Journal of Design Intelligence
https://www.ffspub.com/index.php/jdi
ISSN 3104-9001 (print)
E-mail: jdi.office@ffspub.com

Article

Imitation Learning for Fashion Style Based on Hierarchical Multimodal Representation

Shanglin Yang[1,†], Zhan Shi[2], Shizhu Liu[1], Hui Zhou [1]
[1] *JD.com American Technologies, Mountain View, CA 94043, USA*
[2] *Department of Computer Science and Engineering , Santa Clara University, Santa Clara, CA 95053, USA*
[†] *E-mail: shanglin.yang@jd.com*

Received: December 12, 2025 / Revised: January 14, 2026 / Accepted: January 20, 2026 / Published online: January 28, 2026

Abstract: Fashion is a complex social phenomenon. People follow fashion styles from demonstrations by experts or fashion icons. However, for machine agent, learning to imitate fashion experts from demonstrations can be challenging, especially for complex styles in environments with high-dimensional, multimodal observations. Most existing research regarding fashion outfit composition utilizes supervised learning methods to mimic the behaviors of style icons. These methods suffer from distribution shift: because the agent greedily imitates some given outfit demonstrations, it can drift away from one style to another styles given subtle differences. In this work, we propose an adversarial inverse reinforcement learning formulation to recover reward functions based on hierarchical multimodal representation (HM-AIRL) during the imitation process. The hierarchical joint representation can more comprehensively model the expert composited outfit demonstrations to recover the reward function. We demonstrate that the proposed HM-AIRL model is able to recover reward functions that are robust to changes in multimodal observations, enabling us to learn policies under significant variation between different styles.

Keywords: Fashion Recommendation; Multimodal Representation Learning; Style Modeling; Information Retrieval
https://doi.org/10.64509/jdi.11.62

1 Introduction

Fashion plays important and sophisticated roles in various aspects: social, culture, identity and etc. We can roughly understand a fashion is a type of reaction common to a considerable number of people [1]. It is common that fashion styles suggested by fashion experts, fashion icons or more popularly now by Key Opinion Leaders (KOLs) from the social media are imitated by people.

On internet, a set of outfits with description is a common way to demonstrate a fashion style (Figure 1). People imitate the fast changing fashion styles from these demonstrations without interaction with the experts. However, learning fashion styles from demonstrated outfits by machine is challenging, as it requires how low-level fashion elements can be mapped to high-level styles.

Fashion styles are composed of important design elements such as color, pattern, material, silhouette, and trim [2]. Each outfit item may demonstrate some design elements with corresponding images and attribute information at low level. High level information for style may be available too: relationship between these items and sometime related cultural meaning behind the suggested style. All the items in an outfit should be consistent with the described style, and be compatible with each other as well.

Consider the outfit information as input data, learning-based method can be used if we have a good data set with labels of each item and corresponding style. Most existing outfit composition methods follow this supervised learning approach and concentrate on predicting the compatibility between fashion items. However, compatibility is only part of fashion style knowledge. Compatible items cannot guarantee the consistency with the condition style definition. Besides supervised learning methods based on behavioral cloning (BC) suffer from distribution shift: because the agent greedily imitates demonstrated actions, it can drift away from one style to another styles due to subtle differences.

In this paper, we address this fundamental question from two aspects. First, we design a hierarchical multimodal representation to describe complex latent information of the whole outfit structure. Based on this representation, we further propose to use an inverse reinforcement learning method

[†] Corresponding author: Shanglin Yang
[*] Academic Editor: Li Li

to infer experts' composition reward function and learn the composition value function simultaneously.

Corresponding to the three parts of the outfit demonstration: image, attributes and style description, our outfit encoder network consist of three parts to learn the rich information conveyed in a outfit. At low level, a shared multimodal variational autoencoder is employed to learn the jointly representation from image and attribute information for each item. At high level, we try to cover the whole outfit style by applying two steps: matching strategy is applied to extract the relations between pairs of items once the item vectors are generated [3]; and a pre-trained BERT model [4] is used to encode explanation text as the condition of the whole outfit.

Adversarial inverse reinforcement learning (AIRL) [5] is used in our algorithm to provide the agent the capacity for simultaneous learning of the reward function and value function, which enables us to both make use of the efficient adversarial formulation and recover a generalizable and robust reward function for both item compatibility and style consistency.

Our key contributions can be summarized as follows:

- We propose a learning-based framework for effective fashion imitation. To our best knowledge, our method is the first to address the imitation behavior in the fashion style learning.
- We design a hierarchical multimodal neural network that can effectively encode the rich latent information of whole outfit: at low level, it can capture complementary information from multimodal observations; and interleaving factors are covered at high level.
- We propose an adversarial inverse reinforcement learning method for recovering the style reward function, which can learn more robust reward to avoid the style drift.

For the rest of the paper, we fist discuss the related work in Section 2. We then describe the hierarchical multi-modality fusion model for outfit in Section 3. Next, details of the proposed adversarial training method is introduced. We show our experimental results in Section 5 and give our discussion and future work at the end.

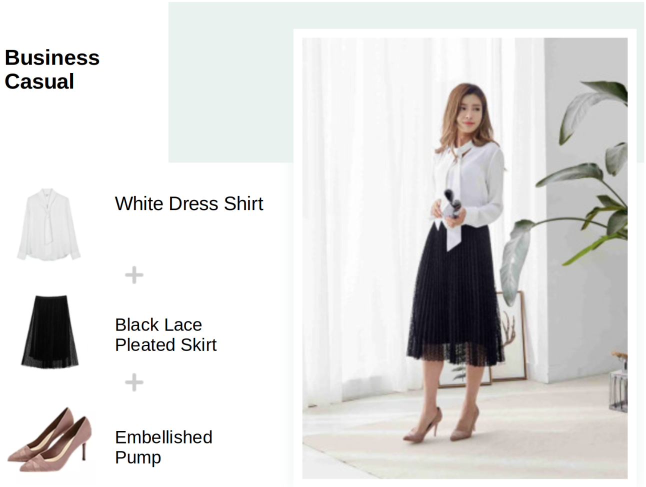

Figure 1: An Outfit Example of ”Business Casual” Style.

