Heterogeneous Feature Fusion and Cross-modal Alignment for Composed Image Retrieval

Gangjian Zhang
Institute of Information Science, Beijing Jiaotong University, Beijing Key Laboratory of Advanced Information Science and Network Technology
Beijing, China
19120317@bjtu.edu.cn

Shikui Wei*
Institute of Information Science, Beijing Jiaotong University, Beijing Key Laboratory of Advanced Information Science and Network Technology
Beijing, China
shkwei@bjtu.edu.cn

Huaxin Pang
Institute of Information Science, Beijing Jiaotong University, Beijing Key Laboratory of Advanced Information Science and Network Technology
Beijing, China
20112005@bjtu.edu.cn

Yao Zhao
Institute of Information Science, Beijing Jiaotong University, Beijing Key Laboratory of Advanced Information Science and Network Technology
Beijing, China
yzhao@bjtu.edu.cn

ABSTRACT

Composed image retrieval aims at performing image retrieval task by giving a reference image and a complementary text piece. Since composing both image and text information can accurately model the users' search intent, composed image retrieval can perform target-specific image retrieval task and be potentially applied to many scenarios such as interactive product search. However, two key challenging issues must be addressed in composed image retrieval occasion. One of them is how to fuse heterogeneous image and text piece in the query into a complementary feature space. The other is how to bridge the heterogeneous gap between text pieces in the query and images in the database. To address the issues, we propose an end-to-end framework for composed image retrieval, which consists of three key components including Multi-modal Complementary Fusion (MCF), Cross-modal Guided Pooling (CGP), and Relative Caption-aware Consistency (RCC). By incorporating MCF and CGP modules, we can fully integrate the complementary information of image and text piece in the query through multiple deep interactions and aggregate obtained local features into an embedding vector. To bridge the heterogeneous gap, we introduce the RCC constraint to align text pieces in the query and images in the database. Extensive experiments on four public benchmark datasets show that the proposed composed image retrieval framework achieves outstanding performance against the state-of-the-art methods.

CCS CONCEPTS

• **Information systems → Image search**; *Top-k retrieval in databases*; *Novelty in information retrieval.*

KEYWORDS

composed image retrieval; modality fusion; modality consistency

ACM Reference Format:
Gangjian Zhang, Shikui Wei, Huaxin Pang, and Yao Zhao. 2021. Heterogeneous Feature Fusion and Cross-modal Alignment for Composed Image Retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3474085.3475659

1 INTRODUCTION

Defined as a fundamental visual task, image retrieval has witnessed the tremendous advancement from SIFT to CNN, which also boosts many related areas, such as face recognition [39, 45, 47], person re-identification [11, 30, 54], and fashion retrieval [31]. However, it is still far from satisfactory for image retrieval, since users' query intent cannot be accurately modeled for single-modal query such as text query and image query. To overcome the limitation of this single-modal search, Vo et al. [50] firstly proposed the so-called composed image retrieval, which composed a reference image and a modification text describing desired modifications to certain parts of the reference image as a query to retrieve the target image. It is very useful for composed image retrieval to support many commercial applications such as product search. In this scenario, users' query intent cannot be met, since it is hard for users to provide an accurate query product image to the search engine. By combining a reference image and natural language text, it is possible to obtain complementary semantics from two modalities and to make the query intent more complete and accurate.

However, we have to take up two hard challenges before performing composed image retrieval: (1) Feature integration of two heterogeneous modalities in the query. For the composed image retrieval, the query consists of a reference image and a text piece, which cooperate to convey the complete users' query intent. Since image and text are totally different modalities, how to integrate the complementary information of image and text piece in the query and aggregate heterogeneous features into an embedding vector is a key issue. (2) Cross-modal alignment. In essence, composed image

*Corresponding author.

MM '21, October 20–24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475659

retrieval performs a matching process between query (text+image) and target (image). Since the matching process crosses two different modalities, how to bridge the heterogeneous gap between image modality and text modality is a key issue. Existing composed image retrieval methods mainly focus on addressing the first issue. For example, TIRG [50] attempted to address the feature integration problem by "modifying" the query image feature. Specifically speaking, it firstly utilized the text feature to modify the image feature in its original feature space and then matched the modified query image feature with the image features in the database. Based on TIRG [50], VAL [6] attempted to change the image feature in different feature maps inside the CNN by using the text feature, and after that measured the similarity at the multi-level with the pure target image features. The process is named "Hierarchical Matching". Locally Bounded method [21] employed a specific cross-modal attention module to fuse words and image regions followed by a feature pooling operation, and then grasped the similarity with the ground truth target image in the original image feature space. However, since cross-modal alignment is not taken into account, it is less possible for these methods to find a metric space for accurately matching queries and targets. To attempt to align two modalities, Eric et al. [9] introduced a weight-sharing self-attention module into composed image retrieval, which not only fused modification words and reference image regions but also embedded both query and target features into a sharing space. In this way, two heterogeneous feature spaces are aligned. However, due to the asymmetry of the query (text+image) and the target (only image), it is not enough to only take a weight-sharing self-attention module for addressing both feature integration and cross-modal alignment. Therefore, it is necessary to find a more effective method to handle the above-mentioned issues.

