Tripartite Graph Regularized Latent Low-rank Representation for Fashion Compatibility Prediction

Peiguang Jing, Jing Zhang, Liqiang Nie, Shu,Ye, Jing Liu, Yuting Su*

***Abstract*—In recent years, an increasing online shopping demand has greatly promoted the innovation and development of the fashion industry. Visual fashion analysis has become a prospective research topic in computer vision and multimedia fields. Among these studies, fashion compatibility analysis is required in many real applications, such as fashion recommendation, matching, and retrieval. However, learning fashion compatibility is nontrivial, not only due to the uncertain and sparse dependencies among fashion items but also the latent and mutual associations among multiple factors such as color, texture, style, and functionality. To better predict fashion compatibility, in this paper, we proposed a tripartite graph regularized latent low-rank representation method, named TGRLLR, for fashion compatibility prediction. In TGRLLR, to learn more low-dimensional and effective representations, we considered the latent low-rank representation by decomposing the original feature matrix in both the column and row directions to tackle the problem of insufficient observations. On this basis, we simultaneously exploited different regularization strategies to encode the structured correlations among features, the high-order relationships among items, and the geometrical structures of outfits for more informative representations. Extensive experiments conducted on a real-world dataset demonstrate the effectiveness of our proposed method compared with state-of-the-art methods.**

***Index Terms*—Fashion compatibility, latent low-rank representation, correlation, graph regularization.**

I. Introduction

In recent years, intelligent fashion analysis has attracted broad attentions due to the flourishing of e-commerce services and the increasing expansion of online shopping demand. Statistics show that the fashion industry generated a worldwide revenue of $481 billion in 2018, and this number is projected to rise to $713 billion by 2022[1]. Additionally, many interactive fashion communities, such as Pureple[2], Looklet[3], and ShopLook[4], have also been developed to enhance connections among shoppers, stylists, and brands or offer personalized fashion experience and suggestions.

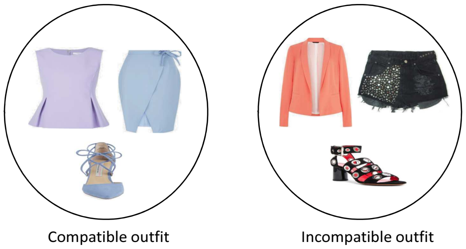


Fig. 1: Examples of compatible and incompatible outfits.

In response to this tendency, emerging research efforts have begun to address the increasingly urgent problems that occur in both the fashion industry and academia. Particularly, several standard fashion benchmarks with rich annotations are presented to promote fashion-related research, including fashion retrieval and recommendations, category classification, outfit composition, and fashion trend forecasting. For example, Hadi et al. [1] presented the Street2Shop dataset to address the matching problem between real-world garment items and the items presented in online shops. Ak et al. [2] developed the Shopping100k dataset to facilitate fashion search and retrieval, where each image consists of only clothing items with a simple background. Zou et al. [3] exploited an iterative process to build the FashionAI dataset for fine-grained attribute recognition tasks. Ge et al. [4] developed the DeepFashion2 dataset to tackle clothes detection, pose estimation segmentation, and retrieval tasks. Among these tasks, fashion compatibility is a fundamental but significant research topic that measures whether two or more fashion items are visually compatible. For example, when designing a fashion recommender system, one major problem that needs to be considered is learning the visual compatibility of fashion items. Examples of compatible and incompatible fashion outfits have been shown in Fig. 1.

An increasing number of fashion communities have emerged to encourage users to compose and share their favorite outfits for various occasions. However, not everyone specializes in matching clothes, making it a tedious and even annoying daily routine. To tackle this issue, several studies have recently been developed to analyze fashion compatibility automatically. For example, Han et

P. Jing, J. Zhang, S.Ye, J. Liu, and Y. Su are with the School of Electrical and Information Engineering, Tianjin University, 300072, China. E-mail: {pgjing@tju.edu.cn, jjzhang_2019@tju.edu.cn, yeshu330@outlook.com, jliu_tju@tju.edu.cn, ytsu@tju.edu.cn}. L. Nie is with the School of Computer Science and Technology, Shandong University, Jinan, 250000, China. E-mail: nieliqiang@gmail.com.

This work was supported in part by Tianjin Municipal Natural Science Foundation (Grant No. 20JCQNJC01210), National Natural Science Foundation of China (Grant No.61802277), and China Postdoctoral Science Foundation Funded Project (Grant No.2019M651038).

Yuting Su is the corresponding author (ytsu@tju.edu.cn).

[1] https://www.datafeedwatch.com
[2] https://purepleapp.wordpress.com
[3] http://www.looklet.com
[4] https://shoplook.io

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3062736, IEEE Transactions on Multimedia

al. [5] developed a fashion compatibility learning method by jointly considering the visual-semantic embedding and the compatibility relationships of fashion items. Vasileva et al. [6] presented an end-to-end type-aware fashion compatibility prediction method to learn an image embedding representation that respects the item type. Yin et al. [7] introduced a fashion compatibility knowledge learning method by taking visual compatibility relationships and fashion style information into account. Song et al. [8] proposed an attentive knowledge distillation-based neural compatibility method for fashion compatibility modeling. Cucurull et al. [9] introduced a context-aware graph neural network by leveraging the relational information between items. Yang et al. [10] developed a novel translation-based neural fashion compatibility modeling framework by exploiting a multi-relational knowledge representation learning strategy.

Previous studies on fashion compatibility aim to obtain visual embedding by leveraging the relationships among fashion items. However, inferring fashion compatibility not only incorporates more complex relationship patterns that can benefit fashion compatibility prediction, including patterns of features, items, and outfits, but also encounters uncertain and sparse dependency problems that are induced by small proportions of groups of items. Moreover, fashion compatibility analysis also involves the latent and mutual associations among multiple factors that are difficult to directly integrate, such as color, texture, style, and functionality. In this paper, we proposed a tripartite graph regularized latent low-rank representation (TGRLLR) method to make better fashion compatibility predictions. Inspired by the recent progress in low-rank representation techniques, we exploited the latent low-rank representation mechanism to decompose the original feature matrix in the column and row directions so that the more low-dimensional intrinsic representations and the impact of the unobserved information can be approached simultaneously. On this basis, we exploited different regularization strategies to explore different correlation patterns embedded in features, items, and outfits for more informative representations. In particular, we estimated a sparse inverse-covariance matrix to encode the latent dependencies among features and constructed a hypergraph regularizer to capture the high-order relationships among items. Furthermore, the local geometrical structure of fashion outfits was preserved by computing the principal angles among outfits. Fig. 2 gives a schematic illustration of our proposed TGRLLR method. Extensive experiments conducted on a real-world dataset demonstrate the effectiveness of our proposed method compared with state-of-the-art methods.

- We proposed a tripartite graph regularized latent low-rank representation method to predict the compatibility scores of fashion outfits, in which the global low-rank structure and the latent effect of insufficiently labeled samples can be jointly considered to ensure more comprehensive and effective representations.
- To make full use of the complex relations embedded in features, items, and outfits, we exploited different strategies, *i.e.,* sparse inverse-covariance estimation, hypergraph regularization, and geometric structure preservation, to address the uncertain and sparse dependency problems.
- We developed an effective optimization algorithm based on the alternating direction method of multipliers (ADMM) to optimize our proposed method. The experimental results demonstrated the effectiveness of our method compared with state-of-the-art methods.

The rest of this paper is organized as follows. In Section II, we briefly review the pioneering efforts related to fashion analysis and low-rank representation learning. In Section III, our proposed method is presented. In Section IV, we report the experimental results, followed by the conclusions and future work in Section V.

II. Related Work

A. Fashion Analysis

The flourishing development of the online fashion industry has attracted increasing research attentions in fashion applications, such as clothes semantic understanding [4] [11] [12], fashion retrieval [2] [13] [14], and fashion recommendation [15] [16] [17].

Regarding clothes semantic understanding, there are many attributes used to characterize middle-level semantics of fashion items, such as style, pattern, texture, and fabric. In the earlier stage of fashion studies, some of the work employed handcrafted features to solve attribute recognition and classification, landmark detection, and clothes segmentation. Inspired by the great success of deep learning techniques, recent research has focused on learning more effective and compact feature representations by leveraging deep neural networks. For example, Ak et al. [2] proposed an attribute manipulation generative adversarial network (AMGAN) to conduct multi-domain image-to-image translation and attribute-relevant region finding. Mall et al. [18] provided an expressive parametric model to automatically understand fashion styles and trends. Fashion retrieval selects similar garments from a gallery of candidate items. For instance, Gu et al. [19] designed a multi-modal embedding learning framework by jointly considering both the homogeneous and heterogeneous similarity constraints on multiple views. Valle et al. [13] presented a semantic compositional network (Comp-Net) in which clothing items are detected from an image and the probability of each item is used to compose a vector representation for the outfit. Kuang et al. [14] developed a graph reasoning network (GRNet) by formulating the local clothing regions as nodes, and the matching result between the query and gallery images can be achieved by reasoning on this graph.

