Leveraging Content-Style Item Representation for Visual Recommendation

Yashar Deldjoo[1], Tommaso Di Noia[1], Daniele Malitesta[1⋆], and Felice Antonio Merra[2] ⋆⋆

[1] Politecnico di Bari, Bari, Italy, name.surname@poliba.it
[2] Amazon Science Berlin, Germany, felmerra@amazon.de

Abstract. When customers' choices may depend on the visual appearance of products (e.g., fashion), visually-aware recommender systems (VRSs) have been shown to provide more accurate preference predictions than pure collaborative models. To refine recommendations, recent VRSs have tried to recognize the influence of each item's visual characteristic on users' preferences, for example, through attention mechanisms. Such visual characteristics may come in the form of content-level item metadata (e.g., image tags) and reviews, which are not always and easily accessible, or image regions-of-interest (e.g., the collar of a shirt), which miss items' style. To address these limitations, we propose a pipeline for visual recommendation, built upon the adoption of those features that can be easily extracted from item images and represent the item content on a stylistic level (i.e., color, shape, and category of a fashion product). Then, we inject such features into a VRS that exploits attention mechanisms to uncover users' personalized importance for each content-style item feature and a neural architecture to model non-linear patterns within user-item interactions. We show that our solution can reach a competitive accuracy and beyond-accuracy trade-off compared with other baselines on two fashion datasets. Code and datasets are available at: https://github.com/sisinflab/Content-Style-VRSs.

Keywords: Visual Recommendation · Attention · Collaborative Filtering.

1 Introduction and Related Work

Recommender systems (RSs) help users in their decision-making process by guiding them in a personalized fashion to a small subset of interesting products or services amongst massive corpora. In applications where visual factors are at play (e.g., fashion [22], food [14], or tourism [33]), customers' choices are highly dependent on the visual product appearance that attracts attention, enhances emotions, and shapes their first impression about products. By incorporating

⋆ Authors are listed in alphabetical order. Corresponding author: Daniele Malitesta (daniele.malitesta@poliba.it).
⋆⋆ Work performed while at Politecnico di Bari, Italy.

this source of information when modeling users' preference, visually-aware recommender systems (VRSs) have found success in extending the expressive power of pure collaborative recommender models [10, 12, 13, 17, 18].

Recommendation can hugely benefit from items' side information [4]. To this date, several works have leveraged the high-level representational power of convolutional neural networks (CNNs) to extract item visual features, where the adopted CNN may be either pretrained on different datasets and tasks, e.g., [3, 11, 18, 26, 29], or trained end-to-end in the downstream recommendation task, e.g., [23,38]. While the former family of VRSs builds upon a more convenient way of visually representing items (i.e., reusing the knowledge of pretrained models), such representations are not entirely in line with correctly providing users' visual preference estimation. That is, CNN-extracted features cannot capture what each user enjoys about a product picture since she might be more attracted by the color and shape of a specific bag, but these features do not necessarily match what the pretrained CNN learned when classifying the product image as a bag.

Recently, there have been a few attempts trying to uncover user's personalized visual attitude towards finer-grained item characteristics, e.g., [7–9, 21]. These solutions disentangle product images at *(i)* `content`-level, by adopting item metadata and/or reviews [9, 31], *(ii)* `region`-level, by pointing the user's interest towards parts of the image [8, 36] or video frames [7], and *(iii)* both `content`- and `region`-level [21]. It is worth mentioning that most of these approaches [7, 8, 21, 36] exploit attention mechanisms to weight the importance of the `content` or the `region` in driving the user's decisions.

Despite their superior performance, we recognize practical and conceptual limitations in adopting both `content`- and `region`-level item features, especially in the fashion domain. The former rely on additional side information (e.g., image tags or reviews), which could be not-easily and rarely accessible, as well as time-consuming to collect, while the latter ignore stylistic characteristics (e.g., color or texture) that can be impactful on the user's decision process [41].

Driven by these motivations, we propose a pipeline for visual recommendation, which involves a set of visual features, i.e., color, shape, and category of a fashion product, whose extraction is straightforward and always possible, describing items' content on a stylistic level. We use them as inputs to an attention- and neural-based visual recommender system, with the following purposes:

- We disentangle the visual item representations on the stylistic content level (i.e., color, shape, and category) by making the attention mechanisms weight the importance of each feature on the user's visual preference and making the neural architecture catch non-linearities in user/item interactions.
- We reach a reasonable compromise between accuracy and beyond-accuracy performance, which we further justify through an ablation study to investigate the importance of attention (in all its configurations) on the recommendation performance. Notice that no ablation is performed on the content-style input features, as we learn to weight their contribution through the end-to-end attention network training procedure.