2 Related Work

Computer vision and recommendation techniques have important and rich applications in fashion domain. The majority of research in this domain focus on fashion image attribute recognition and retrieval, fashion item semantic embedded learning, fashion recommendation, visual compatibility learning and outfit composition.

Fashion Attribute Recognition and Retrieval. Clothing attributes provide a useful tool to assess clothing products as mid-level semantic visual concepts. Recently, more deep networks such as DeepFashion [6] and MTCT [7], were proven to be efficient in attribute recognition on large datasets. Nanoto et al. [8] proposed a multi-label joint learning network to predict cloth attributes from images with minimum human supervision. Several works utilized weakly labeled image-text pairs to discover attributes [9, 10]. Besides, images retrieval and products recommendation tasks can benefit from attributes results. Chen et al. [11] focused on solving the problem of describing people based on fine-grained clothing attributes. AMNet [12] conducted a fashion search after changing an attribute. Mix and match [13] combined a deep network with conditional random fields to explore the compatibility of clothing items and attributes. Wei et al. [14] considered both global and localized attributes as 'words' to describe cloth. Many works above utilized attributes for image retrieval and recommendation tasks, but the attributes can not address the high level information of the whole outfit. In our work, we combine multi-modal information as a global feature. The experiments reveal our feature could also be used for missing attributes inference.

Fashion Visual-semantic Embedding Learning. Fashion item representation learning is the fundamental step of all downstream inference work. There are a lot of works trying to investigate this important problem with different network structure and learning methods. The most common approaches tend to be trained as siamese network [15] or using triplet loss [16], This is extended in Han et al. [17] by feeding the visual representation from each garment within an outfit into an LSTM in order to jointly reason about the outfit as a whole. Simo-Serra et al.[18] trained a classifier network with ranking as the constraint to extract distinctive feature representation and also used high level feature of the classifier network as embedding of fashion style. Many unsupervised representation learning methods also were proposed to learn latent feature from unlabeled data directly. Most of them utilize Variational Auto-Encoder (VAE) [19] and Predictability Minimization (PM) model [20] to learn fashion item embeddings. The encoded embeddings usually contain mixed and unexplainable features of original images. There are several approaches which implemented multi-modal embedding methods to reveal novel feature structures (e.g. [14, 21, 22]). These multimodal methods only to map the image text into the same space without considering the deep correlation between the different modals. In this paper, we try to learn the distributed representation for specific fashion styles. We assume that the complementary representation for fashion style should cover both compatibility between items and common feature shared by whole outfit. To reach this goal, our method learn the joint representation for fashion items from image and attributes information at low level, and capture the relationship between items and condition style.

Fashion Recommendation and Outfit Composition. There are a few approaches for fashion items recommendation. In the context of fashion analysis, visual compatibility measures whether clothing items complement one another across visual categories. For example, “sweat pants” are more compatible with “running shoes” than “high-heeled shoes”. Most existing fashion related research work mainly focus on the compatibility between fashion items, and mainly learning on images data only. Iwata et al. [23] proposed a topic model to recommend ”Tops” for ”Bottoms”. The goal of this work is to compose fashion outfit automatically by building product coordinates from visual features in each fashion item region. Veit et al. [24] built a Siamese Convolutional Neural Network (CNN) architecture to learn clothing matching pair products from the Amazon co-purchase dataset, focusing on the representative problem of learning compatible clothing style. Simo-Serra1 et al. [25] introduced a Conditional Random Field (CRF) to learn the different outfits formula and types of people. The model is further being used to predict how fashionable a person looks on a particular photograph. Li et al. [26] used multi-modal embeddings as features and the quality scores as the label to train a grading model. Xuemeng et al. [27] propose to model the compatibility between fashion items based on Bayesian personalized ranking (BPR). Han et al. [17] jointly learn compatibility relationships among fashion items and employ a Bi-LSTM model to learn the compatibility relationships among fashion items by modeling an outfit as a sequence. [28] used style topic models for compatibility and defined the recommendation as a subset selection problem. Chen et al. [29] proposed an encoder-decoder model to generate personalized fashion outfits. Most of works above utilized supervised learning methods to predict the compatibility between fashion items and validated model performance on validation dataset generated with negative sampling techniques. Thus, these methods are inclined to learn the general patterns rather than capture the subtle difference between fashion styles. To address this problem, our method implemented adversarial learning methods to learns more robust composition policy.

Imitation Learning. Imitation learning techniques aim to mimic human behavior in a given task. An agent, i.e., a learning machine, by learning a mapping between observations and actions, is trained to perform a task from demonstrations [30]. Inverse reinforcement learning (IRL) is a form of imitation learning that accomplishes this by first inferring the expert’s reward function and then training a policy to maximize it [5, 31–35]. Most of imitation learning works are mainly focus on the imitation process in the fields of robotics, adaptive planning, and data-driven animation. In this work, we formalize the outfit composition task as a Markov Decision Process(MDP) and work under the maximum ambiguity causal IRL framework of [36], which allow us to cast the reward learning problem as a maximum likelihood problem. Our IRL algorithm is built upon the adversarial IRL architecture proposed in [37] and [5]. A discriminator is trained to distinguish experts’ selection, while the agent is trained to ”fool” the discriminator into thinking itself is the expert. To our knowledge, this is the first approach that considers the outfit composition as a style imitation learning problem in the fashion domain.

3 Hierarchical Multimodal Representation for Fashion Outfit

3.1 Fusion Representation for Fashion Item

For one fashion item, it is obvious that the corresponding image and attribute tags have the complementary information. For instance, people can easily tell the color and pattern of a garment from image, but need to check the attribute tags to know the garment material and the specific functional usage. Learning from diverse modalities with generative approaches has the potential to yield more generalized joint representations. Inspired by the product-of-experts(PoE) inference network [38], assuming the conditional independence among the modalities and joint posterior from multiple modalities is a product of individual posteriors, we utilize a multimodal variational autoencoder (MVAE) [39] to learn a joint distribution from both image and attribute tags as shown in Figure 2.