Fashion recommendation refers to providing harmoniously matching clothes for the given queries, which can be roughly categorized into scenario-based [16] [20] and style-based methods [17] [21]. In particular, Zhang

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3062736, IEEE Transactions on Multimedia

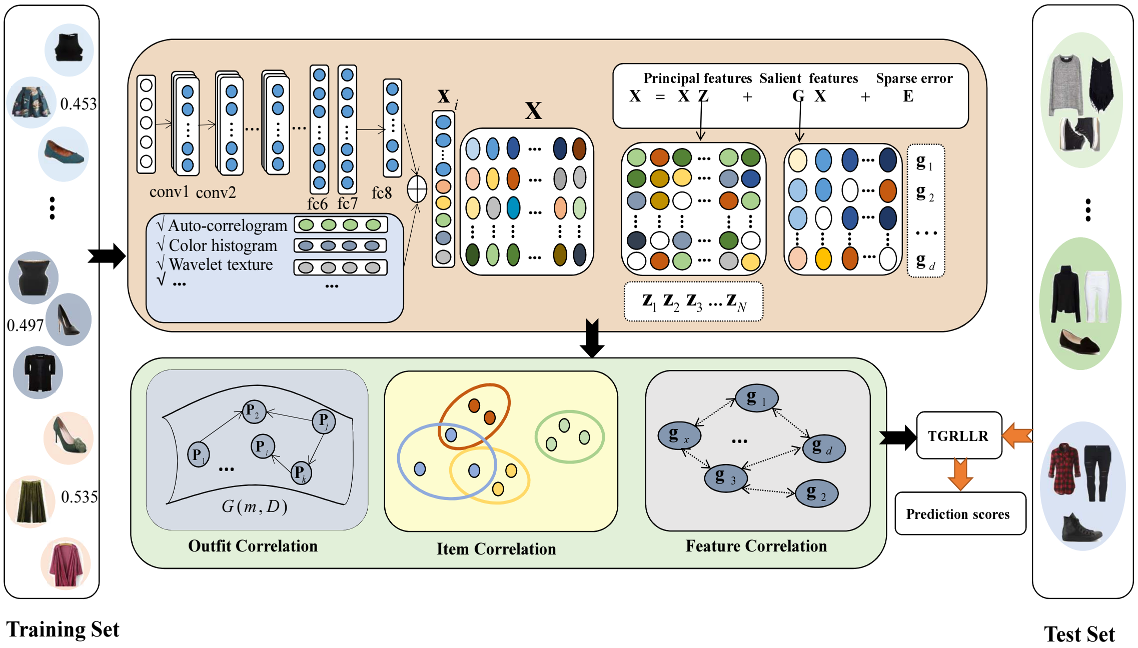


Fig. 2: Schematic illustration of our proposed TGRLLR method for fashion compatibility prediction. TGRLLR first exploits LatLRR to learn the principal feature part $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_N] \in R^{N \times N}$ and the salient feature part $\mathbf{G} = [\mathbf{g}_1^T; \mathbf{g}_2^T; \cdots; \mathbf{g}_D^T] \in R^{D \times D}$ from the original feature matrix $\mathbf{X} \in R^{D \times N}$, and then encodes three different types of latent correlation patterns by mapping outfits as points on a Grassmann manifold $\mathcal{G}(m, D)$, preserving the complex relationship of items through a hypergraph regularizer, and capturing the feature dependency through an inverse covariance matrix estimation term. Specifically, the original feature matrix $\mathbf{X}$ is the concatenated result of 1000D deep features extracted from the output of fc8 layer in VGGNet19 and 634D shallow features involving color, edge, and texture properties.

et al. [16] proposed combining a hybrid multi-label convolutional neural network with a support vector machine to recommend clothes for different travel scenarios. Jo et al. [20] applied a cross-domain generative adversarial network to recommend fashion designs that fit target scenarios. Hou et al. [22] introduced a semantic attribute explainable recommender that incorporates a semantic extraction network to learn the region-specific attribute representations. Remarkably, fashion compatibility plays a fundamental but significant role in fashion recommendation tasks. Although several significant efforts were devoted to exploring fashion compatibility prediction [5] [6] [23] [24], the above approaches only used the relationship of fashion items and neglected the correlation patterns embedded in the features and outfits. In addition, the uncertain and sparse dependency problems induced by very limited labeled samples have not been considered. In contrast, we exploited a latent low-rank representation mechanism to learn more comprehensive representations and take full advantage of inner correlation patterns among features, items, and outfits.

B. Low-rank Representation Learning

Low-rank representation learning has recently attracted extensive research attention due to the intuitively pleasing property in exploring low-dimensional structures, especially for corrupted data. The classical low-rank representation (LRR) [25] method aims to uncover the underlying low-dimensional subspace structures by imposing the nuclear-norm constraint on the latent representation component. Essentially, LRR takes the observed data itself as the dictionary to learn the lowest-rank representation, which is easily affected by insufficient and grossly corrupted observations. To capture the nonlinear geometrical structure that is easily ignored by LRR, Yin et al. [26] presented a non-negative sparse hyper-Laplacian regularized LRR model to solve image classification and clustering tasks. Xie et al. [27] introduced novel low-rank sparse preserving projections (LSPP) by iterating manifold learning and low-rank sparse representation to preserve the intrinsic geometric structure and reduce the negative effects of corruption. Wen et al. [28] presented a low-rank representation with an adaptive graph regularization method to derive a nonnegative graph structure for image clustering. To resolve this deficiency, Liu and Yan [29] proposed an improved version of LRR, named latent low-rank representation (LatLRR), in which the dictionary is constructed by using both observed and hidden data. Ren et al. [30] developed a latent low-rank and sparse embedding method for image feature extraction to ensure that the extracted representations benefit the classification tasks. However, the aforementioned works are generally suitable for classical classification and clustering tasks. Our fashion compatibility prediction task usually contains more complex structures represented by multiple types of group data behaviors. In our proposed method, we not only pursued low-rank intrinsic feature representation learning, but also concentrated on exploring the correlation patterns to distinguish more informative features for performance improvement.

III. Proposed Formulation

A. Problem formulation

We assume that the collection of N fashion items $\{x_1, x_2, \cdots, x_N\}$ is characterized by various types of feature extractors for comprehensively representing images. We normalize each type of features and concatenate them for generating a larger feature representation as input. Without loss of generality, we denote the final feature matrix as $\mathbf{X} = \left[\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\right] \in R^{D \times N}$, where $\mathbf{x}_i \in R^D$ is the feature vector of the ith item and D is the dimension of the concatenated features. In particular, LatLRR factorizes the original feature matrix into a principal feature part, a salient feature part, and a sparse error part by minimizing the following formulation:

$$\begin{aligned} &\min_{\mathbf{Z},\mathbf{G},\mathbf{E}} \|\mathbf{Z}\|_* + \|\mathbf{G}\|_* + \lambda \|\mathbf{E}\|_1 \\ &s.t.\ \mathbf{X} = \mathbf{X}\mathbf{Z} + \mathbf{G}\mathbf{X} + \mathbf{E}, \end{aligned} \tag{1}$$

where $\mathbf{Z} \in R^{N \times N}$ and $\mathbf{G} \in R^{D \times D}$ are used to encode the principal feature part and the salient feature part, respectively; $\mathbf{E} \in R^{D \times N}$ is the sparse noise part that fits the noise; $\|\cdot\|_1$ is the ℓ_1-norm chosen for sparse noise; and $\lambda > 0$ is a weighting parameter; $\|\cdot\|_*$ denotes the nuclear norm, *i.e.,*the sum of the singular values of the target matrix. Specifically, we follow a commonly used practice in rank minimization and approximate the rank constrain by the nuclear norm [25].

An intuitive explanation of LatLRR is that it uncovers the low-rank intrinsic structures by imposing low-rank constraints on both the feature and sample spaces and a sparse constraint on the error space. Although LatLRR is designed to handle specific scenarios such as insufficient samples and missing observations, it cannot cope well with tasks that exhibit complex structures, such as fashion compatibility prediction. For fashion compatibility prediction, more than one factor affects the harmony among items in an outfit. In particular, fashion compatibility prediction can be considered a kind of ternary relation task that connects features, items, and outfits. Motivated by manifold theory, we develop a tripartite graph regularized latent low-rank representation learning model to capture the structures of different kinds of relationships, *i.e.,* the geometric structures in terms of features, items, and outfits.

*1) **Feature correlation modeling:*** In LatLRR, matrix $\mathbf{G}$ is used to encode the salient features from the perspective of row projection. We assume $\mathbf{G} = \left[\mathbf{g}_1, \mathbf{g}_2, ..., \mathbf{g}_D\right]^T$ and each row of $\mathbf{G}$ can be deemed as a latent factor spanned by row space. When the number of factors is large, the pairwise factor relations exhibit sparse patterns since a factor cannot be helpful to all of the other factors. Moreover, sparse relations are beneficial to reduce the risk of overfitting problems compared with dense settings. To achieve this, we present a regularization term to model multiple factors and their latent relations simultaneously as follows:

$$\min_{\mathbf{\Omega} \succeq \mathbf{0}} \underbrace{Tr\left(\mathbf{G}^T \mathbf{\Omega}^{-1} \mathbf{G}\right)}_{\mathcal{D}(\mathbf{\Omega},\mathbf{G})} + \underbrace{\varepsilon \|\mathbf{\Omega} \odot \mathbf{H}\|_1}_{\mathcal{R}(\mathbf{\Omega},\mathbf{H})}, \tag{2}$$

where $\mathcal{D}(\mathbf{\Omega}, \mathbf{G})$ is the data-fidelity term based on the target matrix $\mathbf{\Omega}$, $\mathcal{R}(\mathbf{\Omega}, \mathbf{H})$ is the ℓ_1-norm regularized term obtained by multiplying $\mathbf{\Omega}$ by the prior matrix $\mathbf{H}$ in an elementwise way, ε is a tuning parameter controlling the amount of ℓ_1 shrinkage.

