Deep-based Self-refined Face-top Coordination

HONGLIN LI, Quanzhou Medical College, China
XIAOYANG MAO, University of Yamanashi, Japan
MENGDI XU, State Key Lab of CAD&CG, Zhejiang University, China
XIAOGANG JIN, (Corresponding author), State Key Lab of CAD&CG, Zhejiang University, China

Face-top coordination, which exists in most clothes fitting scenarios, is challenging due to varieties of attributes, implicit correlations, and trade-offs between general preferences and individual preferences. We present a Deep-Based Self-Refined (DBSR) system to simulate face-top coordination based on intuition evaluation. To this end, we first establish a well-coordinated face-top (WCFT) dataset from fashion databases and communities. Then, we use a jointly trained CNN Deep Canonical Correlation Analysis (DCCA) method to bridge the semantic face-top gap based on the WCFT dataset to deal with general preferences. Subsequently, an irrelevance-based Optimum Forest (OPF) method is developed to adapt the results to individual preferences iteratively. Experimental results and user study demonstrate the effectiveness of our method.

CCS Concepts: • **Information systems** → *Information retrieval.*

Additional Key Words and Phrases: Fashion analysis, Personalized face-top coordination, Deep cross-modal learning, Canonical correlation analysis, Relevance Feedback, Optimum-Path Forest.

ACM Reference Format:
Honglin Li, Xiaoyang Mao, Mengdi Xu, and Xiaogang Jin. 2021. Deep-based Self-refined Face-top Coordination. *ACM Trans. Multimedia Comput. Commun. Appl.* , (2021), 23 pages.

1 INTRODUCTION

Overall fashion attractiveness is determined not only by clothes coordination, but also by body shape and facial characteristics, as well as makeup and hairstyle jointly, among which complex latent correlations exist. In the investigation conducted in [42] on world-wide fashion styles, millions of the collected photographs from Instagram [1] contain only faces and torsos because lower garments are often occluded in online photos (especially profile photos), which is in accordance with the fact that top features contribute more than bottom features in ordinary feelings, as was also reported in [54]. During realistic fitting, females always first put selected upper-clothes under their necks in front of mirrors to pre-evaluate the overall fitness prior to trying them on in fitting

[1] www.instagram.com

This work was supported by the National Key R&D Program of China (Grant No. 2017YFB1002600), the Ningbo Major Special Projects of the "Science and Technology Innovation 2025" (Grant No. 2020Z007), and the National Natural Science Foundation of China (Grant Nos. 61972344, 61732015).
Authors' addresses: Honglin Li, Quanzhou Medical College, Quanzhou 362010, Fujian Province, China, lihonglin79@qq.com; Xiaoyang Mao, University of Yamanashi, Yamanashi, Kofu 400-8510, Japan, mao@yamanashi.ac.jp; Mengdi Xu, State Key Lab of CAD&CG, Zhejiang University, Hangzhou 310058, Zhejiang Province, China, 21821045@zju.edu.cn; Xiaogang Jin, (Corresponding author), State Key Lab of CAD&CG, Zhejiang University, Hangzhou 310058, Zhejiang Province, China, jin@cad.zju.edu.cn.

1551-6857/2021/-ART $15.00
https://doi.org/

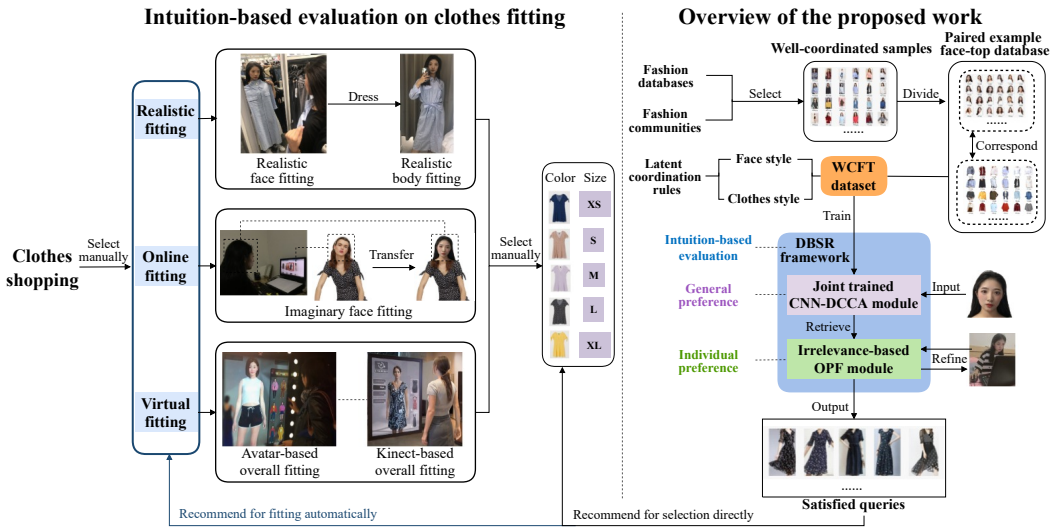

Fig. 1. Intuition-based evaluation on clothes fitting and overview of our proposed work for face-top coordination.

rooms. Regarding online fitting, females usually roughly evaluate their coordination with the desired upper-clothes through observing them worn by fashion models. Virtual fitting is applied in real stores via Kinect-attached devices or on online websites via 3D avatars. As shown in the left part of **Figure 1**, in the above three situations, people always put or transfer the manually selected clothes under faces, and thus imagine whether or not they are coordinated. It will be unnatural and difficult to evaluate the clothes-fitness without the face, since a same top-down garment coordination solution will result in dissimilar impressions on different people. Therefore, the face plays an important role in fashion coordination. Employing a face-top recommendation operation before the traditional top-down coordination operation, can not only enhance overall personal fashion attractiveness effectively, but also relieve users from time-consuming manual coordination. Therefore, such a recommendation operation improves their experiences, especially for upper-clothes selection in situations of separate purchases. Inspired by these motivations, we introduce a novel face-top coordination framework, as shown in the right part of **Figure 1**.

Face styles are jointly represented by hairstyle, face contour, face skin color, facial organ features/positions, etc, and clothes styles are comprised of collar type, color, texture, pattern, etc, as shown in **Figure 2**. The face and clothes styles are generally defined as warm, cool, soft, hard, casual, formal etc, between which various attributes and elements correlate with each other locally or globally. Most of the current fashion studies focus on coordination among garments, and only a few face-related coordination studies were conducted, which focused on local area or element coordination. Face contour, as well as hairstyle, interacts with the collar, which is a crucial component of upper-clothes serving as the frame for the human face [31, 36]. Face skin color, which is similar to body skin color, dominates coordinated upper-clothes color to some extent [2, 63]. Finally, and importantly, facial organ features and positions, which play vital roles in face averageness and symmetry [6, 53], are implicitly correlated with clothes texture and pattern that deeply affect clothes style. However, none of the aforementioned studies considered face-top style coordination integrally. Although extensively discussed in fashion communities, due to varieties of face and clothes characteristics, as well as frequently changing fashion trends, few concrete

face-top coordination rules are provided by the state-of-the-art studies or even fashion experts, which require huge number of annotations and pre-defined rules [36, 61] for training. Therefore, in this paper, our presented framework aims to simulate the real face-top coordination procedure instead of finding detailed face-top coordination patterns. The first challenge to be addressed is how to bridge the semantic gap between face style and clothes style considering general preferences and personal facial characteristics. The second challenge is identification of how to optimally deal with individual preferences for personalized recommendation.

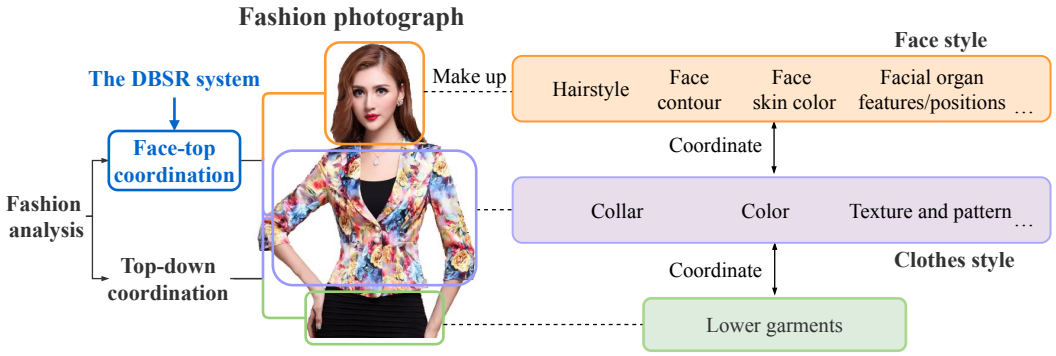

Fig. 2. Latent correlations among fashion attractiveness.

