Neurocomputing 414 (2020) 215–224

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

MGCM: Multi-modal generative compatibility modeling for clothing matching

Jinhuan Liu [a], Xuemeng Song [b,*], Zhumin Chen [b], Jun Ma [b]

[a] College of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266061, China
[b] School of Computer Science and Technology, Shandong University, Qingdao 266237, China

ARTICLE INFO

Article history:
Received 27 September 2019
Revised 10 March 2020
Accepted 9 June 2020
Available online 16 June 2020
Communicated by Xinmei Tian

Keywords:
Multi-modal information
Compatible template generation
Generative compatibility modeling

ABSTRACT

With the recent prevalence of online fashion-oriented communities and advances in multimedia processing, increasing research interests have been paid to the fashion compatibility modeling, where the compatibility between complementary fashion items (e.g., a top and a bottom) can be assessed automatically. Existing fashion compatibility modeling techniques mainly focus on measuring the compatible preference between fashion items with Deep Neural Networks (DNN), but overlook the generative compatibility modeling. Differently, in this paper, we explore the potential of the Generative Adversarial Network (GAN) in fashion compatibility modeling and thus propose a Multi-modal Generative Compatibility Modeling (MGCM) scheme. In particular, we introduce a multi-modal enhanced compatible template generation network, regularized by the pixel-wise consistency and template compatibility, to sketch a compatible template as the auxiliary link between fashion items. Accordingly, MGCM is able to measure the compatibility between complementary fashion items comprehensively from both item-item and item-template perspectives. Experimental results on two real-world datasets demonstrate the superiority of the proposed scheme over state-of-the-art methods.

1. Introduction

According to ShopifyPlus, 67% and 68% of fashion products were purchased online in UK and China in 2019, respectively[1]. With people's tremendous purchase for fashion products in e-commerce, there have been increasing research interests on the automatically fashion analysis techniques, especially the compatibility modeling among complementary fashion items, as it can facilitate many downstream applications, such as the complementary clothing matching [1,2] and the compatibility assessment [3,4]. Essentially, the fashion compatibility modeling works on automatically assessing the compatibility of a given set of complementary fashion items with different categories (e.g., the top, bottom and shoes), helping people avoid the trouble of consulting the professional stylist at great expense.

In a sense, existing fashion compatibility modeling methods mainly focus on learning the latent space with advanced Deep Neural Networks (DNN), where the compatible preference among fashion items can be measured based on their multi-modal representations (i.e., the visual encoding and the textual encoding) [5,6]. Nevertheless, most of them neglected the potential of Generative Adversarial Network (GAN) in fashion compatibility modeling, which has shown remarkable performance in various image translation tasks, such as edges to photos [7], and labels to facade [8]. In fact, GAN can help generating a compatible template (e.g., a bottom template) for a given item (e.g., a top) to enhance the compatibility modeling between fashion items from not only the conventional item-item perspective but also the auxiliary item-template angle. Motivated by this, in this work, to promote the fashion compatibility modeling, we study the generative compatibility modeling, where an auxiliary template generation network is introduced.

Without losing the generality, we focus on the general compatibility modeling between items of the two most essential fashion categories: the top and bottom. Nevertheless, the task is non-trivial due to the following challenges. 1) As the auxiliary template plays an important role in the compatibility modeling, especially from the item-template perspective, how to generate the realistic and compatible template for the given item to guide the compatibility modeling constitutes the first challenge. 2) How to seamlessly integrate the template generation to the fashion compatibility modeling to comprehensively measure the

* Corresponding author.
E-mail addresses: liujinhuan.sdu@gmail.com (J. Liu), sxmustc@gmail.com (X. Song), chenzhumin@sdu.edu.cn (Z. Chen), majun@sdu.edu.cn (J. Ma).

[1] https://www.shopify.com/enterprise/ecommerce-fashion-industry.

https://doi.org/10.1016/j.neucom.2020.06.033
0925-2312/

compatibility and thus boost the model performance is a crucial challenge. And 3) in addition to the visual images, the textual descriptions also convey important semantic features (e.g., the material and style) of fashion items. Accordingly, how to effectively fuse the multi-modal (i.e., the visual image and the textual description) information for both auxiliary template generation and compatibility modeling poses the last challenge.

To address these challenges, we propose a **M**ulti-modal **G**enerative **C**ompatibility **M**odeling (MGCM) scheme as shown in Fig. 1. Our proposed scheme works on enhancing the compatibility modeling between complementary fashion items with the auxiliary template generation. In particular, we introduce a complementary template generation network coupled with the pixel-wise consistency and template compatibility regularization to transfer the given fashion item to its compatible template. Based on the generated auxiliary template, MGCM is enabled to measure the fashion compatibility from both item-item and item-template views. To promote the performance, multiple modalities of fashion items are subtly fused in both template generation and compatibility modeling.

Our main contributions can be summarized in the following three-folds.

1) We propose a Multi-modal Generative Compatibility Modeling (MGCM) scheme, which is able to boost the performance of compatibility modeling with the auxiliary template generation.
2) We design a multi-modal enhanced compatible template generation network, regularized by the pixel-wise consistency and template compatibility regularization, to sketch a compatible template as the auxiliary link between fashion items.
3) Extensive experiments on two real-world datasets show that the generated templates are indeed helpful in guiding the compatibility modeling between complementary fashion items.

2. Related work

2.1. Generative models

In recent years, generative models, such as the Variational Autoencoder (VAE) [9] and GAN [10] have emerged for various image generation tasks. In a sense, variational methods optimize the lower bound of the logarithmic likelihood with probabilistic graphical models by introducing the deterministic bias [11]. Although VAE has shown its great power in various image generation tasks [12,13], it tends to generate blurry samples due to the minimization of the KL divergence between samples and the model [14]. Differently, a typical GAN [15], comprising a generator and a discriminator, works in the min–max optimization strategy. The generator tries to generate realistic samples with random noise, while the discriminator strives to distinguish it from the training data. Then, inspired by GAN, Conditional Generative Adversarial Network (CGAN) [16] was proposed to tackle the image-to-image translation problem, where the image mapping between different domains is learned. Currently, CGAN has received considerable attention from the computer vision research community, such as image synthesis [17], video generation [18] and person re-identification [19]. However, limited efforts have been dedicated to explore its great potential in the field of compatibility modeling, which is the major concern of our work.

2.2. Fashion compatibility modeling

Due to its huge economic value, fashion compatibility modeling has attracted tremendous research attention. For example, Han et al. [1] proposed a Bidirectional Long Short-Term Memory (Bi-LSTM) scheme that is able to sequentially predict the next fashion item conditioned on the existing ones. In order to utilize the rich fashion domain knowledge on clothing matching, Song et al. [5] introduced a knowledge-guided compatibility model for clothing matching. In addition, Vasileva et al. [6] introduced an end-to-end network to learn an embedding subspace, where the pair-wise similarity and compatibility can be jointly measured. Although these studies have achieved compelling success in compatibility modeling, they neglect the generative model, which can generate a compatible template as an auxiliary bridge between complementary fashion items and thus enhance the model performance. In fact, Liu et al. [20] introduced an Attribute-GAN framework to design a compatible bottom for the given top and bottom attributes, and thus make a proper collocation. Different from this work, Lin et al. [21] devised a variational co-supervision outfit recommendation framework in the context of recommending bottoms for a given top,where a bottom would be generated by VAE with the given top image and the desired bottom textural descriptions. Beyond that, to be more flexible in the practical application, we take a step forward and propose the multi-modal generative compatibility modeling framework for the clothing matching, where we devise the auxiliary complementary template generation with GAN simply based on the given top without the desired bottom description, which may be unavailable in practice.