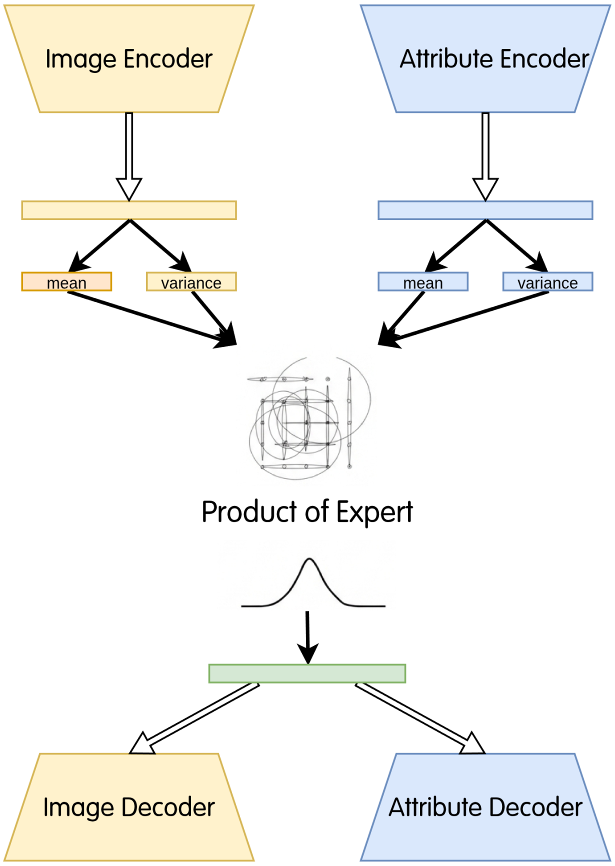

Figure 2: Multimodal Variational Autoencoder Learning on Image and Attributes.

In the multimodal setting, we assume the N modalities, $x_1,...,x_N$, are conditionally independent given the common latent variable, z. Then, we assume that a generative model of the form $p_\theta(x_1,x_2,...,x_N,z) = p(z)p_\theta(x_1|z)p_\theta(x_2|z)\cdots p_\theta(x_N|z)$. If an item is presented by a collection of modalities $X = \{x_i|i^{th}\text{modality present}\}$, then the evidence lower bound(ELBO) becomes:

$$\begin{aligned} ELBO(X) = \mathbb{E}_{q_\phi(z|X)}[\sum_{x_i i \in X} \lambda_i log p_\theta(x_i|z)] \\ -\beta KL[q_\phi(z|X), p(z)] \end{aligned} \quad (1)$$

MVAE can be trained by simply optimizing the evidence lower bound given in Equation (1). The product and quotient distributions are not in general solvable in closed form. However, when $p(z)$ and $\tilde{q}(z|x_i)$ are Gaussian there is a simple analytical solution [40]: a product of Gaussian experts is itself Gaussian with mean $\mu = (\sum_i \mu_i T_i)(\sum_i T_i)^{-1})$ and covariance $V_i = (\sum_i T_i)^{-1}$, where μ_i, V_i are the parameters of the i-th Gaussian expert, and $T_i = V_i^{-1}$ is the inverse of the covariance.

3.2 Language-Conditioned Hierarchical Representation for Outfit Style

Comparing with standard compatibility prediction task, the fashion imitation learning is to learn a reward function that generalizes across different types of styles. While standard supervised methods are typically trained and evaluated without considering style information, we want our language-conditioned reward function to produce correct behavior when generating outfit with given style conditions.

Notice that the description text of the outfits usually explains the common salient features such as occasion, season, trending information with which all the items consistent. We encode the explanation language as the condition of the whole outfit structure. As shown in Figure 3, for any pairs of items in an outfit, a pretrained BERT model [4] is used to encode description text t, and the MVAE encoder convert two items' image and attributes information to an joint representation u, v. The interaction between items is represented as $(|u-v|, u*v)$. The outfit embedding is the concatenation of these two parts, and the outfit style rewards will be learned based on it.

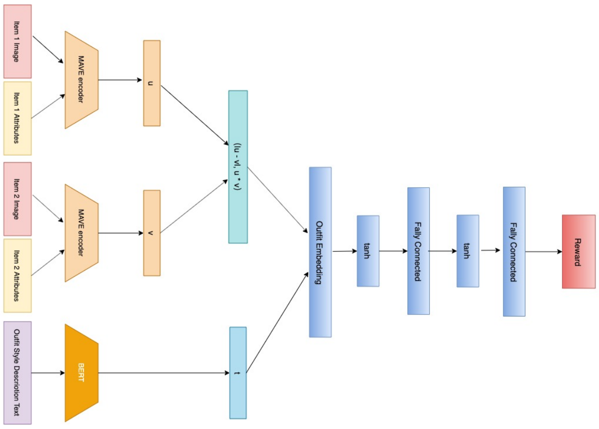

Figure 3: Language-conditioned Hierarchical Multimodal Network for Style Consistency Reward Learning.

4 Adversarial Inverse Reinforcement learning for Style Reward Learning

For the outfit composition task, we formalize the process as follow. Let I denotes the set of all fashion items, O_i denote an outfit, and $x_{i,j} \in I$ denote the items in the outfit O_i, so that $O_i = \{x_{i,1}, x_{i,2}, ..., x_{i,|S_i|}\}$. Each item $x_{i,j}$ belong to a limited number of categories $\{C_1, C_2, ..., C_N\}$. The fashion outfit composition process can be formulated as an iterative item selection process, in which at most one item is selected for each category. For example, a user may want to compose an outfit of "UK Smart Casual Style". Then, he/she needs to select one item from four categories: "Shirt", "Jacket", "Pant" and "Shoes" respectively.

This process can be described with the the maximum causal entropy IRL framework [36], which considers an entropy-regularized Markov decision process (MDP), defined by the tuple $(S, A, T, r, \gamma, \rho_0)$. S, A are the state and action spaces, respectively, $\gamma \in (0,1)$ is the discount factor, the dynamics and transition distribution $T(s'|a,s)$, the initial state distribution $\rho_0(s)$, and the reward function $r(s,a)$ is unknown in the standard reinforcement learning setup and can only be queried through interaction with the MDP. Specifically, S can be presented with the selected fashion items and A refer to the item selection actions during the process. The reward function $r(s,a)$ indicate the compatibility and style consistency between fashion items and the described style.