Towards this end, we propose an end-to-end composed image retrieval framework, which involves three key components for addressing feature integration and cross-modal alignment problems, i.e., Multi-modal Complementary Fusion (MCF), Cross-modal Guided Pooling (CGP), and Relative Caption-aware Consistency (RCC). Firstly, MCF excavates the complementary relationship between the reference image and text piece in the query through multiple deep interactions and then fuses the message from one modality to another. Secondly, CGP obtains two summary embeddings for both image and text by taking the intra-relation of local features (or word vectors) and inter-relation of local features and word vectors as cues followed by a linear addition. MCF and CGP work together to address the feature integration problem. Finally, RCC utilizes a relative caption network to generate a semantic-consistency text by jointly using both query image and target image. By minimizing the generating loss between the generated text piece and the text piece in the query, we can guide the aligning process of heterogeneous modalities. In brief, the main contributions can be summarized as follows:

- We propose an effective framework for composed image retrieval, which can simultaneously address heterogeneous feature integration and cross-modal alignment problems by using the proposed MCF, CGP, and RCC components. Extensive experiments have proved the outstanding performance of the proposed framework.
- A novel heterogeneous feature integration method, i.e., MCF and CGP, is proposed to address the heterogeneous feature integration problem. Different from previous methods, MCF and CGP take advantage of deep interactions to fully integrate complementary information between image and text piece in the query.
- A cross-modal alignment method is proposed by using the Relative Caption-aware Consistency (RCC). By enforcing query and target to embed into a common shared space in a symmetrical manner, we can construct a reliable metric space for matching query and target.

2 RELATED WORK

2.1 Image Retrieval and Product Search

With the development of deep learning, image retrieval as a basic visual task has achieved tremendous advancement. Recently many methods based on deep learning [14, 41] have been proposed to solve the problem in this domain. At the same time, as an important branch, cross-modal image retrieval which makes the retrieval task no longer fix on the form of content-based image retrieval (CBIR) where the query is a single image can flexibly choose other modality forms as input such as a natural sentence [51] or a sketch pictured by the user [42, 57, 58]. Following [50], we investigate the task of composed image retrieval taking a reference image together with a text piece as input. This type of image retrieval usually appears in product search [31]. Obviously, to return a product image that satisfies the user, incorporating user feedback as the auxiliary information into the search query can better understand the true intention of the user and greatly improve the accuracy of the search. Such user feedback may take many forms. These forms include relevance [43], attribute [1, 16, 46], sketch [42, 57, 58], spatial layout [34, 35], and modification text or text piece [6, 21, 50], of which text, as a tool of communication frequently used by people in daily life, is more expressive and more in line with user's habits. So, in this paper, we discuss how to better utilize a modification text or a text piece, this form of user feedback, for image retrieval.

2.2 Multi-modal Learning

Multi-modal learning usually involves two or more modalities (e.g. image and text). Based on this, some tasks are derived, such as image captioning [49, 55] which generates a textual description to depict the content within one or more images [10, 24], visual question answering [2, 3, 33, 53] which attempts to answer a textual question based on a given image, Image-Text retrieval [6, 27, 28, 52, 60] which retrieves an image by a given text or retrieves a text by an image, referring expression [7, 22, 56] which identifies a particular object within a picture according to a natural language description. In the paper, we study the task of composed image retrieval which retrieves a target image by combining a text piece and a reference image, and we consider learning a sharing embedding space where not only can the different modalities compose with each other, but also the similarity of the target and the query can be measured.

2.3 Interactions between Language and Vision

By making the features of language and vision interact with each other, the semantics from these two different modalities can convert, match and fuse better. There are various types of interactions being explored in language and vision tasks. [12, 26, 59] utilize the method of bilinear pooling to interact and fuse the information of language and vision in visual question answering. Co-attention mechanism used in [13, 33, 37, 60] is concerned with the context involved in visual and textual contents to decide how to distribute different weights for image regions and text words. Gated multimodal units [4] adopts in Image-Text retrieval [6, 52] and composed image retrieval [50] normally controls the fusion degree by a gate function. Multimodal residual learning [25] utilized in [40] and [52] elementwisely adds the compositional feature of image and text with the original feature of image or text to increase the frequency of interactions. Currently, several methods [8, 29, 32] building on Transformer [48] architecture can make the features of language and vision interact by adopting a novel self-attention mechanism or cross-attention mechanism.