3. Methodology

In this section, we detail our proposed MGCM, which is able to boost the performance of the compatibility modeling between complementary fashion items (e.g., a top and a bottom) with the auxiliary complementary template generation. We first formally

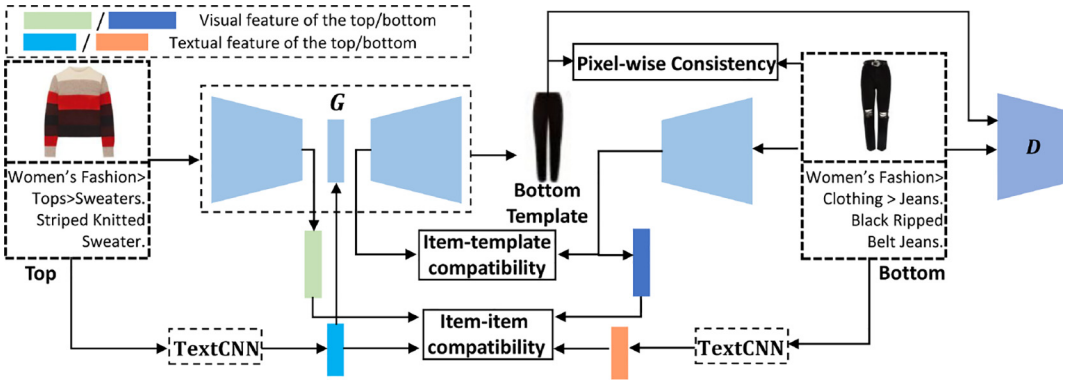


Fig. 1. Illustration of the proposed multi-modal generative compatibility modeling network.

give the *problem formulation* and then introduce the *multi-modal enhanced compatible template generation* network, and finally based on that present the *multi-modal generative compatibility modeling*.

3.1. Problem formulation

Suppose we have two fashion item domains: the top $\mathcal{T}$ and bottom $\mathcal{B}$, and a set of positive top–bottom pairs $\mathcal{P} = \{(t_{i_1}, b_{j_1}), (t_{i_2}, b_{j_2}), \ldots, (t_{i_M}, b_{j_M})\}$, where $t_{i_m} \in \mathcal{T}, b_{j_m} \in \mathcal{B}$, $m = 1, \ldots, M$. M refers to the total number of positive pairs. Each top t_i (bottom b_j) is associated with a visual image $\mathbf{I}_{t_i}$ ($\mathbf{I}_{b_j}$) and textural description $\mathbf{c}_{t_i}$ ($\mathbf{c}_{b_j}$). In this work, we focus on devising an end-to-end multi-modal generative compatibility modeling scheme $\mathcal{C}$ that is able to enhance the compatibility modeling between the top t_i and bottom b_j, by introducing the auxiliary template generation network G as follows:

$$\begin{aligned} &G(\mathbf{I}_{t_i}, \mathbf{c}_{t_i} | \Theta_G) \rightarrow \tilde{\mathbf{I}}_{b_i}; \\ &m_{ij} = \mathcal{C}(\mathbf{I}_{t_i}, \mathbf{c}_{t_i}, \mathbf{I}_{b_j}, \mathbf{c}_{b_j}, \tilde{\mathbf{I}}_{b_i} | \Theta_C), \end{aligned} \tag{1}$$

where m_{ij} denote the compatibility between the top t_i and bottom b_j. Θ_G and Θ_C are the sets of to-be-learned parameters of our scheme.

3.2. Multi-modal enhanced compatible template generation

3.2.1. Complementary template generation

As the compatible template generation is essentially an image-to-image (i.e., top-to-bottom) translation task, we can naturally adopt CGAN as the backbone of our compatible template generation network, which has made remarkable achievements in various image-to-image translation problems [16], such as the attributes-to-images [22], image synthesis [23], and face photos-to-emoji [24]. In our context, the generator $G_{\mathcal{T}\rightarrow\mathcal{B}}$ of CGAN aims to translate the given top $\mathbf{I}_{t_i}$ of the source domain $\mathcal{T}$ to a compatible bottom template $\tilde{\mathbf{I}}_{b_i}$ of the target domain $\mathcal{B}$ as follows:

$$G_{\mathcal{T}\rightarrow\mathcal{B}}(\mathbf{I}_{t_i} | \Theta_G) \rightarrow \tilde{\mathbf{I}}_{b_i}, \tag{2}$$

where Θ_G refers to the set of parameters in the generator $G_{\mathcal{T}\rightarrow\mathcal{B}}$.

In fact, the traditional real-fake discriminator of the standard GAN can only enforce the generator to produce realistic bottom images, which could be incompatible to the given top. In our context, as we expect the generator to synthesize compatible bottom templates as the guidance for compatibility modeling, we introduce the discriminator $D_{\mathcal{B}}$ working on distinguishing the real compatible bottom $\mathbf{I}_{b_j}$ from the generated bottom template $\tilde{\mathbf{I}}_{b_i}$, when a top $\mathbf{I}_{t_i}$ is given as a condition. Therefore, inspired by the CGAN [16], we define the min–max objective function of template generation and discrimination:

$$\begin{aligned} \min_G \max_D \mathcal{L}_{CGAN}(G_{\mathcal{T}\rightarrow\mathcal{B}}, D_{\mathcal{B}}) = &\mathbb{E}_{\mathbf{I}_{t_i}, \mathbf{I}_{\mathbf{b_j}}}[\log D_{\mathcal{B}}(\mathbf{I}_{t_i}, \mathbf{I}_{b_j})] \\ &+ \mathbb{E}_{\mathbf{I}_{\mathbf{t_i}}}[\log(1 - D_{\mathcal{B}}(\mathbf{I}_{t_i}, \tilde{\mathbf{I}}_{b_i}))]. \end{aligned} \tag{3}$$

3.2.2. Multi-modal enhanced compatible template generation

Obviously, the above CGAN-based compatible template generation only takes into account the visual image of the given top, but ignores the great value of the textural description in the compatible template generation that also conveys important cues (e.g., the material and style) regarding the given fashion item. Toward this end, we take one step forward and propose the multi-modal enhanced compatible template generation. Inspired by [7], we redevise a generator $G_{\mathcal{T}_{vc}\rightarrow\mathcal{B}}$ with three components: down-sampling, multi-modal fusion and up-sampling, as shown in Fig. 2.

Specifically, given a top t_i, the down-sampling first learns its visual encoding based on its visual image $\mathbf{I}_{t_i}$ with several convolution layers as follows:

$$\mathbf{H}_k = \phi(\mathbf{W}_k \mathbf{H}_{k-1} + \mathbf{b}_k), \quad k = 0, 1, \ldots, K, \tag{4}$$

where $\Theta_{ds} = \{\mathbf{W}_k, \mathbf{b}_k | \ k = 0, 1, \ldots, K\}$ refers to the parameters for the down-sampling and $\phi(.)$ stands for the Leaky ReLU (LReLU) [25] activation function. In our task, we set $K = 6, \mathbf{H}_0 = \mathbf{I}_{t_i}$ as the input, and $\mathbf{H}_K \in \mathbb{R}^{w* \ h* \ c}$ as the output, where $w* \ h* \ c$ represents the corresponding shape.

