc). This indicates that leveraging the category information may assist to automatically capture the item spatial arrangement in the try-on appearance, and thus boost the performance of our try-on representation learning. 4) CTO-Net (CCM) is superior to w/o-Dis (CCM-w/o-Dis), implying the necessity to explore the fine-grained compatibility between items in the FCM. 5) CTO-Net and CCM significantly outperform w/o-InterP and CCM-w/o-InterP, respectively. This confirms the facility of regarding the compatible information as a passing message rather than the item embedding in the context of outfit compatibility modeling. 6) w/o-Att (CCM-w/o-Att) is inferior to CTO-Net (CCM). The possible reason is that the confidence of the compatibility corresponding to different factors are indeed not the same, while adopting the attention mechanism can flexibly assign the factor importance. And 7) without the mutual learning strategy, w/o-Mut shows inferiority to CTO-Net, which implies that encouraging the two key modules to learn from each other can boost the overall performance of the compatibility evaluation. Meanwhile, we observed that CCM-w/-Mut and TCM-w/-Mut outperform CCM and TCM, respectively. This suggests that both CCM and TCM benefit from mutual learning.

Impact of Latent Factor Number. To study the influence of L, we conducted experiments by ranging L from 1 to 8. As shown

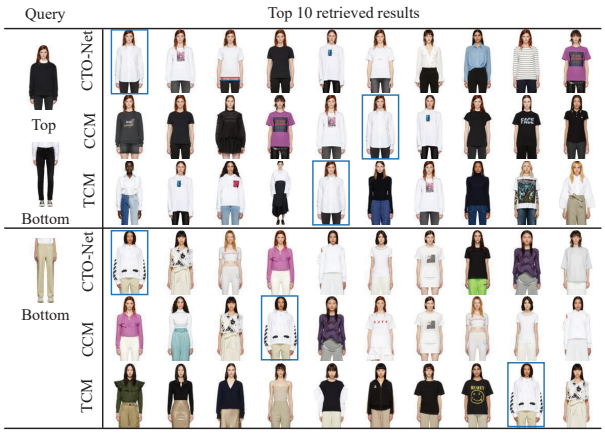

Figure 6: Top-10 retrieval results of different methods.

in Figure 7, the performance grows with L increasing from 1 to 5. The possible reason is that exploring the fine-grained compatibility relationships between items is beneficial to CCM. Nevertheless, the performance drops when L varies from 5 to 8, suggesting that considering too many fine-grained factors may limit the discriminative capability of the disentangled item representation.

Case Study. To get an intuitive understanding on how our model works, we compared the top-10 retrieved results of CCM, TCM, and CTO-Net for two testing queries in Figure 6. The ground truth items are highlighted in the blue boxes. As can be seen, all methods can retrieve the ground truth items in a relative top ranking, indicating that both the outfit collocation and try-on reveal important cues towards FCM. In particular, CTO-Net ranks higher than CCM and TCM, respectively, implying that only exploring either perspective is insufficient for a comprehensive FCM.

4.4 On Try-On Knowledge Distillation Study

To investigate whether the teacher-student knowledge distillation-based TCM shows superiority over the existing method, *i.e.*, [5], we introduced the following two variants based on CTO-Net and TryOn-CM. 1) CCM-TG. In this method, we replaced the try-on knowledge distillation module of CTO-Net with the counterpart in TryOn-CM, which essentially is a try-on template generator working on producing the try-on appearance image based on the discrete composing items of an outfit. And 2) CCM-TG-w/-Sc. Since the original TryOn-CM did not incorporate the item category context in the outfit try-on looking generation, for fair comparison,

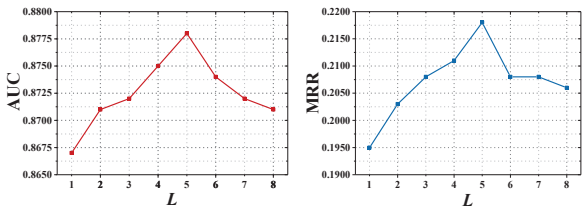

Figure 7: Impact of the latent factor number L.

Table 5: Performance on try-on learning methods.

