

Outfit compatibility prediction with multi-layered feature fusion network[☆]

Shufang Lu^{a,*}, Xiang Zhu^a, Yingying Wu^b, Xianmei Wan^c, Fei Gao^a

^a College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China

^b Department of Interior Design and Fashion Studies, Kansas State University, Manhattan, KS 66502, United States

^c School of Cross-Border E-Commerce, School of Science and Technology, Zhejiang International Studies University, Hangzhou 310023, China



ARTICLE INFO

Article history:

Received 14 October 2020

Revised 24 March 2021

Accepted 11 April 2021

Available online 18 April 2021

Keywords:

Outfit compatibility

Fashion recommendation

Feature fusion model

Non-local operation

ABSTRACT

Clothing plays a critical role in people's daily lives, a perfect styling clothing is able to help people to avoid weaknesses and show their personal temperament, however, not everyone is good at styling. Compatibility is the core of styling, however, determining whether a pair of garments are compatible with each other is a challenging styling issue. Years of research have been devoted to fashion compatibility learning, whereas there are still several drawbacks in visual feature detection and compatibility calculation. In this paper, we propose an end-to-end framework to learn the compatibility among tops and bottoms. In order to improve the effects of visual feature extraction, a Multi-layer Non-local Feature Fusion framework (MNLFF) is developed. Feature fusion model is used to combine both high and low-level features, while non-local block is for global feature detection. We compare our technique with the prior state-of-the-art methods in the outfit compatibility prediction task and extensive experiments on existing datasets demonstrate its effectiveness.

© 2021 Elsevier B.V. All rights reserved.

1. Introduction

Nowadays with the rapid evolution of the fashion industry in the online setting, fashion related computer vision problems have attracted increasing attention. One particularly interesting task is fashion recommendation [1–9], which suggests fashion outfits commensurate with a given item. The key of fashion recommendation is to shape the compatibility between fashion pieces. Outfit compatibility in the domain of computer vision refers to the harmonious and aesthetical assembly of different categories of fashion items as one complete outfit. To achieve maximum compatibility, it not only needs to obtain a comprehensive understanding of the aesthetics behind of each fashion categories as well as the compatibility among different categories, but also to acquire comprehension of the social norms and cultural standards related to dressing behaviors.

This paper is aim to investigate the problem of predicting the outfit compatibility between the top and the bottom items. As shown in Figs. 4 and 5, the compatibility relationship can be further applied in outfit recommendation and match diagnosis. However, due to the challenges of visual feature detection and outfit

compatibility calculation, the modeling of the compatibility relationship of clothing is non-trivial.

On the one hand, there are two challenges in the visual feature detection. First, calculating compatibility between fashion items usually involves color, material, prints, style, and other attribute factors. Most of the existing methods use high-level features to represent clothing. However, high-level features do not cover all the attributes of fashion outfit, and some critical attributes such as color and prints are mostly reflected in low-level features [10]. Therefore, how to retain most of the attributes in visual features and fuse both high and low-level features is one of the problems we need to solve. Second, most of the existing fashion datasets have a large number of deformed and obscured clothes. The deformation and cover of the clothing make it difficult for us to extract visual features. To address the aforementioned problems, we propose a Multi-layer Non-local Feature Fusion (MNLFF) framework to extract the visual features of fashion outfit accurately. This framework includes both non-local operation [11] and feature fusion model. Specifically, the non-local operation is for solving clothing deformation and occlusion problems. It is used to capture the long-distance dependence, which refers to the relationship between two pixels with a certain distance on the image in a variety of dimensions such as time dimension, space dimension and space-time dimension. In this way, we can easily obtain the comprehensive semantic information of the outfit for compatibility detection even

[☆] Editor: Eckart Michaelsen

* Corresponding author.

E-mail address: sflu@zjut.edu.cn (S. Lu).