To facilitate the multi-modal fusion toward the template generation, we reshape the $\mathbf{H}_K$ to a vector $\hat{\mathbf{v}}_{t_i} \in \mathbb{R}^d$, where $d = w \times h \times c$. Regarding the textual modality, we first embed each word with a 300-D vector by applying the pre-trained word2vector [26]. Then, we adopt the TextCNN [27], which has achieved astonishing success in various tasks, like the multi-modal recommendation [28] and Natural Language Processing (NLP) [29]. In particular, we employ 100 kernels for each size of $\{2, 3, 4, 5\}$. Accordingly, we map the textual description of the top t_i to the textual encoding $\hat{\mathbf{c}}_{t_i}(\hat{\mathbf{c}}_{b_j}) \in \mathbb{R}^{400}$.

To fulfill the multi-modal fusion, we first concatenated the visual encoding $\hat{\mathbf{v}}_{t_i}$ and textual encoding $\hat{\mathbf{c}}_{t_i}$. Then, we further employ the fully-connected layer to map the fusion encoding as follows:

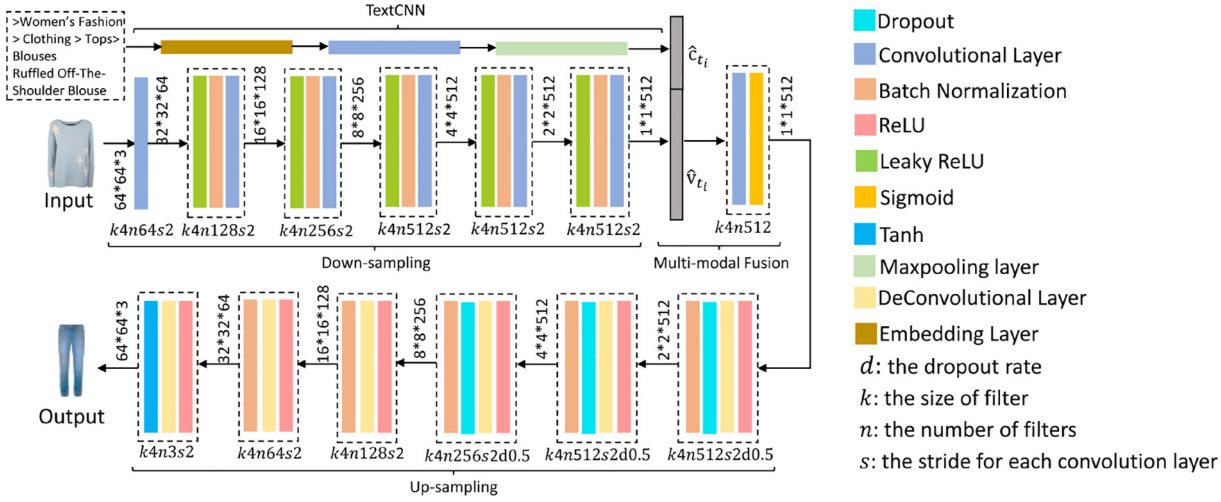


Fig. 2. Illustration of our generator architecture, which is able to generate the complementary bottom template for a given top with multi-modalities.

$$\mathbf{p}_{vc} = \sigma(\mathbf{W}_p[\hat{\mathbf{v}}_{t_i}; \hat{\mathbf{c}}_{t_i}] + \mathbf{b}_p), \tag{5}$$

where $\Theta_p = \{\mathbf{W}_p, \mathbf{b}_p\}$ stands for the parameters of the multi-modal fusion network and $\sigma(.)$ represents the sigmoid activation function. By reshaping the projected feature $\mathbf{p}_{vc} \in \mathbb{R}^d$, we obtain the final encoding of the top $t_i, \mathbf{P}_{vc} \in \mathbb{R}^{w* \ h* \ c}$, for the following up-sampling toward bottom template generation. The up-sampling component is devised to translate the multi-modal feature $\mathbf{P}_{vc}$ to the bottom template $\tilde{\mathbf{I}}_{b_i}$ through multiple deconvolution layers with the parameter Θ_{us}. Ultimately, the generator $G_{\mathcal{T}_{vc}\to\mathcal{B}}$ transforms the given top with multi-modalities in the source domain $\mathcal{T}$ to a bottom template $\tilde{\mathbf{I}}_{b_i}$ in the target domain $\mathcal{B}$ with parameters $\Theta_G = \{\Theta_{ds}, \Theta_p, \Theta_{us}\}$.

Simply applying the cross entropy function mentioned above may encounter the vanishing gradients problem in the process of updating the generator [30]. Therefore, to guarantee the training stability and image generation quality [31], we adopt the least square loss rather than the min–max objective function in the Eqn.(3). Then, we have the objective function for our multi-modal enhanced compatible template generation network as follows:

$$\begin{cases} \min\limits_{D_{\mathcal{B}}} \mathcal{L}(D_{\mathcal{B}}) = \frac{1}{2}\mathbb{E}_{\mathbf{I}_{t_i}, \mathbf{I}_{\mathbf{b_j}}}[(D_{\mathcal{B}}(\mathbf{I}_{t_i}, \mathbf{I}_{b_j}) - 1)^2] \\ \qquad\qquad + \frac{1}{2}\mathbb{E}_{\mathbf{I_{ti}}}[(D_{\mathcal{B}}(\mathbf{I}_{t_i}, \tilde{\mathbf{I}}_{b_i}) - 0)^2], \\ \min\limits_{G_{\mathcal{T}_{vc}\to\mathcal{B}}} \mathcal{L}(G_{\mathcal{T}_{vc}\to\mathcal{B}}) = \frac{1}{2}\mathbb{E}_{\mathbf{I_{ti}}}[(D_{\mathcal{B}}(\mathbf{I}_{t_i}, \tilde{\mathbf{I}}_{b_i}) - 1)^2]. \end{cases} \tag{6}$$

As we expect the generated template would guides the compatibility modeling, we argue that the generated bottom template $\tilde{\mathbf{I}}_{b_i}$ should be compatible with the given top $\mathbf{I}_{t_i}$. Therefore, we introduce the pixel-wise consistency to regularize the low-level difference between the generated bottom $\tilde{\mathbf{I}}_{b_i}$ and the positive one $\mathbf{I}_{b_j}$, which can be defined with the L_1 distance as follows:

$$\mathcal{L}_{pixel} = \left\| \tilde{\mathbf{I}}_{b_i} - \mathbf{I}_{b_j} \right\|_1. \tag{7}$$

3.3. Multi-modal generative compatibility modeling

Based on the above multi-modal enhanced compatible template generation, we can proceed to the compatibility modeling between fashion items, where we take into account the generated bottom template $\tilde{\mathbf{I}}_{b_i}$ and thus model the compatibility between the fashion items from both the item-item and item-template perspectives. Fig. 3 illustrates the workflow of our proposed MGCM scheme.