Method	AUC	MRR	HR			
			@1	@10	@100	@200
CCM	0.848	0.151	0.079	0.309	0.751	0.871
CCM-TG	0.854	0.173	0.103	0.323	0.760	0.874
CCM-TG-w/-Sc	0.864	0.191	0.116	0.346	0.772	0.891
CTO-Net	**0.878**	**0.218**	**0.134**	**0.395**	**0.800**	**0.899**

we additionally fed the category context to CCM-TG in the same manner as TCM. Table 5 shows the performance comparison of different methods. Firstly, as can be seen, with the enhancement of the try-on modeling, both CCM-TG, CCM-TG-w/-Sc, and CTO-Net outperform CCM. This reconfirms the importance of assessing the try-on compatibility of an outfit. Secondly, we noticed that CTO-Net surpasses CCM-TG-w/-Sc, which indicates that the try-on appearance representation learned by our scheme is more reliable than that derived by a try-on template generator. Last but not least, CCM-TG achieves an inferior performance than CCM-TG-w/-Sc, validating the necessity of utilizing the item category to enhance the try-on appearance representation learning.

5 CONCLUSION

In this work, towards outfit compatibility modeling, we present a novel collocation and try-on network, named CTO-Net, which consists of two key components: *CCM* and *TCM*. Particularly, as for the CCM, we inject the disentangled item representations into GCN and devise a novel disentangled compatibility propagation to uncover the fine-grained compatibility relationships in terms of various latent factors. Pertaining to the TCM, we introduce a teacher network to learn a real try-on representation by unsupervised self-encoding, and a student network to imitate the teacher to acquire an accurate try-on representation directly based on composing discrete items, where item category is first studied as the spatial guidance to strengthen the try-on representation learning. Furthermore, we introduce mutual learning to encourage the key two modules to transfer knowledge to each other. Extensive experiments conducted on the real-world dataset demonstrate the superiority of CTO-Net. In addition, we found that propagating the fine-grained relationships between items over the graph does greatly improve the FCM. Meanwhile, employing a teacher-student scheme for try-on representation learning is superior to existing studies. Moreover, transferring knowledge between the collocation and try-on compatibility modules is also helpful to boost the model performance. Currently, we only focus on the visual modality, but overlooking the textual context of each item. In the future, we plan to involve external information of items, such as text descriptions, to enhance the modal performance as well as interpretability.

ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China, No.: 62002090; the Key R&D Program of Shandong (Major scientific and technological innovation projects), No.:2020CXGC010111; the Shandong Provincial Natural Science Foundation, No.:ZR2019JQ23; CCF-Baidu Open Fund, No.: CCF-BAIDU OF2020019; the new AI project towards the integration of education and industry in QLUT.

REFERENCES

[1] Taleb Alashkar, Songyao Jiang, Shuyang Wang, and Yun Fu. 2017. Examples-Rules Guided Deep Neural Network for Makeup Recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI, 941–947.

[2] Suthee Chaidaroon, Yi Fang, Min Xie, and Alessandro Magnani. 2019. Neural Compatibility Ranking for Text-based Fashion Matching. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1229–1232.

[3] Zeyu Cui, Zekun Li, Shu Wu, Xiao-Yu Zhang, and Liang Wang. 2019. Dressing as a whole: Outfit compatibility learning based on node-wise graph neural networks. In *Proceedings of the ACM International Conference on World Wide Web*. ACM, 307–317.

[4] Xue Dong, Xuemeng Song, Fuli Feng, Peiguang Jing, Xin-Shun Xu, and Liqiang Nie. 2019. Personalized Capsule Wardrobe Creation with Garment and User Modeling. In *Proceedings of the 27th ACM International Conference on Multimedia*. ACM, 302–310.

[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] Zan Gao, Yin-ming Li, Wei-li Guan, Wei-zhi Nie, Zhi-yong Cheng, and An-an Liu. 2020. Pairwise view weighted graph network for view-based 3d model retrieval. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 129–138.

[7] 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.

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

[9] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).

[10] Linmei Hu, Siyong Xu, Chen Li, Cheng Yang, Chuan Shi, Nan Duan, Xing Xie, and Ming Zhou. 2020. Graph Neural News Recommendation with Unsupervised Preference Disentanglement. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. ACL, 4255–4264.

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

[12] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[13] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *International Conference on Learning Representations*. ICLR.

[14] Wei-Hong Li and Hakan Bilen. 2020. Knowledge Distillation for Multi-task Learning. In *Proceedings of the European Conference on Computer Vision*. Springer, 163–176.

[15] Yuncheng Li, Liangliang Cao, Jiang Zhu, and Jiebo Luo. 2017. Mining Fashion Outfit Composition Using an End-to-End Deep Learning Approach on Set Data. *IEEE Transactions on Multimedia* 19, 8 (2017), 1946–1955.

[16] Zhi Li, Bo Wu, Qi Liu, Likang Wu, Hongke Zhao, and Tao Mei. 2020. Learning the Compositional Visual Coherence for Complementary Recommendations. In *Proceedings of the International Joint Conference on Artificial Intelligence*. IJCAI, 3536–3543.

