SPIRIT: Style-guided Patch Interaction for Fashion Image Retrieval with Text Feedback

YANZHE CHEN and JIAHUAN ZHOU, Wangxuan Institute of Computer Technology, Peking University, China
YUXIN PENG, Wangxuan Institute of Computer Technology, Peking University, China and Peng Cheng Laboratory, China

Fashion image retrieval with text feedback aims to find the target image according to the reference image and the modification from the user. This is a challenging task, as it requires not only the synergistic understanding of both visual and textual modalities but also the ability to model a wide variety of styles that fashion images contain. Hence, the crucial aspect of addressing this problem lies in exploiting the abundant semantic information inherent in fashion images and correlating it with the textual description of style. Recognizing that style is generally situated at the local level, we explicitly define style as the commonalities and differences between local areas of fashion images. Building upon this, we propose a Style-guided Patch InteRaction approach for fashion Image retrieval with Text feedback (SPIRIT), which focuses on the decisive influence of local details of fashion images on their style. Three corresponding networks are designed pertinently. The Patch-level Style Commonality network is introduced to fully leverage the semantic information among patches and compute their average as the style commonality. Subsequently, the Patch-level Style Difference network employs a graph reasoning network to model the patch-level difference and filter out insignificant patches. By considering the above two networks, mutual information about style is obtained from the interaction between patches. Finally, the Visual Textual Fusion network is utilized to integrate visual features with rich semantic information and textual features. Experimental results on four benchmark datasets demonstrate that our proposed SPIRIT achieves state-of-the-art performance. Source code is available at https://github.com/PKU-ICST-MIPL/SPIRIT_TOMM2024.

CCS Concepts: • **Computing methodologies** → *Artificial intelligence*; *Computer vision*; *Computer vision tasks*; *Visual content-based indexing and retrieval;*

Additional Key Words and Phrases: Fashion image retrieval with text feedback, style modeling, multimodal fusion

ACM Reference Format:
Yanzhe Chen, Jiahuan Zhou, and Yuxin Peng. 2024. SPIRIT: Style-guided Patch Interaction for Fashion Image Retrieval with Text Feedback. *ACM Trans. Multimedia Comput. Commun. Appl.* 20, 6, Article 167 (March 2024), 17 pages. https://doi.org/10.1145/3640345

This work was supported by the grants from the National Natural Science Foundation of China (62132001 and 61925201).
Authors' addresses: Y. Chen and J. Zhou, Wangxuan Institute of Computer Technology, Peking University, Beijing, 100871, China; Y. Peng (Corresponding author), Wangxuan Institute of Computer Technology, Peking University, Beijing, 100871, China and Peng Cheng Laboratory, Shenzhen, 518055, China; e-mail: pengyuxin@pku.edu.cn.

ACM 1551-6857/2024/03-ART167
https://doi.org/10.1145/3640345

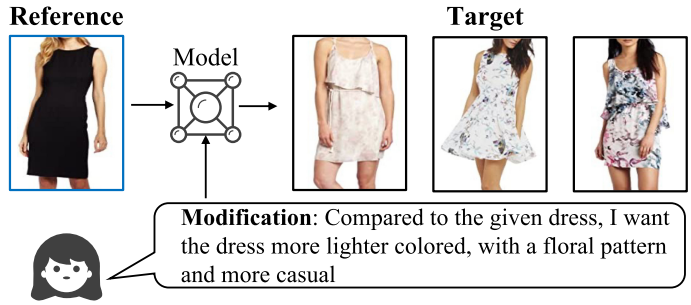
1 INTRODUCTION

As a classic task in the field of computer vision, image retrieval has made great progress in recent years [16, 19, 47]. In more interactive areas such as e-commerce, users have an obvious tendency to retrieve the items they want. At this time, only providing an image as the query cannot well meet the needs of users [8, 21]. Therefore, fashion image retrieval with text feedback, as a branch of conventional image retrieval where a reference image and a modification are used jointly as a query, is thriving in these areas [33, 43, 46]. When image retrieval is applied to the field of e-commerce, the main goal of the task is to find the product that best meets the user's expectations [8, 29, 37], as illustrated in Figure 1. For instance, when a user meets a fashion image, the user may request modifications in terms of color, size, style, and so on.

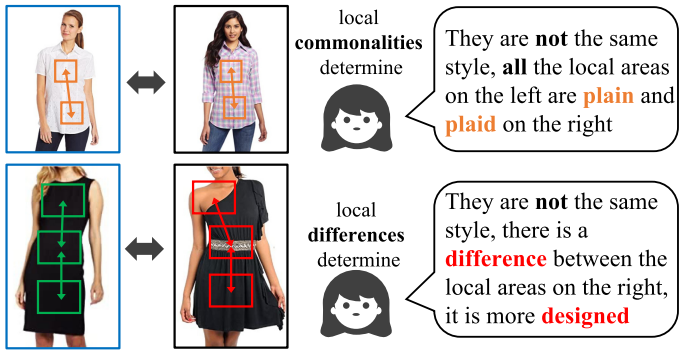
The challenges of the task of fashion image retrieval with text feedback mainly lie in the following two aspects. First, it requires the model to fully correlate the semantic information of two modalities of visual and text and realize the feature fusion between the two modalities [5, 52]. Second, fashion images usually contain a variety of styles, and the model needs to be able to accurately return a fashion image in the specific style that the user wants. Many works have contributed to the task revolving the first challenge [5, 15, 18, 32, 48, 52, 57]. Their main concern is how to fuse the image and text inputs into joint embedding for retrieval. Among the methods, some of the works focus on learning composite representations of reference images and text queries to approximate the embedding of the target image as closely as possible [5, 12, 32]. Several studies revolve around the information contained in the fashion images, which applies fine-grained methods, or the target detection network to extract more details [18, 41].

However, existing methods rarely address the second challenge, but the style of the fashion image is crucial to whether it meets the users' modification requirements. Although the study [37] extends the definition of style from the field of Image Style Transfer to this task, this definition primarily revolves around the global aspects of images, such as colors. Such a definition struggles to encompass the specific and detailed styles present in fashion images, thereby neglecting that style cues are generally manifested on a local level, such as distinct patterns, textures, or design features. These local-level details are frequently identifiable within specific regions of the image rather than uniformly distributed across the entire scene. For instance, styles like "plaid" and "designed," as illustrated in Figure 1(b), are associated with specific local regions of fashion images. For the first row of Figure 1(b), there is a high degree of coherence among the localities of the two shirts, indicating an overall leaning toward a formal style. In the second row of Figure 1(b), the right-side dress exhibits variations among patches, with distinctive local details around the right shoulder and waistband, defining it as a designed dress with stylistic differences from the left-side counterpart. Understanding and effectively capturing these local-level stylistic details are crucial for achieving accurate and meaningful image retrieval in the context of fashion applications.