To solve the first challenge, there are three state-of-the-art techniques which can be adopted according to most up-to-date fashion coordination researches. The first method uses traditional machine-based algorithms for supervised training to evaluate fashion coordination [13, 14, 26, 32, 37, 47, 49, 51, 59, 61, 64] based on pre-defined labels and extracted visual or aesthetic concentrated features, as well as textual attributes, from collected outfit samples. However, few concrete rules are generated. Some associated rule based algorithms like the Apriori algorithm or other rule-based classifiers are adopted in some papers only for specific types of garments, which cannot be generalized. Graph-based [36] and tree-based [61] analyses are also too complex to perform, since manifold marked fashion attributes are required and few face-top coordination rules acting as prior rules exist. Compared to traditional top-down coordination, face-top coordination is a much more subjective issue and difficult to handle by this method, since face-related features and clothes-related features belong to significantly different modalities, which makes it difficult for different modality features to be concentrated and traditional distance-based metrics to be deployed for aesthetic evaluation or compatibility analysis directly. In addition, it is much more challenging to collect unpaired samples for face-top coordination due to personal unwillingness to show inappropriate coordination publicly, which will cause overfitting in supervised training. The Similarity-Based Transfer (SBT) method adopted directly or acting as baselines in [18, 24–26, 29, 30, 60] is the second method for dealing with the paired coordination issue. Nonetheless, this technique has the following three limitations. First, arbitrary information transfer is unreliable because it fails to consider pairwise correlations. Second, lacking a self-learning mechanism makes it impossible to be improved. Finally, updating the pairwise candidate database frequently according to changing fashion trends is almost cost prohibitive. The third popular method is cross-modal learning. It is usually used for learning compatibility across modalities [3, 10, 17, 19], in which underlying interactions between pairwise items from different modalities are analyzed by projecting positive pairs close together and negative pairs far apart. Moreover, it becomes feasible to deploy cross-modal learning on deep-based frameworks [5, 55, 56, 66], which have been adopted in many fashion

coordination research studies [15, 34, 54, 62] based on rapidly increasing fashion communities and fashion model image databases. Therefore, we adopt a jointly trained CNN Deep Canonical Correlation Analysis (DCCA) method to explore overall face-top correlations via our established Well-Coordinated Face-Top dataset (*WCFT*) collected from the state-of-the-art fashion databases, e-commerce websites, and fashion communities.

Individual preference consideration constitutes the second challenge to be addressed in this paper. Traditional Collaborative Filtering (CF) based methods are employed in most personalized outfit recommendation tasks to combine individual preferences with extracted object features for overall evaluation [8, 16, 20, 24, 38] or embed individual preferences into a shared feature space for compatibility analysis [35], in which individual preferences are mainly mined from customers' purchasing or browsing records stored in state-of-the-art e-commerce websites. However, these methods are usually disturbed by the cold-start problem and affected by various non-visual factors, such as prices, substitutable and complementary functions, conformity behavior, etc. In addition, unlike daily consumed products, similar styles of clothes are not purchased repeatedly by younger females due to changing fashion trends and product durability. For the face-top coordination issue to be addressed in this paper, it is difficult for traditional CF methods to consider personal facial characteristics due to personal privacy issues. It is also unreliable to simply consider individual preferences according to personal or group behaviors due to visual-focused factors. Inspired by the interactive techniques for preference-focused issues, we develop a new irrelevance-based Optimum Forest (OPF) method to deal with individual preferences for face-top coordination effectively and efficiently.

In summary, since people determine what clothes they will purchase according to general preferences learned from fashion magazines or communities, and individual preferences jointly, our proposed Deep-Based Self-Refined (*DBSR*) system simulates this procedure by learning general preferences considering personal facial characteristics from the established *WCFT* dataset via the jointly trained CNN-DCCA method, and adapting to individual preferences via the irrelevance-based OPF method. Note that, compared to fashion item coordination, it is much more difficult for the face-top coordination issue to obtain prior rules for training, and our designed system is user-oriented for the intuition-based recommendation instead of being specialist-oriented for generating detailed patterns.

The main contributions of this work are summarized as follows.

- We propose a jointly trained CNN-DCCA method for bridging the semantic face-top gap and a new irrelevance-based OPF method for personalized recommendation, considering general preferences and individual preferences, respectively.
- We present a *DBSR* framework, which integrates the jointly trained CNN-DCCA method with the irrelevance-based OPF method seamlessly and can adapt to frequently changing preferences dynamically.
- We establish the well-coordinated face-top (*WCFT*) dataset constituted of 30,000 pairs of face-top samples by manually selecting and dividing evaluated female fashion photographs from Taobao [2], Mogu [3], Polyvore [4], Chictopia [5], and Lookbook [6] websites, and the DeepFashion [7] database.

[2]https://www.taobao.com
[3]https://www.mogu.com
[4]www.polyvore.com
[5]www.chictopia.com
[6]www.lookbook.nu
[7]http://mmlab.ie.cuhk.edu.hk/projects/DeepFashion.html

2 RELATED WORK

2.1 Fashion recommendation

Fashion recommendation can be regarded as a special type of retrieval with different modalities of query and result, and constitutes the most popular field of fashion-related research study.

Pairwise coordination: Coordination between pairwise items and compatibility within outfits were analyzed in detail for assisting users to wear aesthetically [15, 25, 30, 32, 39, 62] and appropriately according to occasions [37, 52]. Nonetheless, none of the above researches analyzed correlations between face and upper-clothes.

Personalized recommendation: In online shopping, personalized item or outfit recommendations were conducted based on preference-related features or users' behaviors to satisfy users' individual preferences [8, 13, 14, 22, 35, 38, 49, 64, 67]. However, facial characteristics acting as the most important personal features were not considered. In addition, there exist semantic gaps between individual preferences and behaviors as described in **Section 1**. We present the irrelevance-based OPF algorithm based on the relevance feedback mechanism to adapt to individual preferences.

Preference interpretation: Preference interpretation has recently become a highly popular research topic, of which personal tastes were visually highlighted in preferred areas of recommended images or even presented with textual comments [9, 20, 33, 34, 51]. In [61], an attribute-based interpretable compatibility frame-work was proposed for the coordination analysis on fashion items, which could provide the compatibility scores and corresponding interpretable matching patterns of unseen pairs. However, there are few prior rules of face-top coordination exist, especially for considering facial organ features and positions, and clothes textures and patterns. In our work, personal tastes can be reflected in query results directly and iteratively via the irrelevance-based OPF algorithm.

Scene-related recommendation: Scene-related recommendation research studies [27, 37, 47] were performed for evaluating how a person looks in a photograph based on human characteristics, circumstances, outfits, tags, and comments jointly. However, various poses and decorations of faces in the photographs will disturb face-related factors. Furthermore, the above systems might be unreliable since the static photograph backgrounds cannot reflect the dynamic occasions completely. Consequently, our established *WCFT* dataset comprises female fashion photographs taken with constrained poses from the frontal position without decorations, the backgrounds of which should be clear or removed prior to processing.

Body shape related recommendation: Body shape related recommendation [18, 46] is a novel research field focusing on providing users with appropriate clothes according to automatic body shape detection techniques, and used to highlight or cover up certain parts based on customers' preferences automatically. This could be regarded as a complementary part for our work to provide recommendations for overall fitness in the future.

Other related research studies: Fashion analysis described abstract characteristics of fashion styles [21, 28], revealed latent correlations among fashion elements [40] in fashion images, and thus benefitted fashion recommendations. Associated factors on social effects [59] and fashion trends [1, 4, 41, 42, 50] could be integrated with fashion-related features to improve recommendation results. In addition, virtual fitting and image transformation [23, 57] techniques are developing rapidly, in which personal face images or avatar-transformed faces are directly assembled with virtual bodies and manually selected clothes.

2.2 Face-related analysis

Face style contributes greatly to female overall attractiveness, of which face symmetry, face averageness, sexual dimorphism, hairstyle, face skin color, as well as facial organ features and positions,

should be considered in aggregate [6, 7, 12, 45, 53]. Attractiveness of females contributed by face, clothes coordination, or audio was evaluated in [45] jointly or individually without considering mutual correlations. Only one type of face color factor was considered for coordination with one type of Indian clothes in [2], which lacked generalization. An arbitrary SBT method was deployed for item recommendation based on face similarity without considering hairstyle, as well as latent correlations [29]. According to color palette coordinated with color tones extracted from corresponding eyes, skin and hair of input 3D virtual characters, appropriate outfits composed of compatible items were recommended, and thus assembled with the input 3D virtual bodies, which adapted to occasions according to pre-defined coordination rules for color compatibility [63]. In [36], other important factors of face and upper-clothes, such as shape and positions of facial organs, were considered, where quantities of semantic attributes were listed and annotated in their established dataset. These attributes were predicted according to extracted features from given half upper-clothes attached portrait images, and thus used for recommending appropriate hairstyles and makeups jointly based on a proposed multiple tree-structured super-graphs model, where investigated correlations were estimated via mutual information calculation. However, correlations between face and upper-clothes were not analyzed in their paper. In addition, recommended results were presented for user study without concrete coordination rules too. We did not adopt this type of framework for fine-grained attribute analysis due to huge annotation work, too complex graph-based structure establishment, and uncertain face-top coordination results. To the best of our knowledge, we are the first to investigate the overall coordination between face styles and clothes styles.

2.3 Cross-modal learning and interactive techniques

Cross-modal learning is usually employed for compatibility analysis across modalities, in which inter-modal variance should be removed, and intra-modal discrimination should be preserved, as much as possible. Distance-based or probability-associated metrics, such as inner-product, Euclidean distance, probabilistic mixtures of multiple distances, conditional similarity, weighted nearest neighbor, Mahalanobis transform etc., were deployed in researches [15, 43, 62] on compatibility analysis, where the metrics were formulated in the loss functions for iterative optimization. The Siamese network [10] handled two workflows in parallel, which used triplet-based samples for training to deal with cross-modal learning. Multi-view representation learning [17] methods obtained complementary information from multiple sources of the same object, and thus improved the effectiveness of cross-modal learning. Another mainstream is the correlation-based method. The Canonical Correlation Analysis (CCA) [19] method was adopted to generate highly correlated linear transformations of data from different modalities, and Kernel Canonical Correlation Analysis (KCCA) [3] is a non-parametric version of CCA embedded with kernel functions. Compared to traditional CCA methods, DCCA [5] attempted to make non-linear transformations projected from different modalities highly correlated and optimized via deep learning networks, which was proven to achieve better performances. The Deep Canonically Correlated Autoencoders (DCCAE) method [56] attached an autoencoder regularization term to DCCA for achieving better performance. Inspired by Generative Adversarial Networks (GAN), the Adversarial Cross-modal Retrieval (ACMR) [55] method was employed for seeking an effective common subspace based on adversarial learning, which sought to preserve semantic discrimination and modality invariance simultaneously. The Deep Supervised Cross-Modal Retrieval (DSCMR) [66] optimized the objective function in both the label space and the shared feature space to make the learned common representations significantly discriminative. In our paper, we adopt the correlation-based method inspired by [5] for face-top compatibility analysis.