[17] Min Lin, Qiang Chen, and Shuicheng Yan. 2013. Network in network. *arXiv preprint arXiv:1312.4400*.

[18] Yen-Liang Lin, Son Tran, and Larry S. Davis. 2020. Fashion Outfit Complementary Item Retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3311–3319.

[19] Fan Liu, Zhiyong Cheng, Changchang Sun, Yinglong Wang, Liqiang Nie, and Mohan Kankanhalli. 2019. User Diverse Preference Modeling by Multimodal Attentive Metric Learning. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 1526–1534.

[20] Jinhuan Liu, Xuemeng Song, Zhaochun Ren, Liqiang Nie, Zhaopeng Tu, and Jun Ma. 2020. Auxiliary Template-Enhanced Generative Compatibility Modeling. In *Proceedings of the International Joint Conference on Artificial Intelligence*. IJCAI, 3508–3514.

[21] 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.

[22] 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.

[23] Julian 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.

[24] 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).

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

[26] Yehuda Koren Paolo Cremonesi and Roberto Turrin. 2010. Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the ACM Conference on Recommender Systems*. ACM, 39–46.

[27] Xuemeng Song, Fuli Feng, Xianjing Han, Xin Yang, Wei Liu, and Liqiang Nie. 2018. Neural compatibility modeling with attentive knowledge distillation. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 5–14.

[28] 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 International Conference on Multimedia*. ACM, 753–761.

[29] 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, 10373–10382.

[30] Mariya I. Vasileva, Bryan A. Plummer, Krishna Dusad, Shreya Rajpal, Ranjitha Kumar, and David Forsyth. 2018. Learning Type-Aware Embeddings for Fashion Compatibility. In *Proceedings of the European Conference on Computer Vision*. Springer, 390–405.

[31] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. 2019. Neural Graph Collaborative Filtering. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 165–174.

[32] Xionghui Wang, Jian-Fang Hu, Jian-Huang Lai, Jianguo Zhang, and Wei-Shi Zheng. 2019. Progressive Teacher-Student Learning for Early Action Prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3556–3565.

[33] Xiang Wang, Hongye Jin, An Zhang, Xiangnan He, Tong Xu, and Tat-Seng Chua. 2020. Disentangled Graph Collaborative Filtering. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1001–1010.

[34] Xin Wang, Bo Wu, and Yueqi Zhong. 2019. Outfit Compatibility Prediction and Diagnosis with Multi-Layered Comparison Network. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 329–337.

[35] Yinwei Wei, Xiang Wang, Weili Guan, Liqiang Nie, Zhouchen Lin, and Baoquan Chen. 2019. Neural multimodal cooperative learning toward micro-video understanding. *IEEE Transactions on Image Processing* 29 (2019), 1–14.

[36] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2019. MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM International Conference on Multimedia*. ACM, 1437–1445.

[37] Li Xingchen, Wang Xiang, 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.

[38] Xin Yang, Xuemeng Song, Xianjing Han, Haokun Wen, Jie Nie, and Liqiang Nie. 2020. Generative Attribute Manipulation Scheme for Flexible Fashion Search. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 941–950.

[39] Jun Yu, Hao Zhou, Yibing Zhan, and Dacheng Tao. 2021. Deep Graph-neighbor Coherence Preserving Network for Unsupervised Cross-modal Hashing. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI, 4626–4634.

[40] Mingkuan Yuan and Yuxin Peng. 2019. Ckd: Cross-task knowledge distillation for text-to-image synthesis. *IEEE Transactions on Multimedia* 22, 8 (2019), 1955–1968.

[41] Yibing Zhan, Jun Yu, Zhou Yu, Rong Zhang, Dacheng Tao, and Qi Tian. 2018. Comprehensive distance-preserving autoencoders for cross-modal retrieval. In *Proceedings of the ACM international conference on Multimedia*. ACM, 1137–1145.

[42] Liheng Zhang, Guo-Jun Qi, Liqiang Wang, and Jiebo Luo. 2019. AET vs. AED: Unsupervised Representation Learning by Auto-Encoding Transformations Rather Than Data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2547–2555.

[43] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. 2018. Deep mutual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 4320–4328.

[44] Na Zheng, Xuemeng Song, Zhaozheng Chen, Linmei Hu, Da Cao, and Liqiang Nie. 2019. Virtually Trying on New Clothing with Arbitrary Poses. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 266–274.

[45] Junyan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2223–2232.