Neural Computing and Applications (2023) 35:5889–5902
https://doi.org/10.1007/s00521-022-07994-9

ORIGINAL ARTICLE

Dual attention composition network for fashion image retrieval with attribute manipulation

Yongquan Wan[1,3] · Guobing Zou[1] · Cairong Yan[4] · Bofeng Zhang[2,5]

Received: 10 May 2022 / Accepted: 21 October 2022 / Published online: 9 November 2022

Abstract

Due to practical demands and substantial potential benefits, there is growing interest in fashion image retrieval with attribute manipulation. For example, if a user wants a product similar to a query image and has the attribute "3/4 sleeves" instead of "short sleeves" he can modify the query image by entering text. Unlike general items, fashion items are rich in categories and attributes, and some items with different attributes have only very subtle differences in vision. Moreover, the visual appearance of fashion items changes dramatically under different conditions, such as lighting, viewing angle, and occlusion. These pose challenges to the fashion retrieval task. Therefore, we consider learning an attribute-specific space for each attribute to obtain discriminative features. In this paper, we propose a dual attention composition network for image retrieval with manipulation, which addresses two important issues, where to focus and how to modify. The dual attention module aims to capture fine-grained image-text alignment through corresponding spatial and channel attention and then satisfy multi-modal composition through corresponding affine transformation. The TIRG-based semantic composition module combines the query image's attention features and the manipulation text's embedding features to obtain a synthetic representation close to the target image. Meanwhile, we investigate the semantic hierarchy of attributes and propose a hierarchical encoding method, which can preserve the associations between attributes for efficient feature learning. Extensive experiments conducted on three multi-modal fashion-related retrieval datasets demonstrate the superiority of our network.

Keywords Similarity learning · Image retrieval · Attention mechanism · Attribute manipulation · Fashion

1 Introduction

Image retrieval is a fundamental task for a wide range of applications, such as fashion retrieval [1–3], plagiarized detection [4], person re-identification [5, 6], and internet search [7, 8]. Traditional image retrieval systems only allow users to use text or image queries to express their search intent. However, a key limitation of these systems is that it is difficult for users to express their real search intention through a single image or textual query. Therefore, combined query-based image retrieval (CQBIR) has been proposed and gained increasing attention [9–11], which combines images and feedback for image retrieval in order to allow users to flexibly and accurately express their search intent. Several researches have explored incorporating different types of user feedback, such as relative attributes [9–11], sketch [12, 13], or spatial layouts [14], in

✉ Yongquan Wan
wanyq@gench.edu.cn

Guobing Zou
gbzou@shu.edu.cn

Cairong Yan
cryan@dhu.edu.cn

Bofeng Zhang
bfzhang@sspu.edu.cn

[1] School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China

[2] School of Computer and Information Engineering, Shanghai Polytechnic University, Shanghai 201029, China

[3] School of Information Technology, Shanghai Jian Qiao University, Shanghai 201306, China

[4] School of Computer Science and Technology, Donghua University, Shanghai 201620, China

[5] School of Computer Science and Technology, Kashi University, Xinjiang 844006, China

retrieval tasks to help obtain better retrieval results. CQBIR is a new but challenging task.

In this work, we investigate the CQBIR task, explicitly focusing on attribute-manipulated image retrieval in the fashion domain. Attribute-manipulated fashion image retrieval [15–17] has gained extensive research interest in recent years, which refers to retrieving "similar" fashion items but changing some attributes to meet the user's search intent while preserving the rest of the attributes. Fashion semantics can be expressed through clothing attributes, and attribute recognition can help understand high-level semantic concepts. In attribute-manipulated fashion image retrieval, users can describe modifications to query images with these attributes. For example, in Fig. 1a, the user wants longer sleeves and a higher neckline, while in Fig. 1b, the user wants "red" and "more flowers." In Fig. 1c, the user wishes to modify the local attributes and the overall style. It presents a challenging multi-modal learning problem that requires a collaborative understanding of visual and linguistic content at different granularities.

Hierarchical tree structures, such as concept ontology [18] and semantic hierarchy [19, 20], are often used to organize image categories and exploit inter-task correlations to improve visual recognition. These methods are unsuitable for direct application to fashion image recognition because fashion images have more prosperous attributes and different attribute values for the same attribute. Unlike general items, fashion products usually share a lot of similar attributes in appearance, and the differences between products can be subtle, e.g., the different lengths of sleeves such as "long sleeves," "puff sleeves," and "sleeveless." On the other hand, the visual appearance of even the same product can vary significantly with changes in lighting, viewing angles, occlusions, and background conditions which has posed a challenge for extracting more representative features. Even fashion images with the same category have significant intra-class differences between them. Therefore it is attractive to design novel deep learning algorithms to learn attribute-based hierarchical features for fashion discrimination.

However, joint modeling for combined query image retrieval tasks is non-trivial due to the following challenges. (1) Attribute-like text in the query usually describes some expected modifications to the image, including a modification of a specific attribute related to a local area ("long sleeve") or a modification of an abstract attribute related to the overall style of the image ("casual"). It poses the first challenge to our task: how to identify where an image needs attention based on attribute descriptions and images. This challenge involves how identifying local parts and global styles of products. (2) The reference image modified with text attributes should be more similar to the target image, which requires the model to calculate the similarity between the query image's combined features and the target image's features. It presents a second challenge, how to compute salient regions associated with modified attributes to close the modal gap between query and target images.

We propose a dual attention composition network (DACNet) to address these challenges in a unified solution. It learns more representative features for attribute-manipulated fashion image retrieval by integrating hierarchical attribute tree, spatial and channel attention, multi-task

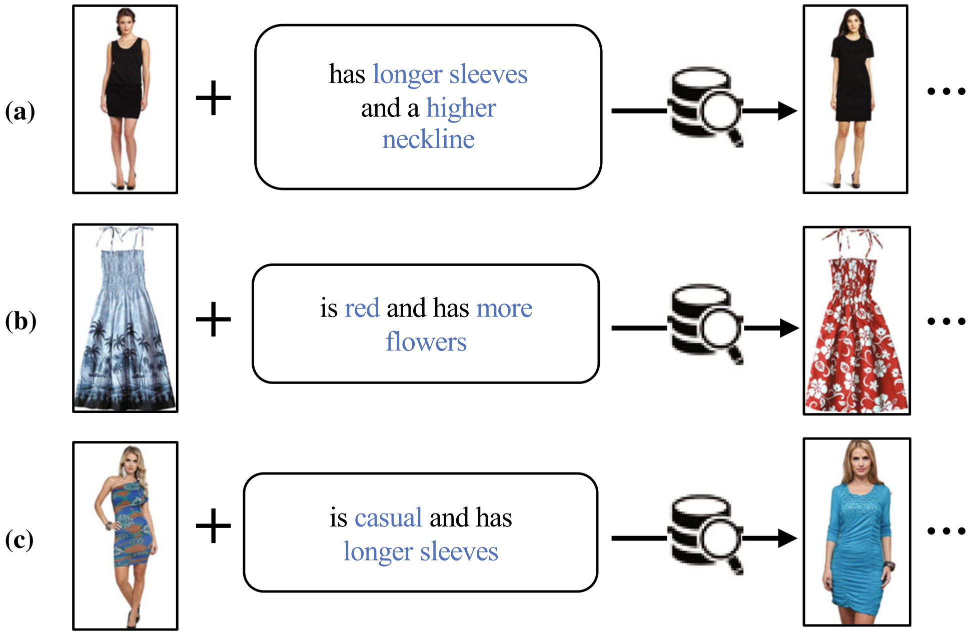

(a)
has longer sleeves and a higher neckline
(b)
is red and has more flowers
(c)
is casual and has longer sleeves
...

