Auxiliary Template-Enhanced Generative Compatibility Modeling

Jinhuan Liu[1] , **Xuemeng Song**[1*] , **Zhaochun Ren**[1] , **Liqiang Nie**[1] , **Zhaopeng Tu**[2] and **Jun Ma**[1]
[1]Shandong University, Qingdao, China
[2]Tencent AI Lab, Shenzhen, China
{liujinhuan.sdu,sxmustc,nieliqiang}@gmail.com, {zhaochun.ren, majun}@sdu.edu.cn,
zptu@tencent.com

Abstract

In recent years, there has been a growing interest in the fashion analysis (e.g., clothing matching) due to the huge economic value of the fashion industry. The essential problem is to model the compatibility between the complementary fashion items, such as the top and bottom in clothing matching. The majority of existing work on fashion analysis has focused on measuring the item-item compatibility in a latent space with deep learning methods. In this work, we aim to improve the compatibility modeling by sketching a compatible template for a given item as an auxiliary link between fashion items. Specifically, we propose an end-to-end Auxiliary Template-enhanced Generative Compatibility Modeling (AT-GCM) scheme, which introduces an auxiliary complementary template generation network equipped with the pixel-wise consistency and compatible template regularization. Extensive experiments on two real-world datasets demonstrate the superiority of the proposed approach.

1 Introduction

According to Statista, the online fashion retail sales of the United States have reached 103 billion dollars in 2018[1], which reflects the great demand for online clothing shopping. Intuitively, people tend to match compatible complementary fashion items (e.g., a shirt and trousers) and make proper outfits. Owing to the recent advances in representation learning, many research efforts have been dedicated to the compatibility modeling between complementary fashion items to assist people in clothing matching.

In a sense, existing methods mainly focus on learning the latent compatibility space [Yang *et al.*, 2019] with advanced neural networks to bridge the gap between complementary fashion items, where the item-item compatibility can be directly measured. In fact, the latest remarkable performance of Generative Adversarial Networks (GAN) in various image generation tasks [Huang *et al.*, 2019] has enabled us to rethink the solution for automatic clothing matching. Imagine that given a top and a bottom, if we first sketch a compatible bottom template for the given top as an auxiliary link between the complementary items, then we can further measure their compatibility from the item-template perspective.

In the light of this, in this work, we aim to boost the performance of compatibility modeling between fashion items with the help of the auxiliary complementary template generation. We argue that the task of template-enhanced compatibility modeling is non-trivial. The main challenge lies in how to seamlessly integrate the auxiliary template generation into the primary item-item compatibility modeling and boost the performance. Also, how to accurately generate a compatible bottom template for the given top to guide the item-template compatibility modeling arises the second challenge. As each fashion item involve multiple modalities (e.g., visual and textual modalities), both of which can convey important message regarding the item features, how to effectively fuse the multi-modal cues poses the last challenge.

To address the aforementioned challenges, we propose an **A**uxiliary **T**emplate-enhanced **G**enerative **C**ompatibility **M**odeling network (abbreviated as AT-GCM) as shown in Figure 1. The scheme comprehensively measures the compatibility between fashion items from the primary item-item perspective and the auxiliary item-template perspective simultaneously. On the one hand, we devise the item-item compatibility modeling component with a dual-path neural network, where each path corresponds to one modality of the fashion item. On the other hand, we introduce an auxiliary complementary template generation network equipped with the pixel-wise consistency and compatible template regularization, working on transferring the given top image to a compatible bottom template image.

Our contributions can be summarized in three-fold. 1) We propose an auxiliary template-enhanced generative compatibility modeling scheme, which seamlessly integrates the primary item-item compatibility modeling and the auxiliary item-template compatibility modeling. To the best of our knowledge, we are the first to explore the potential of GAN in the context of fashion compatibility modeling. 2) We propose an auxiliary complementary template generation network, which is able to sketch a compatible bottom template for a given top and hence facilitate the final compatibility

*Contact Author

[1]https://www.statista.com/topics/3481/fashion-e-commerce-in-the-united-states.