To address the problems above, we propose a **Style-guided Patch InteRaction approach for fashion Image retrieval with Text feedback (SPIRIT)**. The style of a fashion image is explicitly defined as the commonality and difference between its patches, and three networks are designed based on the definition. Given a fashion image, Patch Sampling Strategy is first used to split the multi-scale patches, and **Patch-level Style Commonality (PSC)** generates the style commonality feature by averaging the thoroughly interacted patch features, enabling the model to better distinguish different fashion styles by combining the local information between patches. Then **Patch-level Style Difference (PSD)** further models the differences between patches through a graph reasoning network and filters out unimportant areas through adaptive calculation of patches' weight, so as to distinguish between those with a sense of design that have significant differences in the local areas. Through the interaction of the features extracted from the above two networks, the style features containing mutual information are obtained. After the style features being



(a) Example of the Fashion Image Retrieval with Text Feedback task.



(b) Examples of local areas affect the style of fashion images.

Fig. 1. (a) Task illustration of the Fashion Image Retrieval with Text Feedback task, where a reference image is provided along with a modification to retrieve the most relevant target images. (b) Commonalities and differences among local areas of fashion images reflect their styles.

concatenated with the global features from the whole image, **Visual-Textual Fusion (VTF)** then fully integrates the visual features containing rich semantic information with the textual features, thus improving the mapping ability between the text description of the style and the corresponding images, through the hierarchical operation of information mining within the modalities and mapping between the modalities to the common space.

The main contributions are summarized as follows:

- Recognizing that style is generally situated at the local level, we introduce an explicit definition of style as the commonality and difference between their local parts. Accordingly, we present a tailored approach involving three interconnected networks.
- The Patch-level Style Commonality and Patch-level Style Difference are proposed to model the style commonality and difference features, respectively. Visual Textual Fusion network is proposed to fully integrate the visual features containing rich semantic information with the related textual features in a hierarchical manner.
- Our proposed approach SPIRIT achieves state-of-the-art performance on four benchmark datasets for fashion image retrieval with text feedback.

2 RELATED WORK

2.1 Fashion Image Retrieval with Text Feedback

Image retrieval usually involves a user submitting an image and the system returning the closest alternative [3, 16, 19, 38, 47]. Since natural language is the most fundamental form of interaction

between a user and system, using images and modification to retrieve the target image is often more user-friendly [8, 21, 45, 53]. When such a task is applied to fashion images, it is referred to as Fashion Image Retrieval with Text Feedback. Numerous creative methods revolving around this task have been proposed in recent years. Vo et al. [52] propose a method called TIRG, which contains a residual and gating module to compose image features with textual features. Chen et al. [8] propose a composite transformer plugged in a CNN to transform the visual features conditioned on language semantics. Hosseinzadeh et al. [25], introduce the additional objective function to help the network learn a better representation of the query and target image. Anwaar et al. [2] enable the network to constrain the optimization problem by proposing a rotational symmetry loss function. Graph Convolution Network is introduced by Shin et al. [48] to encode the differences between the source and target images conditioned on the text. Kim et al. [32] also pay attention to the little difference between the reference image and the target image, thus modeling the difference between the reference and target image in the embedding space and matched with the embedding of the text query. Goenka et al. [18] apply VinVL [59] into the proposed method to capture the relationship between the local features of the images and the text. Baldrati et al. [5] propose a combiner network to combine the visual and textual features from the OpenAI CLIP [44] network. However, the above methods fail to include the style of the fashion images as a key attribute in the task, which can significantly affect the experience of the user. In contrast, our approach uses PSC and PSD to excavate the style of fashion images through the commonalities and differences between patches. Then the visual features containing style semantic information, rather than global visual features, are integrated with textual features under the use of VTF.

2.2 Vision-language Model for Fashion

Fashion tasks in the fashion domain encompass various cross-modal activities such as retrieval, matching, and generation, akin to the broader vision-language context [24, 30, 61]. Numerous fashion-related datasets have also been curated and released [10, 17, 22, 54, 58]. KaleidoBERT [64] adopts a multi-stage approach to enhance the salient features of fashion items through several single-task frameworks. Meanwhile, FashionViL [23] employs an end-to-end architecture to pre-train the model across multiple single-task objectives, inspired by the general vision-language paradigm. Hou et al. [26] recognize the diverse attributes present in the e-commerce domain. They put forward a proposition to harness the semantic essence of visual attributes in training convolutional networks. These networks aim to acquire attribute-specific subspaces for individual attributes, leading to the acquisition of disentangled representations. FashionVLP [18] explicitly leverages object detection models to extract the primary regions of fashion images. It also incorporates clothing key points, combining these attributes to enhance retrieval precision. Differing from prior methods, our proposed SPIRIT focuses on the crucial attribute of style in fashion images. We model the style attribute based on local information within the fashion images, without introducing additional information. This approach enhances the alignment between retrieval results and the demands expressed in modifications, thus improving retrieval performance.

2.3 Image Style Modeling

Style plays an essential role to evaluate the performance of the model in tasks such as style transfer [7, 27, 31, 36] and image synthesis [1, 35, 42, 63]. In the tasks above, style is often referred to as the channel statistics that are spatially invariant, while contents are expressed by local features [37]. Huang et al. [27] propose AdaIN, which verified through experience that the two-channel statistics, mean and variance, are highly correlated with image style. Therefore, the style of the image can be changed using instance normalization. On this basis, Li et al. [39] point out that the covariance matrix could better represent the style of an image. Xia et al. [55] introduce the bilateral-space

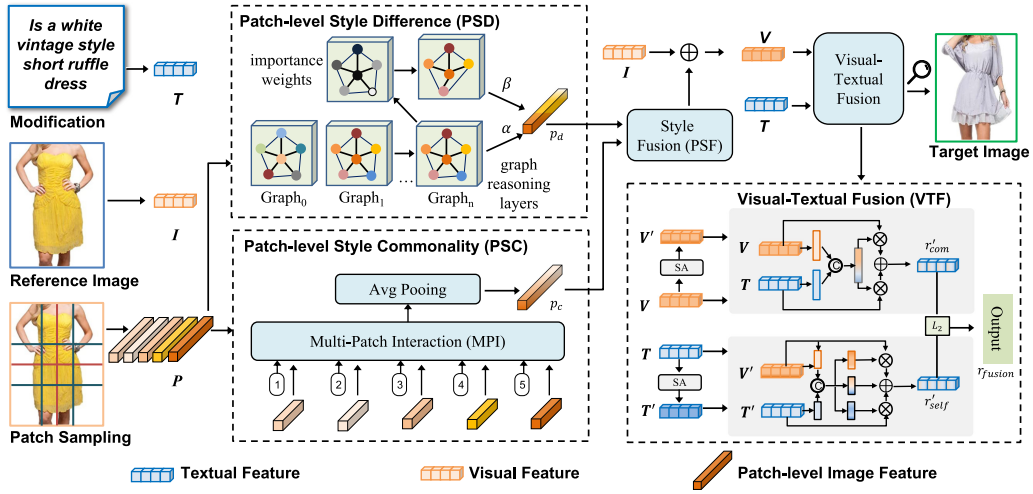


Fig. 2. Overall architecture of our proposed approach SPIRIT, which mainly contains three networks: PSC, PSD, and VTF. The style features containing mutual information are obtained through the interaction of the features extracted from PSC and PSD. VTF then fully integrates the visual features containing rich semantic information with the textual features.