Relevance feedback methods are commonly deployed for users to interact with machine-based systems in real time, and thus reflect their preferences to optimize the query results iteratively. A supervised classification method, called OPF [11], represented each classified category by one or more optimum-path trees rooted at some initial samples called prototypes, which classified samples according to competitions among different types of prototypes. The OPF algorithm was deployed in [31] for relevance feedback application, which was able to optimize upper-clothes retrieval focusing on similar collars iteratively. Only a very small number of training samples are required for initiation to build the OPF classifiers, which can be expanded throughout the training dataset automatically and gradually for superior performance. Moreover, it demonstrates competitive accuracy and significantly better efficiency compared with SVM, and thus can be utilized for real-time interactive prediction. In this paper, we propose a new irrelevance-based OPF method and integrate it with our jointly trained CNN-DCCA method for self-refinement to adapt to individual preferences gradually.

3 DATASET

It can be observed that attractive females dominate most fashion model images. In addition, as described in [59], subjective evaluation is affected by social influence implicitly, as followers tend to give positive evaluations on people that they like in fashion social networks. To address these problems and balance the composition of our dataset, we gathered evaluated female fashion images from e-commerce websites, up-to-date fashion databases, and popular fashion communities jointly to establish the *WCFT* dataset.

We first manually collected well-coordinated female fashion model images without glasses and hats taken under well-controlled environments with constrained human poses in a frontal view from more than 450,000 images of DeepFashion and Taobao. Approximately 40,000 images were selected by a person who was pre-trained via browsing fashion websites for several days. The selected images were subsequently reviewed by five female college students jointly, and those that were evaluated positively by over three students were adopted. A constraint measure was further taken in which that a same model with a same hairstyle should not correspond to more than five upper-clothes in each category. This measure was adopted to reduce many-to-many cases, and thus relieved the overfitting problem, which would confuse the classifiers during training by pushing incompatible items together improperly when they were compatible with the same target, as also described in [32, 62]. As a consequence, 19,237 images remained, and then we supplemented them to reach 21,000 images finally according to the above described principles. Subsequently, another 9,000 fashionable images of ordinary persons from Polyvore, Chictopia, Lookbook and Mogu with high evaluations were collected to balance our dataset. Therefore, a total of 30,000 well-coordinated female fashion images without lower-parts were collected.

These images were divided into two parts by manually drawing bounding boxes over the hair-face area and the area of upper-clothes, respectively, to generate a pairwise face-top dataset, from which 2,000 pairs were used as testing images in the training phase. Since smile expression will cause positive deviations, as described in [7, 12, 45], 50 face images from [58] and 10 face images from our female college students with neutral expression were used in running time for the user study. Finally, another 2,000 upper-clothes images in 2019 comprised of eight categories (long/short T-shirt, long/short shirt, long/short dress, coat, and sweater), were collected as recommendation candidates to follow the latest fashion trends.

It is worth nothing that the backgrounds of the training and testing face and upper-clothes images were removed prior to processing. Moreover, the training dataset covered white Caucasian, yellow race Mongolian, as well as black race melanoderm, but the testing images were constrained within Asian females temporarily. We did not consider male coordination.

Since people do not typically like to show their poor coordination publicly, there were no negative pairs in our dataset for triplet-based training as in [16, 49, 54]. We are making our dataset robust and available for research purposes by expanding and refining the data as well as increasing the evaluators continuously.

4 METHOD

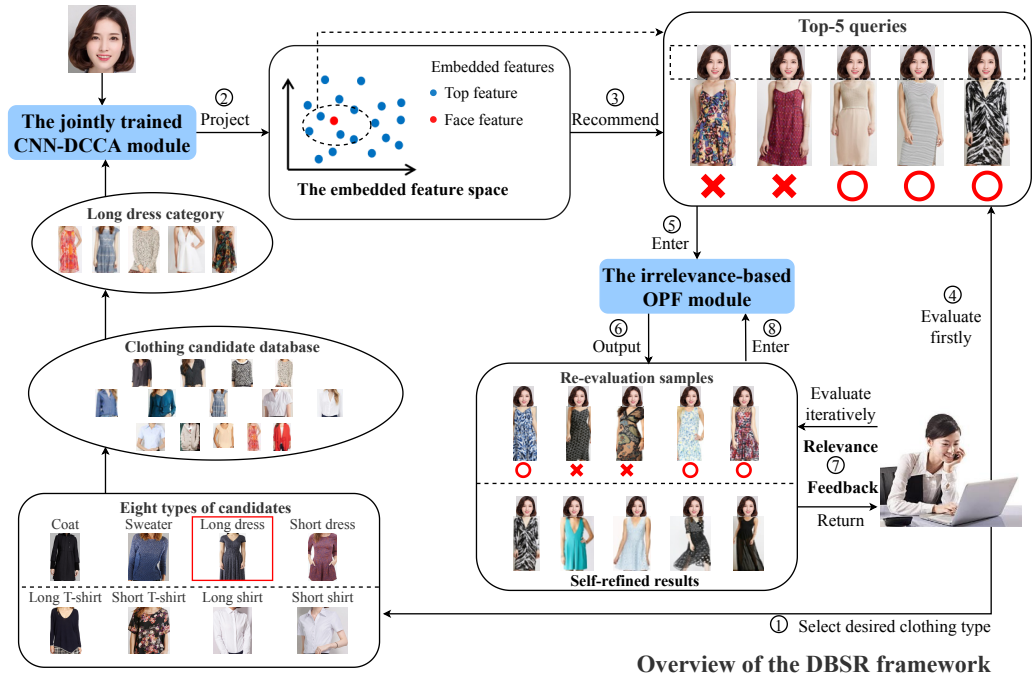

Fig. 3. Overview of the *DBSR* framewok. Details are described in **Section 4.3.**

We propose a *DBSR* framework to support online clothes shopping and virtual fitting for upper-clothes recommendation according to a given face image, in which the jointly trained CNN-DCCA module acting as common feature projectors is deployed for Euclidean distance based retrieval, and the OPF-based relevance feedback module is attached as the optimizer for personalized recommendation. Details are presented in **Figure 3** and **Algorithm 1**. Note that other factors such as body shape, compatibility among clothing items, occasions etc., can be subsequently solved by existing research studies referred in **Section 2.1**.

4.1 The DCCA method

CCA is an important method used for aligning multimodal data in a common feature space, where pairwise data across different modalities are mapped to nearby locations and vice versa. It aims to seek two canonical weights defined as ω_x and ω_y, which should make the correlation between the linear projections $\omega_x^T x$ and $\omega_y^T y$ maximized as follows:

$$(\omega_x, \omega_y) = \underset{(\omega_x, \omega_y)}{\operatorname{argmax}}\, corr\left(\omega_x^T x, \omega_y^T y\right)$$

$$= \underset{(\omega_x,\omega_y)}{\operatorname{argmax}} \frac{\omega_x^T C_{xy} \omega_y}{\sqrt{\omega_x^T C_{xx} \omega_x \cdot \omega_y^T C_{yy} \omega_y}}. \tag{1}$$

Pairwise features across two different modalities are represented as x and y, between which C_{xx}, C_{yy}, and C_{xy} stand for covariance and cross-covariance, respectively. Corresponding weights of the linear projectors denoted as ω_x and ω_y, are optimized to maximize the correlation between pairwise data. Nonetheless, CCA conducts a linear projection, which is difficult to reflect nonlinear relations among various modalities in reality.

DCCA is a combination of CCA and Deep Neural Networks (DNNs), in which DNNs are trained for feature extraction, non-linear mapping, linear mapping and correlation maximization simultaneously in an end-to-end manner. The canonical weights (ω_x and ω_y that model the CCA between $\varphi_x(x)$ and $\varphi_y(y)$), are trained simultaneously to maximize the correlation value after non-linear mapping as follows:

$$(\omega_x, \omega_y, \varphi_x, \varphi_y) = \underset{(\omega_x,\omega_y,\varphi_x,\varphi_y)}{\operatorname{argmax}} corr\left(\omega_x^T \varphi_x(x), \omega_y^T \varphi_y(y)\right). \tag{2}$$

Compared to the CCA formula, φ_x and φ_y functions are attached for nonlinear mapping in the DCCA formula.

4.2 The OPF algorithm

The OPF algorithm conducts the classification task as a graph partition in a given feature space. It is initiated from a complete graph, whose nodes stand for the feature vectors corresponding to candidate images, and arcs are weighted according to the distances among connected nodes. It is worth nothing that the distance of a whole path is the maximum value of the distances of all constituted path arcs, instead of their sum value. We described generation procedures of a simplified OPF and algorithm details of our proposed irrelevance-based OPF as follows, where only relevant and irrelevant types of samples existed.