Fig. 1 Three examples of image retrieval with attribute manipulation. The query conditions are listed on the left, and the results that meet user requirements are on the right

learning, and gated-residual connections. DACNet consists of 4 key components. An image encoder extracts intermediate image representation, and an attribute encoder obtains the embedding of the hierarchical attribute tree. The semantic feature attention captures fine-grained image-text alignment through corresponding spatial and channel attention and then satisfies multi-modal composition through corresponding affine transformation. We adopt a refined structure of TIRG [21] to combine the features of the modified attributes and the attention features of the query image to obtain feature representations similar to the target image. Finally, the output of the network is used as final query representation to retrieval target image.

The main contributions of this paper are summarized as follows:

- We propose a dual attention composition network for the attribute-manipulated fashion image retrieval task, which can learn fine-grained multi-modal composition representation of images and attributes.
- We design two modules, the dual attention module (DAM) and semantic composition module (SCM), to decompose the task of attribute-manipulated image retrieval into two steps: where to focus and how to modify.
- We investigate the semantic hierarchy of attributes and propose a hierarchical encoding method, which can preserve the associations between attributes for efficient feature learning. We demonstrate that learning attribute-driven representation improves the effectiveness of the model.

The remainder of this paper is organized as follows. Section 2 briefly reviews the related work. Section 3 details the proposed network DACNet and its main modules. The experimental results and analysis are presented in Sect. 4, followed by the conclusion and future work in Sect. 5.

2 Related work

This section briefly reviews the most related works on product search and image retrieval, composed query-based image retrieval, and attention mechanism.

2.1 Product search and image retrieval

Content-based image retrieval is an important problem in computer vision and has attracted wide attention due to its convenience and accuracy [1, 22]. Most modern image retrieval methods use deep learning and have made great progress. Depending on the query type, image retrieval can be divided into many categories, such as text and image matching [23, 24] or image to image retrieval [17, 25]. The key point of image-text matching is how to measure the similarity between visual and textual inputs accurately. The key point of image-to-image matching is to compute the similarity between the query image representation and the target image representation. Gao et al. [24] split fashion images into patches and used pre-trained BERT models to learn advanced representations of text and images to solve fine-grained image and text matching problems in cross-modal retrieval. Recently, interactive multi-modal techniques have been applied in image retrieval. Guo et al. [26] proposed a dialogue-based image retrieval method that simulates users providing natural language feedback to improve the ranking of target images.

Unlike existing methods that mainly study image retrieval based on text feedback, in this work, we pursue the problem of image search with attribute feedback, relying on supervised signals of fashion attributes to learn latent embeddings of images.

2.2 Composed query-based image retrieval

CQBIR aims to input a reference image and some additional textual conditions to retrieve semantically related images [27–29]. The main difficulty of CQBIR is effectively manipulating the attributes of the reference image according to the additional text provided while maintaining the image similarity and making the feature difference between the query feature and the target image smaller. Researchers have proposed many methods to learn visual-semantic embeddings to address the above problems. Vo et al. [21] first introduced the CQBIR task into the fashion domain and proposed a residual gating operation to fuse image and text embeddings. Chen et al. [30] leveraged a composite transformer to transform and preserve visual features conditioned on text semantics, which is plugged at multiple layers inside a CNN (convolutional neural network) to learn composite representations of images and texts. Murrugarra-Llerena et al. [13] proposed a reinforcement learning framework where both the user and the system can drive interactions through four forms of input: free-form attribute feedback, attribute-based questions chosen by the system, a sketch of the target image, and labeling the image as relevant or irrelevant. Lee et al. [31] built an image-text compositor that modulates the content and style of a reference image separately based on a given modifier text. Wen et al. [32] designed a comprehensive linguistic-visual composition network that uses local visual descriptors and global reference images to combine images and text and then uses a phase enhancement module to facilitate sharing knowledge. To retrieval image with natural language feedback, Li et al. [33] employed the pre-trained DistilBERT model and the ResNet-50 model to extract the low-level features of images and semantic concepts in text descriptions and fused them by a residual gating mechanism.

Despite the remarkable progress of these efforts, these text conditional models cannot handle various modification needs well, including both concrete and abstract attribute changes. In our task, the study of learning representations involves combining image and attribute inputs, selectively modifying relevant image features, and retaining unmodified features for fine-grained image retrieval.

2.3 Attention mechanism

Attention aims to mimic the human ability to concentrate on salient information selectively. With the development of deep neural networks, attention mechanisms have been widely used in various vision and language tasks, including image captioning [34], image classification [35], visual question answer (VQA) [36], and image retrieval [22, 37, 38]. Li et al. [34] proposed a joint attribute detection and visual attention framework for clothes image captioning. Ferreira et al. [35] built a visual semantic attention model guided by body pose for multi-label fashion classification. Ji et al. [22] used a tag-based attention mechanism and a context-based attention mechanism to improve cross-domain retrieval of fashion images.

The attention mechanism is also applied with the CQBIR task to fuse visual and textual features. Zhang et al. [39] leveraged a text-guided attention mechanism to facilitate the interaction between visual and semantic information to learn attention-based interpretable decisions from clinical knowledge. Most notably, Ma et al. [40] proposed an attribute feature embedding network, which learns attribute-based embedding in an end-to-end manner to measure the attribute-specified fine-grained similarity of fashion items and get the state-of-art performance. The proposed attributes-aware spatial attention (ASA) and attribute-aware channel attention (ACA) are keys to the network.

Inspired by these methods, we propose a visual-attribute attention framework, which learns the interaction of visual features and attribute features. Different from previous work, our method mainly relies on the regional and channel features of the image, avoiding the reliance on pre-trained object detectors, and can be well suited for fine-grained image retrieval.

3 Methodology

Figure 2 presents an overview of our DACNet framework. It contains four components: (a) an image encoder for image representation, (b) an attribute encoder for hierarchical attributes embedding, (c) a dual attention module (DAM) that captures the salient regions in image data at the supervise of attributes, and (d) a semantic composition module (SCM) that compute the composed feature representation using gating and residual functions.

3.1 Problem formulation

We formulate the attribute-manipulated retrieval task as follows. Given a query image I_{ref}, which has associated attribute values $A_{\text{ref}} = (a_q^1, a_q^2, \ldots, a_q^{J_a})$, and a modified attribute T_m, the goal is to find a target image I_{tgt} whose attribute is $A_{\text{tgt}} = (a_t^1, a_t^2, \ldots, a_t^{J_a})$, and the difference between A_{tgt} and A_{ref} is only the attribute T_m. Note that to simplify notation, we combine all attribute values for different attributes into a single list $A = (a^1, a^2, \ldots, a^{J_a})$, where J_a is the total number of all available attribute values. For each $a^j (1 \le j \le J_a)$, we use hierarchical embedding described in the following section.

3.2 Multi-modal input

DACNet first learns semantic associations between visual data (input images) and textual data (input attributes) for encoding. Since two modalities are involved, we first introduce the encoding method for each modality.

3.2.1 Image representation

We use the ResNet50 network [41] pre-trained on the ImageNet dataset as a backbone network. To preserve the spatial information of the image, we remove the last fully connected layer in the CNN. The output of the penultimate layer is a mid-level representation of the image, which will be for attention computation. So the image feature is represented as $I \in R^{c\times h\times w}$, where $h \times w$ is the size of the feature map and c indicates the number of channels.