Laplacian regularizer to achieve an efficient stylized model that can adapt to any style migration. Chiu et al. [11] propose blockwise training to perform coarse-to-fine feature transformations. Lee et al. [37] introduce the ideas from the above methods into their proposed method, with instance normalization changing the style of the image. However, the overall style of the image is not exactly the same as the style of the fashion image. For example, the images of minimalist style and palace style are very similar on the whole. They are both solid colors in large areas, so the channel statistics are also close. What sets the two styles apart, however, are some local details shown in Figure 1(b). In contrast, our approach fully excavates the style of fashion images through the interaction between local patches and the similarities and differences between patches, so as to better fit the task requirements of fashion image retrieval. We explicitly define the style of fashion images as the commonality and difference between their local parts, and we design three networks to better handle the task of fashion image retrieval with text feedback.

3 METHODOLOGY

We propose SPIRIT, shown in Figure 2, which will be elaborated in the following four sections. In the Image and Language Global Embedding section, we present the extraction of global features for the reference image and modification using backbone models. In the Image Patch-level Style Embedding section, we first elaborate on the strategy for sampling patches from the fashion image. Based on these patches, we propose PSC and PSD to extract the style commonality and difference features of the image, respectively. In the Fusion Representation section, we describe how to thoroughly integrate the aforementioned features to achieve precise retrieval. The Model Training section outlines the process details of training these models.

3.1 Image and Language Global Embedding

The features extracted through the image encoder encompass global-level aspects such as color and shape. However, such global features are not enough to characterize the style of fashion images,

just as the design elements of some commodities are shown in the neckline, front, and other small local areas. Therefore, it is crucial to investigate the details of the local area of the image.

3.2 Image Patch-level Style Embedding

3.2.1 Patch Sampling Strategy. Since the style of fashion images is highly related to local details, we apply a Patch Sampling Strategy to obtain the multi-grained patches from a given image. We do not use the pretrained object detection network to extract the specific areas of fashion images. On the one hand, this is set up for a fairer comparison with the existing methods. On the other hand, the areas unrelated to clothing, such as the model's legs and arms, could also reflect the style of fashion images, which is also one of the differences between fashion image retrieval and conventional image retrieval. For instance, when the proportion of the body exposed is large, the clothes are more casual than business. We split the given image into different scales (i.e., 2×2, 3×3, 4×4) similarly to sliding windows. Specific sliding dimensions can be adjusted according to the details of commodities. In this method, we apply 2×2 and 3×3 scales to obtain a total of 13 patches.

3.2.2 Patch-level Style Commonality. The commonality of patch-level style is close to the mean value of patches' features. Compared to simply averaging features, the features obtained by averaging the patch features after semantic information interaction between them better represent the commonalities of a fashion image. Therefore, for patches extracted with the strategy above, we first use the same image encoder to extract patch-level features $P = \{p_1, p_2, \ldots, p_k\}$. Next, we propose a **Multi-Patch Interaction model (MPI)** with a Transformer structure to interact patch features and obtain the interaction feature sequence $p_{in_1}, p_{in_2}, \ldots p_{in_k}$, which can be represented as follows:

$$p_c = \text{Avg}(\text{MPI}(P + e^{pos})), \tag{1}$$

where Avg and e^{pos} denote the average pooling and patch position encoding, respectively. For e^{pos}, we adopt standard learnable absolute position embeddings. The MPI is constructed using the standard multi-head self-attention and feed-forward networks [51].

3.2.3 Patch-level Style Difference. To model the differences in style between patches, a similarity graph is built to exchange the feature information between patches. We first perform a simple but effective self-attention mechanism [51] over the patches, which uses the average of all patches features as the query and obtains the raw average feature p_a by aggregating all the patches. Compared to learning the differences between each pair of patches, we emphasize the differences between patches by learning the differences between each patch and the mean feature rather than defining this parameter based on coordinate positions, as there may be overlaps between patches. We then take all the patch-level features and the average feature p_a as graph nodes $\mathcal{N}$ and follow Reference [14] to build and update the edge e between two nodes $v_{in} \in \mathcal{N}$ to $v_{out} \in \mathcal{N}$, shown as follows:

$$\mathcal{N} = \{p_1, p_2, \ldots, p_k, p_a\}, \tag{2}$$

$$e\left(v_{\text{in}}, v_{\text{out}}; W_{\text{in}}, W_{\text{out}}\right) = \frac{e^{(W_{\text{in}} v_{\text{in}})(W_{\text{out}} v_{\text{out}})}}{\sum_{\text{out}} e^{(W_{\text{in}} v_{\text{in}})(W_{\text{out}} v_{\text{out}})}}, \tag{3}$$

where W_{in} and W_{out} are linear layers that pass information between incoming node v_{in} and outgoing node v_{out}.

After constructing the graph nodes and related edges, we obtain the patch-level style difference feature by updating the nodes together with edges as follows:

$$v_{in}^{L+1} = W_d^L \left(\sum_{\text{out}} e\left(v_{\text{in}}^L, v_{\text{out}}^L; W_{\text{in}}^L, W_{\text{out}}^L\right) \cdot v_{\text{out}}^L \right), \tag{4}$$

where L denotes the number of graph reasoning layers, W_D^L, W_{in}^L, and W_{out}^L are linear layers in each layer. After layer L, the node feature v_{in}^L is replaced with v_{in}^{L+1}.