Because the reward function $r(s,a)$ is unknown, we assume the experts' demonstration outfits are composited with an optimal policy $\pi^*(a|s)$. Inverse reinforcement learning instead seeks inferring the reward function $r(s,a)$ given a set of demonstrations $D = \{\tau_1, ..., \tau_N\}$. Moreover, the dynamics of composition process is known. Instead of using full trajectories, we could focus on the single state and action case. The entire training procedure is detailed in Algorithm 1. During the training process, our algorithm alternate between training a discriminator to classify the expert selection from outfit generated by current policy, and update the policy to confuse the discriminator [5, 37, 41]. The discriminator is trained with the form:

$$D_{\theta,\phi}(s,a,s') = \frac{exp\{f_{\theta,\phi}(s,a,s')\}}{exp\{f_{\theta,\phi}(s,a,s')\} + \pi(a|s)'} \quad (2)$$

where $f_{\theta,\phi}$ is restricted to a reward approximator g_θ and a shaping term h_ϕ as

$$f_{\theta,\phi}(s,a,s') = g_\theta(s,a) + \gamma h_\phi(s') - h_\phi(s) \quad (3)$$

Suppose we are given an expert policy π_E that we wish to rationalize with IRL. r^* is the true reward function. The f^* is the advantage function need to be recoverd. h recovers the optimal value function V^*, which servers as the reward shaping term:

$$f^*(s,a,s') = r^*(s) + \gamma V^*(s') - V^*(s) = A^*(s,a) \quad (4)$$

Algorithm 1 Language-Conditioned Style Reward learning

```
Obtain expert outfit demonstrations $\tau_E$
Initialize policy $\pi$ and discriminator $D_{\theta,\phi}$
for style $t_k \in \{t_1, t_s, ...t_N\}$ do
    for $iteration = 1, 2, \ldots$ do
        Composite outfits $\tau_i = (s_0, a_0, ..., s_T, a_T)$ with given style $t_k$ by executing $\pi$.
        Train $D_{\theta,\phi}$ via binary logistic regression to classify demonstrations $\tau_E$ from generated outfits $\tau_i$
        Update reward $r_{\theta,\phi}(s,a,s') \leftarrow log D_{\theta,\phi}(s,a,s') - log(1 - D_{\theta,\phi}(s,a,s'))$
        Update $\pi$ with respect to $r_{\theta,\phi}$ using policy optimization method.
    end for
end for
```

5 Experimental Results

In this section, we introduce the data collection we used for style imitation learning, the evaluation of the composition agent, and some further analysis.

5.1 Dataset

Different from fashion datasets collected from Polyvore [17, 42] or Lookastic [43] that are suitable for data mining, the dataset with more complete fashion style information works better for fashion imitation. Namely, we like the dataset has an explanation text to describe the style of each outfit, and an optional list of every demonstrated item for the given style. This is natural for the human being imitation and many fashion e-commerce websites use this way to demonstrate fashion items on their platforms. We specifically found a website named Chuanda on wqs.jd.com that is suitable for our need. On Chuanda website, all the outfits are curated by fashion experts and every item is mapped to an identical product for sale.

We collect a fashion style dataset from Chuanda containing 3,557 outfits covering 67 basic fashion styles. In this dataset, each outfit is composed of up to three items, and a short description about the outfit style. Moreover, for each given outfit, there is averagely around 21.04 other candidate items suggested by fashion experts. Every item in this dataset is mapped to an identical sale item on JD.com website. This is convenient and useful as we can collect the product name, images, and attributes for the corresponding fashion item. In another word, in the dataset all these fashion items are labeled with 1,879 distinct fashion related attributes that belong to 5 types: Gender, Season, Style, Material, and Function. More statistics of this dataset is shown in Table 1.

Table 1: Chuanda fashion style dataset statistics

Outfits	Items	Attributes	Basic styles	Avg Opts
3557	18627	1879	67	21.04
Top	**Bottom**	**Skirt**	**Shoes**	**Coat**
6486	5134	2082	2237	1731

5.2 Evaluation

We perform two different evaluations for our proposed learning framework. First, we evaluate the effectiveness of the learned multimodal representation by predicting missing attributes of the given fashion item. Second, we measure style consistency by computing the similarity of composited outfits with the recommendation list provided by fashion experts.

Missing Attribute Imputation. Fashion item representation is critical for the downstream style imitation learning.To verify if our method can learn more complementary information, we conduct the missing attribute imputation task to evaluate the effectiveness of the MVAE. On Chuanda dataset, we simulate incomplete supervision by randomly reserving a fraction of the dataset as multi-modal examples. We examine the effect of supervision on the attribute prediction task $p(x_2|x_1)$, e.g. predict the correct attribute label x_2 from an image x_1. For the MVAE, the total number of examples shown to the model is always fixed, only the proportion of complete bi-modal example is varied. Five important types of attributes (Gender, Season, Style, Material, and Function) are masked first in the input and then predicted with MVAE decoder, the evaluation results are provided in Table 2. We also perform a qualitative analysis of the items representation generated from MVAE and visualize the features space in Figure 4 using t-SNE [44]. Our representation display robustness to background variance and items share similar style and visual appearance can be clustered together. For example, items of ”casual white tops” and ”business style pants” are very close in the space regardless of the background noise.