Generation procedures of a simplified OPF: First, minimum spanning trees (MSTs) are generated from the complete graph, in which the most adjacent nodes that belong to different categories (relevant or irrelevant) are marked as prototypes. Second, the competition processes among prototypes are carried out to partition the graph dynamically and iteratively, which optimize the paths from the prototypes to the remaining nodes. Finally, as the right half part of **Figure 5** shows, all of the non-prototype samples are connected to a prototype with the minimum costs directly or indirectly. Therefore, the optimum-path forest is constituted of the optimum trees, in which the roots represent the prototypes, and the other nodes stand for the non-prototypes.

Algorithm details of the proposed irrelevance-based OPF: We propose a new irrelevance-based OPF algorithm to deal with the relevance feedback issue, in which the number of the non-prototypes (defined as U), the relevant prototypes (defined as R), and the irrelevant prototypes (defined as I) in the OPF space are assumed as i, j, and k, respectively. First, the Euclidean distances from each non-prototype to all the relevant and irrelevant prototypes are calculated and recorded, respectively. Second, the irrelevance value of each non-prototype is obtained via dividing the average distance value from it to the relevant prototypes by that from it to the irrelevant prototypes. Finally, a non-prototype is classified to the relevant type if its irrelevance value is smaller than 1, and vice versa. The computation for irrelevance values is described as follows:

$$Irrelevance_{U_i \to (R_j, I_k)} = \frac{Avg\left(Cost_{U_i \to R_j}\right)}{Avg\left(Cost_{U_i \to I_k}\right)}. \tag{3}$$

4.3 The *DBSR* framework

Prior to starting the personalized face-top aesthetic coordination task, users are required to select a desired upper-clothes type from the eight pre-defined types referred to in **Section 3** for determining the desired clothes category, sleeve type or dress length, as shown in Step 1 of **Figure 3**, since these three factors mainly depend on seasonality and subjective preferences. Note that, in real-time online applications, the pre-defined clothes types can be flexibly changed in accordance with brands, occasions, etc., which should be determined by the corresponding online store stocks.

Initial retrieval based on general preference learning: A given female face image, along with the upper-clothes candidates from the selected category, are sent into the jointly trained CNN-DCCA module to project their extracted features into a latent space for Euclidean distance based retrieval. The upper-clothes candidates with the five smallest distances to the given face are recommended as the Top-5 queries. Users are subsequently required to evaluate the query results, where $\times$ is marked on undesired samples, and $\bigcirc$ is marked on desired examples. The above described procedures are shown from Step 2 to Step 4 in **Figure3**, respectively.

Iterative refinements based on individual preference learning: The evaluated results are then sent into the proposed irrelevance-based OPF module, which is used for initiation to build the OPF classifiers for adapting the recommended results to users' tastes gradually. Ten processed results are sent back to users from the irrelevance-based OPF module, in which the five results in the first row represent the re-evaluation samples for this iteration, and the other five samples in the second row stand for the self-refined results from last iteration. Users check the self-refined results and terminate the iteration of relevance feedback in Step 7 if satisfied. Otherwise, they evaluate the re-evaluation samples again and send the re-evaluated results for reprocessing in Step 8 to conduct the above described procedures iteratively. Details are shown from Step 5 to Step 8 in **Figure3**.

Algorithm 1 Personalized face-top coordination by the *DBSR* framework

Input: Upper-clothes candidate set of selected category T; a given female face image F.
Output: Face-coordinated upper-clothes.

1: Project all items in T along with F into the embedding space S by the jointly trained CNN-DCCA module.
2: Find the five closest upper-clothes features to the input face feature as the $Top-5$ recommended upper-clothes R via Euclidean distance metric.
3: Display the five corresponding upper-clothes images to the user for the subsequent irrelevance-based OPF self-refinement.
4: **repeat**
5: Mark results in the initial R with $\bigcirc$ standing for satisfied, and $\times$ standing for unsatisfied.
6: Send the evaluated results into the irrelevance-based OPF module for dynamic refinement.
7: Output the 10 samples displayed in two rows, in which the first row is the five re-evaluation samples, and the second row is the five self-refined results.
8: **until** The user is satisfied with the self-refined results.
9: The samples marked with the $\bigcirc$ symbols in last self-refined results are provided to the user as face-coordinated upper-clothes.

The right half part of **Figure 4** presents a simplified OPF space, in which only one relevant prototype and one paired irrelevant prototype exist. After calculating the irrelevance values, each non-prototype (displayed as a square) is transformed into a relevant sample (displayed as a black circle) or an irrelevant sample (displayed as a black triangle), and connected to one of the prototypes directly or indirectly. Finally, the space is divided into a relevant and an irrelevant areas. It is worth nothing that the five re-evaluation samples and the five self-refined results in the left half part

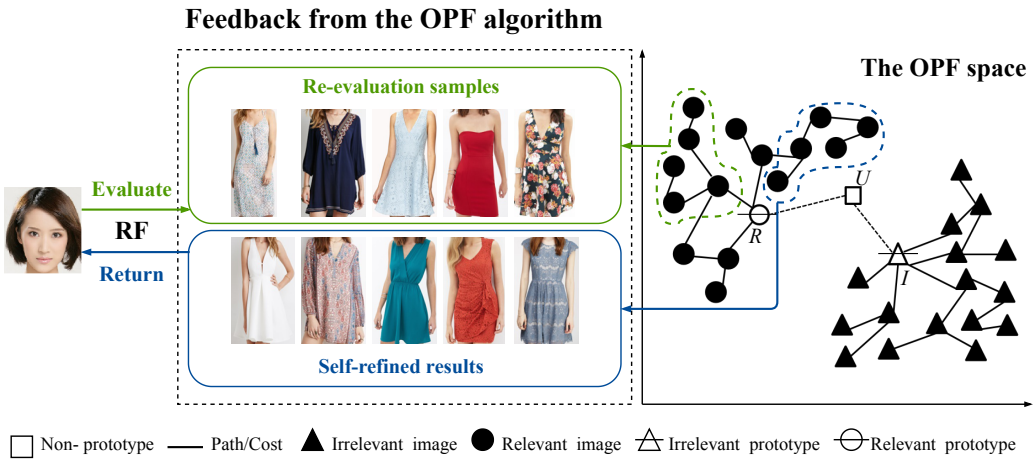

Fig. 4. Self-refining mechanism of the proposed irrelevance-based OPF algorithm.

of **Figure 4** are with the five largest and smallest irrelevance values both from the relevant class, respectively.

In this way, the proposed *DBSR* framework will recommend appropriate face-coordinated upper-clothes to female customers with general preferences and individual preferences considered jointly.

5 EXPERIMENTS AND EVALUATIONS

In this section, we first introduce details of our proposed jointly trained CNN-DCCA module and employ a series of ablation analysis, as well as a baseline comparison, to validate its effectiveness. Subsequently, several experiments are carried out to evaluate the user experience based on general preferences and individual preferences for the proposed *DBSR* framework.

5.1 Analysis on the jointly trained CNN-DCCA module

Implementation details: From the established *WCFT* dataset, we selected 28,000 pairs of face-top samples for training and the other 2,000 pairs for testing. The jointly trained CNN-DCCA module was constituted of feature extractors, feature projectors and a CCA sub-module, in which two workflows were conducted in parallel, and the parameters were defined as H_x and H_y for convolutional kernel matrices, φ_x and φ_y for non-linear projection matrices, and ω_x and ω_y for linear projection matrices, respectively. We illustrate the implementation details in **Figure 5** and **Algorithm 2** as elaborated below.

Firstly, we deployed two parallel VGG16 [48] with pre-trained parameters of ImageNet as feature extractors to generate 4096-d features from pairwise 224×224 face-top RGB images of the training dataset, respectively, which captured overall characteristic features for representing face styles and clothes styles. Compared with concatenating the traditional hand-crafted features of the individual elements to represent the overall objects, deep-based features can capture the overall characteristics seamlessly and be refined in an end-to-end manner. Note that we adopted VGG16, instead of other well developed deep face frameworks, to extract facial features due to the hairstyle-considered factor. Secondly, three non-linear full connected layers (FCLs) deployed in each workflow were leveraged to transform the extracted 4096-d features into 1024-d features three times in sequence. Thirdly, two linear projectors were used to convert the transformed 1024-d features into 10-d components, which were subsequently projected into a shared feature space. Fourthly, a CCA

Algorithm 2 Implementation details of the jointly trained CNN-DCCA module

procedure JOINTTRAIN ($\boldsymbol{T}$, $\boldsymbol{F}$) ▷ $\boldsymbol{T}$:Top, $\boldsymbol{F}$:Face
  Initialize VGG-16 net, FCL, and linear projectors for mapping
  for each epoch **do**
    for each batch (B_T, B_F) of Top and Face **do**
      for each pair $(t, f) \in (B_T, B_F)$ **do**
        $t \rightarrow x$ by VGG-16
        $f \rightarrow y$ by VGG-16
        $x \rightarrow \varphi_x(x)$ by non-linear mapping
        $y \rightarrow \varphi_y(y)$ by non-linear mapping
        $\varphi_x(x) \rightarrow \omega_x^T \varphi_x(x)$ by linear mapping
        $\varphi_y(y) \rightarrow \omega_y^T \varphi_y(y)$ by linear mapping
      end for
      Get converted batch (X, Y)
      Apply CCA method on (X, Y) to compute average correlation value
      Operate negatively to correlation value for obtaining loss value
      Compute the gradient with respect to loss value
      Back propagate to the feature projectors and conduct optimization via the RMSProp gradient descent method
      Back propagate to the feature extractors and conduct optimization via the RMSProp gradient descent method
    end for
  end for
end procedure