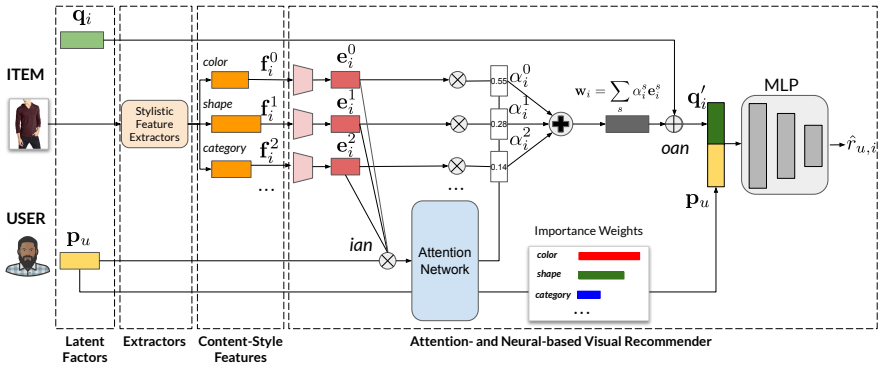

Fig. 1: Our proposed pipeline for visual recommendation, involving content-style item features, attention mechanisms, and a neural architecture.

2 Method

In the following, we present our visual recommendation pipeline (Figure 1).

Preliminaries. We indicate with $\mathcal{U}$ and $\mathcal{I}$ the sets of users and items. Then, we adopt $\mathbf{R}$ as the user/item interaction matrix, where $r_{ui} \in \mathbf{R}$ is 1 for an interaction, 0 otherwise. As in latent factor models such as matrix factorization (MF) [25], we use $\mathbf{p}_u \in \mathbb{R}^{1\times h}$ and $\mathbf{q}_i \in \mathbb{R}^{1\times h}$ as user and item latent factors, respectively, where $h << |\mathcal{U}|, |\mathcal{I}|$. Finally, we denote with $\mathbf{f}_i \in \mathbb{R}^{1\times v}$ the visual feature for item image i, usually the fully-connected layer activation of a pretrained convolutional neural network (CNN).

Content-Style Features. Let $\mathcal{S}$ be the set of content-style features to characterize item images. Even if we adopt $\mathcal{S} = \{\text{color}, \text{shape}, \text{category}\}$, for the sake of generality, we indicate with $\mathbf{f}_i^s \in \mathbb{R}^{1\times v_s}$ the s-th content-style feature of item i. Since all $\mathbf{f}_i^s$ do not necessarily belong to the same latent space, we project them into a common latent space $\mathbb{R}^{1\times h}$, i.e., the same as the one of $\mathbf{p}_u$ and $\mathbf{q}_i$. Thus, for each $s \in \mathcal{S}$, we build an encoder function $enc_s : \mathbb{R}^{1\times v_s} \mapsto \mathbb{R}^{1\times h}$, and encode the s-th content-style feature of item i as:

$$\mathbf{e}_i^s = enc_s(\mathbf{f}_i^s) \tag{1}$$

where $\mathbf{e}_i^s \in \mathbb{R}^{1\times h}$, and enc_s is either trainable, e.g., a multi-layer perceptron (MLP), or handcrafted, e.g., principal-component analysis (PCA). In this work, we use an MLP-based encoder for the color feature, a CNN-based encoder for the shape, and PCA for the category.

Attention Network. We seek to produce recommendations conditioned on the visual preference of user u towards each content-style item characteristic. That is, the model is supposed to assign different importance weights to each encoded feature $\mathbf{e}_i^s$ based on the predicted user's visual preference ($\hat{r}_{u,i}$). Inspired by previous works [7, 8, 21, 36], we use attention. Let $ian(\cdot)$ be the function to aggregate the inputs to the attention network $\mathbf{p}_u$ and $\mathbf{e}_i^s$, e.g., element-wise

multiplication. Given a user-item pair (u, i), the network produces an attention weight vector $\mathbf{a}_{u,i} = [a^0_{u,i}, a^1_{u,i}, \ldots, a^{|\mathcal{S}|-1}_{u,i}] \in \mathbb{R}^{1\times|\mathcal{S}|}$, where $a^s_{u,i}$ is calculated as:

$$a^s_{u,i} = \boldsymbol{\omega}_2(\boldsymbol{\omega_1} ian(\mathbf{p}_u, \mathbf{e}^s_i) + \mathbf{b}_1) + \mathbf{b}_2 = \boldsymbol{\omega}_2(\boldsymbol{\omega_1}(\mathbf{p}_u \odot \mathbf{e}^s_i) + \mathbf{b}_1) + \mathbf{b}_2 \quad (2)$$

where $\odot$ is the Hadamard product (element-wise multiplication), while $\boldsymbol{\omega}_*$ and $\mathbf{b}_*$ are the matrices and biases for each attention layer, i.e., the network is implemented as a 2-layers MLP. Then, we normalize $\mathbf{a}_{u,i}$ through the temperature-smoothed $softmax$ function [20], so that $\sum_s a^s_{u,i} = 1$, getting the normalized weight vector $\boldsymbol{\alpha}_{u,i} = [\alpha^0_{u,i}, \alpha^1_{u,i}, \ldots, \alpha^{|\mathcal{S}|-1}_{u,i}]$. We leverage the attention values to produce a unique and weighted stylistic representation for item i, conditioned on user u:

$$\mathbf{w}_i = \sum_{s\in\mathcal{S}} \alpha^s_{u,i} \mathbf{e}^s_i \quad (3)$$

Finally, let $oan(\cdot)$ be the function to aggregate the latent factor $\mathbf{q}_i$ and the output of the attention network $\mathbf{w}_i$ into a unique representation for item i, e.g., through addition. We calculate the final item representation $\mathbf{q}'_i$ as:

$$\mathbf{q}'_i = oan(\mathbf{q}_i, \mathbf{w}_i) = \mathbf{q}_i + \mathbf{w}_i \quad (4)$$

Neural Inference. To capture non-linearities in user/item interactions, we adopt an MLP to run the prediction. Let $concat(\cdot)$ be the concatenation function and $out(\cdot)$ be a trainable MLP, we predict rating $\hat{r}_{u,i}$ for user u and item i as:

$$\hat{r}_{u,i} = out(concat(\mathbf{p}_u, \mathbf{q}'_i)) \quad (5)$$

Objective Function and Training. We use Bayesian personalized ranking (BPR) [32]. Given a set of triples $\mathcal{T}$ (user u, positive item p, negative item n), we seek to optimize the following objective function:

$$\arg\min_{\boldsymbol{\Theta}} \sum_{(u,p,n)\in\mathcal{T}} -\ln(sigmoid(\hat{r}_{u,p} - \hat{r}_{u,n})) + \lambda||\Theta||^2 \quad (6)$$

where $\boldsymbol{\Theta}$ and λ are the set of trainable weights and the regularization term, respectively. We build $\mathcal{T}$ from the training set by picking, for each randomly sampled (u, p) pair, a negative item n for u (i.e., not-interacted by u). Moreover, we adopt mini-batch Adam [24] as optimizing algorithm.

3 Experiments

Datasets. We use two popular categories from the Amazon dataset [17,28], i.e., Boys & Girls and Men. After having downloaded the available item images, we filter out the items and the users with less than 5 interactions [17, 18]. Boys & Girls counts 1,425 users, 5,019 items, and 9,213 interactions (sparsity is 0.00129), while Men counts 16,278 users, 31,750 items, and 113,106 interactions (sparsity is 0.00022). In both cases, we have, on average, > 6 interactions per user.

Feature Extraction and Encoding. Since we address a fashion recommendation task, we extract color, shape/texture, and fashion category from item images [34, 41]. Unlike previous works, we leverage such features because they are easy to extract and always accessible and represent the content of item images at a stylistic level. We extract the **color** information through the 8-bin RGB color histogram, the **shape/texture** as done in [34], and the **fashion category** from a pretrained ResNet50 [6, 11, 15, 37], where "category" refers to the classification task on which the CNN is pretrained. As for the features encoding, we use a trainable MLP and CNN for color (a vector) and shape (an image), respectively. Conversely, following [30], we adopt PCA to compress the fashion category feature, also to level it out to the color and shape features that do not benefit from a pretrained feature extractor.

Baselines. We compare our approach with pure collaborative and visual-based approaches, i.e., BPRMF [32] and NeuMF [19] for the former, and VBPR [18], DeepStyle [26], DVBPR [23], ACF [7], and VNPR [30] for the latter.