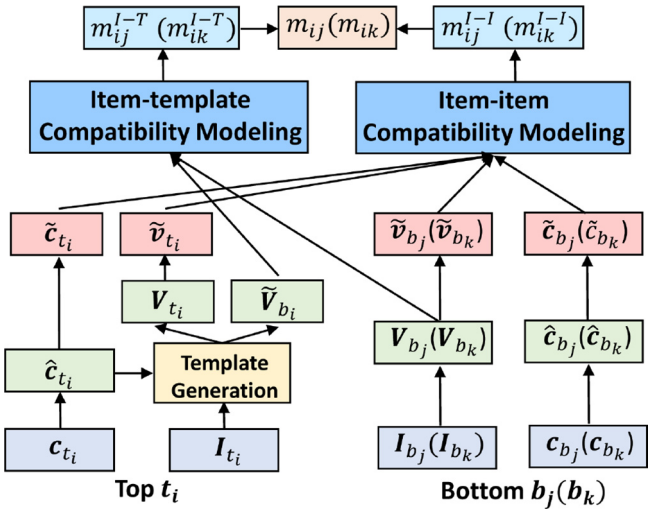


Fig. 3. Workflow of our proposed multi-modal generative compatibility modeling.

3.3.1. Item-item compatibility

To measure the item-item compatibility, we aim to seek the latent representations of fashion items that well support the compatible preference modeling between fashion items. In particular, we first define the output of the $(K-1)$th layer of the down-sampling as the visual representation $\mathbf{V}_{t_i} \in \mathbb{R}^{m* \ n* \ l}$ of top t_i, where $m* \ n* \ l$ represents the shape of the representation. In a similar manner, we can get the visual representation $\mathbf{V}_{b_j} \in \mathbb{R}^{m* \ n* \ l}$ of the bottom b_j. In addition, We adopt the output of the first deconvolution layer of the up-sampling as the visual representation $\tilde{\mathbf{V}}_{b_i} \in \mathbb{R}^{m* \ n* \ l}$ of the generated bottom template $\tilde{\mathbf{I}}_{b_i}$. Moreover, to well exploit the latent feature $\tilde{\mathbf{v}}_{t_i} \in \mathbb{R}^{D_v}$ of top t_i, we first adopt the global average pooling (GAP) to convert $\mathbf{V}_{t_i}$ to $\mathbf{v}_{t_i} \in \mathbb{R}^l$ and further project it as follows:

$$\tilde{\mathbf{v}}_{t_i} = \sigma(\mathbf{W}_v \mathbf{v}_{t_i} + \mathbf{h}_v), \tag{8}$$

where $\mathbf{W}_v \in \mathbb{R}^{D_v \times d}$ and $\mathbf{h}_v \in \mathbb{R}^{D_v}$ refer to the corresponding parameters. In the similar manner, we can get the latent representation $\tilde{\mathbf{v}}_{b_j}$ of the bottom b_j. For the textual features, we can obtain $\tilde{\mathbf{c}}_{t_i}(\tilde{\mathbf{c}}_{b_j}) \in \mathbb{R}^{D_t}$ with Eq. (8). Therefore, the item-item compatibility can be calculated as follows:

$$m_{ij}^{I-I} = \alpha(\tilde{\mathbf{v}}_{t_i})^T \tilde{\mathbf{v}}_{b_j} + (1-\alpha)(\tilde{\mathbf{c}}_{t_i})^T \tilde{\mathbf{c}}_{b_j}, \tag{9}$$

where α is the trade-off parameter to balance the importance of the compatibility measurement with different modalities.

3.3.2. Item-template compatibility

Different from existing fashion compatibility modeling techniques mainly focus on measuring the compatible preference between fashion items with deep neural networks, we further take the generative compatibility modeling into consideration. We argue that the compatible bottoms for the given top should share similar high-level attributes with the generated bottom template. Accordingly, we design the template compatibility regularization to measure the high-level similarity between the bottom b_j and the generated bottom template from the auxiliary item-template perspective:

$$m_{ij}^{I-T} = \left\| \tilde{\mathbf{V}}_{b_i} - \mathbf{V}_{b_j} \right\|_1, \tag{10}$$

where m_{ij}^{I-T} refers to the item-template compatibility. $\tilde{\mathbf{V}}_{b_i}$ and $\mathbf{V}_{b_j}$ represent the high-level visual representation of the generated bottom template and the positive bottom, respectively.

3.3.3. Compatibility modeling

Combining the item-item compatibility and item-template compatibility, the multi-modal template-enhanced compatibility score m_{ij} between the top t_i and bottom b_j can be defined as follows:

$$m_{ij} = m_{ij}^{I-I} + \beta m_{ij}^{I-T}, \tag{11}$$

where β is a hyper-parameter controlling the importance of each compatibility.

In a sense, we can easily derive the positive top–bottom pairs from those have been composed together by fashion experts. However, regarding the non-composed fashion item pairs, we cannot draw the conclusion that they are incompatible as they can be the missing potential positive pairs that can be composed in the future. Toward this end, similar to [32], to accurately model the implicit relationship between fashion items, we adopt the BPR [33] framework by introducing the following training dataset of triplets:

$$\mathcal{E} := \{(i,j,k) | (t_i, b_j) \in \mathcal{P}, b_k \in \mathcal{B} \setminus b_j\}, \tag{12}$$

where the triplet (i,j,k) indicates that the top–bottom pair (t_i, b_j) in the positive top–bottom set $\mathcal{P}$ is more compatible than the pair (t_i, b_k). Notably, the bottom b_k is randomly sampled from the whole set of bottoms $\mathcal{B}$. Then, we define the following objective function:

$$\mathcal{L}_{BPR} = -ln(\sigma(m_{ij} - m_{ik})), \tag{13}$$

where m_{ik} represents the compatibility between the top t_i and bottom b_k corresponding to that defined by the Eq. (11) in our paper. In a sense, we aim to push the given top closer to the positive bottom, but far away from the negative bottom.

3.4. Optimization

Overall, the objective function of our proposed multi-modal generative compatible preference modeling can be defined in an end-to-end manner. Thus we have:

$$\mathcal{L} = \mathcal{L}_{BPR} + \mu\mathcal{L}(G_{\mathcal{T}_{vc}\to\mathcal{B}}) + \nu\mathcal{L}(D_{\mathcal{B}}) + \gamma\mathcal{L}_{pixel} + \delta\|\Theta_C\|^2. \tag{14}$$

where $\Theta_C = \{\Theta_G, \Theta_D, \Theta_{BPR}\}$ refers to the parameters of our proposed network. γ, δ, μ, and ν are the hyper-parameters controlling the strength of different components of the proposed MGCM.

Algorithm 1: Multi-modal Generative Compatibility Modeling (MGCM) training procedure.

Input: A set of paired top–bottom fashion items $\mathcal{P}$ with top in domain $\mathcal{T}$ and bottom in domain $\mathcal{B}$, generator with parameters Θ_G, discriminator with parameters Θ_D, BPR with parameters Θ_{BPR}, learning rate η, hyper-parameters γ, δ, μ, ν.

Output: Parameters $\mathbf{\Theta}_C = \{\Theta_G, \Theta_D, \Theta_{BPR}\}$.