It should be noted that when $\mathbf{\Omega} \propto \mathbf{I}$, where $\mathbf{I}$ denotes an identity matrix, $\mathcal{D}(\mathbf{\Omega}, \mathbf{G})$ is reduced to the Frobenius norm regularization on $\mathbf{G}$, and if $\mathbf{\Omega}$ is set as a diagonal matrix, it becomes the weighted Frobenius norm regularization on $\mathbf{G}$. As pointed out in [31], when $\mathbf{\Omega}$ is restricted to be positive semidefinite, it can be regarded as a covariance matrix to characterize the pairwise relations between factors. Without loss of generality, in our case, we set $\mathbf{\Omega}$ to be positive semidefinite and $\mathbf{H}$ to be an identity matrix. The correlation patterns at the feature level are formulated as follows:

$$\min_{\mathbf{\Omega} \succeq \mathbf{0}} Tr\left(\mathbf{G}^T \mathbf{\Omega}^{-1} \mathbf{G}\right) + \varepsilon \|\mathbf{\Omega}\|_1. \tag{3}$$

*2) **Item correlation modeling:*** Hypergraphs are widely used in various visual tasks, such as clustering, classification, and retrieval. In contrast to a traditional graph, a hypergraph is able to capture group information that incorporates the high-order relationship among three or more vertices. Specific to fashion compatibility prediction, the complex relationship among fashion items can be naturally modeled by hypergraphs since each fashion item belongs to at least one fashion outfit and each fashion outfit contains varying amounts of items.

Hypergraph $G\left(V, E, \mathbf{W}\right)$involves a set of vertices V, hyperedges E, and diagonal hyperedge weight matrix $\mathbf{W}$. Each hyperedge e is a subset of the vertices and is assigned a positive weight $w(e)$. We denote the incident matrix $\mathbf{H} \in R^{|V| \times |E|}$, where $|V|$ and $|E|$ are the number of vertices and edges, respectively. The element $h(v, e) = 1$ if e is said to be incident with a vertex v, *i.e.,* $v \in e$, and $h(v, e) = 0$ otherwise. We also denote $\mathbf{D}_v$ and $\mathbf{D}_e$ as the diagonal matrices of the vertex and the hyperedge degrees, respectively. Based on $\mathbf{H}$, the vertex degree and the edge degree are calculated as follows:

$$d\left(v\right) = \sum_{e \epsilon E} w\left(e\right) h(v, e) \ \text{ and } \ \delta\left(e\right) = \sum_{v \epsilon V} h(v, e). \tag{4}$$

In our case, vertex v_i corresponding to fashion item $\mathbf{x}_i$ represents the i-th column of feature matrix $\mathbf{X}$. Hyperedge e_j is encoded as the j-th fashion outfit and binary incidence matrix $\mathbf{A} \in R^{M \times N}$ is signed to the incident matrix $\mathbf{H}$, characterizes the correspondence between fashion items and outfits, where M denotes the total number of fashion outfits. For the hyperedge weight, we adopt a similar scheme in [32] by considering the hyperedge e as a clique and computing the cumulative heat kernel of pairwise vertices in this clique as follows:

$$w\left(e\right) = \frac{1}{\delta(e)\left(\delta(e) - 1\right)} \sum_{v_i, v_j \in e} \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{2\sigma^2}\right), \tag{5}$$

where radius parameter σ is simply set as the median of the Euclidean distances over all the pairwise vertices.

By denoting $\mathbf{Z} = [\mathbf{z}_1, \cdots, \mathbf{z}_N]$, which uncovers the global low-rank structure of all samples in a latent feature space, normalized hypergraph Laplacian regularizer $\Omega(\mathbf{Z})$ is obtained as follows:

$$\begin{aligned}\Omega(\mathbf{Z}) &= \frac{1}{2}\sum_{e \epsilon E}\sum_{v_i, v_j \epsilon e}\frac{w(e)}{\delta(e)}\left\|\frac{\mathbf{z}_i}{\sqrt{d(v_i)}} - \frac{\mathbf{z}_j}{\sqrt{d(v_j)}}\right\|_2^2. \\ &= Tr\left(\mathbf{Z}\mathbf{L}_A\mathbf{Z}^T\right),\end{aligned} \tag{6}$$

where $\mathbf{L}_A = \mathbf{I} - \mathbf{D}_v^{-1/2}\mathbf{H}\mathbf{W}\mathbf{D}_e^{-1}\mathbf{A}\mathbf{D}_v^{-1/2}$ is defined as the normalized hypergraph Laplacian matrix and $\mathbf{W}$ denotes diagonal matrix of the hyperedge weights

$$W(e, j) = \begin{cases} w(e) & if\ e = j \\ 0 & if\ e \neq j. \end{cases}$$

*3) **Outfit correlation modeling:*** To obtain the latent representations of fashion outfits, a straightforward and efficient strategy is to build a connection between a fashion outfit and fashion items through the prior binary incidence matrix $\mathbf{A}$, that is

$$\min_{\mathbf{C}}\left\|\mathbf{C} - [\mathbf{XZ}; \mathbf{GX}]\,\mathbf{A}^T\right\|_F^2, \tag{7}$$

where $\|\cdot\|_F$ is the Frobenius norm; $\mathbf{C} = [\mathbf{c}_1, \cdots, \mathbf{c}_M] \in R^{2D\times M}$ is the latent low-dimensional representation of the fashion outfits; and $\mathbf{c}_i \in R^{2D}$ is the feature vector of the ith fashion outfit. It is noted that principal feature part $\mathbf{XZ}$ and salient feature part $\mathbf{GX}$ represent the original feature matrix from different perspectives; therefore, we concatenate them together to utilize the complementary information.

To further guide the process of feature representation, it is reasonable to assume that if two fashion outfits are close to each other in the original space, then the representations of these two fashion outfits should be kept close together in a new space. A general scheme is to minimize the following objective function:

$$\min_{\mathbf{C}}\sum_{i,j=1, i\neq j}^{M}\|\mathbf{c}_i - \mathbf{c}_j\|_2^2\,\hat{S}_{ij} = Tr(\mathbf{C}\mathbf{L}_s\mathbf{C}^T), \tag{8}$$

where $\mathbf{L}_s = \mathbf{D} - \hat{\mathbf{S}}$ is the graph Laplacian matrix; $\hat{\mathbf{S}} \in R^{M\times M}$ is a weight matrix obtained by the sum of the canonical correlations and $\mathbf{D}$ is the diagonal degree matrix with $D_{ii} = \sum_j \hat{S}_{ij}$.

Let $\mathbf{Q}_1$ and $\mathbf{Q}_2$ be two orthogonal matrices of size D by m, which represent two points on Grassmann manifold $\mathcal{G}(m, D)$, the distance between them can be measured by principal angles $\theta = [\theta_1, \theta_2, \cdots, \theta_m]^T$, which can be directly obtained from the singular value decomposition (SVD) of matrix $\mathbf{Q}_1^T\mathbf{Q}_2$ as follows:

$$\mathbf{Q}_1^T\mathbf{Q}_2 = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T, \tag{9}$$

where $\mathbf{U}, \mathbf{V} \in R^{D\times m}$ are the left and right singular matrices and $\mathbf{\Lambda} = \mathrm{diag}\{\cos\theta_1, \cdots, \cos\theta_m\}$. The cosines of the principal angles are also known as canonical correlations, and the similarity between two points is calculated by the sum of the canonical correlations.

Specifically, we have M outfits $\{\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_M\}$, where $\mathbf{X}_i \in R^{D\times m_i}$ represents the feature matrix of the ith outfit and m_i is the number of items contained in this outfit. Taking the ith fashion outfit as an example, it is characterized by an orthogonal basis matrix $\mathbf{P}_i \in R^{D\times m}$, *s.t.* $\mathbf{X}_i\mathbf{X}_i^T \simeq \mathbf{P}_i\mathbf{\Lambda}_i\mathbf{P}_i^T$, where $\mathbf{\Lambda}_i$ and $\mathbf{P}_i$ correspond to the top m largest eigenvalues and eigenvectors of $\mathbf{X}_i\mathbf{X}_i^T$, respectively. Based on this approach, the weight matrix $\mathbf{S}$ can be obtained by extending original feature space to an underlying Grassmann manifold.

By combining Eq. (7) and Eq. (8), the correlation pattern embedded in outfits can be modeled as follows:

$$\min_{\mathbf{C}}\left\|\mathbf{C} - [\mathbf{XZ}; \mathbf{GX}]\,\mathbf{A}^T\right\|_F^2 + \alpha Tr(\mathbf{C}\mathbf{L}_s\mathbf{C}^T). \tag{10}$$

where α is a trade-off parameter.

B. Incorporating Supervised Information

The latent feature representations of fashion items can be learned by making full use of the correlation pattern in terms of the feature, item, and outfit. However, a correlation pattern is not sufficient due to the lack of label information. To address this issue, we regard the compatibility prediction of fashion outfits as a regression problem since the fashion compatibility scores are continuous. We adopt the widely used Lasso-type regression approach, which considers linear dependencies $\mathbf{w} \in R^{2D}$ to build the connections between latent feature matrices $\mathbf{C}$ and output score vector $\mathbf{y} \in R^l (l < M)$, where l is the number of labeled fashion outfits. After adding a sparsity regularization to the least squares loss part, we obtain a typical Lasso problem as follows:

$$\min_{\mathbf{w}}\left\|\mathbf{C}^T\mathbf{w} - \mathbf{y}\right\|_2^2 + \varphi\|\mathbf{w}\|_1, \tag{11}$$

where φ is to balance the tradeoff between the empirical loss and the regularization penalty.