3.2.2 Attribute representation

The categories and attributes of fashion products provide clues to perceive the similarity of fashion products. As shown in Fig. 3, tops are associated with attributes such as “neckline” and “sleeve length,” while pants do not have these attributes. Several values are available for the “sleeve length” attribute: “short sleeves,” “3/4 sleeves,” and “long sleeves,” reflecting subtle differences between similar products. Due to the descriptive power of semantic attributes in capturing subtle local differences in fine-grained images, we organize fine-grained fashion classes and attributes hierarchically using a fashion concept tree. The leaf nodes represent attribute values, the parents of the leaf nodes represent attribute concepts, and higher-level

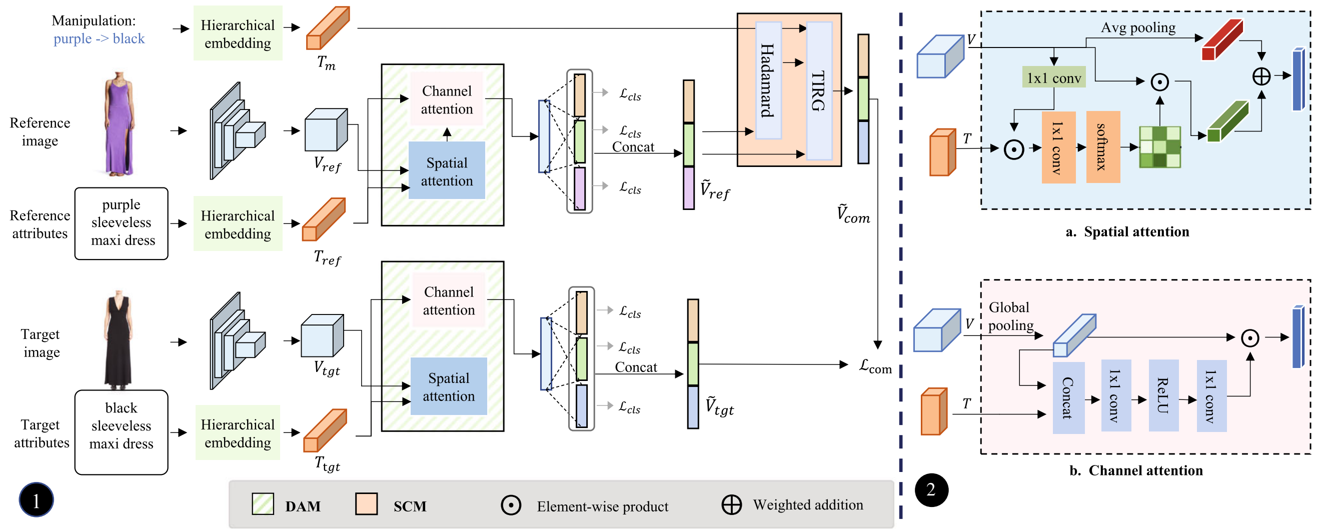


Fig. 2 (1) Overall structure of our proposed DACNet. For a triplet $(I_{\text{ref}}, I_{\text{tgt}}, T_m)$, the image and attribute encoder obtains a representation of images and attributes (Sect. 3.1). **a** DAM computes the attribute-guided spatial attention and channel attention of the reference image I_{ref} and learns the fine-grained representation $\widetilde{V}_{\text{ref}}$ by classification loss L_{cls} (Sect. 3.2). **b** SCM combines the features of $\widetilde{V}_{\text{ref}}$ and modified attribute T_M to get V_{com}, and then matches it with the target image representation $\widetilde{V}_{\text{tgt}}$ (L_{com}) (Sect. 3.3). The whole model is trained with the joint loss of L_{cls} and L_{com} (Section 3.4). (2) DAM consists of two sub-modules: Spatial Attention and Channel Attention

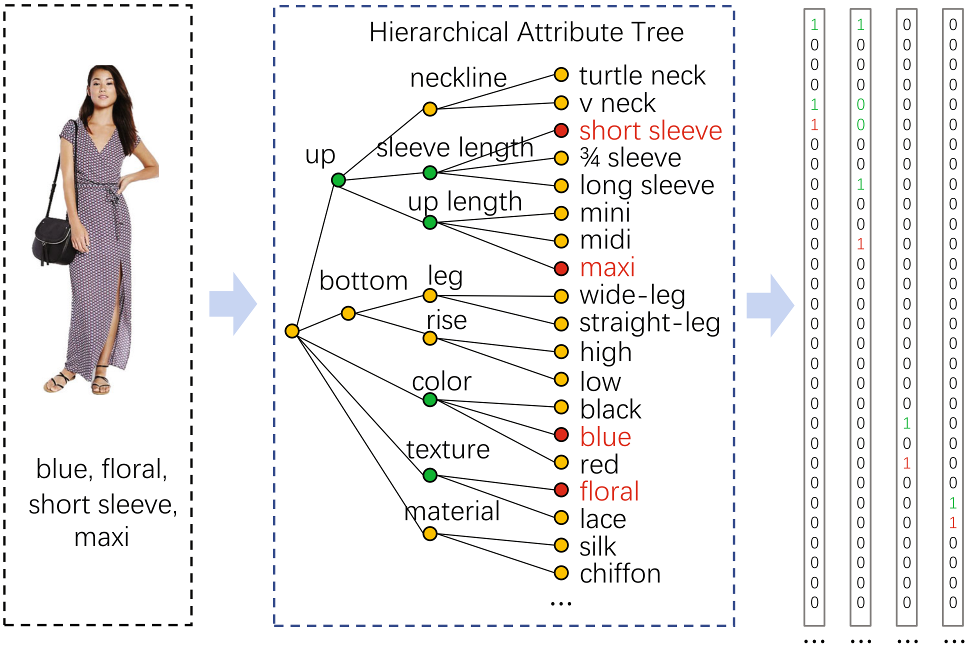
parent nodes represent classes at different granularities. We propose a hierarchical embedding representation of attributes to represent the leaf nodes in the tree and introduce path information from the root node to that leaf node in the embedding representation of each leaf node. The hierarchical embedding representation approach has obvious advantages in fashion recognition. It explicitly represents interrelated categories and attributes. In addition, it achieves a more discriminatory feature representation [25, 42, 43].

Figure 3 shows an example of an image's concept tree and hierarchical embedding. To describe the hierarchical relationship between attributes, we define an indicator "<". Assume that p and s are two nodes in the attribute tree, "$p < s$" means that "s" is the ancestor of "p" in the tree. Then we define $e_{p,s} = 1$ if $p < s$ or $p = s$. The hierarchy embedding of attribute a is defined as :

$$E(a^i) = [e_{a,1}, e_{a,1}, \ldots, e_{a,Q}], \tag{1}$$

where Q is the number of total nodes in the attribute tree.

Fig. 3 An illustration of hierarchical attribute embedding



All attributes of image I can be expressed as $H(A) \in R^{L \times Q}$, where L is the number of all attributes, and $H(A)$ is expressed as:

$$H(A) = [E(a^1), E(a^2), \ldots, E(a^L)]. \tag{2}$$

3.3 Dual attention module

Attention mechanisms according to several implementations using spatial and channel information are widely used in image captioning [44], image classification [45], and image retrieval [40]. In attribute manipulation fashion retrieval, the user's desired modifications may focus on smaller local areas or the global style of the entire image. The spatial attention mechanism can focus on the salient regions in an image. Meanwhile, the channel attention mechanism is considered a process of selecting important global semantic attributes. In this work, we use spatial and channel attention jointly to extract attribute-driven features for classifying and retrieving fashion images.

The dual attention module aims to highlight attribute-guided salient regions in the image. Our DAM module consists of two main parts: (1) spatial attention and (2) channel attention. The first captures the critical areas of images affected by concrete properties, while the second captures the importance of abstract properties.

DACNet adapts the image feature map I_{ref} to the context of the attributes H through spatial attention and channel attention to fuse the features of images and attributes. The attribute-guided feature is demonstrated as $\widetilde{V}_{\text{ref}}$. The attribute-guided visual features $\widetilde{V}_{\text{tgt}}$ of the target image are also computed in the same way:

$$\widetilde{V}_{\text{ref}} = F_{\text{ch}}(S_{\text{ref}}, A_{\text{ref}}), \quad S_{\text{ref}} = F_{\text{sp}}(I_{\text{ref}}, A_{\text{ref}}), \tag{3}$$

$$\widetilde{V}_{\text{tgt}} = F_{\text{ch}}(S_{\text{tgt}}, A_{\text{tgt}}), \quad S_{\text{tgt}} = F_{\text{sp}}(I_{\text{tgt}}, A_{\text{tgt}}), \tag{4}$$

where I_{ref} and I_{tgt} are the intermediate features of the reference image and target image, and $\widetilde{V}_{\text{ref}}$ and $\widetilde{V}_{\text{tgt}}$ are their corresponding visual text features. F_{sp} and F_{ch} are spatial attention and channel attention, which will be introduced below.