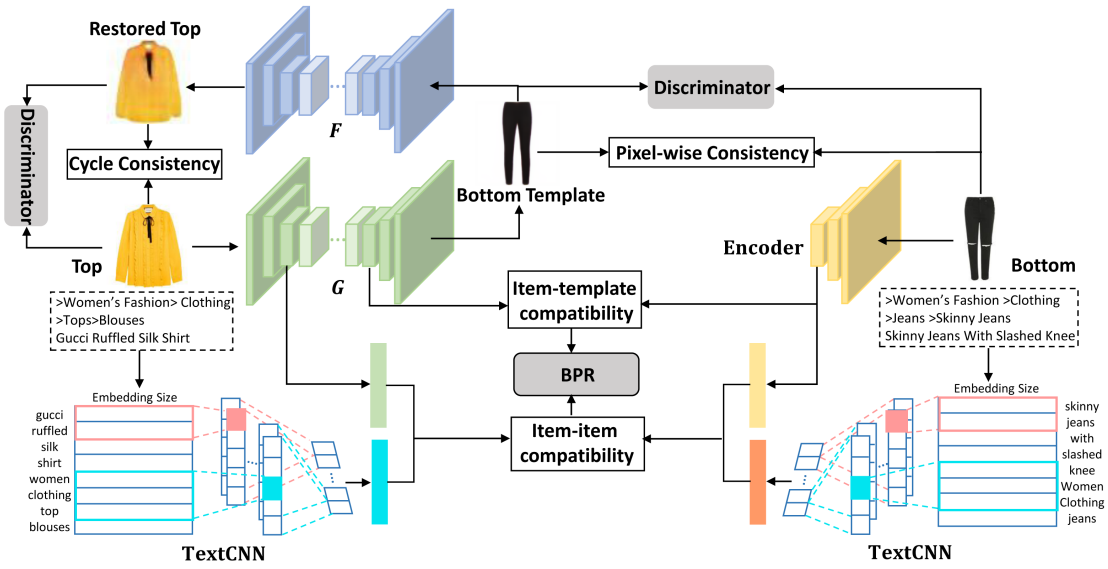

Figure 1: Illustration of the proposed auxiliary template-enhanced generative compatibility modeling scheme.

modeling. And 3) extensive experiments on two real-world datasets demonstrate the superiority of our AT-GCM over the state-of-the-art methods in compatibility modeling, where improvements of 3.56% and 4.87% on AUC and MRR can be achieved by our AT-GCM over the best baseline, respectively.

2 Related Work

2.1 Generative Models

Recent mainstream generative models for automatic image generation include the Variational AutoEncoder (VAE) [Pu *et al.*, 2016] and GAN [Goodfellow *et al.*, 2014]. Variational methods work on introducing the deterministic bias to optimize the lower bound of the logarithmic likelihood with probabilistic graphical models. Despite its compelling performance in various image generation tasks, VAE tends to generate blurry samples due to its KL divergence minimization between the samples and the input data [Jin *et al.*, 2019]. Pertaining to the GAN that usually consists of a generator and a discriminator, the key to its remarkable success lies in its min-max optimization strategy, where the generator tries to synthesize a realistic image to fool the discriminator, while the discriminator attempts to distinguish the generated image from the real one. Despite the huge success in various research tasks [Song *et al.*, 2018; Li *et al.*, 2019; He *et al.*, 2019], the potential of GAN in the compatibility modeling remains largely untapped, which is a major novelty of our work.

2.2 Fashion Compatibility Modeling

Due to the huge economic value of the fashion industry, increasing research attention has been paid to the complementary clothing recommendation [Chen and He, 2018; Wang *et al.*, 2018] and outfit assessment [Cucurull *et al.*, 2019; Ma *et al.*, 2017]. For example, Song et al. [2017] proposed a content-based neural scheme with the Bayesian Personalized Ranking (BPR) framework to model the compatibility between two fashion items. In addition, to handle the outfit compatibility assessment that involves multiple fashion items, Han et al. [2017] presented a Bidirectional Long Short-Term Memory (Bi-LSTM) model that is able to sequentially predict the compatibility among items of an outfit. Later, Vasileva et al. [2018] introduced an end-to-end method working on jointly learning the item similarity and compatibility. In a sense, the above studies focus more on the non-generative item-item compatibility modeling, but overlook the potential of incorporating an auxiliary bridge between fashion items using the generative models. Towards this end, Lin et al. [2019] proposed a novel outfit recommendation framework with the co-supervision of fashion generation, which aims to boost the recommendation performance by generating the auxiliary bottom image with VAE based on the given top and the desired bottom description. Different from it, in this work, we focus on generating an auxiliary complementary template based on GAN rather than VAE to enhance the compatibility modeling from the item-template perspective. Moreover, to promote the model flexibility in practical applications, we devise the generative network to simply take the given top image as the input, making the description of the desired bottom, which is essential to [Lin *et al.*, 2019], unnecessary.

3 Methodology

3.1 Problem Formulation

Suppose we have a set of tops $\mathcal{T} = \{t_1, t_2, \cdots, t_{N_t}\}$ and bottoms $\mathcal{B} = \{b_1, b_2, \cdots, b_{N_b}\}$, where N_t and N_b stand for the number of tops and bottoms, respectively. Let $\mathbf{I}_{t_i}(\mathbf{I}_{b_j})$ and $\mathbf{c}_{t_i}(\mathbf{c}_{b_j})$ represent the visual image and the textual description of the top (bottom) t_i (b_j), respectively. Let $\mathcal{P} = \{(t_{i_1}, b_{j_1}), (t_{i_2}, b_{j_2}), \cdots, (t_{i_M}, b_{j_M})\}$ stands for the set of positive top-bottom pairs, where M refers to the total number of pairs. We define m_{ij} as the compatibility between the top t_i and bottom b_j, based on which we can