1: Initialize parameters $\Theta_G, \Theta_D, \Theta_{BPR}$.
2: **repeat**
3: Randomly draw the set (i,j,k) from $\mathcal{P}$ according to the Eqn. (12).
4: Construct the MGCM according to Eqn. (14).
5: **for** each parameter θ in $\mathbf{\Theta}_C$ **do**
6: Update $\theta_D \leftarrow \theta_D + \eta\nabla_{\theta_D}(\nu\mathcal{L}(D_{\mathcal{B}}) + \gamma\mathcal{L}_{pixel} + \delta\theta_D)$.
7: Update $\theta_G \leftarrow \theta_G + \eta\nabla_{\theta_G}(\mu\mathcal{L}(G_{\mathcal{T}_{vc}\to\mathcal{B}}) + \delta\theta_G)$.
8: Update $\theta_{BPR} \leftarrow \theta_{BPR} + \eta\nabla_{\theta_{BPR}}(\mathcal{L}_{BPR} + \delta\theta_{BPR})$.
9: **end for**
10: **until** Converge
11: Return $\mathbf{\Theta}_C$.

We adopt the Adam [34] and Stochastic Gradient Descent (SGD) [35] optimizer to train the generator and discriminator, respectively. The optimization procedure is shown in Algorithm 1.

4. Experiments

We evaluate our proposed MGCM on the following research questions:

- **RQ1.** How does the proposed MGCM perform as compared to state-of-the-art methods?
- **RQ2.** How do different modalities contribute to the model performance?
- **RQ3.** What is the effect of each component in our framework?

4.1. Dataset and experimental settings

4.1.1. DataSet

To well demonstrate the effectiveness of our proposed MGCM, we experiment on two public real-world datasets: FashionVC [32] and ExpFashion [36], where each item is associated with both a visual image and the textual description. FashionVC consists of 20,726 outfits with 14,870 tops and 13,662 bottoms, while ExpFashion is comprised of 853,991 outfits with 168,682 tops and 117,668 bottoms. Both datasets are crawled from the fashion sharing website (i.e., Polyvore). Notably, for a comparable evaluation, we randomly select 20,000 outfits from ExpFashion instead of using the whole dataset.

4.1.2. Evaluation metric

We adopt the area under the ROC curve (AUC) [37] and the mean reciprocal rank (MRR) [38] as the evaluation metrics to tune hyper-parameters and evaluate the performance. On one hand, we define the AUC metric as follows:

$$AUC = \frac{1}{|\mathcal{P}|}\sum_{(i,j)\in\mathcal{P}}\frac{1}{|\mathcal{E}(i,j)|}\sum_{k\in\mathcal{E}}\delta(m_{ij}, m_{ik}), \tag{15}$$

where $\mathcal{E}(i,j)$ indicates that the top and positive bottom pairs present in $\mathcal{P}$. $\delta(m_{ij}, m_{ik}) = 1$ when the compatibility between the positive top–bottom pair surpasses the negative one (i.e., $m_{ij} > m_{ik}$), and $\delta(m_{ij}, m_{ik}) = 0$ otherwise. On the other hand, we express the MRR metric as follows:

$$MRR = \frac{1}{|\mathcal{P}|}\sum_{n=1}^{|\mathcal{P}|}\frac{1}{R_n}, \tag{16}$$

where R_n refers to the ranking position of the positive bottom for the nth top.

4.1.3. Implementation details

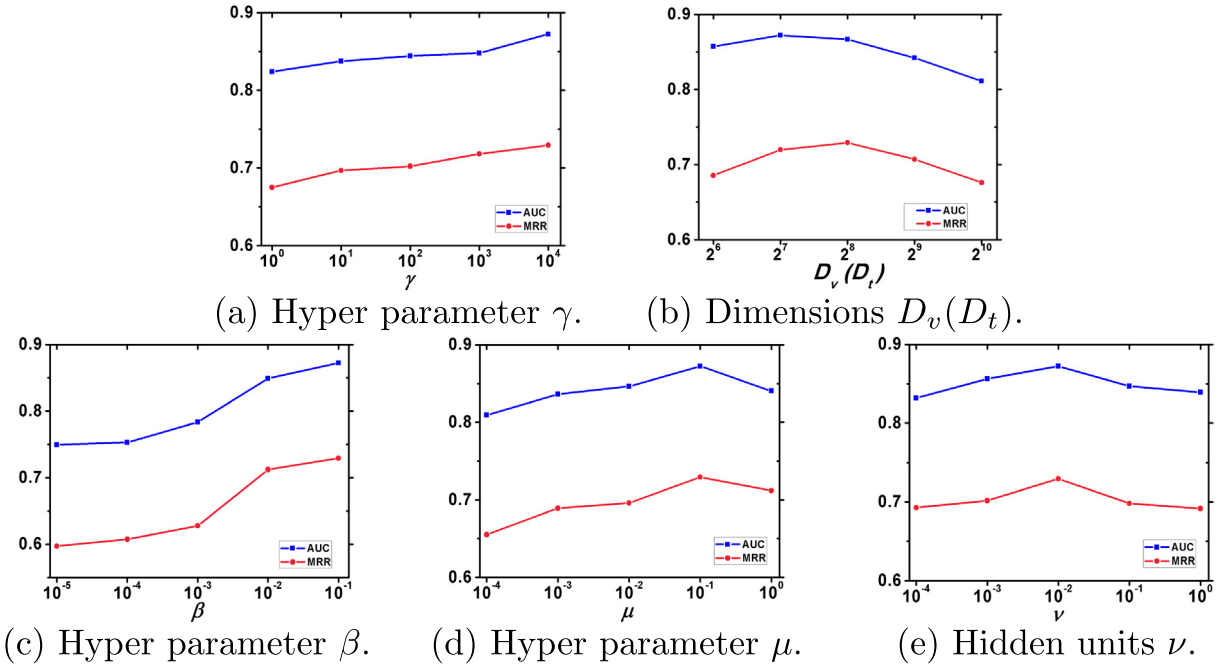
For each dataset, we randomly scramble the outfits (i.e., top–bottom pairs) and leverage the first 80% as the training set, the following 10% as the validation set, and the last 10% as the testing set. For each top–bottom pair in these three subsets, we randomly choose 3 and 9 negative bottoms according to Eq. (12) for the tasks of compatibility modeling and the complementary fashion item retrieval, respectively. To adjust the hyper-parameters, we adopt the gird search strategy with the validation set and obtain the optimal performance on the test set. As shown in Fig. 4, the optimal experimental results are achieved with the $\gamma = 10,000$ (10,000), the dimension of the final representation $D_v = 128$ ($D_t = 256$), $\beta = 0.1$ (0.1), $\mu = 0.1$ (0.1) and $\nu = 0.01$ (0.01) for AUC (MRR). Following with [7], we set η as 0.0002.

We first verified the cost and accuracy convergence of our proposed model, which is also illustrated in many deep learning methods [39,40]. We show the cures of the training loss in Eq. (14) (black solid line) and the training AUC in Eq. (15) (blue dashed line) in Fig. 5. We can see that the two values first change sharply within a few iterations and then tend to be stable, which well validates the convergence of our model.

4.2. Comparison on different models (RQ1)

Our MGCM is compared with following state-of-the-art baselines:

- **POP**: The compatibility between the top t_i and bottom b_j is measured by the number of bottoms that have been matched with the top in the positive top–bottom set $\mathcal{P}$.
- **Bi-LSTM [1]**: Bi-LSTM is designed to sequentially recommend the complementary fashion items for existing items of an outfit. In our context, we adapt this method to deal with the outfit that simply consists of two fashion items (i.e., a top and a bottom).
- **IBR [41]**: This approach models the relationships between fashion items in a latent style space only with the visual information.