Evaluation and Reproducibility. We put, for each user, the last interaction into the test set and the second-to-last into the validation one (i.e., temporal leave-one-out). Then, we measure the model accuracy with the hit ratio ($HR@k$, the validation metric) and the normalized discounted cumulative gain ($nDCG@k$) as performed in related works [7, 19, 39]. We also measure the fraction of items covered in the catalog ($iCov@k$), the expected free discovery ($EFD@k$) [35], and the diversity with the 1's complement of the Gini index ($Gini@k$) [16]. For the implementation, we used the framework Elliot [1, 2].

3.1 Results

What are the accuracy and beyond-accuracy recommendation performance? Table 1 reports the accuracy and beyond-accuracy metrics on top-20 recommendation lists. On Amazon Boys & Girls, our solution and DeepStyle are the best and second-best models on accuracy and beyond-accuracy measures, respectively (e.g., 0.03860 vs. 0.03719 for the HR). In addition, our approach outperforms all the other baselines on novelty and diversity, covering a broader fraction of the catalog (e.g., $iCov \simeq 90\%$). As for Amazon Men, the proposed approach is still consistently the most accurate model, even beating BPRMF, whose accuracy performance is superior to all other visual baselines. Considering that BPRMF covers only the 0.6% of the item catalog, it follows that its superior performance on accuracy comes from recommending the most popular items [5, 27, 40]. Given that, we maintain the competitiveness of our solution, being the best on the accuracy, but also covering about 29% of the item catalog and supporting the discovery of new products (e.g., $EFD = 0.01242$ is the second to best value). That is, the proposed method shows a competitive performance trade-off on accuracy and beyond-accuracy metrics.

How performance is affected by different configurations of attention, *ian*, and *oan*? Following [8, 21], we feed the attention network by exploring three aggregations for the inputs of the attention network (ian), i.e., element-wise multiplication/addition and concatenation, and two aggregations for the

Table 1: Accuracy and beyond-accuracy metrics on top-20 recommendation lists.

Model	*HR*	*nDCG*	*iCov*	*EFD*	*Gini*
Amazon Boys & Girls — `configuration file`					
BPRMF	.01474	.00508	.68181	.00719	.28245
NeuMF	.02386	.00999	.00638	.01206	.00406
VBPR	.03018	.01287	.71030	.02049	.30532
DeepStyle	.03719	.01543	.85017	.02624	.44770
DVBPR	.00491	.00211	.00438	.00341	.00379
ACF	.01544	.00482	.70731	.00754	.40978
VNPR	.01053	.00429	.51584	.00739	.13664
Ours	**.03860**	**.01610**	**.89878**	**.02747**	**.49747**
Amazon Men — `configuration file`					
BPRMF	.01947	.00713	.00605	.00982	.00982
NeuMF	.01333	.00444	.00076	.00633	.00060
VBPR	.01554	.00588	.59351	.01042	.17935
DeepStyle	.01634	.00654	**.84397**	**.01245**	**.33314**
DVBPR	.00123	.00036	.00088	.00069	.00065
ACF	.01548	.00729	.19380	.01147	.02956
VNPR	.00528	.00203	.59443	.00429	.16139
Ours	**.02021**	**.00750**	.28995	.01242	.06451

Table 2: Ablation study on different configurations of attention, *ian*, and *oan*.

Components		Boys & Girls		Men	
ian$(\cdot)$	***oan***$(\cdot)$	*HR*	*iCov*	*HR*	*iCov*
No Attention		.01263	.01136	.01462	.02208
Add	Add	.02316	.00757	.02083	.00076
Add	Mult	.02246	.00458	.00768	.00079
Concat	Add	.01404	.00518	**.02113**	.00076
Concat	Mult	.02456	.00458	.00891	.00085
Mult	*Add*	**.03860**	**.89878**	.02021	**.28995**
Mult	Mult	.02807	.00478	.01370	.01647

output of the attention network (*oan*), i.e., element-wise addition/multiplication. Table 2 reports the *HR*, i.e., the validation metric, and the *iCov*, i.e., a beyond-accuracy metric. No ablation study is run on the content-style features, as their relative influence on recommendation is learned during the training. First, we observe that attention mechanisms, i.e., all rows but *No Attention*, lead to better-tailored recommendations. Second, despite the {Concat, Add} choice reaches the highest accuracy on Men, the {Mult, Add} combination we used in this work is the most competitive on both accuracy and beyond-accuracy metrics.