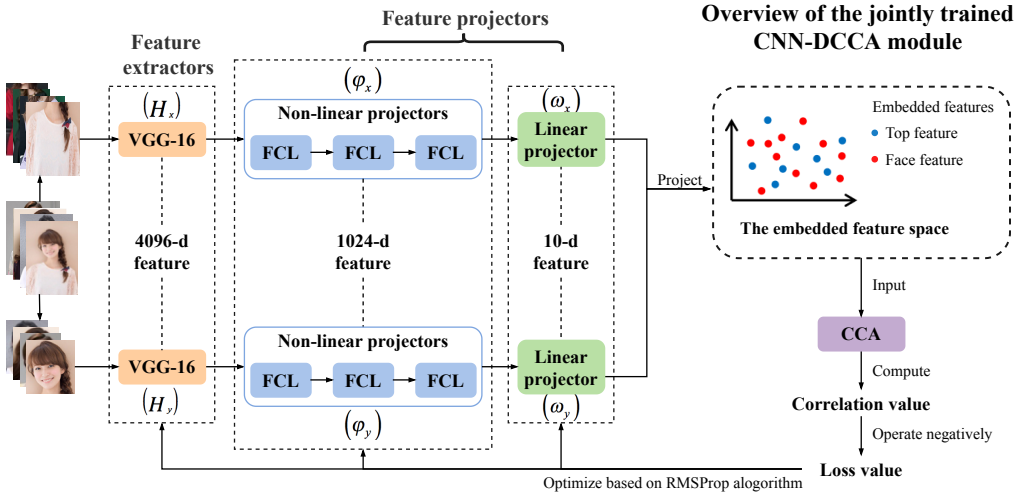

Fig. 5. Implementation details of the proposed jointly trained CNN-DCCA module.

sub-module was employed to generate the correlation values between pairwise components from different modalities, which were the reciprocal of the loss values. Finally, we employed the Root Mean Square Prop (RMSProp) algorithm [44] to perform gradient descent for minimizing the loss value iteratively, which could be deployed on feature projectors and extractors to optimize the corresponding parameters ($H_x, H_y, \varphi_x, \varphi_y, \omega_x$, and ω_y) individually or jointly by formula (4). It

is worth nothing that these parameters of feature extractors and projectors were initiated with the default values, but optimized separately in their respective parallel. It was observed that, by setting the batch size to 200 and the learning rate to 0.001, the proposed jointly trained CNN-DCCA module performed best and converged at approximately the 70th epoch, in which the correlation value was optimized from the initial value of 3.8 to the stable value of 9.8.

Loss function: According to the parameters described above and in **Section 4.1**, we defined $T \triangleq C_{XX}^{-\frac{1}{2}} C_{XY} C_{YY}^{-\frac{1}{2}}$, and thus the loss function is shown below:

$$loss = -argmax\ corr\left(\omega_x^T \varphi_x(top_{H_x}),\ \omega_y^T \varphi_y(face_{H_y})\right) = -tr\left(T^T T\right), \tag{4}$$

where the parameters of convolution kernel matrices, and non-linear and linear matrices were optimized to minimize the loss value jointly.

Evaluation metric: Two common metrics were employed for evaluating retrieval accuracy of the jointly trained CNN-DCCA module, which were mean reciprocal rank (*MRR*) and *Recall q@N*, respectively.

*MRR*1 is used to show the average rank of the most relevant item in all queries, which focuses on the position of the *Top-1* query:

$$MRR1 = \frac{1}{N_q} \sum_{i=1}^{N_q} \frac{1}{rank_i(1)}, \tag{5}$$

where N_q is the number of the queries; and $rank_i(1)$ corresponds to the rank of the most relevant item in the i_{th} query.

Recall q@N is used to describe how often the *Top-q* relevant items are included in the *Top-N* ranked list:

$$Recall\ q@N = \frac{\left|R_q \bigcap L_N\right|}{\left|R_q\right|}, \tag{6}$$

where R_q represents the *Top-q* relevant items; and L_N represents the *Top-N* ranked list. In this work, we set q to 1 as the most relevant item, and N to 5 since in the subsequent user study the *Top-5* queries were provided to users for evaluation. The final *Recall* 1@5 value was averaged over all queries, and used to show how often the most relevant item was included in the *Top-5* ranked list.

The above two metrics were used not only for objective retrieval effectiveness evaluation, but also for coarse pre-evaluation on the subsequent user study. Overall, the bigger are the *MRR*1 and the *Recall* 1@5 values, the better is the performance of the jointly trained CNN-DCCA module proposed in this paper.

Ablation study: Ablation study was conducted to test the effectiveness of different techniques adopted in the CNN DCCA module. The four deployed techniques were Pre-trained CNN for common feature embedding, Pre-trained CNN CCA for CCA-attached common feature embedding, Pre-trained CNN DCCA for DCCA-attached common feature embedding, and jointly trained CNN-DCCA for DCCA-attached common self-refined feature embedding, respectively, whose *MRR*1 and *Recall* 1@5 values were compared based on different component dimensions with an interval of 20 from 10 to 90.

From **Table 1**, several phenomena could be observed. Firstly, low values of the Pre-trained CNN showed that it was difficult to remove the inter-modal variation effectively by just adopting feature embedding, and the employed component dimensions could affect the results to a certain extent. Secondly, improved values of the Pre-trained CNN CCA showed that the CCA-attached

module could diminish inter-modal variation to some degree, which were also affected by the employed component dimensions. Thirdly, significantly improved values of the Pre-trained CNN DCCA revealed that it was effective to improve the performances by self-refining parameters of the linear and non-linear feature projectors with the DCCA method, which were hardly affected by the employed component dimensions. Finally, values of our proposed jointly trained CNN-DCCA demonstrated that joint self-refinement on the feature extractors and projectors could further improve the performances slightly, which were also hardly affected by the employed component dimensions.

Table 1. Ablation study together with comparison with another DCCAE framework.

Experiment metric	Component dimension	Pre-trained CNN	Pre-trained CNN CCA	Pre-trained CNN DCCA	**Jointly trained CNN-DCCA**	DCCAE
MRR1	10	0.007	0.037	0.204	**0.219**	0.224
	30	0.013	0.049	0.199	**0.217**	0.227
	50	0.021	0.058	0.201	**0.223**	0.225
	70	0.025	0.071	0.205	**0.225**	0.224
	90	0.027	0.079	0.206	**0.230**	0.228
Recall 1@5	10	0.011	0.042	0.207	**0.213**	0.211
	30	0.015	0.051	0.205	**0.217**	0.218
	50	0.023	0.067	0.206	**0.215**	0.216
	70	0.028	0.073	0.211	**0.221**	0.219
	90	0.034	0.081	0.209	**0.218**	0.223

Comparison with baseline: The Siamese network [10], MVE [17], DCCAE [56], ACMR [55], and DSCMR [66] are another five state-of-the-art frameworks for cross-modal retrieval tasks, as discussed in **Section 2.3**. Moreover, there are a number of other cross-modal retrieval frameworks referred to in [55], such as Bimodal-AE, Corr-AE, CMDN, CMDL, LCFS, CDLFM, LGCFL, JRL and JFSSL.

However, the Siamese network, and MVE and ACMR frameworks, needed paired and unpaired samples for triplet-based training. The DSCMR framework adopted not only a common representation space, but also a label space for supervised training. Therefore, it required all of the training samples to be equipped with limited corresponding category labels, which was similar to the LCFS, CDLFM, LGCFL, JRL, and JFSSL frameworks referred to in [55]. For our proposed module, we did not prepare unpaired samples and distinguished labels since they were difficult to obtain, as described in **Section 3**. In addition, performances of the Bimodal-AE, Corr-AE, CMDN, and CMDL frameworks adopting correlation loss were similar or inferior to the CCA-based method [55], which was proven to be greatly inferior to our jointly trained CNN-DCCA module, as shown in **Table 1**. Due to the above descriptions, we only adopted the DCCAE framework as the baseline for comparison, which added an autoencoder after the CCA module for reconstruction validation. Experiment results in **Table 1** revealed that it achieved similar performance to our jointly trained CNN-DCCA module since we also adopted feature self-refinement on VGG16.

In summary, compared with those metric values in [65], the moderate values that we obtained were acceptable, in which image-to-image modalities contributed positively. However, some existing many-to-many cases disturbed it negatively to some extent. The four techniques adopted in our jointly trained CNN-DCCA module were both necessary and effective, especially for self-refinement on feature projector parameters via the DCCA method. Moreover, due to similar retrieval performances of different dimension components, we adopted the 10-dimension component as our final output for efficiency of the subsequent relevance feedback procedure.

5.2 User study

In the user study, each recommended top garment image was presented to users with the given face image placed immediately above to improve the cognitive impression, which will be synthesized seamlessly in future work. Participants were divided into three groups, of which the first group comprised 10 participants denoted as the *VO* group for the *viewed-by-others (VBO)* experiments, the second group contained five participants denoted as the *VS* group for the *viewed-by-self (VBS)* experiments, and the third group comprised five participants for review experiments denoted as the *RV* group. This is because that females always buy clothes according to general preferences and individual preferences jointly, which correspond to *viewed-by-others* and *viewed-by-self* modes in fashion coordination, respectively. All of the participants were female college students in their twenties. It is worth nothing that, before conducting experiments, we asked participants to scan another 200 latest well-coordinated fashion model images exclusive of the 2,000 existing candidates for pre-training on face-top aesthetic coordination. In total, four kinds of experiments on the user study were conducted. We deployed the SBT technique as a baseline for comparison in the second and third experiments.