(a) Hyper parameter γ. (b) Dimensions $D_v(D_t)$.

(c) Hyper parameter β. (d) Hyper parameter μ. (e) Hidden units ν.

Fig. 4. Illustration of the AUC and MRR values of MGCM with varying hyper-parameters.

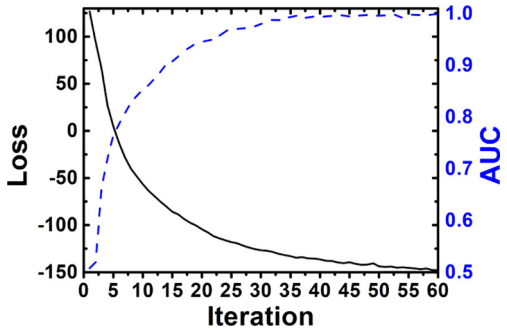


Fig. 5. Illustration of the training cost and accuracy convergence of our proposed MGCM.

- **IBR-VC**: We extend IBR to enable it to measure the compatibility between fashion items with both the visual and textual information. Specifically, we derive IBR-VC from IBR by adding TextCNN, which is the textual feature encoding method used in our MGCM framework.
- **BPR-DAE** [32]: BPR-DAE is a content-based clothing matching scheme, which jointly models the compatible preferences of fashion items with multi-modalities and the coherent relation among different modalities of the same item. To make a fair comparison, we adapt this method to encode the visual and textual representations with Alexnet and TextCNN in an end-to-end manner.
- **FARM** [21]: This baseline aims to fulfil the outfit recommendation with a generated fashion item, where the VAE is adopted to generate a bottom image given a top and desired bottom description. In our context, we remove the input of the bottom text description to accommodate more flexible applications.
- **CycleGAN-CM** [42]: We replace the generative adversarial network in MGCM with the CycleGAN, which is devised to address the unsupervised image-to-image translation problem with unpaired training data based on the forward and backward cycle-consistency networks.
- **Pix2pix-CM** [7]: We substitute the template generation network with Pix2pix, whose generator is constructed with the U-Net [43] and discriminator is devised to distinguish the generated bottom template and the real one for the given top.

Table 1 shows the performance comparison among state-of-the-art baselines in terms of AUC and MRR on FashionVC and ExpFashion, respectively. From this table, we have the following observations. 1) Our approach significantly outperforms all baselines, which verifies the effectiveness of our proposed MGCM. 2) POP achieves the worst performance, which may be due to the fact that this method overlooks the valuable item content, such as the visual and textural features of fashion items. 3) Compared with other non-generative compatibility modeling methods, Bi-LSTM performs worse. One possible reason is that the sequential recommendation method maybe not suitable for the compatibility modeling with only two items. 4) MGCM outperforms IBR, IBR-VC and BPR-DAE, indicating that it is advisable to incorporate the template generation to facilitate the compatibility modeling from the item-template perspective apart from the conventional item-item perspective. 5) Unexpectedly, FARM performs worse even than the non-generative methods, implying that FARM is highly depend on the desired bottom description and cannot well fulfil the compatibility modeling in our context, where the bottom description is

Table 1
Performance comparison of different models in terms of AUC and MRR on FashionVC and ExpFashion.

	FashionVC		ExpFashion	
Approach	AUC	MRR	AUC	MRR
POP	0.4364	0.1989	0.3823	0.2130
Bi-LSTM	0.5464	0.3299	0.5298	0.3261
IBR	0.6189	0.4391	0.6029	0.3715
IBR-VC	0.6807	0.4548	0.6591	0.4159
BPR-DAE	0.7826	0.6214	0.7454	0.5893
FARM	0.5842	0.3710	0.5540	0.3250
CycleGAN-CM	0.8292	0.6884	0.8243	0.6872
Pix2pix-CM	0.8341	0.6932	0.8265	0.6895
MGCM	**0.8724**	**0.7293**	**0.8592**	**0.7169**

unnecessary. 6) Our MGCM surpasses CycleGAN-CM, which may be attributed to the fact that MGCM takes into account both modalities, while CycleGAN-CM neglects the importance of textual modality for the template generation. And 7) MGCM achieves better performance than Pix2pix-CM, where the U-Net structure is considered. One possible explanation is that the U-Net structure is easier to learn the low-level information, which affects the experimental results.

4.3. Comparison of different modalities (RQ2)

To demonstrate the advantages of incorporating the multi-modalities in the compatibility modeling, we introduce two derivatives of our MGCM: MGCM-V and MGCM-T, where only the visual and textual modality is incorporated by MGCM, respectively. Table 2 provides the evaluation results of different modalities in terms of AUC and MRR on the two datasets. From this table, we can make the following observations. 1) Our MGCM can boost the performance with the multi-modal information. Specifically, our method outperforms the MGCM-V by 6.91% and 5.34% in terms of AUC and MRR, respectively. This demonstrates that the visual and textural modalities complement each other toward the compatibility modeling. And 2) MGCM-V is superior to MGCM-T, suggesting that the visual modality captures more intuitive features (e.g., color, pattern and clipping) of fashion items than the textural modality, and hence contributes more in the fashion compatibility modeling.

To further study the effect of different modalities in the item-item compatibility modeling, Fig. 6 shows the performance of our model with respect to the parameter α in Eq. (9) on both the AUC and MRR metrics. As can be seen, MGCM achieves the optimal performance at $\alpha = 0.5$ on AUC and $\alpha = 0.6$ on MRR, which implies that both modalities are comparably important for the item-item compatibility modeling. To some extent, this also reflects that the superior performance of MGCM-V over MGCM-T, which can be attributed to that the visual modality contributes more in the auxiliary template generation component of MGCM as compared with the textural modality.

Table 2
Performance comparison of different modalities in terms of AUC and MRR on FashionVC and ExpFashion.

	FashionVC		ExpFashion	
Approach	AUC	MRR	AUC	MRR
MGCM-V	0.8160	0.6759	0.8126	0.6652
MGCM-T	0.6578	0.5933	0.6179	0.4379
MGCM	**0.8724**	**0.7293**	**0.8592**	**0.7169**

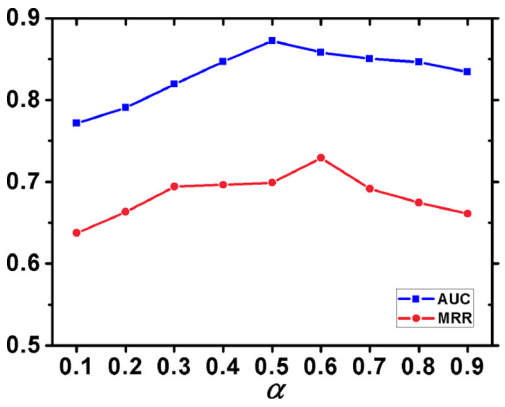


Fig. 6. The performance of our model with respect to the parameter α in terms of AUC and MRR.