Table 2: Missing Attribute Prediction

Attribute Type	#	Accuracy
Gender	4529	89.26%
Season	6928	64.3%
Style	8143	57.71%
Material	7081	65.6%
Function	6957	73.7%

Style Consistency. In every Chuanda outfit, each item have a list of optional alternatives, which are also consistent with the given style. For condition style and query top, we adopted the common strategy [45] that feeds selected top item and conditional style description as query, and randomly selected K bottoms as the candidates. The item in the experts' demonstration and optional alternatives are positive candidates. Thus, we can evaluate the effectiveness of the imitation learning by measuring the average position of the consistent item in the ranking list with the mean average precision (MAP) metric. We have totally 1000 unique tops and styles in test set. We compared out method (HM-AIRL) with two methods: feature based pointwise mutual information (PMI) ranking algorithm and Han et al's BiLSTM based method [17]. Pointwise mutual information is s a measure of association and is used for finding collocations and associations between items. In Chuanda dataset, all the items are labeled with 1,879 attributes. We pre-calculate the the PMI scores between any pair of attribute, and rank the candidate items with the sum of attribute PMI scores. For the Bi-LSTM method, we

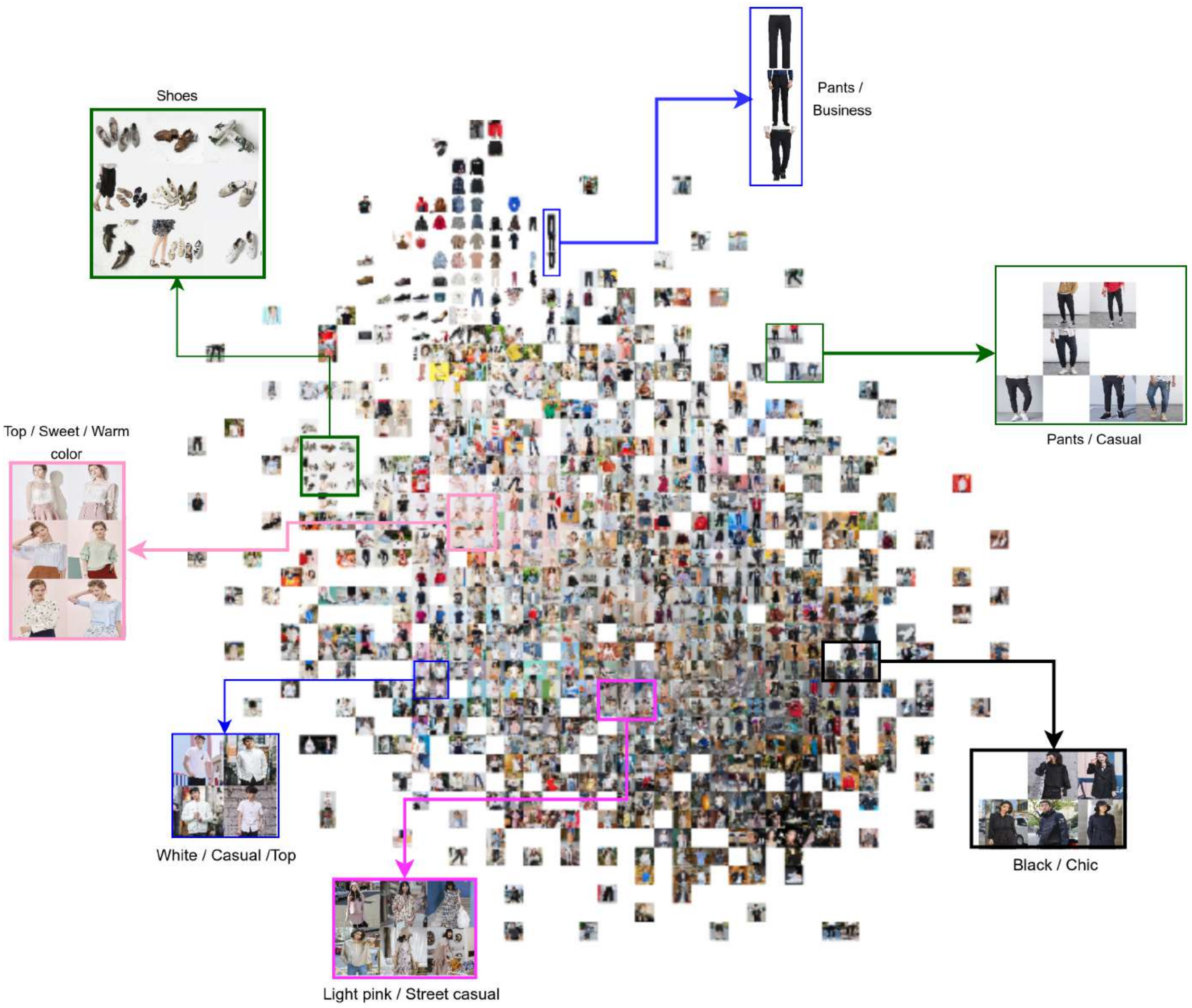

Figure 4: Visualization of the fashion item representation using t-SNE [44]

follow the setting in [17] and retrain the network on Chuanda dataset.

The performance of three methods at different number of top k candidate items is shown in Figure 5. HM-AIRL method get the highest MAP from top 5 to 35 ranking results. We also notice that many items selected by HM-AIRL but not in experts optional list are also compatible with the query top and consistent with the conditional style, which is reasonable in the real application. Bi-LSTM methods get lowest MAP score on this task. By analyzing the ranking list generated by Bi-LSTM model, we think this is mainly caused by three reasons. First, Bi-LSTM failed to consider the style constrain and many selected item are not consistent with condition style. Second, the images in Chuanda dataset is much more noisy. Unlike the clean images on Polyvore website, a lot of images in Chuanda contain irrelevant information such as: price tags, promotion ads etc. Third, Chuanda dataset is a smaller dataset than Polyvore. It is much more challenging to learn complex style concepts on relative small dataset with supervised learning methods.

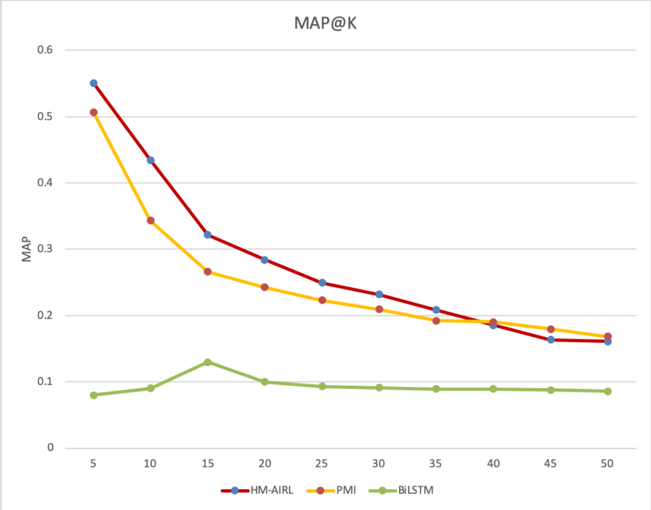

Figure 5: Performance of different methods with respect to MAP at different numbers of top matching items