Experiment I (Top-N hit-rate): In this experiment, we evaluated performances of the jointly trained CNN-DCCA module in *VBO* and *VBS* modes with the *Top-N hit-rate* value to observe whether users could find well face-coordinated upper-clothes for others or themselves in the initial *Top-N* query. Each time a *Top-5* query result was presented to each participant. The left-most image represented the highest ranked result and vice versa. Participants were required to mark satisfactory results with $\bigcirc$ and unsatisfactory results with $\times$ symbols, respectively. The *Top-1*, *Top-3* , and *Top-5* positive hits were added by one as long as one satisfactory image existed in the first one, the first three, and the first five of the *Top-5* query results, respectively. Note that, even if more than one satisfactory result existed, only one positive hit could be recorded in one query. Consequently, the *Top-N hit-rate* value was calculated via dividing the total number of positive hits by the total number of queries. Some of the *Top-5* retrieval results are shown in **Figure 6**.

In *VBO* mode, the 50 testing face images referred to in **Section 3** were sent into the jointly trained CNN-DCCA module in sequence and randomly appointed a desired upper-clothes type from the pre-defined eight types for generating 50 sets of *Top-5* query results. Therefore, the *VO* group should evaluate 50 times per participant.

In *VBS* mode, the five participants in the *VS* group presented their own portrait images to the jointly trained CNN-DCCA module and selected the pre-defined eight upper-clothes types in sequence, respectively. As a consequence, the *VS* group should evaluate eight times per participant.

The *Top-1*, *Top-3* , and *Top-5 hit-rate* values displayed in **Table 2** were averaged from all participants of the *VO* and *VS* groups, respectively.

Table 2. Top-N hit-rate.

	Top-1	Top-3	Top-5
VBO	24.3%	52.7%	84.2%
VBS	13.8%	38.6%	70.6%

From **Table 2**, it could be observed that, although the *Top-1 hit-rate* values were low in the *VBO* and *VBS* modes, both of their *Top-5 hit-rate* values became acceptable. It is also notable that all of the *Top-N hit-rate* values in *VBS* mode were lower than those in *VBO* mode, on average. This might be because, in *VBO* mode, people evaluated fashion attractiveness primarily by aesthetic impression of images; whereas, in *VBS* mode, many latent factors would be considered by themselves, such

Fig. 6. Some examples of the *Top-5* retrieval results from the jointly trained CNN-DCCA module and the *Top-1* binary-competition experiment. Note that, due to different subjective preferences on upper clothes, the retrieval results were evaluated in **Section 5.2** by the first and the second experiments quantitatively and statistically.

as preferred regions [20], desired brands, current fashion trends, etc. This is also in accordance with the fact that fashion model images are appreciated by most people, but worn clothes may not be commonly appreciated by fashion models themselves. In addition, different values between *VBO* mode and *VBS* mode also prove, to some extent, that it is unreliable for female customers to find desired clothes during online shopping. As a consequence, it is indeed necessary to integrate a preference-focused method with the jointly trained CNN-DCCA module to further improve recommendation satisfaction.

Experiment II (Top-5 hit-number): In this experiment, we compared performances among the jointly trained CNN-DCCA method and the SBT method, as well as Randomly Selected (RS) results in *VBO* and *VBS* modes via the numbers of the satisfactory face-coordinated clothes found in their corresponding initial *Top-5* queries. We discarded the traditional score-based metric for subjective evaluation adopted in many researches. This was done because it was unreliable to ask participants to score accurate values for five results each time and repeat many times without deviation, which was proven in practice.

Instead, we used the *Top-5 hit-number* metric, which directly adopted the number of satisfactory results in each *Top-5* query as the corresponding integer score. These values of the jointly trained CNN-DCCA method could be calculated based on the data generated in the first experiment, and those of the SBT method should be calculated by conducting the same procedure as the first experiment again. The average *Top-5 hit-number* values corresponding to the three competitive methods in *VBO* and *VBS* modes are displayed in **Table 3**.

From **Table 3**, it could be seen that the jointly trained CNN-DCCA method achieved results that were competitive with the SBT method and outperformed the RS method, which were validated by one-tailed paired t-test at significance level 5% ($p = 0.05$). The highest average score 2.04 meant

Table 3. Top-5 hit-number.

	Jointly trained CNN-DCCA	SBT	RS
VBO	2.04	1.72	0.97
VBS	1.63	1.47	1.12

that the jointly trained CNN-DCCA method was able to recommend at least two appropriate upper-clothes images to users in a *Top-5* query, on average. The inferior scores of the SBT method might result from the arbitrary paired information transfer without considering latent correlations and dissimilarity of some overall faces caused by mutual interference between hairstyle and facial features. The *Top-5 hit-number* values of the jointly trained CNN-DCCA and SBT methods in *VBS* mode were still lower than those in *VBO* mode, as in the first experiment. The value in the third column in *VBS* mode was higher than that in *VBO* mode, which might be resultant from randomly selected samples.

Experiment III (Top-1 binary-competition): This experiment was carried out to evaluate whether the proposed jointly trained CNN-DCCA method achieved better performances than the other two baselines for the *Top-1* query. The *RV* group was asked to participate in this experiment in *VBO* and *VBS* modes, since some of the provided *Top-1* results were evaluated by the *VO* and *VS* groups. Ten face images selected from the 50 testing images randomly and five self-face images of the *RV* group, along with all of the pre-defined eight upper-clothes types, were adopted as inputs. Therefore, the jointly trained CNN-DCCA method and the SBT method generated 80 *Top-1* results in *VBO* mode and eight in *VBS* mode for each participant, respectively. In addition, a randomly selected 88 upper-clothes images corresponding to the pre-defined eight types for 11 per type were prepared, as well.

Consequently, 88 pairwise jointly trained CNN-DCCA vs. SBT competitors and 88 pairwise jointly trained CNN-DCCA vs. RS competitors were presented to each participant in the *RV* group for evaluation on the *Top-1 binary competition*, as shown in the bottom of **Figure 6**, where the two competitors were placed randomly on both sides of the given face image when tested. The participants were required to select a more satisfactory sample from pairwise competitors, and thus for each pairwise group 400 trials in *VBO* mode along with 40 trials in *VBS* mode were generated. The experiment results are presented in **Table 4**.

Table 4. Top-1 binary-competition.

	Jointly trained CNN-DCCA vs SBT		Jointly trained CNN-DCCA vs RS	
	Jointly trained CNN-DCCA	SBT	Jointly trained CNN-DCCA	RS
VBO	213(53.3%)	187(46.7%)	243(60.7%)	157(39.3%)
VBS	23(57.5%)	17(42.5%)	27(67.5%)	13(33.5%)

From **Table 4**, it could be observed that, the jointly trained CNN-DCCA method outperformed both the SBT method and the RS method, which was in accordance with the second experiment.

Experiment IV (OPF-based optimization): In this experiment, the self-refined performance of the irrelevance-based OPF method to adapt to individual preferences was evaluated. The *Top-5 hit-number* metric referred to previously was adopted in this experiment again. In *VBO* mode, the initial seeds were constituted of 50 sets of *Top-5* queries per participant from the *VO* group generated in the first experiment. In *VBS* mode, the initial seeds were made up of eight sets of *Top-5* queries per participant from the *VS* group in the first experiment as well as from the *RV* group

in the third experiment, respectively. In total, we conducted eight RF phases in this experiment. The *Top-5 hit-number* values in the initial phase and the different RF phases are displayed by a line chart in **Figure 7**, in which the scores were generated from the above described groups on average, respectively.

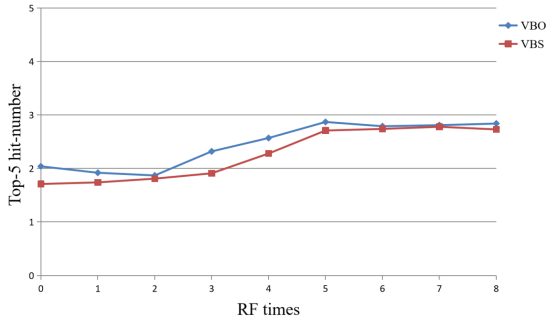

Fig. 7. Top-5 hit-number values of the eight OPF-based RF phases in VBO and VBS modes.

As shown in **Figure 7**, the *Top-5 hit-number* values decreased in *VBO* mode for the first and second phases, but increased from the third phase and became stable from the fifth phase. Furthermore, both of the *VBS* and *VBO* lines reached peak values in the fifth phase and became stable and similar from then on. Therefore, we reach the following conclusions. Firstly, although the performance dropped when starting due to a very small number of training samples, the proposed irrelevance-based OPF method could improve retrieval results from the jointly trained CNN-DCCA module iteratively and finally achieve stable performances within limited RF times. Secondly, integrated with the individual preferences, the OPF-based module could achieve similar performances in both the *VBO* and *VBS* modes. It id worth nothing that this self-refined procedure could be used to supplement traditional online tag-based retrieval, as well.

Some initial query results (displayed in the first row) and iterative self-refined results (displayed in the other rows) generated by the *DBSR* framework with an authorized female face image from the *RV* group and three selected upper-clothes categories (long-dress, short T-shirt, and coat) as inputs are presented in **Figure 8**, in which the owner of the authorized female face image and another two participants from the *RV* group were asked for evaluation, respectively. Note that, in RF phases, the results surrounded by dotted frames are re-evaluation samples for this iteration, and the others are self-refined results from last iteration. If users were satisfied with the self-refined results, the iteration procedure ended. Certain phenomena can be observed as follows. Firstly, the jointly trained CNN-DCCA method is sensitive to hairstyles and poses, even from the same female. Secondly, with the progress of the OPF-based RF phase, the *Top-5* results become increasingly similar either in colors, patterns or local areas (e.g., collars), which proves that, the proposed irrelevance-based OPF method is effective to adapt to different individual preferences gradually.