In addition to the quantitative evaluation, we further visualize the generated bottom templates of different generative models (i.e., FARM, Pix2pix-CM, CycleGAN-CM, and MGCM). As shown in Fig. 7, we find that our MGCM can generate more realistic and compatible bottom templates compared with other generative baselines, especially regarding sketching the color and shape attributes of the templates. Interestingly, although our MGCM also fails to capture the item texture well, it can still improve the performance of compatibility modeling significantly. This suggests that the color and shape are the key factors affecting the compatibility assessment and reconfirms the necessity of the template generation.

Meanwhile, to gain more deep insights of our model, we list several unsuccessful template generation cases in Fig. 8. As can be seen, for the group (a), the generated bottom templates are incorrect or blurry in terms of the item shape, which may be due to the unusually angle of the ground truth image that makes it difficult for our MGCM to generate the bottom template. Pertaining to the group (b), although our MGCM can sketch the item shape properly, it fails to render the complex pattern details for the template. As for the group (c), the ground truth bottoms are folded, which hinders the template generation of our MGCM.

4.4. Comparison of different components (RQ3)

To get a thorough understanding of our model, we study the effects of different components in terms of AUC and MRR. We

Fig. 7. Bottom templates generated by different generative models. GT: ground truth.

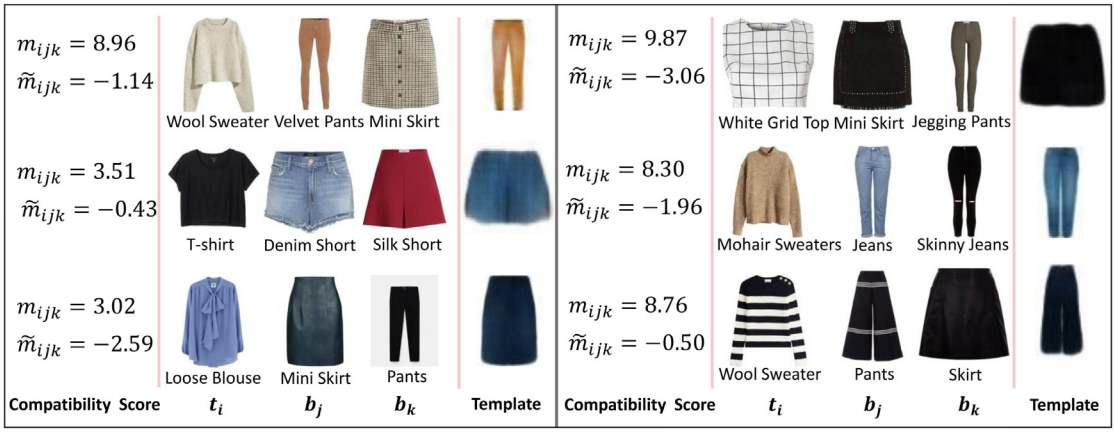
Fig. 8. Failure samples in our dataset.

Table 3
Performance comparison of our MGCM with different component configurations in terms of AUC and MRR on FashionVC and ExpFashion.

	FashionVC		ExpFashion	
Approach	AUC	MRR	AUC	MRR
-noPixel	0.8173	0.6681	0.7920	0.6481
-noTemG	0.7615	0.6546	0.7411	0.6305
MGCM	**0.8724**	**0.7293**	**0.8592**	**0.7169**

adapt our method to -nopixel and -noTemG by setting the hyper-parameters γ to 0 and disabling the multi-modal bottom template generation network, respectively. Table 3 shows the results of our MGCM with different component configurations in terms of AUC and MRR. An interesting observation is that MGCM outperforms -nopixel and -noTemG, indicating both the proposed pixel-wise consistency regularization and the multi-modal enhanced compatible template generation network can boost the performance. In addition, -nopixel performs better than -noTemG, indicating that the multi-modal enhanced template generation plays a more important role in our framework than the pixel-wise consistency regulation.

Fig. 9 visualizes the performance of different models in the tasks of generative compatibility modeling and complementary clothing retrieval, respectively. As shown in Fig. 9(a), the compatible preference of all triplets are correctly identified by MGCM but not -nopixel. In fact, we noticed that for each triplet, the given top seems to be compatible with both the positive and negative bottoms, which may lead the incorrect modeling results of -nopixel. Therefore, incorporating the pixel-wise consistency regularization between the generated template and the bottom candidate,



(a) Compatibility modeling results of -nopix and MGCM, where the corresponding compatibility scores are represented as $\tilde{m}_{ijk} = \tilde{m}_{ij} - \tilde{m}_{ik}$ and $m_{ijk} = m_{ij} - m_{ik}$, respectively. Notably, the preference of all the triplets is that the top t_i should go better with the bottom b_j as compared with the bottom b_k.

(b) Complementary clothing retrieval results of -noTemG and MGCM, where the fashion items highlighted in the red boxes are the ground truth.

Fig. 9. Performance illustration of different models in the tasks of generative compatibility modeling and complementary clothing retrieval, respectively.

namely, taking the generated template as a reference, MGCM is able to distinguish the positive bottom b_j for the given top t_i from the negative one and provides the correct m_{ijk}. In addition, as we can see from Fig. 9(b), -noTemG fails to rank the positive bottoms at the first place, which is corrected by the complete MGCM that takes into account the bottom template generation. Notably, the generated templates do provide the reasonable guidance for ranking the positive bottoms. Ultimately, these observations indeed validate that both the proposed pixel-wise consistency regularization and the auxiliary multi-modal bottom template generation in our MGCM are helpful to improve the model preference in different tasks.

5. Conclusion

In this paper, we propose a multi-modal generative compatibility modeling (MGCM) network, which is able to boost the performance of compatibility modeling between fashion items (e.g., a top and a bottom) with the auxiliary template generation. Specifically, we introduce the multi-modal enhanced compatible template generation network to sketch a compatible template (e.g., a bottom template) for the give fashion item (e.g., a top) with the pix-wise consistency and template compatibility regularization. Our proposed MGCM is able to model the compatibility preference from both the item-item and item-template perspectives. Extensive experiments on two public real-world datasets show that (1) the generated templates are indeed helpful in guiding the compatibility modeling between complementary fashion items; and (2) the pixel-wise consistency regularization does promote the compatibility modeling performance. Currently, our model only measures the compatibility between two fashion items. In the future, we plan to devise more advanced scheme to model the compatibility among multiple fashion items.

CRediT authorship contribution statement

Jinhuan Liu: Conceptualization, Methodology, Software, Formal analysis, Data curation, Investigation, Visualization, Writing - original draft. **Xuemeng Song:** Methodology, Validation, Formal analysis, Investigation, Writing - review & editing, Software, Funding acquisition. **Zhumin Chen:** Validation, Investigation, Resources, Supervision, Funding acquisition. **Jun Ma:** Writing - review & editing, Supervision, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is supported by the National Natural Science Foundation of China, No.: 61702300, 61672324, 61972234, 61902219 and 61672322; the Future Talents Research Funds of Shandong University, No.: 2018WLJH63.