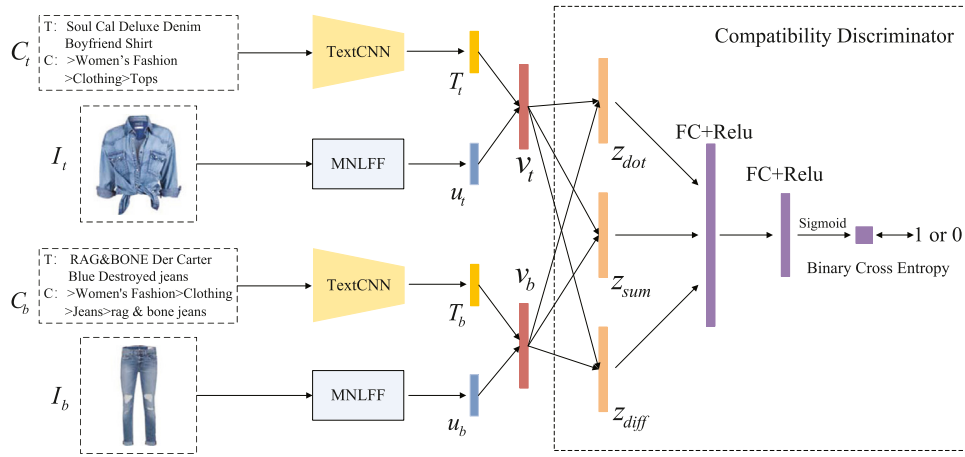


Fig. 1. Overview of the proposed network framework. The framework obtains the visual and semantic features of clothing from MNLFF and TextCNN, and then inputs these features into the compatibility discriminator to obtain the outfit compatibility score.

for a distorted outfit. In addition, the feature fusion model is developed to obtain a variety of clothing attribute factors, which fuses low-level features such as patterns and colors and the high-level features such as style and category.

On the other hand, there are three common methods for calculating compatibility: embedding method [7], sequence method [1] and discriminative method [12]. The embedding method embeds item features into a compatible space, and then calculates their Euclidean distance to learn compatibility. Although this method is simple, it cannot learn the relationship between fashion items effectively. The sequence method treats outfits as a time series, where multiple garments are involved. But it is not suitable for pairwise outfit compatibility prediction with only the top and bottom. The discriminative method we choose is to represent the compatibility prediction task as a simple binary classification problem. The closer the score is to 1, the higher the outfit compatibility is.

In this paper, we also add image-related context metadata to the entire network to provide rich semantic information and effectively implement the compatibility model. Extensive experiments and comparison with the state-of-the-art methods shows the superiority of our technique. Our main contributions are three-fold as follows:

1. We propose a Multi-layer Non-local Feature Fusion (MNLFF) framework to extract the rich visual features of fashion outfit accurately.
2. We add multi-modal information of fashion outfit to the discriminative method to improve the compatibility prediction of fashion items.
3. To the best of our knowledge, the experimental results demonstrate that our method shows superiority over state-of-the-art techniques in pairwise outfit compatibility prediction both in quantitative and qualitative.

2. Related work

Fashion is an important application area for computer vision and multimedia, wherein extensive research work has been carried out, including fashion image retrieval [13,14], clothing recognition [15,16], clothing analysis [17,18], attribute learning [19,20], clothing compatibility [1,8,9,16,21] and fashion recommendation [4,20].

The study focuses on fashion compatibility learning with an objective to score pairs of outfits based on images and contextual metadata of the outfits. Therefore, the review of related work con-

centrates on feature representation of fashion images and fashion compatibility learning.

2.1. Feature representation of fashion outfits

Based on the hierarchical structure of the convolutional neural network, we can obtain the low-level and high-level features of fashion clothing, respectively. Low-level features have higher resolution and contain more position and detail information. However, they have lower semantics and more noise as they have less convolutions. Differently, high-level features have stronger semantic information, but have lower resolution and poor ability to perceive details. For fashion images, the low-level features that represent the color and shape of the outfits are obtained from the earlier layers of the convolutional neural network, while the high-level features for the category and style of the outfits are obtained from the later layers.