In summary, four types of user studies validated the effectiveness of *DBSR*. In the first and second experiments, the *Top-5 hit-rate* and *Top-5 hit-number* values demonstrate that it is possible for users to find at least one or two satisfactory face-coordinated upper-clothes within five query results. In the third experiment, the *Top-1 binary-competition* values confirm that our method outperforms the SBT baseline. In the last experiment, the RF phase line chart and the OPF-based optimized results prove the effectiveness of our self-refined method for adapting to individual preferences. It is worth nothing that, besides better performances in the *Top-5 hit-number* and the *Top-1 binary-competition*

Fig. 8. Some examples from the initial phase, and the first and fifth OPF-based RF phases in *VBO* and *VBS* modes.

experiments, the jointly trained CNN-DCCA method takes latent correlations between face-top pairs into consideration, which could be further developed into dimension-wise analysis in the future. Its performance could be optimized further by expanding the training data and diminishing many-to-many cases based on the deep-based self-learning mechanism. We do not provide the statistical patterns for face-top coordination due to lacking huge number of manual annotations and prior rules, as discussed in [36, 61]. In addition, unlike some studies [14, 43, 54] that visualize the feature spaces containing various fashion items to show their distributions and relationships, face styles and their correlations with top clothing styles are difficult to recognize simply from the visualized space due to the subjectivity issue and their much more different modalities. For example, from a visualized feature-clustered space for face datasets, it is very difficult for ordinary people to distinguish different face styles based on their cluster-based distributions, to say nothing of the face-top combination clusters. Therefore, we validate the effectiveness of the jointly trained CNN-DCCA framework for face-top compatibility learning from the following three folds: the correlation value (it is optimized from the initial value of 3.8 to the stable value of 9.8), the *MRR*1 and the *Recall* 1@5 values (they are significantly increased from the Pre-trained CNN framework to the jointly trained CNN-DCCA framework, which can be correspondingly regarded as the initial phase and the finally optimized phase), and the statistical data in **Table 3** and **Table 4** (the presented method outperforms the RS method, where the RS results can be regarded as the initial results without optimization).

6 CONCLUSION

Our work makes the first effort toward an industry-scale application on personalized aesthetic coordination between female face style and upper-clothes style, which fills the gap of this field for fashion recommendation via the designed *DBSR* framework along with a moderate-scale *WCFT* dataset. Our approach has the following features. First, different from traditional distance-based methods, whose results are fixed according to distance metrics, our approach is flexible for individual preferences along with fashion trends. Second, it can be deployed before the traditional top-down coordination and acts as its supplement to improve the overall fashion impression. Third, it can relieve tremendous applications, such as online 2D clothes shopping, 3D virtual fitting, etc, from time-consuming manual selection processes. Finally, the presented framework can adapt to other coordination-related applications with the *WCFT* dataset replaced.

Our approach possesses certain limitations. Moderate values of the *MRR*1, along with the *Recall* 1@5 values, may cause unreliable recommendation results to some extent, which can be addressed by the irrelevance-based OPF method. A huge scale of candidates would reduce efficiency of the irrelevance-based OPF method. Fortunately, this limitation is not critical, since in online shopping, female customers always focus on some brands of upper-clothes from a certain category with recent popular fashion during a single transaction, of which the scale of the candidates is not large. The given face and recommended upper-clothes are not synthesized seamlessly when displayed, which may diminish the user experience. Finally, we do not provide statistical attribute-based patterns for face-top coordination due to lacking huge annotated attributes and pre-defined rules for training.

In the future, we aim to address the aforementioned limitations and improve our work in the following aspects. First, we will optimize the *WCFT* dataset by expanding the training data along with the evaluators, and reducing many-to-many cases among pairs simultaneously. Second, we will synthesize the given face with the recommended upper-clothes seamlessly for a better user experience, and even make it possible to be assessed by some up-to-date machine evaluation algorithms. Third, other sensitive factors will be considered, such as seasonality, color, price, etc., to make the recommendation results more reasonable. Finally, we will conduct an in-depth

analysis on dimension-level correlations between faces and upper-clothes, as well as comparing with handcrafted features, and elucidate what kinds of visual dimensions interact with each other.

REFERENCES

[1] Kaori Abe, Teppei Suzuki, Shunya Ueta, Akio Nakamura, Yutaka Satoh, and Hirokatsu Kataoka. 2017. Changing fashion cultures. *arXiv preprint arXiv:1703.07920* (2017).

[2] Stuti Ajmani, Hiranmay Ghosh, Anupama Mallik, and Santanu Chaudhury. 2013. An ontology based personalized garment recommendation system. In *Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, Vol. 3. IEEE, 17–20.

[3] Shotaro Akaho. 2006. A kernel method for canonical correlation analysis. *arXiv preprint cs/0609071* (2006).

[4] Ziad Al-Halah, Rainer Stiefelhagen, and Kristen Grauman. 2017. Fashion forward: Forecasting visual style in fashion. In *2017 IEEE International Conference on Computer Vision(ICCV)*. IEEE, 388–397.

[5] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. 2013. Deep canonical correlation analysis. In *2013 International conference on machine learning(ICML)*. ACM, 1247–1255.

[6] Jean-Yves Baudouin and Guy Tiberghien. 2004. Symmetry, averageness, and feature size in the facial attractiveness of women. *Acta psychologica* 117, 3 (2004), 313–332.

[7] Andrea Bottino and Aldo Laurentini. 2010. The analysis of facial beauty: an emerging area of research in pattern analysis. In *International Conference Image Analysis and Recognition*. Springer, 425–435.

[8] Wen Chen, Pipei Huang, Jiaming Xu, Xin Guo, Cheng Guo, Fei Sun, Chao Li, Andreas Pfadler, Huan Zhao, and Binqiang Zhao. 2019. Pog: Personalized outfit generation for fashion recommendation at alibaba ifashion. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2662–2670.

[9] Xu Chen, Hanxiong Chen, Hongteng Xu, Yongfeng Zhang, Yixin Cao, Zheng Qin, and Hongyuan Zha. 2019. Personalized Fashion Recommendation with Visual Explanations based on Multimodal Attention Network: Towards Visually Explainable Recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 765–774.

[10] Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 1. IEEE, 539–546.

[11] Andre Tavares Da Silva, Alexandre Xavier Falcão, and Léo Pini Magalhães. 2011. Active learning paradigms for CBIR systems based on optimum-path forest classification. *Pattern Recognition* 44, 12 (2011), 2971–2978.

[12] Yael Eisenthal, Gideon Dror, and Eytan Ruppin. 2006. Facial attractiveness: Beauty and the machine. *Neural Computation* 18, 1 (2006), 119–142.

[13] Ruining He, Chunbin Lin, Jianguo Wang, and Julian McAuley. 2016. Sherlock: sparse hierarchical embeddings for visually-aware one-class collaborative filtering. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence(IJCAI)*. Morgan Kaufmann, 3740–3746.

[14] Ruining He and Julian McAuley. 2016. VBPR: visual bayesian personalized ranking from implicit feedback. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. AAAI, 403–410.

[15] Ruining He, Charles Packer, and Julian McAuley. 2016. Learning compatibility across categories for heterogeneous item recommendation. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*. IEEE, 937–942.

[16] Tong He and Yang Hu. 2018. FashionNet: Personalized Outfit Recommendation with Deep Neural Network. *arXiv preprint arXiv:1810.02443* (2018).

[17] Wanjia He, Weiran Wang, and Karen Livescu. 2017. Multi-view recurrent neural acoustic word embeddings. In *the 5th International Conference on Learning Representations(ICLR)*.

[18] Shintami Chusnul Hidayati, Cheng-Chun Hsu, Yu-Ting Chang, Kai-Lung Hua, Jianlong Fu, and Wen-Huang Cheng. 2018. What Dress Fits Me Best?: Fashion Recommendation on the Clothing Style for Personal Body Shape. In *Proceedings of the 26th ACM international conference on Multimedia(ACM MM)*. ACM, 438–446.

[19] Harold Hotelling. 1992. Relations between two sets of variates. *Breakthroughs in statistics* (1992), 162–190.

[20] Min Hou, Le Wu, Enhong Chen, Zhi Li, Vincent W Zheng, and Qi Liu. 2019. Explainable Fashion Recommendation: A Semantic Attribute Region Guided Approach. (2019), 4681–4688.

[21] Wei-Lin Hsiao and Kristen Grauman. 2017. Learning the latent "look": Unsupervised discovery of a style-coherent embedding from fashion images. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 4213–4222.

[22] Wei-Lin Hsiao and Kristen Grauman. 2018. Creating capsule wardrobes from fashion images. In *2018 IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*. IEEE, 7161–7170.

[23] Chia-Wei Hsieh, Chieh-Yun Chen, Chien-Lung Chou, Hong-Han Shuai, Jiaying Liu, and Wen-Huang Cheng. 2019. FashionOn: Semantic-guided Image-based Virtual Try-on with Detailed Human and Clothing Information. In *Proceedings of the 27th ACM International Conference on Multimedia(ACM MM)*. ACM, 275–283.