By integrating the functions in Eqs. (1), (3), (8) and (10) with Eq. (11), we propose the following objective function:

$$\begin{aligned}\min_{\mathbf{Z},\mathbf{G},\mathbf{E},\mathbf{\Omega},\mathbf{C},\mathbf{w}} & \|\mathbf{Z}\|_* + \|\mathbf{G}\|_* + \lambda\|\mathbf{E}\|_1 + \varepsilon\|\mathbf{\Omega}\|_1 + \varphi\|\mathbf{w}\|_1 \\ & + \eta Tr(\mathbf{G}^T\mathbf{\Omega}^{-1}\mathbf{G}) + \beta Tr(\mathbf{Z}\mathbf{L}_A\mathbf{Z}^T) + \alpha Tr(\mathbf{C}\mathbf{L}_s\mathbf{C}^T) \\ & + \gamma\left\|\mathbf{C} - [\mathbf{XZ}; \mathbf{GX}]\,\mathbf{A}^T\right\|_F^2 + \phi\left\|\mathbf{C}^T\mathbf{w} - \mathbf{y}\right\|_2^2 \\ & s.t.\ \mathbf{X} = \mathbf{XZ} + \mathbf{GX} + \mathbf{E}, \mathbf{\Omega} \succeq \mathbf{0},\end{aligned} \tag{12}$$

where ϕ, β, η, and γ are trade-off parameters.

C. Optimization

We apply the alternating direction method of multipliers (ADMM) to solve Eq. (12) by dividing a complex problem into easily handled subproblems. We first introduce auxiliary variable $\mathbf{S}$ to relax $\mathbf{G}$. By introducing two Lagrange multipliers, $\mathbf{Y}_1$ and $\mathbf{Y}_2$, and penalty parameter $\mu > 0$, we then obtain the augmented Lagrangian function $\mathcal{L}(\mathbf{Z}, \mathbf{G}, \mathbf{S}, \mathbf{E}, \mathbf{C}, \mathbf{\Omega}, \mathbf{w})$. For simplicity, we merge all of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3062736, IEEE Transactions on Multimedia

the quadratic terms into a single term and formulate the following:

$$\begin{aligned} &\mathcal{L}\left(\mathbf{Z},\mathbf{G},\mathbf{S},\mathbf{E},\mathbf{C},\mathbf{\Omega},\mathbf{w}\right) \\ &= \left\|\mathbf{Z}\right\|_* + \left\|\mathbf{S}\right\|_* + \lambda\left\|\mathbf{E}\right\|_1 + \varepsilon\left\|\mathbf{\Omega}\right\|_1 + \varphi\left\|\mathbf{w}\right\|_1 \\ &\quad + \mathcal{I}_{\mathbf{\Omega}\succeq\mathbf{0}}(\mathbf{\Omega}) + H\left(\mathbf{Z},\mathbf{G},\mathbf{S},\mathbf{E},\mathbf{C},\mathbf{\Omega},\mathbf{w}\right), \end{aligned} \tag{13}$$

where $H\left(\mathbf{Z},\mathbf{G},\mathbf{S},\mathbf{E},\mathbf{C},\mathbf{\Omega},\mathbf{w}\right) = \eta Tr(\mathbf{G}^T\mathbf{\Omega}^{-1}\mathbf{G}) + \alpha Tr(\mathbf{C}\mathbf{L}_s\mathbf{C}^T) + \phi\left\|\mathbf{C}^T\mathbf{w}-\mathbf{y}\right\|_2^2 + \beta Tr(\mathbf{Z}\mathbf{L}_A\mathbf{Z}^T) + \gamma\left\|\mathbf{C}-\left[\mathbf{XZ};\mathbf{GX}\right]\mathbf{A}^T\right\|_F^2 + \mu/2\left\|\mathbf{G}-\mathbf{S}+\mathbf{Y}_2/\mu\right\|_F^2 + \mu/2\left\|\mathbf{X}-\mathbf{XZ}-\mathbf{GX}-\mathbf{E}+\mathbf{Y}_1/\mu\right\|_F^2$.

To better interpret the iteration process, we define $\mathbf{Z}_t$, $\mathbf{G}_t$, $\mathbf{C}_t$, $\mathbf{E}_t$, $\mathbf{\Omega}_t$, $\mathbf{w}_t$, $\mathbf{Y}_{1,t}$, $\mathbf{Y}_{2,t}$, and μ_t as the variables updated in the tth iteration. Under the ADMM framework, problem $\mathcal{L}\left(\mathbf{Z},\mathbf{G},\mathbf{S},\mathbf{E},\mathbf{C},\mathbf{\Omega},\mathbf{w}\right)$ w.r.t. each variable in the $(t+1)$th iteration is optimized with the following scheme:

For $\mathbf{Z}$: We can update $\mathbf{Z}$ by dropping the terms independent of $\mathbf{Z}$ as follows:

$$\begin{aligned} \mathbf{Z}_{t+1} &= \arg\min_{\mathbf{Z}}\left\|\mathbf{Z}\right\|_* + H\left(\mathbf{Z},\mathbf{G},\mathbf{S},\mathbf{E},\mathbf{C},\mathbf{\Omega},\mathbf{w}\right) \\ &= \arg\min_{\mathbf{Z}}\frac{1}{\tau\mu_t}\left\|\mathbf{Z}\right\|_* + \frac{1}{2}\left\|\mathbf{Z}-\mathbf{Z}_t+\frac{1}{\tau}\nabla_{\mathbf{Z}}\mathcal{L}\right\|_F^2, \end{aligned} \tag{14}$$

where $\nabla_{\mathbf{Z}}\mathcal{L}$ is the partial derivative of $\mathcal{L}$ with respect to $\mathbf{Z}$ defined as $\nabla_{\mathbf{Z}}\mathcal{L} = 2\beta\mathbf{Z}_t\mathbf{L}_A + 2\gamma(\mathbf{X}^T\mathbf{X}\mathbf{Z}_t\mathbf{A}^T\mathbf{A} - \mathbf{X}^T\mathbf{C}_t\mathbf{A}) + \mu_t\mathbf{X}^T\mathbf{X}\mathbf{Z}_t - \mu_t\mathbf{X}^T(\mathbf{X}-\mathbf{G}_t\mathbf{X}-\mathbf{E}_t+\mathbf{Y}_{1,t}/\mu_t)$ and $\tau = 1.25\left\|\mathbf{X}\right\|_F^2$.

For $\mathbf{S}$: We can update $\mathbf{S}$ by dropping the terms independent of $\mathbf{S}$ as follows:

$$\begin{aligned} \mathbf{S}_{t+1} &= \arg\min_{\mathbf{S}}\left\|\mathbf{S}\right\|_* + H\left(\mathbf{Z},\mathbf{G},\mathbf{S},\mathbf{E},\mathbf{C},\mathbf{\Omega},\mathbf{w}\right) \\ &= \arg\min_{\mathbf{S}}\frac{1}{\mu_t}\left\|\mathbf{S}_t\right\|_* + \frac{1}{2}\left\|\mathbf{S}_t-\mathbf{G}_t-\frac{\mathbf{Y}_{2,t}}{\mu_t}\right\|_F^2. \end{aligned} \tag{15}$$

Eq. (14) and Eq. (15) are standard nuclear norm minimization problems, which can be approximately solved by the singular value thresholding (SVT) algorithm [33].

For $\mathbf{G}$: By setting the derivative of $\mathcal{L}$ regarding $\mathbf{G}$ to zero, we have

$$\begin{aligned} &2\eta\mathbf{\Omega}_t^{-1}\mathbf{G}_{t+1} + \mathbf{G}_{t+1}(\mu_t\mathbf{I}+\mu_t\mathbf{X}\mathbf{X}^T+2\gamma\mathbf{X}\mathbf{A}^T\mathbf{A}\mathbf{X}^T) \\ &= \mu_t\left[\mathbf{S}_{t+1} + (\mathbf{X}-\mathbf{X}\mathbf{Z}_{t+1}-\mathbf{E}_t+\frac{\mathbf{Y}_{1,t}}{\mu_t})\mathbf{X}^T - \frac{\mathbf{Y}_{2,t}}{\mu_t}\right] \\ &\quad + 2\gamma\mathbf{C}_t\mathbf{A}\mathbf{X}^T. \end{aligned} \tag{16}$$

For $\mathbf{C}$: By setting the derivative of $\mathcal{L}$ regarding $\mathbf{C}$ to zero, we have

$$\begin{aligned} &\left(2\phi\mathbf{w}_t\mathbf{w}_t^T + 2\gamma\mathbf{I}\right)\mathbf{C}_t + 2\alpha\mathbf{C}_t\mathbf{L}_s \\ &= 2\gamma\left[\mathbf{X}\mathbf{Z}_{t+1};\mathbf{G}_{t+1}\mathbf{X}\right]\mathbf{A}^T + 2\phi\mathbf{w}_t\mathbf{y}^T. \end{aligned} \tag{17}$$

Then, Eq. (16) and Eq. (17) can be optimized by solving the Lyapunov equation.