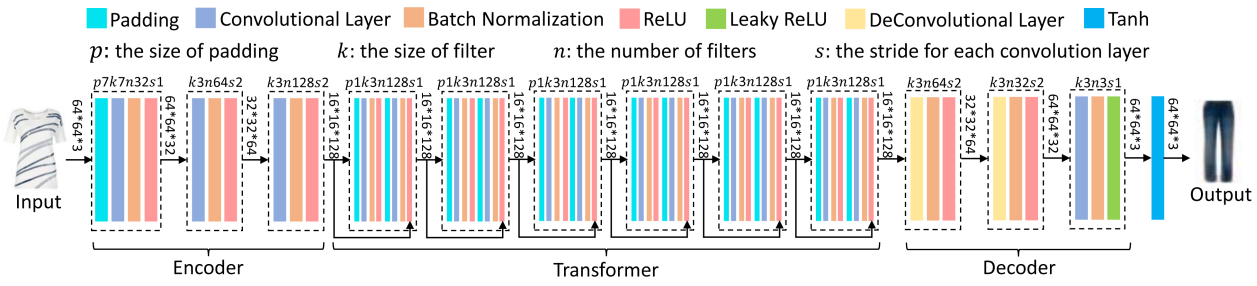

Figure 2: Details of the generator network with three components: encoder, transformer and decoder.

generate a ranking list of bottoms for a given top to make compatible outfits. Essentially, our goal is to devise an end-to-end auxiliary template-enhanced generative compatibility modeling scheme $\mathcal{C}$ with the set of to-be-learned parameters $\mathbf{\Theta}_C$, which is capable of comprehensively measuring the compatibility between the given top and bottom by generating an auxiliary bottom template $\tilde{\mathbf{I}}_{b_i}$.

3.2 Auxiliary Complementary Template Generation

Intuitively, the underlying philosophy for introducing the auxiliary complementary template generation is that for a given top, we can sketch a compatible bottom template to guide the final retrieval of the compatible bottoms. In fact, inspired by the edges-to-photos [Xian *et al.*, 2018], male-to-female [White, 2016], attributes-to-faces [Liu *et al.*, 2017], generating the compatible bottom template for a given top can be cast as an image-to-image translation problem (i.e., top-to-bottom translation). Moreover, GAN has become the mainstream approach for the image-to-image translation task. Accordingly, we adopt GAN as the backbone to transfer the given top image $\mathbf{I}_{t_i}$ from the source domain $\mathcal{T}$ to the bottom template image $\tilde{\mathbf{I}}_{b_i}$ in the target domain $\mathcal{B}$ as follows:

$$G(\mathbf{I}_{t_i}|\mathbf{\Theta}_G) \rightarrow \tilde{\mathbf{I}}_{b_i}, \tag{1}$$

where $\mathbf{\Theta}_G$ refers to the parameters of generator G. Similar to [Zhu *et al.*, 2017], we design the generator with three key components: encoder, transformer and decoder. The detailed network architecture is shown in Figure 2. The encoder, comprising three convolution layers, is employed to learn the visual feature encoding $\mathbf{V}_{t_i}$ of the given top t_i as follows:

$$\mathbf{H}_k = \phi(\mathbf{W}_k\mathbf{H}_{k-1} + \mathbf{b}_k),\ k = 1, \cdots, K, \tag{2}$$

where we compile $\mathbf{\Theta}_{enc} = \{\mathbf{W}_k, \mathbf{b}_k \mid k = 1, \cdots, K\}$ as the parameters of the encoder. We set $K = 3$ and $\mathbf{H}_0 = \mathbf{I}_{t_i}$. We treat the output of the last layer as the visual encoding, i.e., $\mathbf{V}_{t_i} = \mathbf{H}_K \in \mathbb{R}^{w\times h\times c}$, where $w \times h \times c$ refers to the shape of the encoding. In this work, we adopt the ReLU as the activation function $\phi(\cdot)$. In a similar manner, we can derive the visual encoding $\mathbf{V}_{b_j} \in \mathbb{R}^{w\times h\times c}$ for the bottom b_j.

Then, the transformer plays a pivotal role in converting the visual encoding $\mathbf{V}_{t_i}$ of the source top image $\mathbf{I}_{t_i}$ to the visual encoding $\tilde{\mathbf{V}}_{b_i}$ of the target bottom template $\tilde{\mathbf{I}}_{b_i}$. In particular, we deploy the transformer with the deep residual networks (ResNets) [He *et al.*, 2016] as follows:

$$\mathbf{R}_l = \mathbf{R}_{l-1} + \mathcal{F}(\mathbf{R}_{l-1}, \mathbf{\Theta}_{trans}),\ l = 1, \cdots, L, \tag{3}$$

where $\mathbf{R}_{l-1}(\mathbf{R}_0 = \mathbf{V}_{t_i})$ is the input of the l-th residual block and $\mathcal{F}$ refers to the residual function. $\mathbf{\Theta}_{trans}$ represents the corresponding parameters. L refers to the total number of blocks in the transformer, where we set $L = 6$. Similarly, the output $\mathbf{R}_L$ is treated as the transformed visual encoding $\tilde{\mathbf{V}}_{b_i} \in \mathbb{R}^{w\times h\times c}$ for the generated bottom template $\tilde{\mathbf{I}}_{b_i}$.