[24] Yang Hu, Xi Yi, and Larry S Davis. 2015. Collaborative fashion recommendation: A functional tensor factorization approach. In *Proceedings of the 23rd ACM international conference on Multimedia(ACM MM)*. ACM, 129–138.
[25] Tomoharu Iwata, Shinji Wanatabe, and Hiroshi Sawada. 2011. Fashion coordinates recommender system using photographs from fashion magazines. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence(IJCAI)*, Vol. 3. Morgan Kaufmann, 2262–2267.
[26] Vignesh Jagadeesh, Robinson Piramuthu, Anurag Bhardwaj, Wei Di, and Neel Sundaresan. 2014. Large scale visual recommendations from street fashion images. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1925–1934.
[27] Wang-Cheng Kang, Eric Kim, Jure Leskovec, Charles Rosenberg, and Julian McAuley. 2019. Complete the Look: Scene-based Complementary Product Recommendation. In *2019 IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*. IEEE, 10532–10541.
[28] M Hadi Kiapour, Kota Yamaguchi, Alexander C Berg, and Tamara L Berg. 2014. Hipster wars: Discovering elements of fashion styles. In *2014 European conference on computer vision(ECCV)*. Springer, 472–488.
[29] Dmitry Kornilov. 2019. Recommendation system based on a user's physical features. US Patent App. 15/826,533.
[30] Sudhir Kumar and Mithun Das Gupta. 2019. *c+* GAN: Complementary Fashion Item Recommendation. *arXiv preprint arXiv:1906.05596* (2019).
[31] Honglin Li, Masahiro Toyoura, Kazumi Shimizu, Wei Yang, and Xiaoyang Mao. 2016. Retrieval of clothing images based on relevance feedback with focus on collar designs. *The Visual Computer* 32, 10 (2016), 1351–1363.
[32] Yuncheng Li, Liangliang Cao, Jiang Zhu, and Jiebo Luo. 2017. Mining fashion outfit composition using an end-to-end deep learning approach on set data. *IEEE Transactions on Multimedia* 19, 8 (2017), 1946–1955.
[33] Lizi Liao, Xiangnan He, Bo Zhao, Chong-Wah Ngo, and Tat-Seng Chua. 2018. Interpretable multimodal retrieval for fashion products. In *Proceedings of the 26th ACM international conference on Multimedia(ACM MM)*. ACM, 1571–1579.
[34] Yujie Lin, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Jun Ma, and Maarten De Rijke. 2020. Explainable Outfit Recommendation with Joint Outfit Matching and Comment Generation. *IEEE Transactions on Knowledge and Data Engineering* 32, 8 (2020), 1502–1516.
[35] Jian Liu, Pengpeng Zhao, Yanchi Liu, Victor S Sheng, Fuzheng Zhuang, Jiajie Xu, Xiaofang Zhou, and Hui Xiong. 2019. Deep Cross Networks with Aesthetic Preference for Cross-domain Recommendation. *arXiv preprint arXiv:1905.13030* (2019).
[36] Luoqi Liu, Junliang Xing, Si Liu, Hui Xu, Xi Zhou, and Shuicheng Yan. 2014. Wow! you are so beautiful today! *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 11, 1s (2014), 1–20.
[37] Si Liu, Jiashi Feng, Zheng Song, Tianzhu Zhang, Hanqing Lu, Changsheng Xu, and Shuicheng Yan. 2012. Hi, magic closet, tell me what to wear!. In *Proceedings of the 20th ACM international conference on Multimedia(ACM MM)*. ACM, 619–628.
[38] Zhi Lu, Yang Hu, Yunchao Jiang, Yan Chen, and Bing Zeng. 2019. Learning Binary Code for Personalized Fashion Recommendation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 10562–10570.
[39] Yihui Ma, Jia Jia, Suping Zhou, Jingtian Fu, Yejun Liu, and Zijian Tong. 2017. Towards better understanding the clothing fashion styles: A multimodal deep learning approach. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. AAAI, 38–44.
[40] Yunshan Ma, Xun Yang, Lizi Liao, Yixin Cao, and Tat-Seng Chua. 2019. Who, Where, and What to Wear? Extracting Fashion Knowledge from Social Media. In *Proceedings of the 27th ACM International Conference on Multimedia(ACM MM)*. ACM, 257–265.
[41] Utkarsh Mall, Kevin Matzen, Bharath Hariharan, Noah Snavely, and Kavita Bala. 2019. GeoStyle: Discovering Fashion Trends and Events. In *2019 IEEE/CVF International Conference on Computer Vision(ICCV)*. IEEE, 411–420.
[42] Kevin Matzen, Kavita Bala, and Noah Snavely. 2017. Streetstyle: Exploring world-wide clothing styles from millions of photos. *arXiv preprint arXiv:1706.01869* (2017).
[43] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 43–52.
[44] Mahesh Chandra Mukkamala and Matthias Hein. 2017. Variants of rmsprop and adagrad with logarithmic regret bounds. In *the 34th International Conference on Machine Learning(ICML)*, Vol. 70. ACM, 2545–2553.
[45] Tam V Nguyen, Si Liu, Bingbing Ni, Jun Tan, Yong Rui, and Shuicheng Yan. 2012. Sense beauty via face, dressing, and/or voice. In *Proceedings of the 20th ACM international conference on Multimedia(ACM MM)*. ACM, 239–248.
[46] Hosnieh Sattar, Gerard Pons-Moll, and Mario Fritz. 2019. Fashion is taking shape: Understanding clothing preference based on body shape from online sources. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 968–977.
[47] Edgar Simo-Serra, Sanja Fidler, Francesc Moreno-Noguer, and Raquel Urtasun. 2015. Neuroaesthetics in fashion: Modeling the perception of fashionability. In *2015 IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*.

IEEE, 869–877.
[48] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
[49] Guang-Lu Sun, Zhi-Qi Cheng, Xiao Wu, and Qiang Peng. 2018. Personalized clothing recommendation combining user social circle and fashion style consistency. *Multimedia Tools and Applications* 77, 14 (2018), 17731–17754.
[50] Moeko Takagi, Edgar Simo-Serra, Satoshi Iizuka, and Hiroshi Ishikawa. 2017. What makes a style: Experimental analysis of fashion prediction. In *2017 IEEE International Conference on Computer Vision Workshops(ICCVW)*. IEEE, 2247–2253.
[51] Pongsate Tangseng and Takayuki Okatani. 2020. Toward explainable fashion recommendation. In *The IEEE Winter Conference on Applications of Computer Vision*. IEEE, 2153–2162.
[52] Kristen Vaccaro, Sunaya Shivakumar, Ziqiao Ding, Karrie Karahalios, and Ranjitha Kumar. 2016. The elements of fashion style. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. 777–785.
[53] Dario Riccardo Valenzano, Andrea Mennucci, Giandonato Tartarelli, and Alessandro Cellerino. 2006. Shape analysis of female facial attractiveness. *Vision research* 46, 8-9 (2006), 1282–1291.
[54] Andreas Veit, Balazs Kovacs, Sean Bell, Julian McAuley, Kavita Bala, and Serge Belongie. 2015. Learning visual clothing style with heterogeneous dyadic co-occurrences. In *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 4642–4650.
[55] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. 2017. Adversarial cross-modal retrieval. In *Proceedings of the 25th ACM international conference on Multimedia(ACM MM)*. ACM, 154–162.
[56] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. 2015. On deep multi-view representation learning. In *Proceedings of the 32nd International Conference on Machine Learning(ICML)*, Vol. 37. ACM, 1083–1092.
[57] Zhonghua Wu, Guosheng Lin, Qingyi Tao, and Jianfei Cai. 2019. M2e-try on net: Fashion from model to everyone. In *Proceedings of the 27th ACM International Conference on Multimedia(ACM MM)*. ACM, 293–301.
[58] Duorui Xie, Lingyu Liang, Lianwen Jin, Jie Xu, and Mengru Li. 2015. SCUT-FBP: A benchmark dataset for facial beauty perception. In *2015 IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, 1821–1826.
[59] Kota Yamaguchi, Tamara L Berg, and Luis E Ortiz. 2014. Chic or social: Visual popularity analysis in online fashion networks. In *Proceedings of the 22nd ACM international conference on Multimedia(ACM MM)*. ACM, 773–776.
[60] Wei Yang, Masahiro Toyoura, and Xiaoyang Mao. 2012. Hairstyle suggestion using statistical learning. In *International Conference on Multimedia Modeling(MMM)*. Springer, 277–287.
[61] Xun Yang, Xiangnan He, Xiang Wang, Yunshan Ma, Fuli Feng, Meng Wang, and Tat-Seng Chua. 2019. Interpretable Fashion Matching with Rich Attributes. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 775–784.
[62] Xun Yang, Yunshan Ma, Lizi Liao, Meng Wang, and Tat-Seng Chua. 2019. Transnfcm: Translation-based neural fashion compatibility modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. AAAI, 403–410.
[63] Lap-Fai Yu, Sai Kit Yeung, Demetri Terzopoulos, and Tony F Chan. 2012. DressUp!: outfit synthesis through automatic optimization. *ACM Trans. Graph.* 31, 6 (2012).
[64] Wenhui Yu, Huidi Zhang, Xiangnan He, Xu Chen, Li Xiong, and Zheng Qin. 2018. Aesthetic-based clothing recommendation. In *Proceedings of the 2018 World Wide Web Conference(WWW)*. ACM, 649–658.
[65] Yi Yu, Suhua Tang, Francisco Raposo, and Lei Chen. 2019. Deep cross-modal correlation learning for audio and lyrics in music retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 15, 1 (2019), 1–16.
[66] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. 2019. Deep supervised cross-modal retrieval. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 10386–10395.
[67] Zhengzhong Zhou, Xiu Di, Wei Zhou, and Liqing Zhang. 2018. Fashion sensitive clothing recommendation using hierarchical collocation model. In *Proceedings of the 26th ACM international conference on Multimedia(ACM MM)*. ACM, 1119–1127.