For $\mathbf{w}$: We can optimize $\mathbf{w}$ by dropping the terms independent of $\mathbf{w}$ as follows:

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}}\left\|\mathbf{C}_{t+1}^T\mathbf{w}_t-\mathbf{y}\right\|_2^2 + \frac{\varphi}{\phi}\left\|\mathbf{w}_t\right\|_1. \tag{18}$$

The optimization of Eq. (18) can be solved by the well-known soft-shrinkage operator [34].

For $\mathbf{E}$: We can optimize $\mathbf{E}$ by dropping the terms independent of $\mathbf{E}$ as follows:

$$\mathbf{E}_{t+1} = \arg\min_{\mathbf{E}}\frac{1}{2}\left\|\mathbf{E}-\hat{\mathbf{E}}_t\right\|_F^2 + \frac{\lambda}{\mu_t}\left\|\mathbf{E}\right\|_1, \tag{19}$$

where $\hat{\mathbf{E}}_t = \mathbf{X}-\mathbf{X}\mathbf{Z}_{t+1}-\mathbf{G}_{t+1}\mathbf{X}+\mathbf{Y}_{1,t}/\mu_t$. The optimization of Eq. (19) can be solved by using the shrinkage operator.

For $\mathbf{\Omega}$: We can optimize $\mathbf{\Omega}$ by dropping the terms independent of $\mathbf{\Omega}$ as follows:

$$\mathbf{\Omega}_{t+1} = \arg\min_{\mathbf{\Omega}\succeq\mathbf{0}}\eta Tr(\mathbf{G}_{t+1}^T\mathbf{\Omega}^{-1}\mathbf{G}_{t+1}) + \varepsilon\left\|\mathbf{\Omega}\right\|_1. \tag{20}$$

Similar to the optimization of $\mathbf{Z}$, we solve Eq. (20) by applying the fast iterative shrinkage thresholding algorithm (FISTA) [35], which minimizes a combination of two convex functions. By defining $f(\mathbf{\Omega}) = \eta Tr\left(\mathbf{G}_{t+1}^T\mathbf{\Omega}^{-1}\mathbf{G}_{t+1}\right)$, Eq. (20) can be equivalently solved as follows:

$$\min_{\mathbf{\Omega}}\frac{l}{2}\left\|\mathbf{\Omega}-(\mathbf{\Omega}_t-\frac{1}{l}\nabla_{\mathbf{\Omega}}f(\mathbf{\Omega}))\right\|_F^2 + \varepsilon\left\|\mathbf{\Omega}_t\right\|_1, \tag{21}$$

where $\nabla_{\mathbf{\Omega}}f(\mathbf{\Omega}) = -\eta\mathbf{\Omega}_t^{-1}\mathbf{G}_{t+1}\mathbf{G}_{t+1}^T\mathbf{\Omega}_t^{-1}$ is the partial derivative of $f(\mathbf{\Omega})$ with respect to $\mathbf{\Omega}_t$ and $l = \eta\|\mathbf{\Omega}_t^{-2}\mathbf{G}_{t+1}^T\|_F^2$.

The optimization process of our proposed method is summarized in Algorithm 1.

Algorithm 1 The optimization procedure of the proposed TGRLLR

```
Input: Feature matrix $\mathbf{X}$, fashion compatibility score $\mathbf{y}$
Initialize: $\mathbf{G}_0 = \mathbf{S}_0 = \mathbf{C}_0 = \mathbf{E}_0 = \mathbf{Y}_{1,0} = \mathbf{Y}_{2,0} = \mathbf{0}$,
        $\mathbf{w}_0 = \mathbf{0}$, $\mu_0 = 0.1$, $\mu_{max} = 10^{10}$, $\rho = 1.1$.
While not converged do ;
  1) Fixed others and update $\mathbf{Z}_{t+1}$ using Eq. (14);
  2) Fixed others and update $\mathbf{S}_{t+1}$ using Eq. (15);
  3) Fixed others and update $\mathbf{G}_{t+1}$ using Eq. (16);
  4) Fixed others and update $\mathbf{C}_{t+1}$ using Eq. (17);
  5) Fixed others and update $\mathbf{w}_{t+1}$ using Eq. (18);
  6) Fixed others and update $\mathbf{E}_{t+1}$ using Eq. (19);
  7) Fixed others and update $\mathbf{\Omega}_{t+1}$ using Eq. (21);
  8) Update Lagrangian multipliers $\mathbf{Y}_{1,t+1}, \mathbf{Y}_{2,t+1}$ by:
     $\mathbf{Y}_{1,t+1} = \mathbf{Y}_{1,t} + \mu_t(\mathbf{X} - \mathbf{X}\mathbf{Z}_{t+1} - \mathbf{G}_{t+1}\mathbf{X} - \mathbf{E}_{t+1})$
     $\mathbf{Y}_{2,t+1} = \mathbf{Y}_{2,t} + \mu_t(\mathbf{G}_{t+1} - \mathbf{S}_{t+1})$
  9) Update parameter $\mu_{t+1} = \min(\rho\mu_t, \mu_{max})$;
End while
Output: Z, G, w.
```

D. Computational Complexity

To analyze the complexity of the proposed model, we consider that the number of samples is much larger than the dimensionality of the data. We find that the complexity of the algorithm mainly comes from the following aspects: 1) the calculation of the nuclear norm in steps 1 and 2; 2) solving the Lyapunov equation in steps 3 and 4; 3) the calculation of the inverse matrix in step 7. The computational complexity of the nuclear norm steps 1 and 2 are $O(N^3)$ and $O(D^3)$, respectively. The total computational complexity of solving the Lyapunov equation in steps 3 and

4 are $O(2D^3)$. The computational complexity of the inverse matrix in step 7 is $O(D^3)$. If the algorithm converges after t iterations, the computational complexity is approximately $O(4tD^3 + tN^3)$.

IV. Experiments and Results

A. *Experimental settings*

We investigated the effectiveness of our proposed method on the Polyvore dataset [5], which contains a total of 21,889 fashion outfits that were formed with 164,379 fashion items. Many high-quality fashion outfits were constructed by fashion experts and are favored by Polyvore users. Particularly, to evaluated the predicted performance of our proposed method with a small number of samples, we constructed a refined version of this dataset by selecting 4,800 outfits with uniform probability and split them into 3 parts. Fig. 3 shows three fashion outfits with different compatibility scores. For each of the parts, 1,200 fashion outfits were selected for training, and the remaining were selected for testing. Each outfit had several fashion items, and we removed the interference items and only kept 3 items, most of which included tops, bottoms and shoes. We considered the normalized ratio of the number of likes to views as the compatibility score. To obtain comprehensive descriptions for each fashion item, we not only extracted 1,000D middle-level features from the output of fc8 layer in VGGNet19 [36] but also considered five types of low-level visual features, including 144D color auto-correlogram features [37] capturing the spatial correlation of colors, 225D block-wise color moment features [38] representing color distributions of images through three statistics, 64D color histogram features quantized in LAB color space [39], 73D edge direction histogram features [40] encoding the distribution of the directions of edges, and 128D wavelet texture features [41] characterizing texture prosperities at multiple resolutions. Before training our model, we first exploited the ℓ_2-norm to address each type of feature and then concatenated all the features to generate new 1,634D feature vectors. Eventually, all feature vectors were further normalized to the same unit length under the same strategy. We empirically set the parameters as $\lambda = 1e{-}3$, $\varepsilon = 1e{-}2$, $\gamma = 1e-1$, $\phi = 1e-7$, $\varphi = 1e-5$, and $\mu = 1e1$. The trade-off parameters η, θ, and β in our model were selected by a grid-search approach. We first performed a grid search at a coarse level and then conducted a finer grid search to identify an ideal parameter. Finally, we set $\alpha = 1e-5$, $\eta = 25$, and $\beta = 30$ by default.

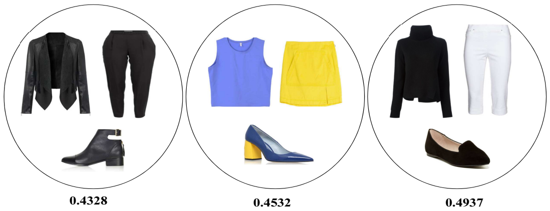


Fig. 3: Three samples of fashion outfits randomly selected from the Polyvore datasets.

To measure the consistency between the prediction results and ground truths, we employed the typical normalized mean squared error (nMSE) [42] as the measurement. The nMSE value, which is equal to the mean squared error divided by the variance of the ground truth, is defined as

$$nMSE = \frac{1}{M\sigma^2}\sum_{i=1}^{M}\left(y_i - \hat{y}_i\right)^2, \tag{22}$$

where M is the total number of fashion outfits; σ is the standard deviation from the ground truth; y_i and $\hat{y}_i$ are the real and the predicted score of the i-th outfits, respectively. The smaller the value of the nMSE is, the better the performance of the model.

B. *Experimental results*

In our experiments, we evaluated our proposed method with respect to convergence, component analysis, parameter sensitivity, a case study, and a comparison with state-of-the-art methods. According to their predicted results, we sorted all fashion outfits in descending order and selected the top $\{50, 100, 150, 200\}$ and bottom $\{50, 100, 150, 200\}$ fashion outfits to report their averaged fashion compatibility scores.