4 Conclusion and Future Work

Unlike previous works, we argue that in visual recommendation scenarios (e.g., fashion), items should be represented by easy-to-extract and always accessible visual characteristics, aiming to describe their content from a stylistic perspective (e.g., color and shape). In this work, we disentangled these features via attention to assign users' personalized importance weights to each content-style feature. Results confirmed that our solution could reach a competitive accuracy and beyond-accuracy trade-off against other baselines, and an ablation study justified the adopted architectural choices. We plan to extend the content-style features for other visual recommendation domains, such as food and social media. Another area where item content visual features can be beneficial is in improving accessibility to extremely long-tail items (distant tails), for which traditional CF or hybrid approaches are not helpful due to the scarcity of interaction data.

Acknowledgment. The authors acknowledge partial support of the projects: CTE Matera, ERP4.0, SECURE SAFE APULIA, Servizi Locali 2.0.

References

1. Anelli, V.W., Bellogín, A., Ferrara, A., Malitesta, D., Merra, F.A., Pomo, C., Donini, F.M., Noia, T.D.: Elliot: A comprehensive and rigorous framework for reproducible recommender systems evaluation. In: SIGIR. pp. 2405–2414. ACM (2021)
2. Anelli, V.W., Bellogín, A., Ferrara, A., Malitesta, D., Merra, F.A., Pomo, C., Donini, F.M., Noia, T.D.: V-elliot: Design, evaluate and tune visual recommender systems. In: RecSys. pp. 768–771. ACM (2021)
3. Anelli, V.W., Deldjoo, Y., Noia, T.D., Malitesta, D., Merra, F.A.: A study of defensive methods to protect visual recommendation against adversarial manipulation of images. In: SIGIR. pp. 1094–1103. ACM (2021)
4. Anelli, V.W., Noia, T.D., Sciascio, E.D., Ferrara, A., Mancino, A.C.M.: Sparse feature factorization for recommender systems with knowledge graphs. In: RecSys. pp. 154–165. ACM (2021)
5. Boratto, L., Fenu, G., Marras, M.: Connecting user and item perspectives in popularity debiasing for collaborative recommendation. Inf. Process. Manag. **58**(1), 102387 (2021)
6. Chen, J., Ngo, C., Feng, F., Chua, T.: Deep understanding of cooking procedure for cross-modal recipe retrieval. In: ACM Multimedia. pp. 1020–1028. ACM (2018)
7. Chen, J., Zhang, H., He, X., Nie, L., Liu, W., Chua, T.: Attentive collaborative filtering: Multimedia recommendation with item- and component-level attention. In: SIGIR. pp. 335–344. ACM (2017)
8. Chen, X., Chen, H., Xu, H., Zhang, Y., Cao, Y., Qin, Z., Zha, H.: Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation. In: SIGIR. pp. 765–774. ACM (2019)
9. Cheng, Z., Chang, X., Zhu, L., Kanjirathinkal, R.C., Kankanhalli, M.S.: MMALFM: explainable recommendation by leveraging reviews and images. ACM Trans. Inf. Syst. **37**(2), 16:1–16:28 (2019)
10. Chong, X., Li, Q., Leung, H., Men, Q., Chao, X.: Hierarchical visual-aware minimax ranking based on co-purchase data for personalized recommendation. In: WWW. pp. 2563–2569. ACM / IW3C2 (2020)
11. Deldjoo, Y., Noia, T.D., Malitesta, D., Merra, F.A.: A study on the relative importance of convolutional neural networks in visually-aware recommender systems. In: CVPR Workshops. pp. 3961–3967. Computer Vision Foundation / IEEE (2021)
12. Deldjoo, Y., Schedl, M., Cremonesi, P., Pasi, G.: Recommender systems leveraging multimedia content. ACM Computing Surveys (CSUR) **53**(5), 1–38 (2020)
13. Deldjoo, Y., Schedl, M., Hidasi, B., He, X., Wei, Y.: Multimedia recommender systems: Algorithms and challenges. In: Recommender Systems Handbook. Springer US (2022)
14. Elsweiler, D., Trattner, C., Harvey, M.: Exploiting food choice biases for healthier recipe recommendation. In: SIGIR. pp. 575–584. ACM (2017)
15. Gao, X., Feng, F., He, X., Huang, H., Guan, X., Feng, C., Ming, Z., Chua, T.: Hierarchical attention network for visually-aware food recommendation. IEEE Trans. Multim. **22**(6) (2020)
16. Gunawardana, A., Shani, G.: Evaluating recommender systems. In: Recommender Systems Handbook, pp. 265–308. Springer (2015)
17. He, R., McAuley, J.J.: Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In: WWW. pp. 507–517. ACM (2016)