Most techniques use only high-level features to predict the compatibility of fashion outfits. Han et al. [1] employed CNN features derived from the InceptionV3 model as the image representation. Song et al. [22] adopt the 50-layer residual network (ResNet-50) for fashion image representation learning. They ignored the attributes contained in the low-level features and the high-level features loss caused by multiple convolutions. Zou et al. [23] proposed that low-level features may directly determine the aesthetic feeling of fashion. Wang et al. [10] extracted features from different layers of CNN to build multiple comparison modules, and demonstrate the results of outfits compatibility prediction from different levels of visual features. However, their method does not combine high- and low-level features to improve the accuracy of outfits compatibility prediction. In this paper, we make comprehensive use of various levels of visual features, and fuse low-level features and high-level features to achieve the complementary advantages of multiple features, and obtain more robust and accurate prediction results.

2.2. Fashion compatibility learning

There are three methods of modeling outfits compatibility: embedding method, sequence method and discriminative method.

Some researchers calculate the Euclidean distance of item features in the embedding space to predict outfits compatibility. Veit et al. [9] used Siamese CNN architecture to predict compatibility between co-purchasing items. McAuley et al. [16] proposed a parametric distance transformation to predict the outfit compatibility. Cui et al. [24] introduced a content-based neural scheme that uses

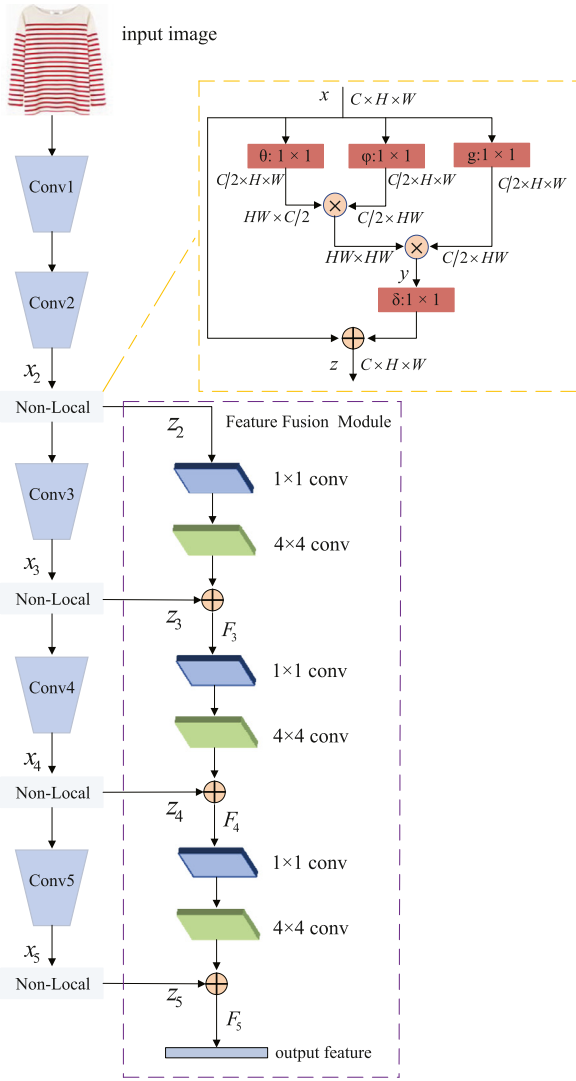


Fig. 2. The multi-layer non-local feature fusion network. The non-local block is inserted between the last four blocks of the ResNet network, and then non-local features are fused by 1×1 and 4×4 convolutions in the feature fusion module.

the Bayesian Personalized Ranking framework for predicting compatibility between fashion-based multi-modal data. However, the distance calculation in the embedding method is not an effective way to learn the interaction between item features.

Other existing methods attempt to directly model outfits as a time series. Li et al. [5] used a recurrent neural network to predict fashion outfits compatibility. Han et al. [1] regarded outfits as an orderly sequence, and predicted the next outfit under the condition of previous outfit through Bi-directional Long Short-Term Memory (Bi-LSTM). Based on Bi-LSTM, Nakamura and Goto [25] built an auto-encoder to learn the style embedding of outfits. However, an outfit is more like a set rather than a sequence because it is disordered. The sequence method is more suitable for predicting the compatibility of multiple garments rather than only top and bottom.