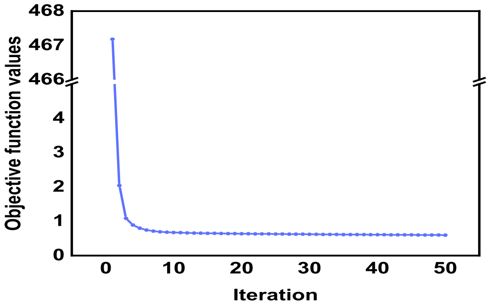


Fig. 4: The convergence curve of our proposed TGRLLR method. (The horizontal axis represents the number of iterations, and the vertical axis shows the objective function values).

1) Evaluation of Convergence: It is essential to investigate the convergence of optimization algorithm to guarantee the reliability of the experimental results. We randomly selected one trial to display the results. Since $\mathbf{Z}$ and $\mathbf{G}$ span the latent low-rank subspaces along the row and column directions to provide complementary feature representation, we calculated the variance between two sequential concatenated feature matrices with the following scheme:

$$D(t) = \left\|[\mathbf{X}\mathbf{Z}_t; \mathbf{G}_t\mathbf{X}] - [\mathbf{X}\mathbf{Z}_{t-1}; \mathbf{G}_{t-1}\mathbf{X}]\right\|_F. \tag{23}$$

Fig. 4 shows the convergence curve with an increasing number of iterations. From the figure, we can observe that the curve has a dramatic drop during the first few iterations and tends to be steady after 40 iterations, meaning that the feature representations become increasingly insensitive to the number of iterations. Thus, we used the relative change falling below a threshold of 0.6084 and a maximum of 40 iterations as the stopping criteria for our proposed method.

TABLE I: Prediction performance comparison of the involved components in our proposed method.

	noGC	**noHR**	**noFC**	**noReg**	**TGRLLR**
Top 50	0.272	0.278	0.266	0.291	0.272
Top 100	0.282	0.276	0.268	0.276	0.275
Top 150	0.266	0.262	0.251	0.261	0.268
Top 200	0.259	0.259	0.250	0.260	0.263
Bottom 200	0.210	0.210	0.218	0.209	0.206
Bottom 150	0.204	0.204	0.210	0.207	0.201
Bottom 100	0.200	0.199	0.207	0.203	0.201
Bottom 50	0.186	0.186	0.202	0.187	0.182
nMSE	0.347	0.351	0.369	1.027	**0.347**

2) Evaluation of Components: To illustrate the effectiveness of each component involved in our proposed method, we conducted experiments from the following perspectives:

- **noGC**: We considered the influence of the local geometric structure consistency constraint on fashion outfits by setting $\alpha = 0$.
- **noHR**: We considered the influence of the hypergraph regularization term by setting $\beta = 0$.
- **noFC**: We considered the effect of feature correlation pattern learning by setting $\eta = 0$.
- **noReg**: We considered separating feature representation learning and Lasso-type regression learning into two separate steps.

TABLE I shows the prediction performance comparison of the involved components in terms of the nMSE. From the table, we can see that the predicted compatibility scores over different ranges satisfy Top 50>Top 100>Top 150>Top 200> Bottom 200>Bottom 150>Bottom 100> Bottom 50, which is in accordance with our expectation. Moreover, we sorted the values of the nMSE and found **noReg**> **noFC**>**noHR**> **noGC**. **noReg** generates the greatest impact on the prediction performance, illustrating that the colearning of feature representation and Lasso-type regression is necessary to improve the effectiveness of our model. **noFC** achieves unsatisfactory results, indicating that exploiting latent correlation patterns embedded in features is beneficial for more robust and intrinsic feature representation. The prediction results of **noHR** and **noGC** are inferior to those of TGRLLR, illustrating that the relationship information among items and outfits plays an indispensable role in fashion compatibility prediction tasks.

TABLE II: Performance comparison with various values of α on our proposed method.

	1e-7	1e-6	1e-5	1e-4	1e-3	1e-1
Top 50	0.287	0.273	0.272	0.268	0.263	0.287
Top 100	0.279	0.274	0.275	0.278	0.257	0.268
Top 150	0.267	0.267	0.268	0.266	0.258	0.259
Top 200	0.259	0.261	0.263	0.262	0.252	0.255
Bottom 200	0.210	0.208	0.206	0.207	0.217	0.215
Bottom 150	0.206	0.203	0.201	0.206	0.209	0.207
Bottom 100	0.200	0.199	0.201	0.201	0.200	0.202
Bottom 50	0.186	0.182	0.182	0.197	0.189	0.188
nMSE	0.348	0.348	**0.347**	0.357	0.358	0.387

3) Evaluation of Parameters: In this part, we evaluated the influence of parameters α, β, and η on our proposed method. Particularly, parameter α controls the strength of the local geometric structure consistency, which ranges from $1e-7$ to $1e-1$ with an interval of 10 times. Table II reports the nMSE results for various values of α. From the table, we can see that the prediction results are sensitive to α. The best prediction performance is achieved when $\alpha = 1e-5$. Too large or too small values of α easily lead to suboptimal nMSE results. Similarly, parameters β and η are also investigated in the same way and are selected from $\{15, 20, 25, 30, 35, 45\}$ and $\{5, 15, 20, 25, 30, 35\}$, respectively. TABLE III and IV show the corresponding prediction results. From the table, we can observe that the best prediction performance is obtained when $\beta = 30$ and $\eta = 25$. Furthermore, when β and η are set to 0, our proposed method becomes to discard the item correlation and feature correlation terms, which easily causes unsatisfactory results.

TABLE III: Performance comparison with various values of β on our proposed method.

	15	20	25	30	35	45
Top 50	0.281	0.277	0.276	0.272	0.268	0.280
Top 100	0.279	0.265	0.272	0.275	0.276	0.276
Top 150	0.266	0.266	0.268	0.268	0.270	0.259
Top 200	0.261	0.263	0.261	0.263	0.265	0.257
Bottom 200	0.208	0.206	0.207	0.206	0.204	0.212
Bottom 150	0.202	0.203	0.204	0.201	0.204	0.208
Bottom 100	0.198	0.200	0.201	0.201	0.200	0.204
Bottom 50	0.188	0.184	0.184	0.182	0.180	0.194
nMSE	0.355	0.352	0.348	**0.347**	0.356	0.356

TABLE IV: Performance comparison with various values of η on our proposed method.

	5	15	20	25	30	35
Top 50	0.267	0.291	0.251	0.272	0.276	0.289
Top 100	0.259	0.279	0.247	0.275	0.275	0.271
Top 150	0.260	0.267	0.253	0.268	0.266	0.256
Top 200	0.250	0.259	0.254	0.263	0.260	0.256
Bottom 200	0.218	0.210	0.214	0.206	0.209	0.213
Bottom 150	0.222	0.204	0.220	0.201	0.200	0.201
Bottom 100	0.212	0.204	0.222	0.201	0.204	0.206
Bottom 50	0.192	0.183	0.196	0.182	0.179	0.196
nMSE	0.440	0.362	0.351	**0.347**	0.347	0.462

4) Comparison with state-of-the-art methods: We compared our proposed method with several state-of-the-art methods, including ridge regression (Ridge), Lasso, support vector regression (SVR) [43], low-rank linear regressions (LLR) [44], multi-feature learning via hierarchical regression (MLHR) [45], supervised regularization-based robust subspace (SRRS) [46], discriminative elastic-net regularized linear regression (DENLR) [47], regularized label relaxation linear regression (RLRLR) [48], interclass sparsity-based discriminative least square regression (IC-S_DLSR) [49], supervised approximate low-rank projection learning (SALPL) [50], bidirectional LSTM (Bi-LSTM) [5], and bidirectional GRU (Bi-GRU) [5].

TABLE V shows the prediction results of TGRLLR and state-of-the-art methods. From the table, we can derive the following conclusions: 1) Our proposed TGRLLR

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3062736, IEEE Transactions on Multimedia

TABLE V: Performance comparison of our proposed method and state-of-the-art methods (p-value$<$0.001:$\star\star\star$, p-value$<$0.05:$\star\star$, p-value$<$0.1:$\star$).