3 METHOD

In this section, we will present a detailed description of our proposed method. As explained above, we aim to propose a method that can embed both the text+image query and target image to a sharing space where the matched pairs (query and target) are as close as possible and the mismatch pairs are far away.

3.1 Images and Texts Representation

Image Representation: To fully explore the visual information, we use ResNet50 pretrained on the ImageNet dataset as the image feature encoder. For each image, we extract the output feature maps of the third and fourth blocks. These feature maps are then projected to the same 512 channels respectively using two learnable 1×1 convolution layers with the width and height keeping. After concatenating the features from two different blocks, the reference image is encoded as $f_{img}(\mathbf{I}_r) = \phi_r \in \mathbb{R}^{d \times N}$, where f_{img} is the encoding operation, d is the feature dimension whose size is 512, and N is the number of grid features and calculated as $N = 7*7+14*14 = 245$. Similarly, the target image is encoded as $f_{img}(\mathbf{I}_t) = \phi_t \in \mathbb{R}^{d \times N}$.
Text Representation: In terms of text piece, We use a single-layer LSTM to encode it. While the other advanced encoders such as bi-LSTM and LSTM attention which are more powerful can be adopted as well, but it is not the point that we focus on in our paper. Formally, the text is tokenized into a token sequence which is then encoded by the LSTM to obtain the hidden state output. We define hidden state output as $f_{text}(\mathbf{T}) = \phi_m \in \mathbb{R}^{d \times L}$, where d is the feature dimension whose size is 512, and L is the length of the current text.

3.2 Baseline Model

Our method is based on what is proposed in [9]. Having finished the feature extracting operation mentioned above, MAAF [9] treats the 245 image local feature representations ϕ_r as tokens and concatenates them with the text tokens ϕ_m to obtain a joint token sequence $\Phi = [\phi_r, \phi_m] \in \mathbb{R}^{d \times (N+L)}$. Then the joint token sequence Φ is fed into two Transformer blocks [48] including self-attention layers and feed-forward layers. Self-attention layer firstly projects the joint token sequence into three matrices then uses a specific self-attention operation to obtain the output:

$$Self(\Phi) = Attn(\Phi^T W_Q, \Phi^T W_K, \Phi^T W_V), \tag{1}$$

where Q is query matrix, K is key matrix, and V is value matrix. Following this, a feed-forward layer is used to project the hidden sequence. Consequently, image tokens coming from different blocks and text tokens are average pooled separately. The final embedding eq is obtained by averaging these three vectors. As for the obtaining of the target image, the same weight-sharing image feature extractor and Transformer blocks are taken to get the final target embedding et except that the text piece is absent.

Given the mini-batch of query embeddings and their only pair target embeddings, the training objective is to push the embeddings of the matching pairs closer, while pulling mismatch images away. So a batch-based classification loss is used which can be written as:

$$L_{bat} = \frac{1}{B} \sum_{i=1}^{B} -log \left\{ \frac{exp\left\{\kappa\left(eq_i, et_i\right)\right\}}{\sum_{j=1}^{B} exp\left\{\kappa\left(eq_i, et_j\right)\right\}} \right\}, \tag{2}$$

where κ is a similarity kernel, and B is the size of the training mini-batch which is also the number of pairs (query and their corresponding target). The whole network is trained with standard backpropagation.

Our model employs the same architecture including ResNet50 for image encoder, one layer LSTM for text encoder, and Transformer block except that we reduce the number of Transformer blocks from 2 to 1. Also, to facilitate the narrative later, we define the output of the Transformer block as $\Psi = Tr(\Phi) = [\psi_r, \psi_m]$. where Tr is the operation of the Transformer block, Ψ is regarded as the hidden state output of Φ with the identical sequence order and length. $\psi_r \in \mathbb{R}^{d \times N}$ and $\psi_m \in \mathbb{R}^{d \times L}$ are the image and text feature outputs respectively. Same for the target image features, after weight-sharing Transformer block the hidden state output of ϕ_t can be obtained, which is defined as $\psi_t \in \mathbb{R}^{d \times N}$.

3.3 Proposed Method

The Transformer block allows each token to attend to all tokens including itself so that the features from image modality would be modified by the features of text modality or vice versa. Meanwhile, sharing the same weight can enforce the features from query and target simultaneously into a common space. However, taking the discrepancy and asymmetry of query (image features plus text features) and target (only image feature) into consideration, it is difficult to solve these two problems by one shot. So to cope with the issues above, we propose two modules to facilitate the fusion process between two distinct modalities in query and propose one effective constraint to prompt both the query and target into a common space.