Also, there are some methods that cascade the feature of multiple clothing items, and then use discriminative method to predict outfit compatibility. Li et al. [12] used three conversion functions to transform the features of the top and bottom, and then used a multi-layer perceptron to calculate the compatibility score after aggregating the transformed features. Tangseng et al. [26] considered an outfit as a bag of fashion items and utilized deep neu-

	Top	$b_{positive}$	$b_{negative}$		Top	$b_{positive}$	$b_{negative}$
Example 1							
	Ours		BPR-DAE		Ours		BPR-DAE
Example 3							
	Ours		BPR-DAE		Ours		BPR-DAE
Example 5							
	Ours		BPR-DAE		Ours		BPR-DAE

Fig. 3. The comparison between our method and BPR-DAE on testing triples. The circle and the cross represent correct and wrong judgments, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

ral networks to produce a score for a fixed-length representation of outfits. Their costume representation is an ordered concatenation of term embeddings from a convolutional network, and non-existent terms are expressed by means of mean term representation to handle variable-length input. Lin et al. [27] developed a new negative sampling scheme to choose outfits that individuals are unlikely to like based on what they previously liked, then cascade all image features to predict. In this method, we use discriminative method to predict the pairwise outfits compatibility.

3. Methodology

3.1. Network architecture

The proposed network is composed of three components: a Multi-layer Non-Local Feature Fusion framework (MNLFF), a text semantic extractor, and a compatibility discriminator. Fig. 1 illustrates the architecture of the proposed network. Two input images, a top (I_t) and a bottom (I_b) are first adjusted to a size of 224×224 and sent to the proposed multi-layer non-local feature fusion network to extract the visual feature vectors u_t and u_b . Meanwhile, we input the context metadata (C_t , C_b) related to the two images to a text semantic extractor to obtain the semantic feature vectors T_t and T_b . Then, the visual feature vectors and the semantic feature vectors are connected at the channel level to obtain connection vectors v_t and v_b . Finally, v_t and v_b are sent to a compatibility discriminator to calculate the compatibility score that determines whether the input images I_t and I_b are compatible or not.

3.2. Multi-layer non-local feature fusion framework

As shown in Fig. 2, the proposed MNLFF framework includes three parts: a resnet-18 network, non-local blocks and a feature fusion module.

3.2.1. Non-local block

The fashion analysis is a challenging task due to the large variation and non-rigid deformation of clothes in images. To deal with this issue, we add non-local blocks to fashion outfit compatibility prediction model to exploit rich knowledge for clothes. The non-local operation captures long-range global dependencies within a clothing image, and enhances the representation power of output

features. With this operation, the output features are well facilitated to have global contextual information for input clothing image.

The non-local block [28] is shown in the orange dotted box of Fig. 2. It is inserted to each layer from conv2 to conv5. Let x denote the input feature of non-local block, which is the feature map of each layer from conv2 to conv5. Compared with the traditional convolution operation, the non-local block focuses on calculating the remote dependence relationship between any two different points and summing the weighted input features. It is formulated as:

$$y = \frac{1}{C(x)} \exp(\theta(x)^T \varphi(x)) g(x), \quad (1)$$

where $\theta(\cdot)$, $\varphi(\cdot)$ and $g(\cdot)$ are all 1×1 convolutions, and $C(x)$ is the normalize factor of x . By utilizing the non-local operation with a residual connection, we obtain z :

$$z = \delta(y) + x, \quad (2)$$

where $\delta(\cdot)$ is a 1×1 convolution.

3.2.2. Feature fusion

It is conceivable that the low-level features of the convolutional neural network carry information about outfit colors and silhouettes, while the high-level features mainly include outfit style and category information [10]. Therefore, in order to obtain as many features as possible, we fuse the output features of each non-local block.