Methods	nMSE	p-value
Ridge	0.999± 1.05E-02	$\star\star\star$
Lasso	0.982± 4.99E-02	$\star\star\star$
SVR	0.362 ± 6.70E-03	$\star\star\star$
LLR	0.355 ± 1.60E-03	$\star\star$
MLHR	0.354 ± 2.90E-03	$\star\star$
SRRS	0.653 ± 3.35E-02	$\star\star$
DENLR	0.757 ± 1.92E-01	$\star\star\star$
RLRLR	0.540 ± 3.69E-02	$\star\star$
ICS_DLSR	0.353 ± 1.21E-03	$\star\star$
SALPL	0.377 ± 7.13E-03	$\star\star$
Bi-LSTM	0.443 ± 1.12E-03	$\star\star$
Bi-GRU	0.436 ± 1.94E-03	$\star\star$
TGRLLR	**0.347** ± **4.52E-03**	-

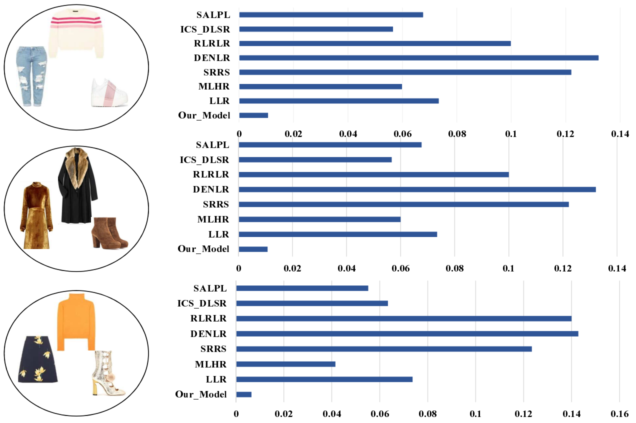


Fig. 5: Qualitative comparison of fashion compatibility prediction results using different methods.

outperforms the compared methods in terms of the nMSE. 2) The Ridge and Lasso methods perform the worst, indicating that the simple combination of feature selection and regression analysis is insufficient for fashion compatibility prediction tasks. 3) SRRS, DENLR, and RLRLR obtain unsatisfactory results, indicating that more robust and intrinsic feature representation is of vital importance to fashion compatibility prediction. 4) After exploiting the radial basis function (RBF) kernel to the original feature space, SVR achieves a better prediction performance. 5) SALPL, LLR, MLHR, and ICS_DLSR are low-rank-based methods and achieve prediction results that are superior to the results of other methods. In particular, although SALPL learns two projection matrices from different directions but is still inferior to our proposed method due to its deficiency in exploiting complex correlation patterns. MLHR uses a multifeature fusion strategy to explore the structural information. ICS_DLSR exploits the assumption that the transformed samples have a common sparsity structure. These results show that our proposed method is more suitable than these other methods for characterizing the complex relationships in fashion compatibility prediction tasks; 6) We compared our proposed model with Bi-LSTM and Bi-GRU, which only consider a bidirectional LSTM and GRU without incorporating any semantic information in an end-to-end fashion. The results indicate that our proposed method outperforms both deep models since the limited labeled samples are not enough for robust and effective model training. 7) Furthermore, we used a P-value [51] to assess the differences between TGRLLR and the other methods. We discovered that the P-values are smaller than the significance level of 0.05, which indicates that the null hypothesis is clearly rejected and that the improvements of our proposed method are statistically significant. Furthermore, we conducted a qualitative comparison by randomly selecting three fashion outfits. Fig. 5 shows the qualitative comparison of various methods, in which we calculated the absolute deviations between their predicted and the real fashion compatibility scores. From the figure, we can see that our proposed TGRLLR method exhibits higher consistency with the ground truth than the other methods. Moreover, we selected two classical matching problems, *i.e.,bottom-to-top* and *top-to-bottom*, and showed the ranking results in descending order according to predicted compatibility scores in Fig. 6. From the figure, we found that the outfits that have been matched in this dataset are visually compatible.

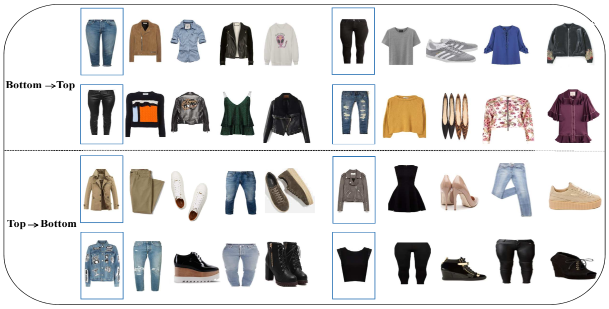
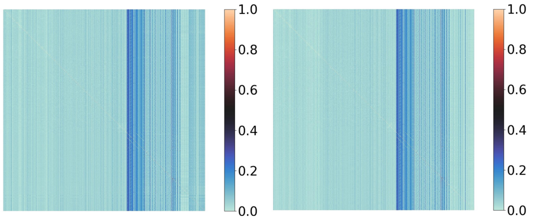


Fig. 6: Illustration of the ranking results of our proposed method.

5) Feature correlation analysis: In Fig.7, we visualized the normalized the feature correlation matrix $\mathbf{\Omega}^{-1}$ learned from the overall method and the feature correlation matrix $\hat{\mathbf{\Omega}}^{-1}$ learned by discarding the local geometric structure consistency constraint imposed on fashion outfits. From the figure, we can see that Fig.7(a) exhibits more distinct correlations among features, especially at the right of the diagram in comparison with Fig.7(b), indicating that more intrinsic representation of data can be benefited from the preservation of correlation between outfits.

C. Case study

In this section, we reported findings from a user study. We constructed 5 sets of fashion samples, and each set involves four fashion outfits that are sorted in ascending order relying on the predicted fashion compatibility scores. We conducted a survey of 200 participants, including 119 males and 81 females ranging in age from 20 to 30 and reported their subjective preference with the sorting results of five groups. A participant's subjective preference



(a) $\mathbf{\Omega}^{-1}$ (b) $\hat{\mathbf{\Omega}}^{-1}$

Fig. 7: A visual comparison of two feature correlation matrices. The higher values correspond to higher correlations.

was assigned a score of 0 to 3 (“0” is disapproval, “1” is borderline, “2” is approval, and “3” is full approval) to reflect their relative satisfaction. Table VI shows the participants’ satisfaction-based ratings of various groups. From the table, we observe that the average rating scores are greater than 2, indicating that the ranking lists produced by our proposed method are acceptable to the participants.

TABLE VI: Participant’ satisfaction-based rating.

	set1	**set2**	**set3**	**set4**	**set5**	**Average**
Male	2.234	2.326	2.225	2.332	2.124	2.248
Female	2.197	2.278	2.029	2.438	2.057	2.200
Average	2.219	2.307	2.146	2.374	2.097	2.229

V. Conclusions and Future Work

In this paper, we proposed TGRLLR to predict the compatibility scores of fashion outfits. TGRLLR firstly exploited a latent low-rank representation mechanism to tackle insufficient training samples and sparsity problems. And then TCRLLR constructed three graph regularization terms to capture the correlation patterns embedded in features, items, and outfits. The experiments demonstrated the positive effect of the involved graph regularization terms. Despite this, we found that the proposed method has insufficient generalization ability when used for large-scale scenarios and the sparsity problem is still a challenge for high performance. In the future, we will focus on developing an end-to-end convolutional neural network to better deal with fashion compatibility prediction tasks. Moreover, we will emphasize more on the exploration of transmission of relationship among fashion items.

References

[1] M. Hadi Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg, “Where to buy it: Matching street clothing photos in online shops,” in *Proceedings of IEEE International Conference on Computer Vision*, 2015, pp. 3343–3351.

[2] K. E. Ak, J. H. Lim, J. Y. Tham, and A. A. Kassim, “Efficient multi-attribute similarity learning towards attribute-based fashion search,” in *Proceedings of IEEE Winter Conference on Applications of Computer Vision*, 2018, pp. 1671–1679.

[3] X. Zou, X. Kong, W. Wong, C. Wang, Y. Liu, and Y. Cao, “Fashionai: A hierarchical dataset for fashion understanding,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[4] Y. Ge, R. Zhang, X. Wang, X. Tang, and P. Luo, “Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5337–5345.

[5] X. Han, Z. Wu, Y.-G. Jiang, and L. S. Davis, “Learning fashion compatibility with bidirectional lstms,” in *Proceedings of ACM Conference on Multimedia*, 2017, pp. 1078–1086.

[6] M. I. Vasileva, B. A. Plummer, K. Dusad, S. Rajpal, R. Kumar, and D. Forsyth, “Learning type-aware embeddings for fashion compatibility,” *arXiv preprint arXiv:1803.09196*, 2018.

[7] R. Yin, K. Li, J. Lu, and G. Zhang, “Enhancing fashion recommendation with visual compatibility relationship,” in *Proceedings of the World Wide Web Conference*, 2019, pp. 3434–3440.

[8] X. Song, F. Feng, X. Han, X. Yang, W. Liu, and L. Nie, “Neural compatibility modeling with attentive knowledge distillation,” in *Proceedings of International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018, pp. 5–14.

[9] G. Cucurull, P. Taslakian, and D. Vazquez, “Context-aware visual compatibility prediction,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 609–12 618.

[10] X. Yang, . Y. Ma, . L. Liao, . M. Wang, . T.-S. Chua, S. of Computing, N. U. of Singapore, and Singapore, “Transnfcm: Translation-based neural fashion compatibility modeling,” in *Proceedings of AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 403–410.

[11] A. Saha, M. Nawhal, M. M. Khapra, and V. C. Raykar, “Learning disentangled multimodal representations for the fashion domain,” in *Proceedings of IEEE Winter Conference on Applications of Computer Vision*, 2018, pp. 557–566.

[12] H. Lee and S.-g. Lee, “Fashion attributes-to-image synthesis using attention-based generative adversarial network,” in *Proccedings of IEEE Winter Conference on Applications of Computer Vision*, 2019, pp. 462–470.

[13] D. Valle, N. Ziviani, and A. Veloso, “Effective fashion retrieval based on semantic compositional networks,” in *Proceedings of International Joint Conference on Neural Networks*, 2018, pp. 1–8.

[14] Z. Kuang, Y. Gao, G. Li, P. Luo, Y. Chen, L. Lin, and W. Zhang, “Fashion retrieval via graph reasoning networks on a similarity pyramid,” in *Proceedings of IEEE International Conference on Computer Vision*, 2019, pp. 3066–3075.

[15] W. Kang, C. Fang, Z. Wang, and J. McAuley, “Visually-aware fashion recommendation and design with generative image models,” in *Proceedings of IEEE International Conference on Data Mining*, 2017, pp. 207–216.