3.3.1 Spatial attention

Since spatial location information generally contains important features for accurate object identification, spatial attention mechanisms attempt to pay more attention to semantic-related regions. Without loss of generality, we discard the subscript ref and tgt in the following description. We engage spatial attention to capture long-range correlations within the visual embeddings I guided by attributes H to obtain the fused representation. First, we transform the image features and attribute embeddings to have the same dimensionality. For the image representation, we apply a convolutional layer followed by a nonlinear activation function to obtain the mapped image features $p(x) \in \mathrm{R}^{c' \times h \times w}$.

$$p(I) = \tanh(\text{conv}_{c'}(I)), \tag{5}$$

where $\text{conv}_{c'}$ refers to a convolutional layer with c' 1×1 convolution kernels. For attribute, we first project the hierarchical attribute embedding $H(I)$ into a c' dimensional vector implemented by a fully connected (FC) layer and then expand it to the same dimension as the image feature. Hence, the mapped attribute $p(A) \in \mathrm{R}^{c' \times h \times w}$ is expressed as:

$$p(A) = W_{s2} \tanh(W_{s1} H(A)), \tag{6}$$

where $W_{s1} \in \mathrm{R}^{c' \times Q}$ is a transformation matrix and $W_{s2} \in \mathrm{R}^{c' \times h \times w}$ is a all one matrix. Then the spatial attended weight $\alpha_s \in \mathrm{R}^{h \times w}$ is computed as:

$$\alpha_s = \text{softmax}(\tanh(\text{conv}_1(p(a) \odot p(x)))), \tag{7}$$

where $\otimes$ is the Hadamard element-wise product; conv_1 is a 1×1 convolution kernel. Finally, through the following equation, we can get the spatial attention vector $V_s \in \mathrm{R}^c$.

$$V_s[j] = \alpha_s[j] \odot I[j]. \tag{8}$$

We apply the global average pool to the visual feature I to obtain the pooled attention visual feature. Finally the attribute-conditioned visual embedding S_{ref} is obtained by the weight addition of the pooled feature and the attribute-guided image feature V_s.

$$S_{\text{ref}} = \text{GAP}(I) + V_s. \tag{9}$$

3.3.2 Channel attention

Each convolutional filter in the CNN network acts as a pattern detector. The corresponding convolutional filter activates each feature map channel in the CNN. Thus the channel attention mechanism can be thought of as selecting important semantic attributes. To measure the importance of abstract properties, we further apply channel attention to the output features of spatial attention. This module is jointly driven by features from the spatial attention module and hierarchical attribute embeddings. Considering the different purposes of the two attentions, we use a separate attribute embedding layer for channel attention. Channel attended vector $V_c \in \mathrm{R}^c$ is obtained through the channel attention module. An FC layer embeds A into an embedding vector $q(A)$ with the same dimensionality of S_{ref}.

$$q(A) = \delta(W_c H(A)), \tag{10}$$

where $W_c \in \mathrm{R}^{c\times Q}$ represents the transformation matrix, and δ refers to ReLU activation function. Two consecutive FC layers are employed in sequence to obtain channel attention weight $\alpha_c \in \mathrm{R}^c$.

$$\alpha_c = \sigma(W_{c2}\delta(W_{c1}[q(A), S_{\text{ref}}])), \tag{11}$$

where [,] refers to concatenation operation, $W_{c1} \in \mathrm{R}^{\frac{c}{r}\times c}$ and $W_{c2} \in \mathrm{R}^{c\times\frac{c}{r}}$ refer to the transformation matrices, r is the reduction rate, σ is ReLU function. To limit the complexity of the model, following the approach of [46], we implement two FC layers by a dimensionality-reduction layer with parameters W_{c1} with reduction ratio r, a ReLU and dimensionality-increasing layer with parameter W_{c2}, which have fewer parameters than one FC layer. The channel attention vector $\widetilde{V}_{\text{ref}} \in \mathrm{R}^c$ is obtained by the element-wise multiplication between α_c and S_{ref} shown as:

$$\widetilde{V}_{\text{ref}} = \alpha_c \otimes S_{\text{ref}}. \tag{12}$$

The semantic feature embedding $\widetilde{V}_{\text{tgt}}$ of the target image is also computed by the same attention mechanism. The obtained query image feature $\widetilde{V}_{\text{ref}}$ is passed to the semantic modification module, which is expected to be closer to $\widetilde{V}_{\text{tgt}}$ in the embedding space.

3.4 Semantic composition module

Here, we combine the image's semantic feature $\widetilde{V}_{\text{ref}}$ and the modification attribute's feature T_m to make the combined feature as similar as possible to the correct target image's feature I_{tgt}. Inspired by TIRG [21], we perturb the gated features to obtain combined features with a residual connection.

$$\tilde{t}_m = \widetilde{V}_{\text{ref}} \odot T_m, \tag{13}$$

$$V_{\text{com}} = W_g f_{\text{gate}}\left(\widetilde{V}_{\text{ref}}, \tilde{t}_m\right) + W_r f_{\text{res}}\left(\widetilde{V}_{\text{ref}}, \tilde{t}_m\right), \tag{14}$$

where $\odot$ is Hadamard element-wise product, $f_{\text{gate}}, f_{\text{res}} \in \mathrm{R}^{c\times h\times w}$ are the gating and residual connections, and W_g and W_r are learnable parameters. The gating connection is computed by:

$$f_{\text{gate}}\left(\widetilde{V}_{\text{ref}}, \tilde{t}_m\right) = \sigma\left(W_{g2} * \mathrm{ReLU}\left(W_{g1} * [\widetilde{V}_{\text{ref}}; \tilde{t}_m]\right) \odot I_{\text{ref}}\right), \tag{15}$$

where W_{g1}, W_{g2} are 3×3 convolution filters, σ is the sigmoid function, $*$ represents 2d convolution with batch normalization. The residual connection is computed by :

$$f_{\text{res}}\left(\widetilde{V}_{\text{ref}}, \tilde{t}_m\right) = W_{r2} * \mathrm{ReLU}\left(W_{r1} * \left[\widetilde{V}_{\text{ref}}; \tilde{t}_m\right]\right). \tag{16}$$

3.5 Loss function

We construct triples $(I_{\text{ref}}, T_m, I_{\text{tgt}})$ from the training dataset. Correspondingly, V_{com} represents the combined text-conditioned image embedding of (I_{ref}, T_m), and $\widetilde{V}_{\text{tgt}}$ represents the latent embedding of I_{tgt}. Consider another negative image I_{neg} sampled from the training set, s.t. $I_{\text{neg}} \notin I_{\text{ref}} \cup I_{\text{tgt}}$, where $\widetilde{V}^{-}_{\text{tgt}}$ represents its latent visual embedding generated using the same pipeline as $\widetilde{V}_{\text{tgt}}$. We adopt semi-hard mining method [47] to select the negative pair I_{neg}. We next explain the different loss functions for training the model.

3.5.1 Triplet loss

Our ultimate goal is to align the composite output V_{com} with the target image $\widetilde{V}_{\text{tgt}}$, and we formulate a triplet loss to encourage the combined representation V_{com} to be close to the correct target representation with the desired properties. The triplet loss is defined as:

$$L_{\text{com}} = \max\left(0, d(V_{\text{com}}, \widetilde{V}^{+}_{\text{tgt}}) - d(V_{\text{com}}, \widetilde{V}^{-}_{\text{tgt}}) + m\right), \tag{17}$$

where m is the margin parameter, empirically set to 0.2 in our experiments, and $d(\cdot)$ is the L2 distance.