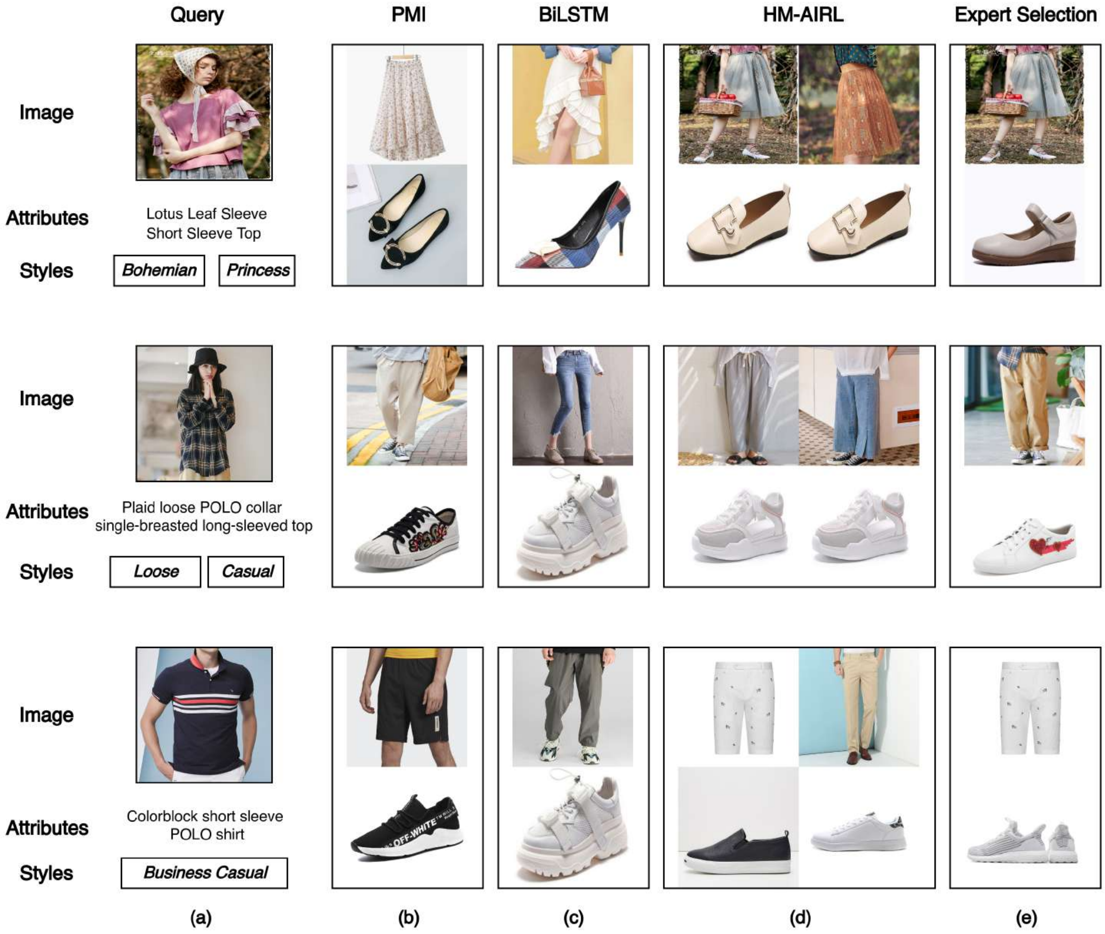

Figure 6: Outfit composition results from different methods with the same condition style and the same querying top item. Three different condition styles and the querying items are listed in column (a). Outfits generated by PMI and BiLSTM methods are listed in column (b) and column (c), respectively. Column (d) shows top 2 outfits by the proposed method. Experts' selections in the demonstration outfits are shown in column (e).

In Figure 6, we demonstrate the outfits generated with attributes pointwise mutual information ranking, Han et al's BiLSTM method, our HM-AIRL and fashion experts' selections for query tops under three distinct condition styles. Compared with experts' selection, only HM-AIRL guarantee both compatibility between items and consistency with condition style description. In the outfit generated by Bi-LSTM and PMI based algorithms, the selected matching items actually belong to other styles.

In the experiment, we use Adam optimizer with a batch size of 256, learning rate of 0.00005 for the HM-ARIL optimization. For the MVAE model used to learn fashion item fusion representation, image encoder and decoder follow the standard DCGAN architecture [46]. Attribute encoder and decoder is a standard 3-layer fully connected network VAE architecture. The two VAEs share the identical latent variables size of 256. The outfit style description encoder uses the base 12-layer BERT model that was pre-trained on Chinese Wikipedia corpus with 21128 unique Chinese characters, and we fix it during the training. The entire framework is implemented with Pytorch [47].

6 Remarks and Future Work

Fashion experts keep proposing novel styles that are appreciated and imitated by individuals sharing the same taste or preference. In this work, we propose a framework to imitate fashion styles from outfit demonstrations. A hierarchical multimodal network is introduced to represent the whole outfit structure. Comparing with other work, our method captures the latent contextual information behind the fashion style by learning both the joint representation from image and attributes for each item and the compatibility and style consistency between items.

Relying on this hierarchical multimodal representation, we train the agent with an inverse reinforcement learning algorithm based on adversarial learning. Our approach builds upon a vast line of work on IRL. Hence, our approach, just like IRL, does not interact with the expert during training and adapt training samples to improve learning efficiency. Our experiment shows that HM-AIRL can learn the value function for imitating fashion styles and is robust to style shift.

In recommendation, user behaviour data such as browsing history is very important. Content-based analysis such as style suggestion is only one factor. For the future, we like to integrate our framework into a full recommendation system and evaluate its performance. Note that the framework presented in this paper is not limited to fashion. Design artifacts in many domains contain latent concepts that can be expressed with sets of human-interpretable features capturing different levels of granularity [48, 49]. This model also offers attractive capabilities: it can infer latent abstract concepts, and imitate experts from their demonstrations. In the future, we like to explore how this framework can empower applications in other domains such as interior design, architecture and etc.