18. He, R., McAuley, J.J.: VBPR: visual bayesian personalized ranking from implicit feedback. In: AAAI. pp. 144–150. AAAI Press (2016)
19. He, X., Liao, L., Zhang, H., Nie, L., Hu, X., Chua, T.: Neural collaborative filtering. In: WWW. pp. 173–182. ACM (2017)
20. Hinton, G.E., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. CoRR **abs/1503.02531** (2015)
21. Hou, M., Wu, L., Chen, E., Li, Z., Zheng, V.W., Liu, Q.: Explainable fashion recommendation: A semantic attribute region guided approach. In: IJCAI. pp. 4681–4688. ijcai.org (2019)
22. Hu, Y., Yi, X., Davis, L.S.: Collaborative fashion recommendation: A functional tensor factorization approach. In: ACM Multimedia. pp. 129–138. ACM (2015)
23. Kang, W., Fang, C., Wang, Z., McAuley, J.J.: Visually-aware fashion recommendation and design with generative image models. In: ICDM. pp. 207–216. IEEE Computer Society (2017)
24. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (Poster) (2015)
25. Koren, Y., Bell, R.M., Volinsky, C.: Matrix factorization techniques for recommender systems. Computer **42**(8), 30–37 (2009)
26. Liu, Q., Wu, S., Wang, L.: Deepstyle: Learning user preferences for visual recommendation. In: SIGIR. pp. 841–844. ACM (2017)
27. Mansoury, M., Abdollahpouri, H., Pechenizkiy, M., Mobasher, B., Burke, R.: Feedback loop and bias amplification in recommender systems. In: CIKM. pp. 2145–2148. ACM (2020)
28. McAuley, J.J., Targett, C., Shi, Q., van den Hengel, A.: Image-based recommendations on styles and substitutes. In: SIGIR. ACM (2015)
29. Meng, L., Feng, F., He, X., Gao, X., Chua, T.: Heterogeneous fusion of semantic and collaborative information for visually-aware food recommendation. In: ACM Multimedia. pp. 3460–3468. ACM (2020)
30. Niu, W., Caverlee, J., Lu, H.: Neural personalized ranking for image recommendation. In: WSDM. pp. 423–431. ACM (2018)
31. Packer, C., McAuley, J.J., Ramisa, A.: Visually-aware personalized recommendation using interpretable image representations. CoRR **abs/1806.09820** (2018)
32. Rendle, S., Freudenthaler, C., Gantner, Z., Schmidt-Thieme, L.: BPR: bayesian personalized ranking from implicit feedback. In: UAI. pp. 452–461. AUAI Press (2009)
33. Sertkan, M., Neidhardt, J., Werthner, H.: Pictoure - A picture-based tourism recommender. In: RecSys. pp. 597–599. ACM (2020)
34. Tangseng, P., Okatani, T.: Toward explainable fashion recommendation. In: WACV. pp. 2142–2151. IEEE (2020)
35. Vargas, S.: Novelty and diversity enhancement and evaluation in recommender systems and information retrieval. In: SIGIR. p. 1281. ACM (2014)
36. Wu, Q., Zhao, P., Cui, Z.: Visual and textual jointly enhanced interpretable fashion recommendation. IEEE Access (2020)
37. Yang, X., He, X., Wang, X., Ma, Y., Feng, F., Wang, M., Chua, T.: Interpretable fashion matching with rich attributes. In: SIGIR. pp. 775–784. ACM (2019)
38. Yin, R., Li, K., Lu, J., Zhang, G.: Enhancing fashion recommendation with visual compatibility relationship. In: WWW. pp. 3434–3440. ACM (2019)
39. Zhang, Y., Zhu, Z., He, Y., Caverlee, J.: Content-collaborative disentanglement representation learning for enhanced recommendation. In: RecSys. pp. 43–52. ACM (2020)

40. Zhu, Z., Wang, J., Caverlee, J.: Measuring and mitigating item under-recommendation bias in personalized ranking systems. In: SIGIR. pp. 449–458. ACM (2020)
41. Zou, Q., Zhang, Z., Wang, Q., Li, Q., Chen, L., Wang, S.: Who leads the clothing fashion: Style, color, or texture? A computational study. CoRR **abs/1608.07444** (2016)