3.5.2 Auxiliary classification loss

To minimize the intra-class distance and maximize the inter-class distance, we propose an auxiliary task to predict attributes in a hierarchical manner. The attribute tree we created is based on the analysis of the dataset. To generate an attributes tree, we sorted all attribute labels from different datasets (e.g., FashionIQ and Fashion200k). The fashion attribute tree organizes attribute concepts from general to specific to form a hierarchical structure. The non-leaf nodes in the tree are product categories and attribute concepts, and the leaf nodes are attribute values. We prune the attribute tree to fit most datasets, keeping eight common attribute concepts: sleeve length, collar type, neckline, top length, color, texture, shape, and material. The concept of the attribute tree can be generalized to other objects (such as shoes) with multi-class attributes. The feature map output by SCM is divided into 8 equal parts, and each branch represents an attribute concept. The corresponding attribute concept layer is followed by a fully connected layer and a softmax loss layer for classification. If there is no professional definition of fine-grained clothing semantics such as sleeve length and waist height, there will be no corresponding bifurcation parallel prediction and structure.

The classification of each attribute sub-space is done as an independent multi-class attribute classification task, and we adopt a cross-entropy loss to supervise the training of these sub-spaces as follows:

$$L_{\text{cls}} = \frac{1}{N}\sum_{n=1}^{N}\sum_{a=1}^{J_a}(l_a), \tag{18}$$

$$l_a = -\left(y_a \cdot \log \hat{y}_a + (1-y_a)\cdot \log(1-\hat{y}_a)\right), \tag{19}$$

where N is the number of samples in the training set, J_a is the number of attributes, y_a is the actual label of the a-th attribute for n-th image, and $\hat{y}_a$ is the prediction score.

3.5.3 Total loss

It is computed as a weighted sum of them :

$$L_{\text{total}} = L_{\text{com}} + \lambda L_{\text{cls}}, \tag{20}$$

where λ is a trade-off hyperparameter. By jointly optimizing our network with the triplet loss and classification loss, we could finally learn a modality-invariant representation for the composed query-based image retrieval task.

4 Experimental results

We aim to answer the following research questions:

RQ1: Can our proposed DACNet perform better than other competitive methods on the attribute-manipulated fashion retrieval task?

RQ2: Are the key components in DACNet helpful for improving retrieval results?

RQ3: What is the visual effect of the attention mechanism in the model?

4.1 Experimental settings

4.1.1 Datasets

We evaluate the performance of DACNet on attribute-manipulated image retrieval tasks on three public fashion datasets: FashionIQ [48], Fashion200k [9], and Shoes [26].

FashionIQ [48] is a widely used interactive fashion product retrieval dataset. It contains 77,684 images crawled from https://www.amazon.com, that are split into three distinct categories: *Dress*, *Toptee*, and *Shirt*. Among the 46,609 images in the training set, there are 18,000 image pairs. Each pair is accompanied by two human-written natural language sentences describing different attributes between the reference image and the target image. For example, "*is shorter and has longer sleeves*" and "*has a v-neck and is lighter colored.*" The test set in FashionIQ is unavailable, so we run evaluations on the validate set. We report results on individual categories and the averaged results over three categories.

Fashion200k [9] is a large-scale dataset with over 200,000 fashion images. Each image is accompanied by descriptive information such as product descriptions and bounding boxes of clothes. The description is a list of attributes, such as "*gray kneelength dress*" or "*red high-rise satin trousers.*" Following [21], we pair two images with a one-word difference in the description as reference and target images. The training dataset contains about 172,000 images, and the test set contains 33,480 queries.

Shoes [49] is a dataset proposed for attribute discovery task and is further annotated by [26] with relative caption for dialog-based interactive retrieval. Following [26], we train the model with 10,000 images and evaluate the model with 4658 images.

4.1.2 Baseline methods

To verify the effectiveness of our proposed model on the task of attribute-manipulated image retrieval, we compare it with the following representative multi-modal learning methods.

- *Image_only* treats only the features of the reference image as the combined query representation.
- *Text_only* treats only modification attribute embedding as combined query representation.
- *Concatenation* concatenates the query image and the representation of the attribute and sends it to a 2-layer MLP (follow details from TIRG) to get the combined query representation.
- *TIRG* [21] is the first to propose image retrieval combining images and text, using gating and residual connections to fuse the two to obtain combined query representation.
- *VAL* [30] applies the attention mechanism to combine the visual and textual representations of multiple CNN layers.
- *COSMO* [31] introduces a disentangled multi-modal non-local block for content modification and a style modulator for style modification.

4.1.3 Implementation details

We use ResNet50 [41] trained on ImageNet [50] as the backbone network of the image encoder and remove the final pooling layer to preserve the spatial information of the feature maps. Accordingly the mid-level representation size of the image is 2048 × 14 × 14. We apply L2 normalization to the image and attribute embeddings. Each training batch contains 32 triples and is shuffled at the beginning of each training epoch. We use the SGD optimizer with a learning rate of 0.0001 and a decay of 0.95 at every epoch. The parameter λ in Eq. 20 is set to 0.2. We conduct all the experiments in PyTorch and fix the random seeds to ensure repeatability.

4.1.4 Evaluation metric

Following previous work on these datasets, we report the recall (recall@k) at rank K as an evaluation metric, specifically recall@10 and recall@50. Each method employs the same basic encoder. We run all experiments three times independently and report the average.

4.2 Attribute-manipulated retrieval results (RQ1)

Tables 1, 2, and 3 show quantitative comparisons with the latest benchmarks in terms of R@1, R@10, and R@50 on the FashionIQ, Fashion200k, and Shoes datasets, respectively. Our results are shown at the bottom of the tables. For convenience, we highlight the best figures in bold in the table. First, our proposed method, DACNet, outperforms all baseline methods on all datasets. For FashionIQ, DacNet improves over the previous best method, COSMO, by 7% and 10% in R@10 and R@50. The results on Fashion200k further validate the effectiveness of DACNet in the attribute-manipulated image retrieval task. For Fashion200k, we are constructing two images with only a one-word difference in the two descriptions as query-target pairs, so the target images in Fashion200k are not unique. DACNet improves over TIRG by 21% and 17% on R@10 and R@50 metrics. We improve 2% and 3% over the second best method at R@10 and R@50. On the Shoes dataset, the retrieval results show a similar pattern to the FashionIQ results, but with a slightly lower increase. Specifically, DACNet improves 3% and 4% over the previous best method for R@10 and R@50.

Second, these methods perform similarly to FashionIQ on the Fashion200k and shoes datasets. The worst retrieval results are obtained for image_only and text_only, which are trained by only one modal data in the query. Although concatenation only combines features from two modalities through a simple MLP, it also achieves improvement on all datasets. The other baseline models (TIRG, VAL, and COSMO) all obtain a representation of the query through different feature composition methods. It suggests that exploring a learning method that can effectively combine different modal features is necessary for the attribute-manipulated image retrieval task. Compared to these methods, our method achieves better results, which we believe is due to the simultaneous performance of attribute manipulation and modality alignment learning. On the one hand, it learns the relationship between images and attributes well through spatial and channel attention. On the other hand, it learns the interaction between heterogeneous data features of images and text through gating and residual connectivity.

4.3 Ablation study (RQ2)

In this section, we evaluate the feasibility of each component in DACNet, including the model structure and loss function. We also study the effect of hyperparameters in DACNet. For all our ablation experiments, we limit the scope to the FashionIQ dataset for ease of presentation and analysis.

4.3.1 The importance of spatial attention and channel attention

In this experiment, we compare several different arrangements of spatial attention and channel attention. (1) SA + SCM: training the model with only spatial attention and SCM. (2) CA + SCM: training the model with only channel attention and SCM. (3) SCA (parallel) + SCM: training the model with a spatial-channel pipeline structured in parallel and SCM. (4) SCA (Channel-spatial) + SCM: training the model with a sequential SCA pipeline (channel-spatial) and SCM. (5) SCA (Spatial-channel) + SCM: training the model with a sequential SCA pipeline (spatial-channel). Each module has a different role, with spatial attention focusing on the local and channel attention focusing more on the global, and the order may affect the overall performance of the model. For the CQBIR task and our approach, it is illogical to remove the SCM to analyze the network, so we omit that in our study.