As shown in the purple dotted box in Fig. 2, Let z_i denote the input feature of the feature fusion module, which is obtained by the i th convolution block feature x_i through non-local operation. It is formulated as:

$$z_i = NL(x_i), \quad i = 3, 4, 5 \quad (3)$$

where $NL(\cdot)$ represents the non-local operation mentioned in the previous section. In the feature fusion module, z_i from different layers are adjusted to the same size by convolution for fusion. The formula of the feature fusion module is defined as:

$$F_i = \begin{cases} z_2 & i = 2 \\ C_4(ReLu(C_1(F_{i-1}))) + z_i & i = 3, 4, 5 \end{cases} \quad (4)$$

where F_i is a fusion feature map of the i th layer, F_5 is the final visual output feature. C_1 is a 1×1 convolution and C_4 is a 4×4 convolution.

3.3. Compatibility discriminator

In the compatibility discriminative method [12], we learn a discriminative function $p(\cdot)$, which takes the joint feature representation (v_t, v_b) of a group of compatible or incompatible pairs of clothing items as input. Then, the function generates a score to evaluate the compatibility of the top and bottom. For the input feature vectors (v_t, v_b) , we get three feature transformation vectors z_{dot} , z_{diff} , z_{sum} by following corresponding transformation functions:

$$\begin{aligned} z_{dot}(v_t, v_b) &= [v_{t1}v_{b1}, v_{t2}v_{b2}, \dots, v_{td}v_{bd}]^T, \\ z_{diff}(v_t, v_b) &= [(v_{t1} - v_{b1})^2, (v_{t2} - v_{b2})^2, \dots, (v_{td} - v_{bd})^2]^T, \\ z_{sum}(v_t, v_b) &= [v_{t1} + v_{b1}, v_{t2} + v_{b2}, \dots, v_{td} + v_{bd}]^T. \end{aligned} \quad (5)$$

$z_{dot}(v_t, v_b)$ is an element-wise multiplication of two vectors, $z_{diff}(v_t, v_b)$ is the operation of squared element-wise difference between two vectors, and $z_{sum}(v_t, v_b)$ is an element-wise sum operation of two vectors. d represents the length of the vector.

Table 1

Performance comparison among ablation research and different methods in terms of AUC. “A” is the basic framework of the Discriminative method. “B”, “C”, “D” respectively add new operations on the basis of the previous layer. “D” is our complete framework.

Approach	AUC
POP	0.4206
RAND	0.5094
Bi-LSTM [3]	0.5502
BPR-DAE [7]	0.6026
ExIBR [16]	0.6366
A: Discriminative Method (only image)	0.6672
B: A + text description	0.7135
C: B + feature fusion	0.7528
D: C + non-local block(complete framework)	0.7707

Then, z_{dot} , z_{diff} , z_{sum} are put into the discriminant function $p(\cdot)$ simultaneously. The operation of the entire compatibility discriminator is defined as:

$$p(I_i, I_j) = \text{sigmoid}(\text{Relu}(f_{c2}(\text{Relu}(f_{c1}(c(z_{dot}, z_{diff}, z_{sum})))))), \quad (6)$$

where $c(\cdot, \cdot, \cdot)$ is the cascade eigenvector of z_{dot} , z_{diff} , z_{sum} , and f_{c1} and f_{c2} are two fully connected layers.

4. Experimental results

4.1. Dataset

In order to provide the model with rich data, facilitate the experiment, and ensure the quality of the evaluation, we adopt the public real-world dataset FashionVC [7] by crawling outfits from Polyvore. This dataset includes 20,726 outfits with 14,871 tops and 13,663 bottoms. The outfits of the FashionVC are composed by on-line fashion experts. Each fashion item is associated with a product image of white background, a list of category keywords and the title description.

The original outfits are regarded as a positive dataset with 20,726 positive samples. Based on the original dataset, we create a negative dataset by randomly selecting bottoms. Specifically, for each original outfit, we fix the top in an outfit and then randomly select a different bottom to replace the original bottom. For both positive and negative datasets, each dataset is split into 80% for training and 20% for testing, based on the number of outfits.