After the similarity graph reasoning, the node feature p_a fully exchanges comparison information with other patches and combines the style differences among patches, denoted as p_d^{origin}. However, our Patch Sampling Strategy will inevitably capture areas unrelated to clothing. Although some patches of the body are beneficial to the model's learning of style, there are still patches that are meaningless to the results, such as blank patches. Therefore, we replace p_a with p_a^{origin} in the original $\mathcal{N}$, and the importance weight w_i of each patch feature is calculated to obtain the filtered feature p_d^{filter} as follows:

$$w_i = \frac{\sigma(BN(W_f v_i))}{\sum_{v_j \in \mathcal{N}} \sigma(BN(W_f v_j))}, \tag{5}$$

$$p_d^{filter} = \sum_{v_i \in \mathcal{N}} w_i v_i, \tag{6}$$

where σ denotes the sigmoid function, BN denotes the batch normalization, and W_f denotes a fully connected layer. To improve the universality of Patch-level Style Difference, it is also necessary to consider the case that for some fashion images with rich details, each patch can be crucial in determining its style. For instance, consider an e-commerce short-sleeve try-on image. If the display includes a model wearing the item, then patches containing the model's head or lower body may be filtered out. However, for a flat-laid image of the short sleeve, the details of each patch contribute to the overall style of the entire garment. When exploring the differences between patches, it is essential to consider these diverse possibilities to enhance its retrieval performance in real-world scenarios. Therefore, a residual-like mechanism is used to select the most important features automatically with two self-learning parameters α and β for style differences as follows:

$$p_d = \alpha \cdot p_d^{origin} + \beta \cdot p_d^{filter}. \tag{7}$$

3.3 Fusion Representation

3.3.1 Patch-level Style Fusion. The outputs from PSC (i.e., commonality features $P_c = [p_{c1}, p_{c2}, \ldots, p_{cn}] \in \mathbb{R}^{n \times d}$) and from PSD (i.e., difference features $P_d = [p_{d1}, p_{d2}, \ldots, p_{dn}] \in \mathbb{R}^{n \times d}$) contain attended information about patches' style. Therefore, we need an effective method to integrate these two local style features from patches. Following Reference [5], we use a feature fusion method that has a simple structure but can achieve better results than many complex ones. We first map P_c and P_d to common spaces using projection layers f_c and f_d (i.e., FC-ReLU-Dropout (0.5)). The low-level interactive feature P_l is generated by a similar projection layer as follows:

$$P_l = f_l([f_c(P_c); f_d(P_d)]), \tag{8}$$

where $[[f_c(P_c); f_d(P_d)]]$ denotes the concatenation of features $f_c(P_c)$ and $f_d(P_d)$.

For fashion images of different styles, the commonality and differences between patches may be of different importance. Therefore, we assign their weights by learning a dynamic parameter θ, and combine them with P_l to obtain the high-level interactive feature P_h as follows:

$$\theta = f_\theta([f_c(P_c); f_d(P_d)]), \tag{9}$$

$$P_h = f_h(P_l) + \theta \cdot P_c + (1 - \theta) \cdot P_d, \tag{10}$$

where f_θ denotes the mapping function, which consists of a two-layer MLP (i.e., FC-ReLU-Dropout (0.5)-FC1(1)) and a Sigmoid layer after two MLP layers. The feature P_h contains high-level style clues from the patch-level features that pay attention to the commonality and difference of styles between patches, respectively.

3.3.2 Visual Textual Fusion. We concatenate the global image feature and the patch-level style feature P_h together as total visual feature $V = [v_1, v_2, \ldots, v_m] \in \mathbb{R}^{m \times 2d}$ and concatenate the same textual feature extracted from the modifications to align with V, denoted as $T = [t_1, t_2, \ldots, t_m] \in \mathbb{R}^{m \times 2d}$.

Next, the problem to be addressed is how to efficiently fuse V and T and use the fusion feature to retrieve the closest target visual feature. Different from the method in Reference [5], we also hold the view that features V and T can be mapped to the same common space for feature fusion, but for the visual features with more vivid styles, the information contained is more than the information contained in the corresponding modification of the triplet, which is difficult to align in the same representation space. Hence, we introduce a two-way feature fusion. In addition to the direct interaction between modalities, the other path first thoroughly explores semantic information within each modality before further engaging in cross-modal interaction. This hierarchical approach is conducive to improve the mapping ability between the text description of the style and the corresponding images. As a result, the proposed method can make full use of the features of each modality and fuse them in a similar way to the Patch-level Style Fusion. We first use two fully connected layers (i.e., FC-ReLU-Dropout (0.5)-FC (1)) for V and T and generate self-attended features V' and T',

$$V' = \sum_{i=1}^{m} \text{Softmax}(\text{MLP}(V))v_i, \tag{11}$$

$$T' = \sum_{i=1}^{m} \text{Softmax}(\text{MLP}(T))t_i. \tag{12}$$

After obtaining rich attended features from each modality, we follow the same architecture of Patch-level Style Fusion to get the low-level self-attended interactive feature r_{self}^{low} and interactive feature based on the same common space r_{com}^{low}, together with dynamic parameters θ_{self} and θ_{com}. The formula of the approach to get the final fusion feature r_{fusion} is as follows:

$$r'_{self} = f_{self}\left(r_{self}\right) + \theta_{self} \cdot V' + \left(1 - \theta_{self}\right) T', \tag{13}$$

$$r'_{com} = f_{com}\left(r_{com}\right) + \theta_{com} \cdot V' + \left(1 - \theta_{com}\right) T', \tag{14}$$

$$r_{fusion} = r'_{self} + r'_{com}, \tag{15}$$

where f_{self} and f_{com} are fully connected layers.

3.4 Model Training

We apply the PSC and PSD networks to the reference-modification pairs to extract the style features from the reference image. Afterward, we utilize **Patch-level Style Fusion (PSF)** to obtain the fusion feature r_{fusion}. For the target image, we utilize the PSC and PSD networks to extract style features and concatenate them with the image global features to obtain the target feature r_{target}. Following existing methods [5, 18], we employ a batch-based classification loss as the training loss function for the PSC, PSD, and PSF networks. In this loss function, every entry within a batch serves as a negative sample for all other entries. For a batch of B reference-modification pairs, the loss is defined as follows:

$$\mathcal{L} = \frac{1}{B} \sum_{i=1}^{B} -\log \frac{\exp \kappa(r_{fusion}^{i}, r_{target}^{i})}{\sum_{j=1}^{B} \exp \kappa(r_{fusion}^{i}, r_{target}^{j})}. \tag{16}$$

During the inference stage of the model, we generate feature r_{fusion} using the aforementioned approach for the reference-modification pairs. For all the test images, we compute feature r_{target} and then rank the test images based on the similarity between r_{fusion} and r_{target}.