Funding

This research received no funding.

Author Contributions

Conceptualization, Shanglin Yang and Zhan Shi; methodology, Shanglin Yang; software, Shanglin Yang; validation, Zhan Shi; formal analysis, Zhan Shi; investigation, Shanglin Yang and Shizhu Liu; resources, Shanglin Yang; data curation, Shanglin Yang; writing—original draft preparation, Shanglin Yang; writing—review and editing, Zhan Shi and Hui Zhou; visualization, Shanglin Yang; supervision, Shanglin Yang; project administration, Shanglin Yang. All authors have read and agreed to the published version of the manuscript.

Conflict of Interest

All the authors declare that they have no conflict of interest.

Data Avaliable

The dataset was constructed by collecting fashion outfit demonstrations curated by professional stylists from the Chuanda platform. Each outfit is associated with corresponding commercial products. No personal user data or personally identifiable information is involved. The dataset is used solely for academic research and is available upon reasonable request.

References

[1] Bogardus, E.S.: Fashion Imitation in Fundamentals of Social Psychology, 1st edn., pp. 151–167. Century, New York, USA (1924)

[2] Sorger, R., Udale, J.: The fundamentals of fashion design. Bloomsbury Publishing, New York, USA (2006)

[3] Conneau, A., Kiela, D., Schwenk, H., Barrault, L., Bordes, A.: Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In Conference on Empirical Methods on Natural Language Processing (EMNLP), pp. 670–680 (2017). https://doi.org/10.18653/v1/D17-1070

[4] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), pp. 4171–4186 (2019). https://doi.org/10.18653/v1/N19-1423

[5] Fu, J., Luo, K., Levine, S.: Learning Robust Rewards with Adversarial Inverse Reinforcement Learning. In International Conference on Learning Representations (ICLR), pp. 1–15 (2018)

[6] Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: DeepFashion: Powering Robust Clothes Recognition and Retrieval With Rich Annotations. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1096–1104 (2016). https://doi.org/10.1109/CVPR.2016.124

[7] Dong, Q., Gong, S., Zhu, X.: Multi-task Curriculum Transfer Deep Learning of Clothing Attributes. 2017 IEEE Winter Conference on Applications of Computer Vision (WACV) (2017). https://doi.org/10.1109/wacv.2017.64

[8] Inoue, N., Simo-Serra, E., Yamasaki, T., Ishikawa, H.: Multi-label Fashion Image Classification with Minimal Human Supervision. In 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), pp. 2261–2267 (2017). https://doi.org/10.1109/ICCVW.2017.265

[9] Berg, T.L., Berg, A.C., Shih, J.: Automatic Attribute Discovery and Characterization from Noisy Web Data. In 11th European Conference on Computer Vision-ECCV 2010, pp. 663–676 (2010). https://doi.org/10.1007/978-3-642-15549-9_48

[10] Vittayakorn, S., Umeda, T., Murasaki, K., Sudo, K., Okatani, T., Yamaguchi, K.: Automatic Attribute Discovery with Neural Activations. In Computer Vision - ECCV 2016 - 14th European Conference, pp. 252–268 (2016). https://doi.org/10.1007/978-3-319-46493-0_16

[11] Chen, Q., Huang, J., Feris, R., Brown, L.M., Dong, J., Yan, S.: Deep Domain Adaptation for Describing People Based on Fine-Grained Clothing Attributes. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5315–5324 (2015). https://doi.org/10.1109/CVPR.2015.7299169

[12] Zhao, B., Feng, J., Wu, X., Yan, S.: Memory-Augmented Attribute Manipulation Networks for Interactive Fashion Search. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6156–6164 (2017). https://doi.org/10.1109/CVPR.2017.652

[13] Yamaguchi, K., Okatani, T., Sudo, K., Murasaki, K., Taniguchi, Y.: Mix and Match: Joint Model for Clothing and Attribute Recognition. In Proceedings of the British Machine Vision Conference (BMVC), pp. 51–15112

(2015). https://doi.org/10.5244/C.29.51

[14] Hsiao, W.-L., Grauman, K.: Learning the Latent "Look": Unsupervised Discovery of a Style-Coherent Embedding from Fashion Images. In IEEE International Conference on Computer Vision, (ICCV), pp. 4203–4212 (2017). https://doi.org/10.1109/ICCV.2017.451

[15] Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R.: Signature Verification Using a "Siamese" Time Delay Neural Network. In Proceedings of the 6th International Conference on Neural Information Processing Systems, pp. 737–744 (1993)

[16] Schultz, M., Joachims, T.: Learning a Distance Metric from Relative Comparisons. In Proceedings of the 16th International Conference on Neural Information Processing Systems, pp. 41–48 (2003)

[17] Han, X., Wu, Z., Jiang, Y.-G., Davis, L.S.: Learning Fashion Compatibility with Bidirectional LSTMs. In Proceedings of the 25th ACM International Conference on Multimedia, pp. 1078–1086 (2017). https://doi.org/10.1145/3123266.3123394

[18] Simo-Serra, E., Ishikawa, H.: Fashion style in 128 floats: Joint ranking and classification using weak data for feature extraction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 298–307 (2016). https://doi.org/10.1109/CVPR.2016.39

[19] Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013). https://doi.org/10.48550/arXiv.1312.6114

[20] Schmidhuber, J.: Learning Factorial Codes by Predictability Minimization. Neural Computation **4**(6), 863–879 (1992). https://doi.org/10.1162/neco.1992.4.6.863

[21] Al-Halah, Z., Stiefelhagen, R., Grauman, K.: Fashion Forward: Forecasting Visual Style in Fashion. In IEEE International Conference on Computer Vision, ICCV 2017, pp. 388–397 (2017). https://doi.org/10.1109/ICCV.2017.50

[22] Shih, Y.-S., Chang, K.-Y., Lin, H.-T., Sun, M.: Compatibility Family Learning for Item Recommendation and Generation. arXiv preprint arXiv:1712.01262 (2017). https://doi.org/10.48550/arXiv.1712.01262