4.2. Implementation details

We implement our method using Pytorch 1.1.0 framework and all experiments are run on a commodity workstation with a NVIDIA GTX TITAN Xp graphics card. The memory of our GPU is 12 GB. We use 512-dimensional convolution features derived from the MNLFF framework as image features. The text representations are 64-dimensional feature vectors derived from TextCNN [29]. We set $d_v = 576$ as the final size of the input features for the compatibility discriminator. Then, we use Binary Cross Entropy loss to learn the parameters of the fashion outfit compatibility prediction model. Each method is trained for 30 epochs using Adam optimizer with a learning rate of $10e - 4$. Our training process costs about 10 h.

4.3. Quantitative results

In order to evaluate our proposed model, we utilize the Area Under Curve (AUC) as evaluation criteria to measure the quality of outfits compatibility predictions. Table 1 shows the quantitative results on FashionVC dataset.









Query	Matching List									
	1	2	3	4	5	6	7	8	9	10
	Ours									
	B-D									
	Ours									
	B-D									

Fig. 4. The outfit compatibility match comparison between BPR-DAE (B-D) and our method.

In order to analyze the effects of the components, we add one new component to our baseline Discriminative Method at each time. Table 1 gives AUC for several configurations. The baseline configuration (A) corresponds to the basic setup of Discriminative method, which employs only outfit images as input for the network. It achieves competitive results compared with other state-of-the-art methods. We first improve our baseline by adding text description as extra input to the network as shown in configuration (B). Compared with the configuration (A), the AUC score increases by about 5% by adding text features. The combination of visual and textual features of outfit is beneficial to enrich the details of outfit and improve the experimental results. On the basis of configuration (B), we further improve the experimental performance by constructing a new baseline (C), which fuses the low-level and high-level features of the outfit images. In order to obtain global features, we add a non-local block before each fused feature of the configuration (C) and the final experimental results are shown in configuration (D).

To evaluate the performance of our method, we compare our proposed model against other state-of-the-art deep models, such as POP, RAND, Bi-LSTM [1], ExIBR [16] and BPR-DAE [7]. As shown in Table 1, the accuracy of the proposed method is much higher than the results obtained by other existing methods in terms of AUC scores. POP method aims to recommend outfits with high popularity. But the popularity is not equal to compatibility. The most fashionable and popular clothing is not necessarily suitable for all other clothing. The RAND method randomly selects clothes for recommendation, which is full of uncertainty. Bi-LSTM is a time series method, which is more suitable for predicting the compatibility of multiple garments, rather than just tops and bottoms. Although ExIBR and BPR-DAE outperform previous methods by considering the contextual information, they still lack sufficient visual features for fashion compatibility prediction. Compared with them, our method achieve better results by non-local operations and feature fusion in the MNLFF network.

To further demonstrate its effectiveness of method, we compare it with the method of Wang et al. [10]. However, their approach is a typical compatibility modeling method for multiple clothing items and the dataset they use is Polyvore-T. It includes 19,835 outfits with each has at least 3 items and up to 5 items, and 15,089 of them have both top and bottom items. Since our method is to predict the compatibility between paired clothing items, in order to apply it on their dataset, we select these 15,089 outfits and only keep the top and bottom items in each outfit. The AUC score of our model on this dataset is 0.7612. At the same time, we also train the method of Wang et al. [10] on this revised dataset, and its AUC score is 0.7607, which is much lower than 0.919 on Polyvore-T dataset as shown in their paper. As a re-

sult, our method achieves competitive results compared with Wang et al. [10].

4.4. Qualitative analysis

In addition to quantitative research, we also conduct qualitative analysis of well-trained model and other methods to evaluate the effectiveness of our method.

Outfit compatibility prediction To intuitively illustrate the impact of low-level features, the comparison between BPR-DAE and our method with several testing triples are shown in Fig. 3. Notably, as aforementioned before, each testing triplet (top, positive, negative) indicates that the positive bottom is preferred than negative bottom to match the given top. As shown in Fig. 3, the green circle and the red cross represent correct and wrong judgments, respectively. When given two bottoms - positive and negative bottom, our method achieves better results because of rich low-level features such as color and shape (see example 1 and example 6). In addition, as can be seen from the example 2 and example 4, BPR-DAE cannot predict positive bottoms because they are unable to extract visual features of deformed or obscured clothes. By contrast, thanks to the non-local block, our method improves the prediction results by capturing long-range global features.