4 EXPERIMENTS

4.1 Experimental Setup

4.1.1 Datasets. FashionIQ [54]: This is a widely used benchmark for text-conditioned image retrieval. It consists of 77,684 fashion images crawled from the web divided into three distinct categories: Dress, Toptee and Shirt. The dataset is organized by triplets, with a reference image, a target image, and two crowd-sourced captions that describe the differences between the two images. There may be a gap between these two captions, as different people have different aesthetic interpretations. We follow the experimental setting as in References [5, 18], which constructs the candidate set by unifying all reference and target images in the test set. *Shoes* [6]: This dataset was originally proposed for attribute study from web pages, and the images have been tagged with captions for fashion image retrieval task [21]. We use the original split in Reference [21], which provides 10,000 training queries and 4,658 validation queries. *CIRR* [40]: The dataset contains 21,552 real-world images from NLVR2 [49]. There are 36,554 triplets in total, divided into three subsets with 80% in training, 10% in validation, and 10% in testing. *Fashion200k* [22]: The dataset has 172,000 training and 33,000 testing images. The process used to generate textual feedback involves comparing attributes between image pairs and follows a simple format of "replace [sth] with [sth]."

4.1.2 Evaluation Metrics. Following the evaluation metrics in References [5, 59], we evaluate the performance of SPIRIT using the standard top-K recall metric for image retrieval, denoted as R@K. In particular, we use Recall@10 (R@10) and Recall@50 (R@50).

4.1.3 Experimental Details. We choose Reference [4] as our baseline and keep the same experimental setup as our baseline method [4]. It involves first pretraining the encoders and then freezing them during the fusion phase to train the multimodal fusion model. This method enhances batch size efficiency on the same device. We keep our setup aligned with CLIP4Cir's [4] by maintaining encoders freezing while training PSC, PSD, and PSF. We use Adam [34] to optimize the network. The batch size for the CIRR dataset is set to 2048, while the rest are set to 1024. The initial leaning rate is 4e-5 and we adopt a cosine annealing strategy to adjust it. The total number of training epochs is 150.

4.2 Comparison with State-of-the-art Methods

Experiments on the widely used datasets are conducted to evaluate our approach and recent state-of-the-art methods. The details of the datasets are as follows.

FashionIQ. We first evaluate SPIRIT and compare it with state-of-the-art methods on the FashionIQ dataset. The results are summarized in Table 1. Compared with the previous methods in the table, we contribute a new state-of-the-art approach in every metric. Specifically, the average results of R@10 and R@50 in the VAL evaluation protocol and the original evaluation protocol are 62.54 and 55.54, which is 11.86% and 4.19% higher than the previous state-of-the-art methods [24, 56] and 5.51% higher than the baseline method [4]. Due to the use of our proposed Patch-level Style Commonality and Patch-level Style Difference networks, our proposed model leverages the patch interaction to fully exploit the commonality and differences between patches and utilizes the mutual information between them to effectively represent the style of fashion images. Visual Textual Fusion is further used to align rich visual information containing style with textual features, thus establishing a mapping between fashion images containing multiple styles and modifications.

Shoes, CIRR, and Fashion200k. Tables 2, 3, and 4 show the quantitative results on the Shoes, CIRR, and Fashion200k datasets. The results on the three datasets show a similar trend to the results

Table 1. Results on the FashionIQ Dataset

Method	Venues	Dress		Toptee		Shirt		Overall		
		R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50	**Mean**
		VAL [8] Evaluation Protocol								
TIRG [52]	CVPR 2019	14.87	34.66	18.26	37.89	19.08	39.62	17.40	37.39	27.40
VAL [8]	CVPR 2020	21.12	42.19	25.64	49.49	21.03	43.44	22.60	45.04	33.82
CosMo [37]	CVPR 2021	25.64	50.30	29.21	57.46	24.90	49.18	26.58	52.31	39.45
DCNet [32]	AAAI 2021	28.95	56.07	30.44	58.29	23.95	47.30	27.78	53.89	40.84
SAC [28]	WACV 2022	28.02	51.86	26.52	51.01	32.70	61.23	29.08	54.70	41.89
ARTEMIS [13]	ICLR 2022	27.16	52.40	29.20	43.64	21.78	54.83	26.05	50.29	38.17
MGUR [9]	ArXiv 2022	30.60	57.46	37.37	68.41	31.54	58.29	33.17	61.39	47.28
FashionVLP [18]	CVPR 2022	32.42	60.29	38.51	68.79	31.89	58.44	34.27	62.51	48.39
ComqueryFormer [56]	TMM 2023	33.86	61.08	42.07	69.30	35.57	62.19	37.17	64.19	50.68
SPIRIT (Ours)	—	**43.83**	**68.86**	**56.60**	**79.25**	**52.50**	**74.19**	**50.98**	**74.10**	**62.54**
		Original Evaluation Protocol								
TIRG [52]	CVPR 2019	14.13	34.61	14.79	34.37	13.10	30.91	14.01	33.30	23.66
CosMo [37]	CVPR 2021	21.39	44.45	21.32	46.02	16.90	37.49	19.87	42.62	31.25
MGUR [9]	ArXiv 2022	24.54	50.12	29.06	55.63	20.70	45.53	24.77	50.43	37.60
FashionVLP [18]	CVPR 2022	26.77	53.20	28.51	57.47	22.67	46.22	25.98	52.30	39.14
FashionViL [23]	ECCV 2022	33.47	59.94	34.98	60.79	25.17	50.39	31.21	57.04	44.12
CLIP4Cir [4]	CVPR 2022	33.81	59.40	41.41	65.37	39.99	60.45	38.32	61.74	50.03
FashionSAP [24]	CVPR 2023	33.71	60.43	41.91	70.93	33.17	61.33	36.26	64.23	50.25
SPIRIT (Ours)	—	**39.86**	**64.30**	**47.68**	**71.70**	**44.11**	**65.60**	**43.88**	**67.20**	**55.54**

Best scores are highlighted in bold and underlined formats.

Table 2. Results on the Shoes Dataset

Methods	Venues	R@10	R@50	Mean
TIRG [52]	CVPR 2019	45.45	69.39	57.32
VAL [8]	CVPR 2020	49.12	73.53	61.32
CosMo [37]	CVPR 2021	48.36	75.64	62.00
FashionVLP [18]	CVPR 2022	49.08	77.32	63.20
SAC [28]	WACV 2022	51.73	77.28	64.51
MGUR [9]	Arxiv 2022	53.63	79.84	66.74
ARTEMIS [13]	ICLR 2022	53.11	79.31	66.21
AMC [62]	TOMM 2023	56.89	79.27	68.08
SPIRIT (Ours)	—	**56.90**	**81.49**	**69.19**

Best scores are highlighted in bold and underlined formats.