Table 4 summarizes the experimental results of different attention arrangement methods. From the results, we can find that all the arranged methods outperform the methods using only spatial attention or channel attention, indicating that both types of attention are crucial. We also note that sequentially arranged attention performs better than parallel arranged attention. In addition, spatial-first order performs slightly better than channel-first order. It can be explained as the channel attention module loses a lot of regional image information in extracting global features, resulting in the subsequent spatial attention module being unable to extract accurate features. So we use spatial-channel in the final DACNet model to represent the corresponding model.

4.3.2 The effects of feature combination operators

Here, we study the use of attributes to modify image features based on attribute features. Accordingly, we validate

Table 1 Performance comparison on FashionIQ dataset

Model	Dress		Shirt		Toptee		Average	
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
Image_only	4.86	12.73	5.60	14.21	5.20	13.36	5.22	13.43
Text_only	7.58	21.89	8.64	27.52	8.85	27.84	8.36	25.75
Concat	9.27	27.32	10.00	28.77	8.67	26.28	9.31	27.46
TIRG [21]†	14.87	34.66	18.26	37.89	19.08	39.62	17.40	37.39
VAL ($\mathcal{L}_{vv}$) [30]†	21.12	42.19	21.03	43.44	25.64	49.49	22.60	45.04
VAL (Glove) [30]†	22.53	44.00	22.38	44.15	27.53	51.58	24.15	46.61
COSMO [31]‡	25.64	50.30	24.90	48.18	29.21	57.46	26.58	51.31
DACNet (Ours)	**29.64**	**56.49**	**28.97**	**55.04**	**32.80**	**58.45**	**30.47**	**56.66**

† indicates the results are cited from [30], while ‡ denotes the results from [31]

Table 2 Performance comparison on Fashion200k dataset

Model	R@1	R@10	R@50
Image_only	3.5	22.7	38.3
Text_only	1.0	12.3	21.8
Concat	11.9	39.7	62.6
TIRG†	14.1	42.5	63.8
VAL($\mathcal{L}_{vv}$)†	21.2	49.0	68.8
VAL(Glove)†	22.9	50.8	72.7
COSMO‡	23.3	50.4	69.3
DACNet (Ours)	**25.1**	**51.7**	**74.9**

† indicates the results are cited from [30], while ‡ denotes the results from [31]

Table 3 Performance comparison on Shoes dataset

Model	R@1	R@10	R@50
Image_only	6.07	25.61	47.87
Text_only	0.60	6.20	19.4
Concat	5.70	20.32	39.97
TIRG†	12.60	45.45	69.39
VAL ($\mathcal{L}_{vv}$)†	16.49	49.12	73.53
VAL (Glove)†	17.18	51.52	75.83
COSMO‡	16.72	48.36	75.64
DACNet (Ours)	**25.1**	**54.22**	**78.89**

† indicates the results are cited from [30], while ‡ denotes the results from [31]

Table 4 The effect of combining methods of channel and spatial attention

Method	R@10	R@50
SA + SCM	27.45	52.64
CA + SCM	26.73	51.33
SCA (parallel) + SCM	28.28	54.41
SCA (channel-spatial) + SCM	29.35	55.52
SCA (spatial-channel) + SCM	**30.47**	**56.66**

Table 5 The effect of feature combination operators

Method	R@10	R@50
Concatenation	21.65	46.37
Hadamard product	22.79	47.81
Residual gating	**30.47**	**56.66**

Table 6 The importance of attribute hierarchy

Method	R@10	R@50
One-hot	27.17	50.25
Hierarchy embedding	**30.47**	**56.66**

this design choice by comparing the following operators: concatenation, Hadamard product, and residual gating (defined in Eq. 14). The results are shown in Table 5, which highlights that our operator significantly outperforms other alternatives.

4.3.3 The importance of attribute hierarchy embedding

We also investigate the effect of hierarchical attribute embedding compared to one-hot. Table 6 shows how hierarchical embedding leads to better performance. The natural language description corresponding to each image

Table 7 The impact of loss functions

Composition	R@10	R@50
L_{trip}	25.35	50.42
$L_{\text{trip}} + L_{\text{cls}}$	**30.47**	**56.66**

pair in the FashionIQ dataset involves one or more attributes and, in most cases, no more than four. These text descriptions are not unified. For example, two different descriptions of "is long sleeved" and "has longer sleeves" express the same meaning, i.e., the same attribute value. We build a vocabulary, map natural language descriptions to terms in the vocabulary, and finally represent them as hierarchical embedding representations.

4.3.4 The effect of the loss function

In this section we investigate DACNet in an ablation experiment to demonstrate the effectiveness of different loss functions. Table 7 shows that all losses are required to improve performance. This demonstrates the importance of the image attribute recognition auxiliary task for attribute-manipulated retrieval, as it can help to learn more discriminative image features.

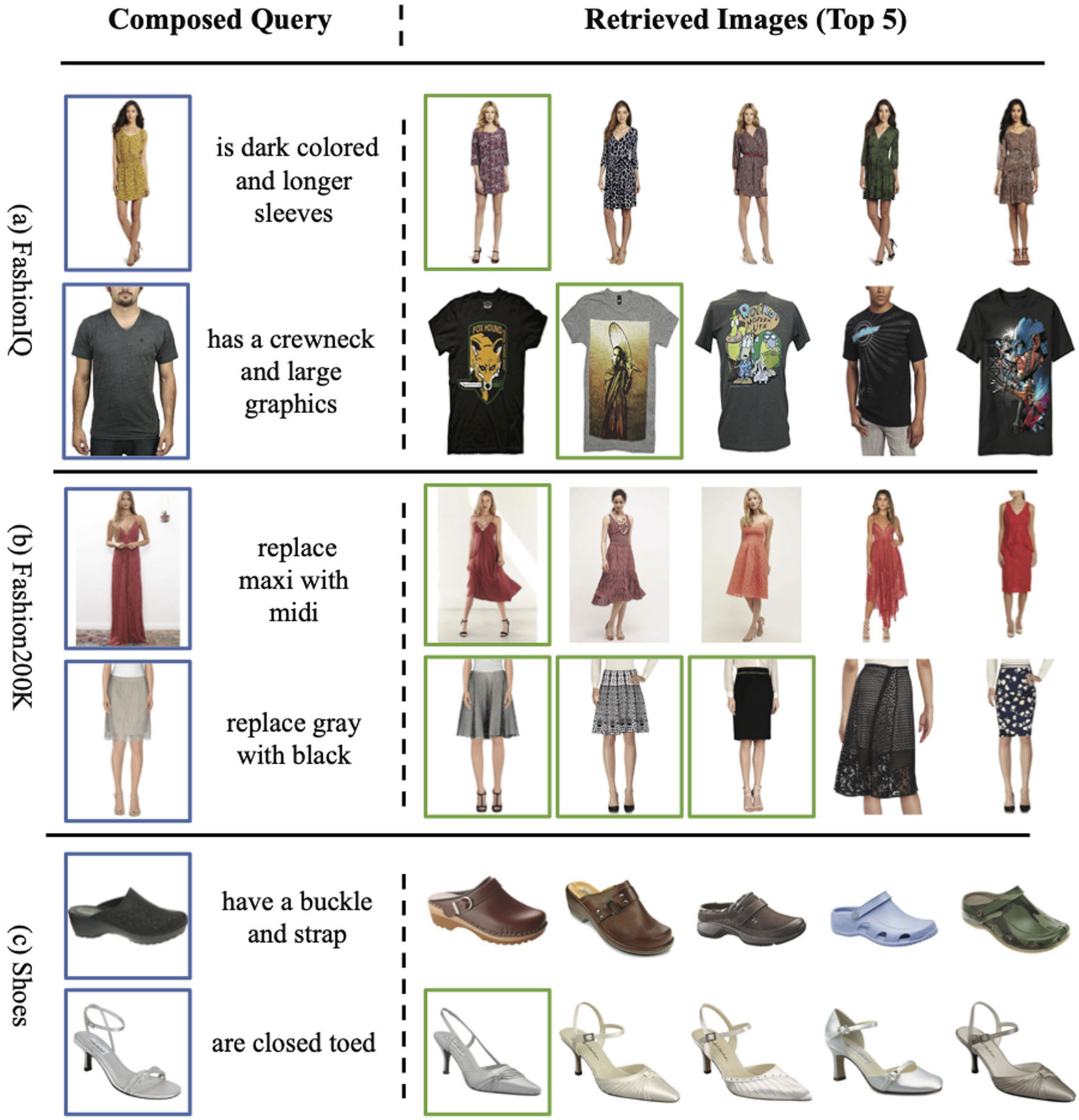


Fig. 4 Quantitative results of image retrieval with attribute manipulation on three public datasets. The query image and modification attribute are on the left. The retrieved images are on the right and ranked from left to right (ground truth is in green contour) (Color figure online)