Outfit compatibility match and rank Fig. 4 is the outfit compatibility match comparison between BPR-DAE and our method. We select a top/bottom as the query, and randomly select T bottoms/tops as the ranking candidates. Then, we calculate the compatibility score between each candidate and the query item by using our trained model, and select top ten with the highest scores. As shown in Fig. 4, both BPR-DAE and our method consider the similarity for evaluating compatibility. Most of the recommended items and query items are similar in some aspects, such as materials, colors and patterns. However, BPR-DAE does not fully consider the color and style between the top and bottom, for example, a black skirt is selected for the blue denim jacket. As our method incorporates low-level features of the image, our matching items are not only similar in color and texture, but also similar in visual style and patterns.

Moreover, we choose query items (top/bottom) from test dataset and randomly select ten candidates (bottoms/tops) with one positive item (ground truth) shown in the red box of Fig. 5. Then, we sort the candidate items according to the compatibility scores calculated between each candidate and the query item. Fig. 5 shows the outfit recommendation ranking comparison between BPR-DAE and our method. For these four query items, our method ranks the positive bottom at first places. Although BPR-DAE can rank the positive items in the top three in most cases,

Query	Ranking List									
	1	2	3	4	5	6	7	8	9	10
1 	Ours 									
	B-D 									
2 	Ours 									
	B-D 									
3 	Ours 									
	B-D 									
4 	Ours 									
	B-D 									

Fig. 5. The outfit recommendation ranking comparison between BPR-DAE (B-D) and our method. The clothes with the bounding box is the ground truth that actually matches the query item. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the performance is not stable enough. In the query 3, it ranks the positive top in the seventh place.

5. Conclusion

In this paper, we propose a Multi-layer Non-Local Feature Fusion network, which uses non-local operations to capture global information on the image, and fuses low-level features and high-level features as visual representation of the image. We have applied it into pairwise fashion outfit compatibility prediction, which is essential for fashion recommendation and outfit match diagnose. However, our method still has several limitations. First, although we have improved the performance of fashion outfit compatibility prediction, the cause of whether the outfit is compatibility or not is still unknown. Therefore, how to explain the fashion compatibility is one of the future researches. Meanwhile, customer personalized fashion preferences is important for fashion recommendation, we will consider it when predicting fashion compatibility in the future.

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.

Acknowledgments

This work is supported by Zhejiang Provincial Natural Science Foundation of China (No. LY19F020027).