[16] X. Zhang, J. Jia, K. Gao, Y. Zhang, D. Zhang, J. Li, and Q. Tian, “Trip outfits advisor: Location-oriented clothing recommendation,” *IEEE Transactions on Multimedia*, vol. 19, no. 11, pp. 2533–2544, 2017.

[17] W.-L. Hsiao and K. Grauman, “Creating capsule wardrobes from fashion images,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7161–7170.

[18] U. Mall, K. Matzen, B. Hariharan, N. Snavely, and K. Bala, “Geostyle: Discovering fashion trends and events,” in *Proceedings of IEEE International Conference on Computer Vision*, 2019, pp. 411–420.

[19] X. Gu, Y. Wong, L. Shou, P. Peng, G. Chen, and M. S. Kankanhalli, “Multi-modal and multi-domain embedding learning for fashion retrieval and analysis,” *IEEE Transactions on Multimedia*, vol. 21, no. 6, pp. 1524–1537, 2018.

[20] S. Jo, S. Jang, H. Cho, and J. Jeong, “Scenery-based fashion recommendation with cross-domain geneartive adverserial networks,” in *Proceedings of IEEE International Conference on Big Data and Smart Computing*, 2019, pp. 1–4.

[21] S. Verma, S. Anand, C. Arora, and A. Rai, “Diversity in fashion recommendation using semantic parsing,” in *Proceedings of IEEE International Conference on Image Processing*, 2018, pp. 500–504.

[22] M. Hou, L. Wu, E. Chen, Z. Li, V. W. Zheng, and Q. Liu, “Explainable fashion recommendation: a semantic attribute region guided approach,” pp. 4681–4688, 2019.

[23] X. Song, F. Feng, J. Liu, Z. Li, L. Nie, and J. Ma, “Neurostylist: Neural compatibility modeling for clothing matching,” in *Proceedings of ACM International Conference on Multimedia*, 2017, pp. 753–761.

[24] Y. H. Long Chen, “Dress fashionably: Learn fashion collocation with deep mixed-category metric learning,” *Proceedings of AAAI Conference on Artificial Intelligence*, pp. 2103–2110, 2018.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3062736, IEEE Transactions on Multimedia

[25] G. Liu, Z. Lin, and Y. Yu, “Robust subspace segmentation by low-rank representation,” in *Proceedings of International Conference on Machine Learning*, 2010, pp. 663–670.

[26] M. Yin, J. Gao, and Z. Lin, “Laplacian regularized low-rank representation and its applications,” *IEEE Transacons on Pattern Analysis and Machine Intelligence*, vol. 38, no. 3, pp. 504–517, 2015.

[27] L. Xie, M. Yin, X. Yin, Y. Liu, and G. Yin, “Low-rank sparse preserving projections for dimensionality reduction,” *IEEE Transactions on Image Processing*, vol. 27, no. 11, pp. 5261–5274, 2018.

[28] J. Wen, X. Fang, Y. Xu, C. Tian, and L. Fei, “Low-rank representation with adaptive graph regularization,” *Neural Networks*, vol. 108, pp. 83–96, 2018.

[29] G. Liu and S. Yan, “Latent low-rank representation for subspace segmentation and feature extraction,” in *Proceedings of IEEE International Conference on Computer Vision*, 2011, pp. 1615–1622.

[30] Z. Ren, Q. Sun, B. Wu, X. Zhang, and W. Yan, “Learning latent low-rank and sparse embedding for robust image feature extraction,” *IEEE Transactions on Image Processing*, vol. 29, no. 1, pp. 2094–2107, 2019.

[31] Y. Zhang and D.-Y. Yeung, “A convex formulation for learning task relationships in multi-task learning,” *arXiv preprint arXiv:1203.3536*, 2012.

[32] S. Huang, M. Elhoseiny, A. Elgammal, and D. Yang, “Learning hypergraph-regularized attribute predictors,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 409–417.

[33] J.-F. Cai, E. J. Candès, and Z. Shen, “A singular value thresholding algorithm for matrix completion,” *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[34] Z. Lin, M. Chen, and Y. Ma, “The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices,” *Journal of Structural Biology*, vol. 181, pp. 116–127, 2013.

[35] A. Beck and M. Teboulle, “A fast iterative shrinkage-thresholding algorithm for linear inverse problems,” *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[36] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.

[37] J. Huang, S. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih, “Image indexing using color correlograms,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 1997, pp. 762–768.

[38] M. A. Stricker and M. Orengo, “Similarity of color images,” in *Storage and Retrieval for Image and Video Databases III*, vol. 2420, 1995, pp. 381–393.

[39] S. L. G. and S. G. C., *Computer Vision*, Prentice Hall, 2003.

[40] D. K. Park, Y. S. Jeon, and C. S. Won, “Efficient use of local edge histogram descriptor,” in *Proceedings of ACM workshops on Multimedia*, 2000, pp. 51–54.

[41] B. S. Manjunath and W.-Y. Ma, “Texture features for browsing and retrieval of image data,” *IEEE Transacons on Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 837–842, 1996.

[42] L. Nie, L. Zhang, Y. Yang, M. Wang, R. Hong, and T.-S. Chua, “Beyond doctors: Future health prediction from multimedia and multimodal observations,” in *Proceedings of ACM International Conference on Multimedia*, 2015, pp. 591–600.

[43] H. Drucker, C. J. Burges, L. Kaufman, A. J. Smola, and V. Vapnik, “Support vector regression machines,” in *Proceedings of Advances in Neural Information Processing Systems*, 1997, pp. 155–161.

[44] X. Cai, C. Ding, F. Nie, and H. Huang, “On the equivalent of low-rank linear regressions and linear discriminant analysis based regressions,” in *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2013, pp. 1124–1132.

[45] Y. Yang, J. Song, Z. Huang, Z. Ma, N. Sebe, and A. G. Hauptmann, “Multi-feature fusion via hierarchical regression for multimedia analysis,” *IEEE Transactions on Multimedia*, vol. 15, no. 3, pp. 572–581, 2013.

[46] S. Li and Y. Fu, “Learning robust and discriminative subspace with low-rank constraints,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 11, pp. 2160–2173, 2016.

[47] Z. Zhang, Z. Lai, Y. Xu, L. Shao, J. Wu, and G.-S. Xie, “Discriminative elastic-net regularized linear regression,” *IEEE Transactions on Image Processing*, vol. 26, no. 3, pp. 1466–1481, 2017.

[48] X. Fang, Y. Xu, X. Li, Z. Lai, W. K. Wong, and B. Fang, “Regularized label relaxation linear regression,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 4, pp. 1006–1018, 2018.

[49] J. Wen, Y. Xu, Z. Li, Z. Ma, and Y. Xu, “Inter-class sparsity based discriminative least square regression,” *Neural Networks*, vol. 102, pp. 36–47, 2018.

[50] X. Fang, N. Han, J. Wu, Y. Xu, J. Yang, W. K. Wong, and X. Li, “Approximate low-rank projection learning for feature extraction,” *IEEE Transactions on Neural Networks and Learning Systems*, no. 99, pp. 1–14, 2018.

[51] L. Nie, Y.-L. Zhao, M. Akbari, J. Shen, and T.-S. Chua, “Bridging the vocabulary gap between health seekers and healthcare knowledge,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 2, pp. 396–409, 2015.

Peiguang Jing is currently an associate professor with the School of Electrical and Information Engineering, Tianjin University, Tianjin, China. He received his M.S. degree and Ph.D. degree from Tianjin University in 2013 and 2018, respectively. From 2014 to 2015, He was a visiting student with the National University of Singapore. His current research interests include multimedia computing, signal processing, and machine learning.

Jing Zhang received the B.E. degree from Changchun University Of Science And Technology in 2018. She is currently pursuing the M.S. degree with the School of Electrical and Information Engineering, Tianjin University, Tianjin, China. Her research interests are machine learning and multimedia analysis.

Liqiang Nie is currently a professor with the School of Computer Science and Technology, Shandong University. Meanwhile, he is the adjunct dean with the Shandong AI institute. He received his B.Eng. and Ph.D. degree from Xi’an Jiaotong University in July 2009 and National University of Singapore (NUS) in 2013, respectively. After PhD, Dr. Nie continued his research in NUS as a research follow for more than three years. His research interests lie primarily in multimedia computing and information retrieval. Dr. Nie has co-/authored more than 180 papers, received more than 7700 Google Scholar citations as of May 20120. He is an AE of Information Science, an area chair of ACM MM 2019.

Shu Ye received the B.S. degree from Nanchang University in 2017. She is currently pursuing the M.S. degree with the School of Electrical and Information Engineering, Tianjin University, Tianjin, China. Her research interests are machine learning and multimedia computing.

Jing Liu Jing Liu received the B.E. and Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2011 and 2017, respectively. She is currently an Associate Professor with the Multimedia Institute of Tianjin University. From 2014 to 2015, she was a visiting student with the Department of Computer Science and Engineering, State University of New York at Buffalo, U.S. Her research interests include multimedia signal processing and perceptual visual processing.

Yuting Su received the B.S. degree, the M.S. degree and the Ph.D. degree from Tianjin university, Tianjin, China, in 1995, 1998 and 2001 respectively. He is currently a professor with the School of Electrical and Information Engineering, Tianjin University, Tianjin, China. His research interests include computer vision, multimedia content analysis, information security, and tensor decomposition.