Pertaining to the decoder, opposite to the encoder, it devotes to restoring the low-level representation $\tilde{\mathbf{I}}_{b_i}$ from the transformed visual encoding $\tilde{\mathbf{V}}_{b_i}$ with two deconvolution layers and one convolution layer, where we use $\mathbf{\Theta}_{dec}$ to denote the decoder parameters. Ultimately, the generator G is able to transform the given top $\mathbf{I}_{t_i}$ to a bottom template $\tilde{\mathbf{I}}_{b_i}$ with parameters $\mathbf{\Theta}_G = \{\mathbf{\Theta}_{enc}, \mathbf{\Theta}_{trans}, \mathbf{\Theta}_{dec}\}$.

Simply applying the standard GAN can only ensure the generated bottom to be realistic. We take a step forward to constrain the generated bottom template $\tilde{\mathbf{I}}_{b_i}$ to be not only natural but also compatible with the given top to facilitate the following retrieval of compatible bottoms. To guarantee the training stability and avoid the vanishing gradient [Xian *et al.*, 2018], we adopt the least square loss [Mao *et al.*, 2017]:

$$\begin{aligned}\mathcal{L}_{GAN}(G_{\mathcal{T}\rightarrow\mathcal{B}}, D_{\mathcal{B}}) = \min_{\mathbf{\Theta}}\{&\frac{1}{2}(1-t)(D_{\mathcal{B}}(\mathbf{I}_{b_j})-1)^2 \\ &+ \frac{1}{2}(D_{\mathcal{B}}(\tilde{\mathbf{I}}_{b_i}) - t)^2\},\end{aligned} \tag{4}$$

where $t = 1$ refers to training the generator $G_{\mathcal{T}\rightarrow\mathcal{B}}$ with $\mathbf{\Theta} = \mathbf{\Theta}_{G_{\mathcal{T}\rightarrow\mathcal{B}}}$, while $t = 0$ corresponds to optimizing the discriminator $D_{\mathcal{B}}$ with $\mathbf{\Theta} = \mathbf{\Theta}_{D_{\mathcal{B}}}$. $\mathbf{\Theta}_{G_{\mathcal{T}\rightarrow\mathcal{B}}}$ and $\mathbf{\Theta}_{D_{\mathcal{B}}}$ represent the parameters of the generator and the discriminator, respectively. Meanwhile, we adopt the pixel-wise L1 loss to prevent the blurring [Isola *et al.*, 2017] of the generated compatible bottom and further encourage the generated bottom template to be compatible with the given top by minimizing the pixel-level differences between the bottom template $\tilde{\mathbf{I}}_{b_i}$ and the positive bottom $\mathbf{I}_{b_j}$ as follows:

$$\mathcal{L}_{pixel} = \left\|\tilde{\mathbf{I}}_{b_i} - \mathbf{I}_{b_j}\right\|_1. \tag{5}$$

To alleviate the mode collapse problem [Li *et al.*, 2017] of the typical GAN, similar to [Zhu *et al.*, 2017], we adopt the cycle consistency loss to enforce that the source top image $\tilde{\mathbf{I}}_{t_i}$ should be able to be restored by the generated bottom template $\tilde{\mathbf{I}}_{b_i}$. Towards this end, we introduce an extra inverse generator $F_{\mathcal{B}\rightarrow\mathcal{T}}$ as follows:

$$F_{\mathcal{B}\rightarrow\mathcal{T}}(\tilde{\mathbf{I}}_{b_i}|\mathbf{\Theta}_{F_{\mathcal{B}\rightarrow\mathcal{T}}}) \rightarrow \tilde{\mathbf{I}}_{t_i} \approx \mathbf{I}_{t_i}, \tag{6}$$

where $\mathbf{\Theta}_{F_{\mathcal{B}\to\mathcal{T}}}$ represents the generator parameters. Then we have the following adversarial loss $\mathcal{L}_{GAN}(F_{\mathcal{B}\to\mathcal{T}}, D_{\mathcal{T}}) =$:

$$\min_{\mathbf{\Theta}}\{\frac{1}{2}(1-z)(D_{\mathcal{T}}(\mathbf{I}_{t_i})-1)^2+\frac{1}{2}(D_{\mathcal{T}}(\tilde{\mathbf{I}}_{t_i})-z)^2\}, \quad (7)$$

where $z = 1$ corresponds to $\mathbf{\Theta} = \mathbf{\Theta}_{F_{\mathcal{B}\to\mathcal{T}}}$, while $z = 0$ refers to $\mathbf{\Theta} = \mathbf{\Theta}_{D_{\mathcal{T}}}$. $\mathbf{\Theta}_{D_{\mathcal{T}}}$ represents the parameters of the discriminator $D_{\mathcal{T}}$. As the restored top $\tilde{\mathbf{I}}_{t_i}$ should be consistent with the source input $\mathbf{I}_{t_i}$, we have the following cycle consistency loss:

$$\mathcal{L}_{cycp} = \left\|\tilde{\mathbf{I}}_{t_i} - \mathbf{I}_{t_i}\right\|_1. \quad (8)$$

3.3 Template-Enhanced Generative Compatibility Modeling

We proceed to detail our proposed template-enhanced generative compatibility modeling component. Accordingly, the compatibility m_{ij} can be measured from both the primary item-item and auxiliary item-template perspectives.