Table 3. Results on the CIRR Dataset

Methods	Recall@K				R_{sub}@K			Mean
	K = 1	K = 5	K = 10	K = 50	K = 1	K = 2	K = 3	
TIRG [52]	14.61	48.37	64.08	90.03	22.67	44.97	65.14	35.52
MAAF [15]	10.31	33.03	48.30	80.06	21.05	41.81	61.60	27.04
CIRPLANT [40]	19.55	52.55	68.39	92.38	39.20	63.03	79.49	45.88
ARTEMIS [13]	16.96	46.10	61.31	87.73	39.99	62.20	75.67	43.05
CLIP4Cir [4]	39.75	73.71	83.90	96.87	70.92	87.42	94.19	72.32
CompoDiff [20]	39.99	73.63	**86.77**	96.55	68.41	86.12	94.80	71.02
SPIRIT (Ours)	**40.23**	**75.10**	84.16	**96.88**	**73.74**	**89.60**	**95.93**	**74.42**

Best scores are highlighted in bold and underlined formats. Mean = (R@5 + R_{sub}@1)/2.

Table 4. Results on the Fashion200k Dataset

Methods	Venues	R@10	R@50	Mean
TIRG [52]	CVPR 2019	42.5	63.8	53.2
VAL [8]	CVPR 2020	49.0	68.8	58.9
DCNet [32]	AAAI 2021	46.9	67.6	57.3
CosMo [37]	CVPR 2021	50.4	69.3	59.8
FashionVLP [18]	CVPR 2022	49.9	70.5	60.2
ARTEMIS [13]	ICLR 2022	51.1	70.5	60.8
Css-Net [60]	Arxiv 2023	50.5	69.7	60.1
ComqueryFormer [56]	TMM 2023	<u>52.2</u>	<u>72.2</u>	<u>62.2</u>
SPIRIT (Ours)	—	**55.2**	**73.6**	**64.4**

Best scores are highlighted in bold and underlined formats.

Table 5. Ablation Study on the FashionIQ Dataset of Different Components

Methods	R@10	R@50	Mean
Baseline [4]	38.32	61.74	50.03
Baseline + PSC	38.39	63.82	52.11
Baseline + PSC + PSD	42.96	66.28	54.62
Baseline + PSC + PSD + PSF	43.45	66.71	55.08
Baseline + PSC + PSD + PSF + VTF (Ours)	**43.88**	**67.20**	**55.54**

on the FashionIQ dataset. Our proposed approach also outperforms the existing state-of-the-art methods [4, 9, 56] in each metric, achieving an average improvement of 1.11%, 2.10%, and 2.20% on the Shoes, CIRR, and Fashion200k datasets, respectively. This demonstrates the effectiveness and generalizability of the proposed SPIRIT.

4.3 Ablation Studies

4.3.1 Effects of Key Designs. To investigate the effectiveness of SPIRIT, we evaluate the key designs on the FashionIQ dataset, respectively. We add each component incrementally to verify the effect of each component, shown in Table 5.

In the table, the first row shows the results of baseline method [4]. When PSF is not adopted, we concatenate features from PSC and PSD directly. When VTF is not adopted, the original combiner from the baseline method is applied.

By comparing the experimental results from the tables, each component adopted in the experiment plays a role in improving the final results. Specifically, PSC generates the style commonality feature between split patches by thoroughly interacting between patches, increasing the model's ability to better distinguish different fashion styles by combining the local information between patches. PSD further models the differences among patches through a graph reasoning network and filters out unimportant areas through adaptive calculation of patches' weight to improve the model's ability to distinguish between those with a sense of design that have significant differences in the local areas. PSF complements the mutual style information from PSC and PSD and gets the visual features containing rich information. Finally, VTF fully integrates the visual features containing rich semantic information with the textual features through the hierarchical operation of information mining within the modalities and mapping between the modalities to the common space. Therefore, the mapping ability between the text description of the style and the corresponding images is further improved.

Table 6. Parameter Analysis on the FashionIQ Dataset on Different Layers in PSD, Denoted as N

Methods	R@10	R@50	Mean
N = 2	42.71	66.38	54.55
N = 3	**43.88**	67.20	**55.54**
N = 4	43.42	**67.31**	55.37
N = 5	43.70	66.98	55.34

Fig. 3. Qualitative results on FashionIQ and Shoes. We show reference images with blue boxes on the left and top-10 retrievals with descending scores on the right. Ground truths are shown with green boxes, and others are shown in red boxes.

4.3.2 Effects of Key Parameters. SPIRIT has a key parameter, the number of graph layers in PSD. The results are shown in Table 6. For the number of graph layers, when the number of layers is low, the model does not own the ability to fully learn the differences between patches, and the commonality between patches dominates. When the number of layers increases, the differences between patches become more important. Therefore, an appropriate number of layers can simultaneously consider the differences and commonalities between patches. From the results shown in Table 6, the most appropriate layer number is 3.

4.4 Qualitative Analysis

4.4.1 Qualitative Results. We show the references images, related textual feedback and top-10 retrievals predicted by our approach in Figure 3. Ground truths are shown with green boxes. From the experimental results, our approach can not only retrieve the target image in the low-rank recall but also the other retrieval results are highly related to the textual feedback. The first, second,

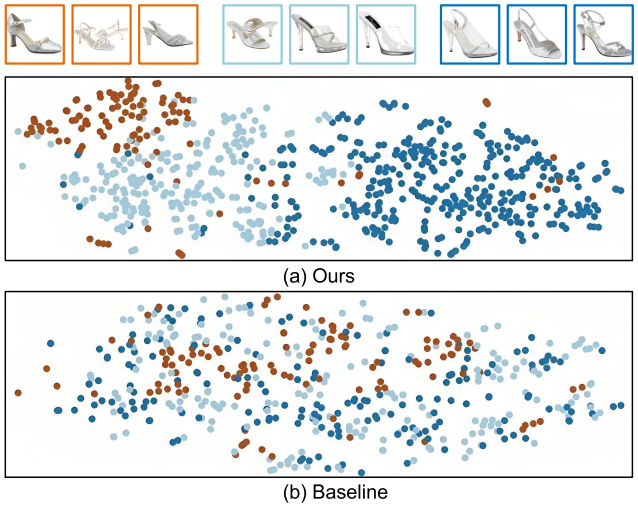


Fig. 4. t-SNE visualizations with the predictions of SPIRIT and the baseline method [4].

and fourth rows of feedback shown in the figure are all related to style, and our approach also shows good performance. Among them, the modification requirements of the fourth row are relatively simple. Our PSC can complete the target image retrieval according to the common features between patches, which are the basic patterns of plaid shirts. However, the first two rows define style in a relatively abstract way. The country style in the first row not only needs PSC to find out the commonness in patches, which are the colors like the same earth color but also needs PSD to ensure that there are no large design differences among patches. In the second-row demand, whimsical style is characterized by different partial design differences. Therefore, our PSD plays a dominant role and can identify target images in a small range.