[23] Iwata, T., Watanabe, S., Sawada, H.: Fashion coordinates recommender system using photographs from fashion magazines. In International Joint Conferences on Artificial Intelligence (IJCAI), pp. 2262–2267 (2011)

[24] Veit, A., Kovacs, B., Bell, S., McAuley, J., Bala, K., Belongie, S.: Learning Visual Clothing Style with Heterogeneous Dyadic Co-Occurrences. In 2015 IEEE International Conference on Computer Vision (ICCV), pp. 4642–4650 (2015). https://doi.org/10.1109/ICCV.2015.527

[25] Simo-Serra, E., Fidler, S., Moreno-Noguer, F., Urtasun, R.: Neuroaesthetics in fashion: Modeling the perception of fashionability. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 869–877 (2015). https://doi.org/10.1109/CVPR.2015.7298688

[26] Li, Y., Cao, L., Zhu, J., Luo, J.: Mining Fashion Outfit Composition Using an End-to-End Deep Learning Approach on Set Data. IEEE Transactions on Multimedia **19**(8), 1946–1955 (2017). https://doi.org/10.1109/TMM.2017.2690144

[27] Song, X., Feng, F., Liu, J., Li, Z., Nie, L., Ma, J.: NeuroStylist: Neural Compatibility Modeling for Clothing Matching. In Proceedings of the 25th ACM International Conference on Multimedia, pp. 753–761 (2017). https://doi.org/10.1145/3123266.3123314

[28] Hsiao, W.-L., Grauman, K.: Creating Capsule Wardrobes from Fashion Images. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7161–7170 (2018). https://doi.org/10.1109/CVPR.2018.00748

[29] Chen, W., Zhao, B., Huang, P., Xu, J., Guo, X., Guo, C., Sun, F., Li, C., Pfadler, A., Zhao, H.: POG: Personalized Outfit Generation for Fashion Recommendation at Alibaba iFashion. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD'19, pp. 2662–2670 (2019). https://doi.org/10.1145/3292500.3330652

[30] Billard, A.G., Calinon, S., Dillmann, R.: Learning from Humans in Springer Handbook of Robotics, pp. 1995–2014. Springer, Cham, Switzerland (2016). https://doi.org/10.1007/978-3-319-32552-1_74

[31] Ziebart, B.D., Maas, A., Bagnell, J.A., Dey, A.K.: Maximum Entropy Inverse Reinforcement Learning. In Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, pp. 1433–1438 (2008)

[32] Finn, C., Levine, S., Abbeel, P.: Guided cost learning: deep inverse optimal control via policy optimization. In Proceedings of the 33rd International Conference on International Conference on Machine Learning, pp. 49–58 (2016)

[33] Argall, B.D., Chernova, S., Veloso, M., Browning, B.: A survey of robot learning from demonstration. Robotics and Autonomous Systems **57**(5), 469–483 (2009). https://doi.org/10.1016/j.robot.2008.10.024

[34] Abbeel, P., Ng, A.Y.: Apprenticeship Learning via

Inverse Reinforcement Learning. In Proceedings of the Twenty-first International Conference on Machine Learning, p. 1 (2004). https://doi.org/10.1145/1015330.1015430

[35] Shi, Z., Yang, S.: Integrating Domain Knowledge into Large Language Models for Enhanced Fashion Recommendations (2025). https://doi.org/10.48550/arXiv.2502.15696

[36] Ziebart, B.: Modeling purposeful adaptive behavior with the principle of maximum causal entropy. PhD thesis, Carnegie Mellon University (2010)

[37] Finn, C., Christiano, P., Abbeel, P., Levine, S.: A Connection between Generative Adversarial Networks, Inverse Reinforcement Learning, and Energy-Based Models. ArXiv preprint arXiv:1611.03852 (2016). https://doi.org/10.48550/arXiv.1611.03852

[38] Hinton, G.E.: Training Products of Experts by Minimizing Contrastive Divergence. Neural Computation **14**(8), 1771–1800 (2002). https://doi.org/10.1162/089976602760128018

[39] Wu, M., Goodman, N.: Multimodal Generative Models for Scalable Weakly Supervised Learning. In 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), pp. 1–11 (2018)

[40] Cao, Y., Fleet, D.J.: Generalized product of experts for automatic and principled fusion of gaussian process predictions. arXiv preprint arXiv:1410.7827 (2018). https://doi.org/10.48550/arXiv.1410.7827

[41] Ho, J., Ermon, S.: Generative Adversarial Imitation Learning. In Proceedings of the 30th International Conference on Neural Information Processing Systems, pp. 4572–4580 (2016)

[42] Song, X., Feng, F., Liu, J., Li, Z., Nie, L., Ma, J.: NeuroStylist: Neural Compatibility Modeling for Clothing Matching. In Proceedings of the 25th ACM International Conference on Multimedia, pp. 753–761 (2017). https://doi.org/10.1145/3123266.3123314

[43] Yang, X., He, X., Wang, X., Ma, Y., Feng, F., Wang, M., Chua, T.-S.: Interpretable Fashion Matching with Rich Attributes. In Proceedings of the 42Nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 775–784 (2019). https://doi.org/10.1145/3331184.3331242

[44] Maaten, L., Hinton, G.: Visualizing Data using t-SNE. Journal of Machine Learning Research **9**(86), 2579–2605 (2008)

[45] Koren, Y.: Factorization meets the neighborhood: a multifaceted collaborative filtering model. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 426–434 (2008). https://doi.org/10.1145/1401890.1401944

[46] Radford, A., Metz, L., Chintala, S.: Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In The International Conference on Learning Representations (ICLR) (2016)

[47] Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic Differentiation in PyTorch. In 31st Conference on Neural Information Processing Systems (NIPS 2017), pp. 1–4 (2017)

[48] Adar, E., Dontcheva, M., Laput, G.: CommandSpace: modeling the relationships between tasks, descriptions and features. In Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology, pp. 167–176 (2014). https://doi.org/10.1145/2642918.2647395

[49] Michailidou, E., Harper, S., Bechhofer, S.: Visual complexity and aesthetic perception of web pages. In Proceedings of the 26th Annual ACM International Conference on Design of Communication, pp. 215–224 (2008). https://doi.org/10.1145/1456536.1456581