References

- [1] X. Han, Z. Wu, Y.-G. Jiang, L.S. Davis, Learning fashion compatibility with bidirectional LSTMs, in: *Proceedings of the 25th ACM International Conference on Multimedia*, 2017, pp. 1078–1086.
- [2] R. He, C. Packer, J. McAuley, Learning compatibility across categories for heterogeneous item recommendation, in: *2016 IEEE 16th International Conference on Data Mining (ICDM)*, IEEE, 2016, pp. 937–942.
- [3] W.-L. Hsiao, K. Grauman, Creating capsule wardrobes from fashion images, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7161–7170.
- [4] Y. Hu, X. Yi, L.S. Davis, Collaborative fashion recommendation: a functional tensor factorization approach, in: *Proceedings of the 23rd ACM International Conference on Multimedia*, 2015, pp. 129–138.
- [5] Y. Li, L. Cao, J. Zhu, J. Luo, Mining fashion outfit composition using an end-to-end deep learning approach on set data, *IEEE Trans. Multimed.* 19 (8) (2017) 1946–1955.
- [6] J. Oramas, T. Tuytelaars, Modeling visual compatibility through hierarchical mid-level elements, *arXiv preprint arXiv:1604.00036* (2016).
- [7] X. Song, F. Feng, J. Liu, Z. Li, L. Nie, J. Ma, NeuroStylist: neural compatibility modeling for clothing matching, in: *Proceedings of the 25th ACM International Conference on Multimedia*, 2017, pp. 753–761.
- [8] M.I. Vasileva, B.A. Plummer, K. Dusad, S. Rajpal, R. Kumar, D. Forsyth, Learning type-aware embeddings for fashion compatibility, in: *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 390–405.
- [9] A. Veit, B. Kovacs, S. Bell, J. McAuley, K. Bala, S. Belongie, Learning visual clothing style with heterogeneous dyadic co-occurrences, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4642–4650.
- [10] X. Wang, B. Wu, Y. Zhong, Outfit compatibility prediction and diagnosis with multi-layered comparison network, in: *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 329–337.
- [11] A. Buades, B. Coll, J.-M. Morel, A non-local algorithm for image denoising, in: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2, IEEE, 2005, pp. 60–65.
- [12] K. Li, C. Liu, R. Kumar, D. Forsyth, Using discriminative methods to learn fashion compatibility across datasets, *arXiv preprint arXiv:1906.07273* (2019).
- [13] M. Hadi Kiapour, X. Han, S. Lazebnik, A.C. Berg, T.L. Berg, Where to buy it: matching street clothing photos in online shops, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3343–3351.
- [14] K. Yamaguchi, M. Hadi Kiapour, T.L. Berg, Paper doll parsing: retrieving similar styles to parse clothing items, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 3519–3526.
- [15] Z. Liu, P. Luo, S. Qiu, X. Wang, X. Tang, DeepFashion: powering robust clothes recognition and retrieval with rich annotations, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1096–1104.
- [16] J. McAuley, C. Targett, Q. Shi, A. Van Den Hengel, Image-based recommendations on styles and substitutes, in: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2015, pp. 43–52.

- [17] X. Liang, L. Lin, W. Yang, P. Luo, J. Huang, S. Yan, Clothes co-parsing via joint image segmentation and labeling with application to clothing retrieval, *IEEE Trans. Multimed.* 18 (6) (2016) 1175–1186.
- [18] K. Yamaguchi, M.H. Kiapour, L.E. Ortiz, T.L. Berg, Parsing clothing in fashion photographs, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 3570–3577.
- [19] J. Huang, R.S. Feris, Q. Chen, S. Yan, Cross-domain image retrieval with a dual attribute-aware ranking network, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1062–1070.
- [20] S. Liu, J. Feng, Z. Song, T. Zhang, H. Lu, C. Xu, S. Yan, Hi, magic closet, tell me what to wear!, in: Proceedings of the 20th ACM International Conference on Multimedia, 2012, pp. 619–628.
- [21] I. Tautkute, T. Trzciński, A.P. Skrupa, L. Brocki, K. Marasek, DeepStyle: multimodal search engine for fashion and interior design, *IEEE Access* 7 (2019) 84613–84628.
- [22] X. Song, X. Han, Y. Li, J. Chen, X.-S. Xu, L. Nie, GP-BPR: personalized compatibility modeling for clothing matching, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 320–328.
- [23] Q. Zou, Z. Zhang, Q. Wang, Q. Li, L. Chen, S. Wang, Who leads the clothing fashion: style, color, or texture? A computational study, *arXiv preprint arXiv:1608.07444* (2016).
- [24] Z. Cui, Z. Li, S. Wu, X.-Y. Zhang, L. Wang, Dressing as a whole: outfit compatibility learning based on node-wise graph neural networks, in: The World Wide Web Conference, 2019, pp. 307–317.
- [25] T. Nakamura, R. Goto, Outfit generation and style extraction via bidirectional LSTM and autoencoder, *arXiv preprint arXiv:1807.03133* (2018).
- [26] P. Tangseng, K. Yamaguchi, T. Okatani, Recommending outfits from personal closet, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017, pp. 2275–2279.
- [27] Y. Lin, M. Moosaei, H. Yang, Learning personal tastes in choosing fashion outfits, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019 0–0.
- [28] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7794–7803.
- [29] Y. Kim, Convolutional neural networks for sentence classification, *arXiv preprint arXiv:1408.5882* (2014).