As for the primary item-item compatibility modeling, we aim to seek the latent representations for items that enable us to accurately measure the compatible preference between items. In a sense, it is natural to argue that compatible items should follow certain visually distinguished patterns. For example, the "Chiffon Blouse with Bow Detail" matches well with the "High-waisted Pleated Design Midi Skirt", while the "Striped Shirt" goes well with the "Fray Hem Denim Wide Leg Pants". To well capture the distinguished features of items, we adopt the global average pooling (GAP) [Lin *et al.*, 2013] for its powerful capability in locating the discriminant areas of an image. According to GAP, each feature map with the shape of $w \times h$ would be averaged to one value. Therefore, we can derive the global visual feature $\mathbf{v}_{t_i}(\mathbf{v}_{b_j}) \in \mathbb{R}^c$ from the visual encoding $\mathbf{V}_{t_i}(\mathbf{V}_{b_j}) \in \mathbb{R}^{w\times h\times c}$ for the top (bottom) t_i (b_j). Moreover, to enhance the nonlinear compatibility modeling, we project the global visual feature with a fully-connected layer to get the final latent visual representation for each item. Taking the top as an example, we fed $\mathbf{v}_{t_i}$ to the following layer:

$$\tilde{\mathbf{v}}_{t_i} = \sigma(\mathbf{W}_v\mathbf{v}_{t_i} + \mathbf{h}_v), \quad (9)$$

where $\tilde{\mathbf{v}}_{t_i} \in \mathbb{R}^{D_v}$ represents the final visual representation of the top t_i. σ denotes the sigmoid activation function.

Apart from the visual cue, the textual information may also convey important features (e.g., the category and style) of fashion items and hence also merit our attention. Accordingly, we define $\tilde{\mathbf{c}}_{t_i}(\tilde{\mathbf{c}}_{b_j}) \in \mathbb{R}^{D_t}$ as the latent textual representation for the top (bottom), which can be obtained in a similar manner with the visual representation $\tilde{\mathbf{v}}_{t_i}(\tilde{\mathbf{v}}_{b_j})$. Then, we define the item-item compatibility as follows:

$$m_{ij}^{I-I} = \mu(\tilde{\mathbf{v}}_{t_i})^T\tilde{\mathbf{v}}_{b_j} + (1-\mu)(\tilde{\mathbf{c}}_{t_i})^T\tilde{\mathbf{c}}_{b_j}, \quad (10)$$

where μ is used to balance the importance of the visual and textual modalities.

Pertaining to the auxiliary item-template compatibility, we argue that the compatible bottom candidates for a given top should be semantically similar to the latent generated bottom template. Accordingly, we define the auxiliary compatibility as the similarity between the high-level visual encodings of the generated bottom template and the given bottom. Then,

$$m_{ij}^{I-T} = \left\|\tilde{\mathbf{V}}_{b_i} - \mathbf{V}_{b_j}\right\|_1. \quad (11)$$

Ultimately, the compatibility score m_{ij} between the top and bottom can be defined as follows:

$$m_{ij} = m_{ij}^{I-I} + \alpha m_{ij}^{I-T}, \quad (12)$$

where α is the trade-off non-negative hyper-parameter. Towards the final compatibility modeling, we build the following triplet dataset:

$$\mathcal{E} := \{(i,j,k)|(t_i,b_j) \in \mathcal{P}, b_k \in \mathcal{B}\setminus b_j\}. \quad (13)$$

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), where b_k is randomly sampled from the bottom set $\mathcal{B}$. Adopting the BPR [He and McAuley, 2016], we model the compatible relationship between fashion items as follows:

$$\mathcal{L}_{BPR} = -ln(\sigma(m_{ij} - m_{ik})), \quad (14)$$

where m_{ik} can be derived according to Eqn.(12). Essentially, we expect the given top would share the higher compatibility with the positive bottom as compared to the negative one.

Optimization. To boost the performance, we seamlessly integrate the auxiliary complementary template generation with the compatibility modeling in an end-to-end manner:

$$\begin{aligned}\mathcal{L} =& \mathcal{L}_{BPR} + \mathcal{L}_{GAN}(G_{\mathcal{T}\to\mathcal{B}}, D_{\mathcal{B}}) + \mathcal{L}_{GAN}(F_{\mathcal{B}\to\mathcal{T}}, D_{\mathcal{T}}) \\ &+ \beta\mathcal{L}_{cycp} + \gamma\mathcal{L}_{pixel} + \delta\left\|\mathbf{\Theta}_C\right\|^2, \end{aligned} \quad (15)$$

where β, γ, δ are the non-negative hyper-parameters controlling the strength of different components of AT-GCM. Then, we adopt the back-propagation algorithm to learn the network parameter $\mathbf{\Theta}_C$.