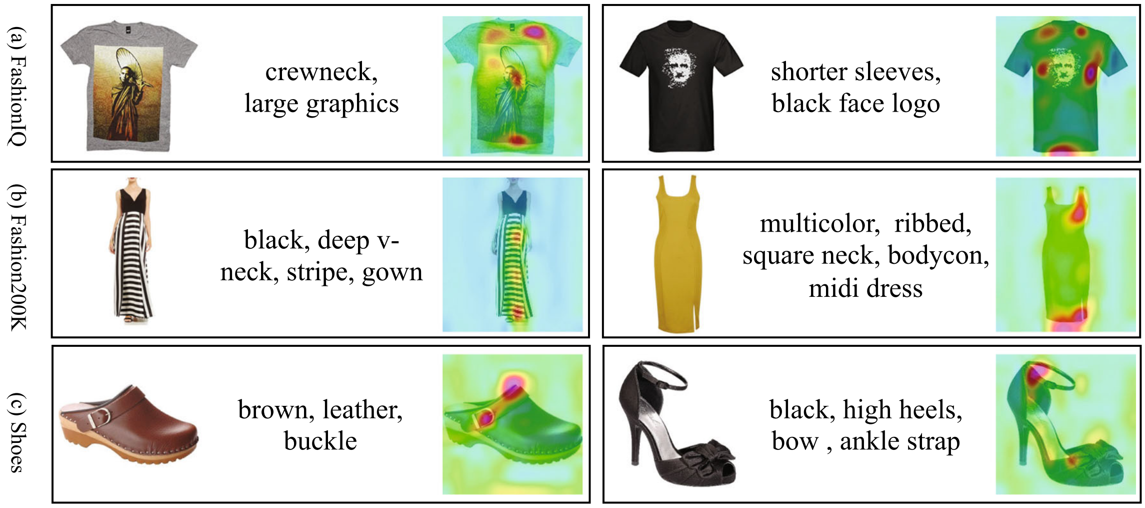

(a) FashionIQ
crewneck, large graphics
shorter sleeves, black face logo
(b) Fashion200K
black, deep v-neck, stripe, gown
multicolor, ribbed, square neck, bodycon, midi dress
(c) Shoes
brown, leather, buckle
black, high heels, bow , ankle strap

Fig. 5 Heatmaps of different attributes on different models

4.3.5 The effect of λ

We evaluate the effect of different settings of the hyperparameter λ in Eq. 20 on top-10 recall. We try λ with values ranging from 0 to 1 with an interval of 0.1. On FashionIQ, the R@10 of DACNet peaks when λ reaches 0.2. With a further increase in λ, R@10 starts to decrease. It suggests that the classification task contributes to good retrieval results and should also be balanced with the retrieval task.

4.4 Case study (RQ3)

Below we observe the effect of DACNet in terms of retrieval visualization and attention visualization.

4.4.1 Image retrieval visualization

We performed a qualitative analysis to demonstrate our quantitative observations, and Fig. 4 illustrates the attribute-manipulated image retrieval results obtained by our DACNet on the three datasets. We report the top 5 images retrieved and indicate the target images in green. We observe that DACNet can merge multiple semantic transformations from attribute descriptions into the visual representation at the same time. The retrieval results show that: (1) DACNet can retrieve new images and modify some attributes based on attribute feedback. For example, the first line of Fig. 4b DACNet modifies skirt length to midi, and the second line of Fig. 4b modifies color to black. (2) DACNet can extract multiple fine-grained semantic concepts from attribute feedback for image retrieval. For example, in the first row of Fig. 4a, "dark colored" and "longer sleeves" represent different granular semantics of the image, and DACNet captures all the semantics in the image and aggregates them effectively. (3) DACNet can jointly understand the concrete and abstract attributes of the image. For example, in the second row of Fig. 4a, DACNet preserves the image's overall appearance, and the view finds variations that satisfy "crewneck" and "graphic." At the same time, we also observe some examples of failure. For example, in the first row of Fig. 4c, DACNet found no target images in the top-5 search results. However, all top-5 retrieval results match the modified text in the query input. Overall, these observations validate the usefulness of DACNet.

4.4.2 Attention visualization

To verify the ability of the attention modules in DACNet to locate areas based on attributes, we visualize the attribute-guided attention function on fashion images. We provide attention maps from the spatial attention module of DACNet in Fig. 5.

As seen from Fig. 5a, the T-shirts pay more attention to the attribute "crewneck," while "large graphics" and "black face logo" focus on the pattern part of the T-shirts. In Fig. 5b, "stripe," "square neck," and "midi" are considered as the most informative attributes corresponding to the images. In Fig. 5c, "buckle" and "ankle strap" are most related to the local area of the shoes. Overall, we can confirm the effectiveness of DACNet in capturing fine-grained attribute-image alignment.

5 Discussion and conclusion

We put forward DACNet, a novel approach to tackle the challenging task of image search with attribute manipulation. DACNet features three critical components, the hierarchical encoding for attributes preserves the associations between attributes, the dual attention module captures fine-grained image-text alignment, and the semantic composition module obtains modified image representation. We validate the efficacy of DACNet on three datasets and demonstrate its superiority in handling various attributes. We also explore the impact of different modules on DACNet performance. Since features of abstract attributes such as style are difficult to align with image features, we will study the feature mapping and combination of abstract attributes and images in future work.

Data availability The data that support the findings of this study are derived from the following public domain resources. 1. FashionIQ can be downloaded from https://github.com/hongwang600/fashion-iq-metadata. 2. Fashion200k can be downloaded from https://github.com/xthan/fashion-200k. 3. Shoes can be downloaded from https://github.com/XiaoxiaoGuo/fashion-retrieval/tree/master/dataset.