References

[1] X. Han, Z. Wu, Y.-G. Jiang, L. S. Davis, Learning fashion compatibility with bidirectional lstms, in: MM, 2017, pp. 1078–1086.
[2] C.P. Huynh, A. Ciptadi, A. Tyagi, A. Agrawal, Craft: complementary recommendations using adversarial feature transformer, arXiv preprint arXiv:1804.10871.
[3] G. Cucurull, P. Taslakian, D. Vazquez, Context-aware visual compatibility prediction, arXiv preprint arXiv:1902.03646.
[4] Y. Li, L. Cao, J. Zhu, J. Luo, Mining fashion outfit composition using an end-to-end deep learning approach on set data, TMM 19 (8) (2017) 1946–1955.
[5] X. Song, F. Feng, X. Han, X. Yang, W. Liu, L. Nie, Neural compatibility modeling with attentive knowledge distillation, SIGIR (2018) 5–14.
[6] M.I. Vasileva, B.A. Plummer, K. Dusad, S. Rajpal, R. Kumar, D. Forsyth, Learning type-aware embeddings for fashion compatibility, ECCV (2018) 390–405.
[7] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, CVPR (2017) 1125–1134.
[8] Z. Yi, H. Zhang, P. Tan, M. Gong, Dualgan: unsupervised dual learning for image-to-image translation, ICCV (2017) 2849–2857.
[9] Y. Pu, Z. Gan, R. Henao, X. Yuan, C. Li, A. Stevens, L. Carin, Variational autoencoder for deep learning of images, labels and captions, NIPS (2016) 2352–2360.
[10] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein generative adversarial networks, ICML (2017) 214–223.
[11] D.P. Kingma, M. Welling, Auto-encoding variational bayes, arXiv preprint arXiv:1312.6114.
[12] T. Ma, J. Chen, C. Xiao, Constrained generation of semantically valid graphs via regularizing variational autoencoders, NIPS (2018) 7113–7124.
[13] M. Wu, N. Goodman, Multimodal generative models for scalable weakly-supervised learning, in: NIPS, 2018, pp. 5575–5585.
[14] I. Goodfellow, Nips 2016 tutorial: generative adversarial networks, arXiv preprint arXiv:1701.00160.
[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, NIPS (2014) 2672–2680.
[16] M. Mirza, S. Osindero, Conditional generative adversarial nets, arXiv preprint arXiv:1411.1784.
[17] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, B. Catanzaro, High-resolution image synthesis and semantic manipulation with conditional gans, CVPR (2018) 8798–8807.
[18] Y. Balaji, M.R. Min, B. Bai, R. Chellappa, H.P. Graf, Conditional gan with discriminative filter generation for text-to-video synthesis, IJCAI (2019) 1995–2001.
[19] V.V. Kniaz, V.A. Knyaz, J. Hladuvka, W.G. Kropatsch, V. Mizginov, Thermalgan: multimodal color-to-thermal image translation for person re-identification in multispectral dataset, in: ECCV, 2018.
[20] L. Liu, H. Zhang, Y. Ji, Q.J. Wu, Toward ai fashion design: an attribute-gan model for clothing match, Neurocomputing 341 (2019) 156–167.
[21] Y. Lin, P. Ren, Z. Chen, Z. Ren, J. Ma, M. de Rijke, Improving outfit recommendation with co-supervision of fashion generation, WWW (2019) 1095–1105.
[22] X. Yan, J. Yang, K. Sohn, H. Lee, Attribute2image: conditional image generation from visual attributes, ECCV (2016) 776–791.
[23] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, H. Lee, Generative adversarial text to image synthesis, arXiv preprint arXiv:1605.05396.
[24] Y. Taigman, A. Polyak, L. Wolf, Unsupervised cross-domain image generation, arXiv preprint arXiv:1611.02200.
[25] A.L. Maas, A.Y. Hannun, A.Y. Ng, Rectifier nonlinearities improve neural network acoustic models, in: Proc. icml, 2013, p. 3.
[26] C. McCormick, Word2vec tutorial-the skip-gram model, 2016.
[27] Y. Kim, Convolutional neural networks for sentence classification, arXiv preprint arXiv:1408.5882.
[28] J. Liu, X. Song, Z. Chen, J. Ma, Neural fashion experts: I know how to make the complementary clothing matching, Neurocomputing.
[29] A. Severyn, A. Moschitti, Twitter sentiment analysis with deep convolutional neural networks, SIGIR (2015) 959–962.
[30] X. Mao, Q. Li, H. Xie, R.Y. Lau, Z. Wang, S. Paul Smolley, Least squares generative adversarial networks, ICCV (2017) 2794–2802.
[31] W. Xian, P. Sangkloy, V. Agrawal, A. Raj, J. Lu, C. Fang, F. Yu, J. Hays, Texturegan: controlling deep image synthesis with texture patches, CVPR (2018) 8456–8465.
[32] X. Song, F. Feng, J. Liu, Z. Li, L. Nie, J. Ma, Neurostylist: Neural compatibility modeling for clothing matching, in: MM, 2017, pp. 753–761.
[33] S. Rendle, C. Freudenthaler, Z. Gantner, L. Schmidt-Thieme, Bpr: bayesian personalized ranking from implicit feedback, UAI (2009) 452–461.
[34] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, arXiv preprint arXiv:1412.6980.
[35] E. Moulines, F.R. Bach, Non-asymptotic analysis of stochastic approximation algorithms for machine learning, in: NIPS, 2011, pp. 451–459.
[36] Y. Lin, P. Ren, Z. Chen, Z. Ren, J. Ma, M. De Rijke, Explainable outfit recommendation with joint outfit matching and comment generation, TKDE.
[37] T. Fawcett, An introduction to roc analysis, Pattern Recogn. Lett. (2006) 861–874.
[38] Y. Koren, R. Bell, C. Volinsky, Matrix factorization techniques for recommender systems, Computer (2009) 30–37.
[39] X. Han, X. Song, J. Yin, Y. Wang, L. Nie, Prototype-guided attribute-wise interpretable scheme for clothing matching, SIGIR (2019) 785–794.
[40] S. Qi, X. Wang, X. Zhang, X. Song, Z.L. Jiang, Scalable graph based non-negative multi-view embedding for image ranking, Neurocomputing 274 (2018) 29–36.
[41] J. McAuley, C. Targett, Q. Shi, A. Van Den Hengel, Image-based recommendations on styles and substitutes, SIGIR (2015) 43–52.
[42] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, ICCV (2017) 2223–2232.
[43] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, in: MICCAI, Springer, 2015, pp. 234–241.

Jinhuan Liu received the Ph.D. degree from the school of Computer Science and Technology at Shandong University, Qingdao, China. She is currently working in Qingdao University of Science and Technology. Her research interests focus on information retrieval, recommendation systems and fashion analysis. She has published several papers in the international conference and journals, such as IJCAI, TOMM and Neurocomputing.

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

Zhumin Chen is an associate professor in School of Computer Science and Technology of Shandong University. He is a member of the Chinese Information Technology Committee, Social Media Processing Committee, China Computer Federation Technical Committee (CCF) and ACM. He received his Ph.D. from Shandong University. His research interests mainly include information retrieval, big data mining and processing, as well as social media processing.

Jun Ma received the B.E., M.S., and Ph.D. degrees from Shandong University in China, Ibaraki University, and Kyushu University in Japan, respectively. He is currently a professor at Shandong University. He was a senior researcher in Ibaraki Univsity in 1994 and German GMD and Fraunhofer from 1999 to 2003. His research interests include information retrieval, Web data mining, recommendation systems and machine learning. He has published more than 150 International Journal and conference papers, including SIGIR, MM, TOIS and TKDE. He is a member of the ACM and IEEE.