4 Experiments

4.1 Experimental Settings

To evaluate the proposed AT-GCM, we conduct extensive experiments on public datasets: FashionVC [Song *et al.*, 2017] and ExpFashion [Lin *et al.*, 2019], which consist of $20,726$ and $853,991$ outfits, respectively. Each fashion item involves an image and the textual context (i.e., corresponding item categories and title description). The image is directly fed into the generator network. Pertaining to the textual information, we adopt TextCNN [Kim, 2014] and derive the textual representation $\mathbf{c}_{t_i}(\mathbf{c}_{b_j}) \in \mathbb{R}^{400}$ for each fashion item. Notably, to balance the two datasets, we randomly sample $20,000$ outfits instead of using the whole ExpFashion dataset.

For each dataset, we randomly select 80% for training, 10% for validation, and the rest for testing. To comprehensively show the effectiveness of our AT-GCM in compatibility modeling, we adopt the area under the ROC curve (AUC) and mean reciprocal rank (MRR) to evaluate its performance in both triplet-wise compatibility modeling and list-wise complementary item retrieval tasks, respectively. We randomly sample 3 negative bottoms according to Eqn.(13) to build the triplet dataset for the former task, while 9 negative bottoms together with the positive one to comprise the bottom retrieval candidates for the latter one.

Approach	FashionVC		ExpFashion	
	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
Pix2pix	0.8208	0.6579	0.8165	0.6253
CycleGAN	0.8292	0.6884	0.8243	0.6872
AT-GCM	**0.8587**	**0.7219**	**0.8395**	**0.7058**

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

Figure 3: Bottom templates generated by different generative models. GT: ground truth.

4.2 Comparisons

To demonstrate the effectiveness of our proposed AT-GCM for the task of compatibility clothing matching, we compare it with the following state-of-the-art baselines.

POP: We measure the compatibility between top t_i and bottom b_j by the number of bottoms that have been matched with the given top t_i in the dataset.

Bi-LSTM [Han *et al.*, 2017]: This approach models the outfit compatibility by exploring the sequential relationships among fashion items in an outfit. We adapt this method to deal with outfits that consist of only two fashion items (i.e., the top and bottom).

IBR [McAuley *et al.*, 2015]: IBR models the relationships between items with a linear latent style space, which is learned simply based on the visual modality.

IBR-VC: We extend IBR to measure the compatibility between fashion items with both the visual and textual information. Specifically, we employ the TextCNN that is also used in AT-GCM to encode the textual features.

BPR-DAE [Song *et al.*, 2017]: This baseline is originally designed to jointly model the compatibility between fashion items and the coherent modality consistency of items. For the sake of fairness, we adapt it in an end-to-end fashion, where the Alexnet and TextCNN are used to extract the visual and textual representations, respectively.

FARM [Lin *et al.*, 2019]: This model aims to boost the compatibility modeling by the item generation, where VAE is adopted to generate the auxiliary bottom image based on the given top and the bottom description. Notably, we disable the input of the bottom text description, which is unnecessary in our context.

CycleGAN [Zhu *et al.*, 2017]: We replace the template generative network in our model with 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 [Isola *et al.*, 2017]: We utilize the pix2pix to fulfill the auxiliary complementary template generation of our scheme, which adopts the U-Net [Ronneberger *et al.*, 2015] architecture for its generator.

Table 1 shows the performance comparison with the state-of-the-art methods in terms of AUC and MRR both on the FashionVC and ExpFashion datasets, respectively. From this table, we have the following observations. 1) Our AT-GCM significantly outperforms all the other methods, verifying the effectiveness of our proposed generative compatibility modeling scheme. 2) The most naive baseline POP achieves the worst performance, which is reasonable as it may be inappropriate to match fashion items without considering the item contents that intuitively capture the item's various features. 3) AT-GCM surpasses all the non-generative content-based methods (i.e., IBR, IBR-VC, Bi-LSTM and BPR-DAE), indicating the advantage of taking into account the auxiliary compatible template generation in the compatibility modeling. 4) AT-GCM achieves more superior performance than Pix2pix and CycleGAN. This may be due to the fact that the U-Net structure of Pix2pix tends to learn the low-level information [Yi *et al.*, 2017], while the CycleGAN is more suitable for the image translation tasks with the narrow domain gap (e.g., zebra to horse, summer to winter) rather than the task with a large domain gap like ours. 5) Surprisingly, the generative model FARM performs worse than the conventional non-generative methods. The possible explanations are twofold: a) the variational method adopted by FARM may produce samples with the lower quality [Goodfellow *et al.*, 2014] and hence hurt the model performance; and b) FARM highly relies on the given bottom text description, which hinders its adaptability in our more general and practical context.

To gain more deep insights regarding the effectiveness of our proposed framework, we intuitively compare the generated bottom templates of AT-GCM, FARM, pix2pix, and CycleGAN with several examples in Figure 3. As we can see, our AT-GCM outperforms the baselines by generating more natural and compatible bottom templates for the given top. In particular, we observe that the bottom templates yielded by our AT-GCM can capture the shape of the desired compatible bottom better than the texture. Interestingly, even with the auxiliary bottom templates with fuzzy texture, AT-GCM can boost the compatibility modeling performance significantly (as shown in Table 1). One possible reason lies in that shape is an important attribute of a fashion item, highly correlated to other attributes, like the item category, style design, sleeve length as well as the clothing fitness, and thus plays an essential role in compatibility modeling.