4.4.2 Qualitative Results on Style. We select similar products from categories wedding shoes, stiletto, and high heels, where the category wedding shoes is labeled brown, the stiletto is labeled light blue, and high heels are labeled dark blue. We visualize the features extracted from our approach and the baseline method using t-SNE [50]. From the results shown in Figure 4, in (a) there are three distinct cluster centers, while in (b), the points are mixed together without distinguishing the boundary. Since these three categories are highly similar in terms of global features such as color and overall appearance, the baseline method is not good at distinguishing these three categories. Our approach not only considers the local details such as the thickness and length of the heel, so as to distinguish the two types of stiletto and high heels but also makes use of the local decorative details to subdivide the two similar types of wedding shoes and stiletto.

4.4.3 Failure Cases Analysis. Figure 5 shows typical failure cases using our approach. From the results shown in the figure, we find that our failure cases mainly come from the modification, which does not match the target image well. As shown in the first line of the figure, feedback contains too little information compared with the style of the target image and cannot accurately reflect the characteristics of the target. The top several retrieval results returned by our model are also consistent in feedback. It is worth noting that the feedback of the western style in the second line of the figure is too broad, and the image with the highest score retrieved by our method contains a cross, proving that it connects the cross, a typically western object, with the word *western*.

Fig. 5. Failure cases of our method. Qualitative results on FashionIQ and Shoes. We show reference images with blue boxes on the left, ground truths with yellow boxes in the middle and wrong retrievals with red boxes on the right.

5 CONCLUSION

In this work, we tackle the challenge of underutilized style information in fashion images by explicitly defining style as the commonalities and differences between patches. Building upon this, we introduce a SPIRIT. Our proposed SPIRIT consists of three components. PSC and PSD help to model the style commonality and difference features with the interaction with the patches, respectively. By fusing the complementary information contained in the two features, the style feature is obtained, and the rich information within and between modalities is fused by VTF.

The future work lies in two aspects: First, we will improve the interpretability of the approach and model more possible causes of influence on style. Second, we will further enhance the integration ability of vision and short text, and improve the reasoning ability of the model. Both of them will be employed to further boost the accuracy of the task of fashion image retrieval with text feedback.

REFERENCES

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. 2019. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4432–4441.

[2] Muhammad Umer Anwaar, Egor Labintcev, and Martin Kleinsteuber. 2021. Compositional learning of image-text query for image retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1140–1149.

[3] Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. 2023. Zero-shot composed image retrieval with textual inversion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15338–15347.

[4] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. 2022. Conditioned and composed image retrieval combining and partially fine-tuning CLIP-based features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4959–4968.

[5] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. 2022. Effective conditioned and composed image retrieval combining clip-based features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21466–21474.

[6] Tamara L. Berg, Alexander C. Berg, and Jonathan Shih. 2010. Automatic attribute discovery and characterization from noisy web data. In *Proceedings of the 11th European Conference on Computer Vision (ECCV '10), Part I 11*. Springer, 663–676.

[7] Prashanth Chandran, Gaspard Zoss, Paulo Gotardo, Markus Gross, and Derek Bradley. 2021. Adaptive convolutions for structure-aware style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7972–7981.

[8] Yanbei Chen, Shaogang Gong, and Loris Bazzani. 2020. Image search with text feedback by visiolinguistic attention learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3001–3011.

[9] Yiyang Chen, Zhedong Zheng, Wei Ji, Leigang Qu, and Tat-Seng Chua. 2022. Composed image retrieval with text feedback via multi-grained uncertainty regularization. arXiv:2211.07394. Retrieved from https://arxiv.org/abs/2211.07394

[10] Yanzhe Chen, Huasong Zhong, Xiangteng He, Yuxin Peng, and Lele Cheng. 2023. Real20M: A large-scale e-commerce dataset for cross-domain retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*. 4939–4948.

[11] Tai-Yin Chiu and Danna Gurari. 2022. Photowct2: Compact autoencoder for photorealistic style transfer resulting from blockwise training and skip connections of high-frequency residuals. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2868–2877.

[12] Guillaume Couairon, Matthijs Douze, Matthieu Cord, and Holger Schwenk. 2022. Embedding arithmetic of multimodal queries for image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4950–4958.

[13] Ginger Delmas, Rafael Sampaio de Rezende, Gabriela Csurka, and Diane Larlus. 2022. Artemis: Attention-based retrieval with text-explicit matching and implicit similarity. arXiv:2203.08101. Retrieved from https://arxiv.org/abs/2203.08101

[14] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. 2021. Similarity reasoning and filtration for image-text matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 1218–1226.

[15] Eric Dodds, Jack Culpepper, Simao Herdade, Yang Zhang, and Kofi Boakye. 2020. Modality-agnostic attention fusion for visual search with text feedback. arXiv:2007.00145. Retrieved from https://arxiv.org/abs/2007.00145

[16] Yujie Fu, Pengju Zhang, Bingxi Liu, Zheng Rong, and Yihong Wu. 2022. Learning to reduce scale differences for large-scale invariant image matching. *IEEE Trans. Circ. Syst. Vid. Technol.* 33, 3 (2022), 1335–1348.

[17] Yuying Ge, Ruimao Zhang, Xiaogang Wang, Xiaoou Tang, and Ping Luo. 2019. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5337–5345.

[18] Sonam Goenka, Zhaoheng Zheng, Ayush Jaiswal, Rakesh Chada, Yue Wu, Varsha Hedau, and Pradeep Natarajan. 2022. Fashionvlp: Vision language transformer for fashion retrieval with feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14105–14115.

[19] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. 2016. Deep image retrieval: Learning global representations for image search. In *Proceedings of the 14th European Conference on Computer Vision (ECCV '16), Part VI 14*. Springer, 241–257.

[20] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, HeeJae Jun, Yoohoon Kang, and Sangdoo Yun. 2023. CompoDiff: Versatile composed image retrieval with latent diffusion. arXiv:2303.11916. Retrieved from https://arxiv.org/abs/2303.11916

[21] Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Feris. 2018. Dialog-based interactive image retrieval. In *Advances in Neural Information Processing Systems*, Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2018/file/a01a0380ca3c61428c26a231f0e49a09-Paper.pdf

[22] Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis. 2017. Automatic spatially-aware fashion concept discovery. In *Proceedings of the IEEE International Conference on Computer Vision*. 1463–1471.

[23] Xiao Han, Licheng Yu, Xiatian Zhu, Li Zhang, Yi-Zhe Song, and Tao Xiang. 2022. Fashionvil: Fashion-focused vision-and-language representation learning. In *European Conference on Computer Vision*. Springer, 634–651.