Declarations

Conflict of interest We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

References

1. Liu Z, Luo P, Qiu S, Wang X, Tang X (2016) Deepfashion: powering robust clothes recognition and retrieval with rich annotations. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). pp 1096–1104
2. Gu X, Wong Y, Shou L, Peng P, ChenG Kankanhalli MS (2018) Multi-modal and multi-domain embedding learning for fashion retrieval and analysis. IEEE Trans Multimed 21(6):1524–1537
3. D'Innocente A, Garg N, Zhang Y, Bazzani L, Donoser M (2021) Localized triplet loss for fine-grained fashion image retrieval. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 3910–3915
4. Lang Y He Y Yang F, Dong J, Xue H (2020) Which is plagiarism: fashion image retrieval based on regional representation for design protection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp 2595–2604
5. Mansouri N, Ammar S, Kessentini Y (2021) Re-ranking person re-identification using attributes learning. Neural Comput Appl 33(19):12827–12843
6. Li S, Yu H, Hu R (2020) Attributes-aided part detection and refinement for person re-identification. Pattern Recogn 97:107016
7. Li X, Yang J, Ma J (2021) Recent developments of content-based image retrieval (CBIR). Neurocomputing 452(10):675–689
8. Zhang F, Xu M, Xu C (2022) Geometry sensitive cross-modal reasoning for composed query based image retrieval. IEEE Trans Image Process 31:1000–1011
9. Han X, Wu Z, Huang PX, Zhang X, Zhu M, Li Y, Zhao Y, Davis LS (2017) Automatic spatially-aware fashion concept discovery. In: Proceedings of the IEEE international conference on computer vision (ICCV). pp 1463–1471
10. Kovashka A, Devi P, Kristen G (2012) Whittlesearch: image search with relative attribute feedback. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). pp 2973–2980
11. Yu A, Kristen G (2019) Thinking outside the pool: active training image creation for relative attributes. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp 708–718
12. Jifei S, Yi-Zhe S, Tao X, Timothy H, Xiang R (2016) Deep multi-task attribute-driven ranking for fine-grained sketch-based image retrieval. In: Proceedings of the British machine vision conference (BMVC). pp 132–113211
13. Murrugarra-Llerena N, Kovashka A (2021) Image retrieval with mixed initiative and multimodal feedback. Comput Vis Image Underst 207:103204
14. Mai L, Jin H, Lin Z, Fang C, Brandt J, Liu F (2017) Spatial-semantic image search by visual feature synthesis. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 4718–4727
15. Cheng W, Song S, Chen C, Hidayati SC, Liu J (2021) Fashion meets computer vision: a survey. ACM Comput Surv 54(4):1–41
16. Huang J, Feris RS, Chen Q, Yan S (2015) Cross-domain image retrieval with a dual attribute-aware ranking network. In: Proceedings of the IEEE international conference on computer vision (ICCV). pp 1062–1070
17. Kuang Z, Gao Y, Li G, Luo P, Chen Y, Lin L, Zhang W (2019) Fashion retrieval via graph reasoning networks on a similarity pyramid. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV), pp 3066–3075
18. Barz B, Denzler J (2019) Hierarchy-based image embeddings for semantic image retrieval. In: 2019 IEEE winter conference on applications of computer vision (WACV). pp 638–647
19. Zhao J, Peng Y, He X (2020) Attribute hierarchy based multi-task learning for fine-grained image classification. Neurocomputing 395:150–159
20. Narayana P, Pednekar A, Krishnamoorthy A, Sone K, Basu S (2019) Huse: Hierarchical universal semantic embeddings. arXiv:1911.05978
21. Vo N, Jiang L, Sun C, Murphy K, Li L-J, Fei-Fei L, Hays J (2019) Composing text and image for image retrieval-an empirical odyssey. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp 6439–6448
22. Ji X, Wang W, Zhang M, Yang Y (2017) Cross-domain image retrieval with attention modeling. In: Proceedings of the 25th ACM international conference on multimedia (MM). pp 1654–1662
23. Zhang Y, Lu H (2018) Deep cross-modal projection learning for image-text matching. In: Proceedings of the European conference on computer vision (ECCV). pp 686–701
24. Gao D, Jin L, Chen B, Qiu M, Li P, Wei Y, Hu Y, Wang H (2020) Fashionbert: text and image matching with adaptive loss for cross-modal retrieval. In: Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval (SIGIR). pp 2251–2260
25. Liao L, He X, Zhao B, Ngo C-W, Chua T-S (2018) Interpretable multimodal retrieval for fashion products. In: Proceedings of the 26th ACM international conference on multimedia (MM). pp 1571–1579
26. Guo X, Wu H, Cheng Y, Rennie S, Tesauro G, Feris R (2018) Dialog-based interactive image retrieval. In: Proceedings of the conference on advances in neural information processing systems (NIPS). pp 678–688
27. Liu H, Wang R, Shan S, Chen X (2019) What is a tabby? Interpretable model decisions by learning attribute-based classification criteria. IEEE Trans Pattern Anal Mach Intell 43(5):1791–1807

28. Xu Y, Bin Y, Wang G, Yang Y (2021) Hierarchical composition learning for composed query image retrieval. In: ACM multimedia Asia. pp 1–7
29. Zhang F, Xu M, Xu C (2021) Geometry sensitive cross-modal reasoning for composed query based image retrieval. IEEE Trans Image Process 31:1000–1011
30. Chen Y, Gong S, Bazzani L (2020) Image search with text feedback by visiolinguistic attention learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp 3001–3011
31. Lee S, Kim D, Han B(2021) Cosmo: content-style modulation for image retrieval with text feedback. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp 802–812
32. Wen H, Song X, Yang X, Zhan Y, Nie L(2021) Comprehensive linguistic-visual composition network for image retrieval. In: Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval (SIGIR). pp 1369–1378
33. Li X, Rong Y, Zhao M, Fan J (2021) Interactive clothes image retrieval via multi-modal feature fusion of image representation and natural language feedback. In: International conference on neural computing for advanced applications. Springer, pp 578–589
34. Li X, Ye Z, Zhang Z, Zhao M (2021) Clothes image caption generation with attribute detection and visual attention model. Pattern Recogn Lett 141:68–74
35. Quintino Ferreira B, Costeira JP, Sousa RG, Gui L-Y, Gomes JP (2019) Pose guided attention for multi-label fashion image classification. In: Proceedings of the IEEE/CVF international conference on computer vision workshops (ICCVW). pp 3125–3128
36. Peng L, Yang Y, Wang Z, Huang Z, Shen HT (2020) Mra-net: Improving vqa via multi-modal relation attention network. IEEE Trans Pattern Anal Mach Intell 44(1):318–329
37. Wu J, Weng W, Fu J, Liu L, Hu B (2022) Deep semantic hashing with dual attention for cross-modal retrieval. Neural Comput Appl 34(7):5397–5416
38. Su H, Wang P, Liu L, Li H, Li Z, Zhang Y (2020) Where to look and how to describe: fashion image retrieval with an attentional heterogeneous bilinear network. IEEE Trans Circuits Syst Video Technol 31(8):3254–3265
39. Zhang Z, Chen P, Shi X, Yang L (2019) Text-guided neural network training for image recognition in natural scenes and medicine. IEEE Trans Pattern Anal Mach Intell 43(5):1733–1745
40. Ma Z, Dong J, Long Z, Zhang Y, He Y, Xue H, Ji S (2020) Fine-grained fashion similarity learning by attribute-specific embedding network. Proc AAAI Conf Artif Intell (AAAI) 34:11741–11748
41. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). pp 770–778
42. Kuang Z, Zhang X, Yu J, Li Z, Fan J (2021) Deep embedding of concept ontology for hierarchical fashion recognition. Neurocomputing 425:191–206
43. Yan C, Ding A, Zhang Y, Wang Z (2021) Learning fashion similarity based on hierarchical attribute embedding. In: Proceedings of 2021 IEEE 8th international conference on data science and advanced analytics (DSAA). pp 1–8
44. Chen L, Zhang H, Xiao J, Nie L, Shao J, Liu W, Chua T-S (2017) Sca-cnn: spatial and channel-wise attention in convolutional networks for image captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). pp 5659–5667
45. Shajini M, Ramanan A (2021) An improved landmark-driven and spatial-channel attentive convolutional neural network for fashion clothes classification. Vis Comput 37(6):1517–1526
46. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). pp 7132–7141
47. Schroff F, Kalenichenko D, Philbin J (2015) Facenet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp. 815–823
48. Wu H, Gao Y, Guo X, Al-Halah Z, Rennie S, Grauman K, Feris R (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 (CVPR). pp 11307–11317
49. Berg TL, Berg AC, Shih J (2010) Automatic attribute discovery and characterization from noisy web data. In: Proceedings of the European conference on computer vision (ECCV). pp 663–676
50. Krizhevsky A, Sutskever I, Hinton GE (2017) Imagenet classification with deep convolutional neural networks. Commun ACM 60(6):84–90

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.