Approach	FashionVC		ExpFashion	
	AUC	MRR	AUC	MRR
AT-GCM-V	0.8124	0.6680	0.7963	0.6534
AT-GCM-T	0.6413	0.5735	0.6092	0.4202
AT-GCM	**0.8587**	**0.7219**	**0.8395**	**0.7058**

(a) Performance on different modality configurations.

Approach	FashionVC		ExpFashion	
	AUC	MRR	AUC	MRR
-noCyc	0.8308	0.6766	0.8172	0.6586
-noPixel	0.8133	0.6678	0.7840	0.6403
-noTemG	0.7599	0.6472	0.7362	0.6265
AT-GCM	**0.8587**	**0.7219**	**0.8395**	**0.7058**

(b) Performance on different component configurations.

Table 2: The ablation study of AT-GCM in terms of AUC and MRR on FashionVC and ExpFashion.

Figure 4: Illustration of the ranking results in term of MRR with -noTemG and AT-GCM. The clothing items highlighted in the red boxes are the ground truth.

4.3 The Ablation Study

To comprehensively verify our AT-GCM, we further conduct the ablation study, where the model performance with different modality and component configurations are evaluated.

On Modality Comparison. To demonstrate the advantages of incorporating the multiple modalities of fashion items, we introduce two derivatives of our model: AT-GCM-V and AT-GCM-T, where only the visual and textual modality is adopted in our framework, respectively. Table 2(a) shows the evaluation results with different modality configurations on FashionVC and ExpFashion. As can be seen, our AT-GCM significantly outperforms AT-GCM-V and AT-GCM-T. Even for the better derivative AT-GCM-V, improvements of 5.69% and 8.06% can be achieved by AT-GCM regarding the AUC and MRR, respectively. This well demonstrates the benefits of incorporating both the visual and textual modality into the compatibility modeling. Moreover, we notice that AT-GCM-V performs better than AT-GCM-T, confirming the visual modality conveys more intuitive features (e.g., the color and shape) of fashion items and is more reliable for compatibility modeling. In addition, this may be attributed to the fact that the visual information contributes to the auxiliary template generation more than the textual modality.

On Component Comparison. To gain deep insights regarding our AT-GCM, we study the effects of its several key components. In particular, to evaluate the auxiliary template generation component, we introduce the baseline -noTemG, where the bottom template generation network is disabled, resulting in the model to measure the compatibility simply from the item-item perspective. In addition, to further check the impacts of the pixel-wise L1 regularization and the cycle generative network in the auxiliary template generation component, we adapt our method to -noPixel and -noCyc by setting γ to 0 and removing the cycle generative network $F_{\mathcal{B}\to\mathcal{T}}$, respectively. As can be seen from Table 2(b), AT-GCM significantly outperforms -noTemG, demonstrating the necessity of the auxiliary template generation component for AT-GCM. In addition, both -noPixel and -noCyc surpass -noTemG, implying that both the pixel-wise consistency and the cycle generative network are essential to the auxiliary template generation component of AT-GCM. Figure 4 visualizes the ranking results of -noTemG and AT-GCM in the task of complementary item retrieval with two examples. As can be seen from the first example, given the top "Short Sleeve T-shirt", the positive bottom "Ripped Light Jeans" is ranked at the third place by -noTemG, but promoted to the first place by AT-GCM taking the bottom template generation into account. Indeed, the auxiliary bottom template generated by AT-GCM does help to identify the jeans as the positive compatible bottom rather than the "Stripe Mini Dress" and the "Petal Black Skirts". Similar observations can be also found in the second example, where the generated bottom template help to promote the ranking of the positive item.

5 Conclusion

In this work, we present an end-to-end auxiliary template-enhanced generative compatibility modeling scheme (AT-GCM), which is able to comprehensively model the compatibility between fashion items from both the item-item and item-template perspectives. As a major novelty, we introduce an auxiliary complementary template generation network to help sketch a template and enhance the compatibility modeling, where the pixel-wise consistency and compatible template regularization are jointly modeled. Extensive experiments on two real-world datasets demonstrate the advantage of taking into account the item-template compatibility modeling. Currently, we sketch the auxiliary complementary template simply by the visual cue but overlook the textual message. In the future, we plan to further consider the textual context to promote the template generation quality and enhance the compatibility modeling performance.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (61702300, 61672322, 61972234, 61902219), the Future Talents Research Funds of Shandong University (No.: 2018WLJH63), the Innocation Teams in Colleges and Universities in Jinan (No.:2018GXRC014), the Tencent AI Lab Rhino-Bird Focused Research Program (JR201932), the Foundation of State Key Laboratory of Congitive Intelligence, iFLYTEK, P.R.China (COGOSC-20190003), and the Fundamental Research Funds of Shandong University.