[24] Yunpeng Han, Lisai Zhang, Qingcai Chen, Zhijian Chen, Zhonghua Li, Jianxin Yang, and Zhao Cao. 2023. FashionSAP: Symbols and attributes prompt for fine-grained fashion vision-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15028–15038.

[25] Mehrdad Hosseinzadeh and Yang Wang. 2020. Composed query image retrieval using locally bounded features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3596–3605.

[26] Yuxin Hou, Eleonora Vig, Michael Donoser, and Loris Bazzani. 2021. Learning attribute-driven disentangled representations for interactive fashion retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 12147–12157.

[27] Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*. 1501–1510.

[28] Surgan Jandial, Pinkesh Badjatiya, Pranit Chawla, Ayush Chopra, Mausoom Sarkar, and Balaji Krishnamurthy. 2022. SAC: Semantic attention composition for text-conditioned image retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 4021–4030.

[29] Surgan Jandial, Ayush Chopra, Pinkesh Badjatiya, Pranit Chawla, Mausoom Sarkar, and Balaji Krishnamurthy. 2020. Trace: Transform aggregate and compose visiolinguistic representations for image search with text feedback. arXiv:2009.01485. Retrieved from https://arxiv.org/abs/2009.01485

[30] Yang Jin, Yongzhi Li, Zehuan Yuan, and Yadong Mu. 2023. Learning instance-level representation for large-scale multi-modal pretraining in e-commerce. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11060–11069.

[31] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the 14th European Conference on Computer Vision (ECCV '16), Part II 14*. Springer, 694–711.

[32] Jongseok Kim, Youngjae Yu, Hoeseong Kim, and Gunhee Kim. 2021. Dual compositional learning in interactive image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 1771–1779.

[33] Jin-Hwa Kim, Sang-Woo Lee, Donghyun Kwak, Min-Oh Heo, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. 2016. Multimodal residual learning for visual qa. *Adv. Neural Inf. Process. Syst.* 29 (2016).
[34] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv:1412.6980.
[35] Umut Kocasari, Alara Dirik, Mert Tiftikci, and Pinar Yanardag. 2022. StyleMC: Multi-channel based fast text-guided image generation and manipulation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 895–904.
[36] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. 2018. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European Conference on Computer Vision (ECCV '18)*. 35–51.
[37] Seungmin Lee, Dongwan Kim, and Bohyung Han. 2021. Cosmo: Content-style modulation for image retrieval with text feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 802–812.
[38] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. 2023. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 23390–23400.
[39] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. 2017. Universal style transfer via feature transforms. *Adv. Neural Inf. Process. Syst.* 30 (2017).
[40] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. 2021. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2125–2134.
[41] Zhe Ma, Jianfeng Dong, Zhongzi Long, Yao Zhang, Yuan He, Hui Xue, and Shouling Ji. 2020. Fine-grained fashion similarity learning by attribute-specific embedding network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11741–11748.
[42] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2085–2094.
[43] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.
[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.
[45] Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. 2023. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19305–19314.
[46] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. 2017. A simple neural network module for relational reasoning. *Adv. Neural Inf. Process. Syst.* 30 (2017).
[47] Rishab Sharma and Anirudha Vishvakarma. 2019. Retrieving similar e-commerce images using deep learning. arXiv:1901.03546. Retrieved from https://arxiv.org/abs/1901.03546
[48] Minchul Shin, Yoonjae Cho, Byungsoo Ko, and Geonmo Gu. 2021. Rtic: Residual learning for text and image composition using graph convolutional network. arXiv:2104.03015. Retrieved from https://arxiv.org/abs/2104.03015
[49] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2018. A corpus for reasoning about natural language grounded in photographs. arXiv:1811.00491. Retrieved from https://arxiv.org/abs/1811.00491
[50] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 11 (2008).
[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Adv. Neural Inf. Process. Syst.* 30 (2017).
[52] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. 2019. Composing text and image for image retrieval-an empirical odyssey. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6439–6448.
[53] Haokun Wen, Xian Zhang, Xuemeng Song, Yinwei Wei, and Liqiang Nie. 2023. Target-guided composed image retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*. 915–923.
[54] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. 2021. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11307–11317.
[55] Xide Xia, Meng Zhang, Tianfan Xue, Zheng Sun, Hui Fang, Brian Kulis, and Jiawen Chen. 2020. Joint bilateral learning for real-time universal photorealistic style transfer. In *Proceedings of the 16th European Conference on Computer Vision (ECCV '20), Part VIII 16*. Springer, 327–342.
[56] Yahui Xu, Yi Bin, Jiwei Wei, Yang Yang, Guoqing Wang, and Heng Tao Shen. 2023. Multi-modal transformer with global-local alignment for composed query image retrieval. *IEEE Trans. Multimedia (TMM'23)*.

[57] Youngjae Yu, Seunghwan Lee, Yuncheol Choi, and Gunhee Kim. 2020. Curlingnet: Compositional learning between images and text for fashion iq data. arXiv:2003.12299. Retrieved from https://arxiv.org/abs/2003.12299

[58] Xunlin Zhan, Yangxin Wu, Xiao Dong, Yunchao Wei, Minlong Lu, Yichi Zhang, Hang Xu, and Xiaodan Liang. 2021. Product1m: Towards weakly supervised instance-level product retrieval via cross-modal pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11782–11791.

[59] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5579–5588.

[60] Xu Zhang, Zhedong Zheng, Xiaohan Wang, and Yi Yang. 2023. Relieving triplet ambiguity: Consensus network for language-guided image retrieval. arXiv:2306.02092. Retrieved from https://arxiv.org/abs/2306.02092

[61] Xiaoyang Zheng, Zilong Wang, Sen Li, Ke Xu, Tao Zhuang, Qingwen Liu, and Xiaoyi Zeng. 2023. Make: Vision-language pre-training based product retrieval in taobao search. In *Companion Proceedings of the ACM Web Conference 2023*. 356–360.

[62] Hongguang Zhu, Yunchao Wei, Yao Zhao, Chunjie Zhang, and Shujuan Huang. 2023. Amc: Adaptive multi-expert collaborative network for text-guided image retrieval. *ACM Trans. Multimedia Comput. Commun. Appl.* 19, 6 (2023), 1–22.

[63] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. 2020. Sean: Image synthesis with semantic region-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5104–5113.

[64] Mingchen Zhuge, Dehong Gao, Deng-Ping Fan, Linbo Jin, Ben Chen, Haoming Zhou, Minghui Qiu, and Ling Shao. 2021. Kaleido-bert: Vision-language pre-training on fashion domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12647–12657.

Received 26 August 2023; revised 18 November 2023; accepted 28 December 2023