References

[Chen and He, 2018] Long Chen and Yuhang He. Dress fashionably: Learn fashion collocation with deep mixed-category metric learning. In *AAAI*, 2018.

[Cucurull *et al.*, 2019] Guillem Cucurull, Perouz Taslakian, and David Vazquez. Context-aware visual compatibility prediction. *arXiv preprint arXiv:1902.03646*, 2019.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.

[Han *et al.*, 2017] Xintong Han, Zuxuan Wu, Yu-Gang Jiang, and Larry S Davis. Learning fashion compatibility with bidirectional lstms. In *MM*, pages 1078–1086, 2017.

[He and McAuley, 2016] Ruining He and Julian McAuley. Vbpr: Visual bayesian personalized ranking from implicit feedback. In *AAAI*, pages 144–150, 2016.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *ECCV*, pages 630–645, 2016.

[He *et al.*, 2019] Tianyu He, Yingce Xia, Jianxin Lin, Xu Tan, Di He, Tao Qin, and Zhibo Chen. Deliberation learning for image-to-image translation. *IJCAI*, 2019.

[Huang *et al.*, 2019] Zhiwu Huang, Jiqing Wu, and Luc Van Gool. Manifold-valued image generation with wasserstein generative adversarial nets. In *AAAI*, 2019.

[Isola *et al.*, 2017] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[Jin *et al.*, 2019] Di Jin, Bingyi Li, Pengfei Jiao, Dongxiao He, and Weixiong Zhang. Network-specific variational auto-encoder for embedding in attribute networks. In *IJCAI*, pages 2663–2669, 2019.

[Kim, 2014] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[Li *et al.*, 2017] Chunyuan Li, Hao Liu, Changyou Chen, Yuchen Pu, Liqun Chen, Ricardo Henao, and Lawrence Carin. Alice: Towards understanding adversarial learning for joint distribution matching. In *NIPS*, 2017.

[Li *et al.*, 2019] Qintong Li, Hongshen Chen, Zhaochun Ren, Zhumin Chen, Zhaopeng Tu, and Jun Ma. Empgan: Multi-resolution interactive empathetic dialogue generation. *arXiv preprint arXiv:1911.08698*, 2019.

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

[Lin *et al.*, 2019] Yujie Lin, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Jun Ma, and Maarten de Rijke. Improving outfit recommendation with co-supervision of fashion generation. In *WWW*, pages 1095–1105, 2019.

[Liu *et al.*, 2017] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *NIPS*, pages 700–708, 2017.

[Ma *et al.*, 2017] Yihui Ma, Jia Jia, Suping Zhou, Jingtian Fu, Yejun Liu, and Zijian Tong. Towards better understanding the clothing fashion styles: A multimodal deep learning approach. In *AAAI*, pages 38–44, 2017.

[Mao *et al.*, 2017] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *ICCV*, pages 2794–2802, 2017.

[McAuley *et al.*, 2015] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. Image-based recommendations on styles and substitutes. In *SIGIR*, pages 43–52, 2015.

[Pu *et al.*, 2016] Yunchen Pu, Zhe Gan, Ricardo Henao, Xin Yuan, Chunyuan Li, Andrew Stevens, and Lawrence Carin. Variational autoencoder for deep learning of images, labels and captions. In *NIPS*, pages 2352–2360, 2016.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

[Song *et al.*, 2017] Xuemeng Song, Fuli Feng, Jinhuan Liu, Zekun Li, Liqiang Nie, and Jun Ma. Neurostylist: Neural compatibility modeling for clothing matching. In *MM*, pages 753–761, 2017.

[Song *et al.*, 2018] Jingkuan Song, Tao He, Lianli Gao, Xing Xu, Alan Hanjalic, and Heng Tao Shen. Binary generative adversarial networks for image retrieval. In *AAAI*, 2018.

[Vasileva *et al.*, 2018] Mariya I Vasileva, Bryan A Plummer, Krishna Dusad, Shreya Rajpal, Ranjitha Kumar, and David Forsyth. Learning type-aware embeddings for fashion compatibility. In *ECCV*, pages 390–405, 2018.

[Wang *et al.*, 2018] Zihan Wang, Ziheng Jiang, Zhaochun Ren, Jiliang Tang, and Dawei Yin. A path-constrained framework for discriminating substitutable and complementary products in e-commerce. In *WSDM*, pages 619–627, 2018.

[White, 2016] Tom White. Sampling generative networks. *arXiv preprint arXiv:1609.04468*, 2016.

[Xian *et al.*, 2018] Wenqi Xian, Patsorn Sangkloy, Varun Agrawal, Amit Raj, Jingwan Lu, Chen Fang, Fisher Yu, and James Hays. Texturegan: Controlling deep image synthesis with texture patches. In *CVPR*, 2018.

[Yang *et al.*, 2019] Xun Yang, Yunshan Ma, Lizi Liao, Meng Wang, and Tat-Seng Chua. Transnfcm: translation-based neural fashion compatibility modeling. In *AAAI*, 2019.

[Yi *et al.*, 2017] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*, pages 2849–2857, 2017.

[Zhu *et al.*, 2017] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2223–2232, 2017.