Multi-modal Complementary Fusion (MCF): Inspired by [26, 37, 52], we propose MCF module to explore the complementary relationship between text piece and reference image through multiple feature interactions of these two modalities. With the help of such a complementary relationship, the complementary information of one modality can be converted into another modality. Concretely, for the procedure of converting text piece into reference image, we

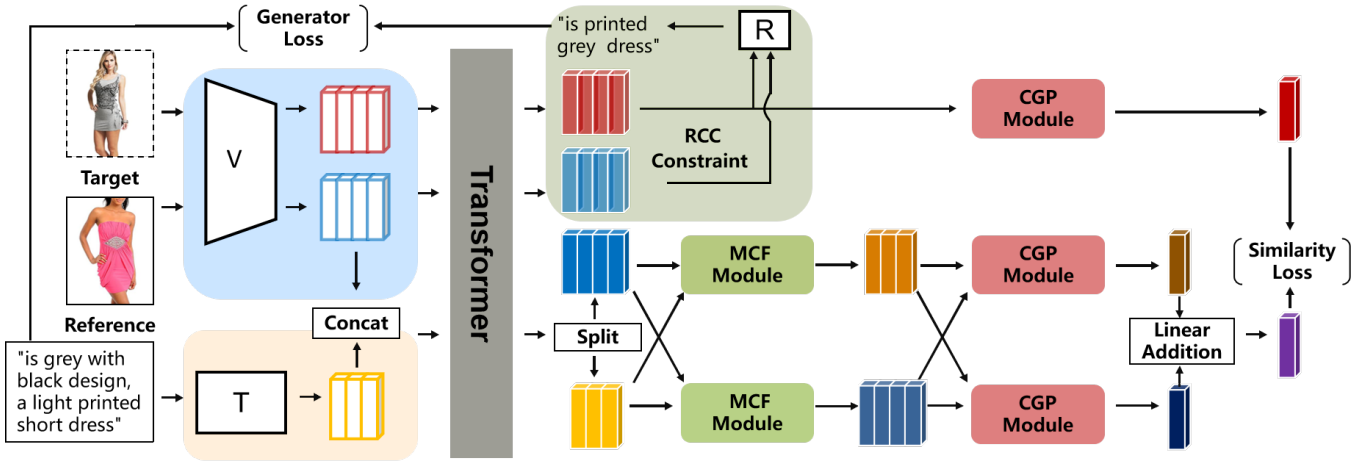

Figure 1: Overview of the proposed method. The proposed framework is composed of a visual encoder V, a language encoder T, a weight-sharing Transformer module, a relative caption generator R, two Multi-modal Complementary Fusion (MCF) modules (Figure 2), and three Cross-modal Guided Pooling (CGP) modules (Figure 3). The loss function includes the similarity loss of embedding and generator loss of sentence.

compute the interaction map as follows:

$$M_{r1} = \psi_r^T V_r \psi_m, \tag{3}$$

$$M_{r2} = W_r M_{r1}, \tag{4}$$

where $V_r \in \mathbb{R}^{d\times d}$, $W_r \in \mathbb{R}^{N\times N}$ are learned weight matrix, $M_{r1} \in \mathbb{R}^{N\times L}$ is the first interaction map which is also regarded as the inter-modal correlation matrix in many attention methods but in our method, the matrix is used to capture the complementary relationship between modalities. $M_{r2} \in \mathbb{R}^{N\times L}$ is the second interaction map which is computed through the multiplication of learned weight matrix and first interaction map. It is used not only to excavate a more complex relationship between two modalities but also to increase the capacity of the model. In addition to that, $M_{r2,ij}$ represents the complementary relationship between ith image spatial region and jth text word. Second, with the assistance of an interaction map, we can naturally project the features of text into image modality as follows:

$$M_r = softmax\left(M_{r2}\right), \tag{5}$$

$$\hat{\psi}_m^T = M_r \psi_m^T, \tag{6}$$

where $softmax$ is normalization over the image feature dimension, $\hat{\psi}_m \in \mathbb{R}^{d\times N}$ is the feature matrix projecting from language modality to visual modality. Having obtained the representation of text feature in image space, we need to accomplish fusion procedure with image feature. Here, a modified gate operation is proposed as follows:

$$O_r = ReLU\left(W_{r1}\left[\psi_r^T, \hat{\psi}_m^T\right]^T + b_{r1}\right) + \hat{\psi}_m, \tag{7}$$

$$G_r = signmoid\left(\left(W_{r2}\psi_r + b_{r2}\right) \odot \left(W_{r3}\hat{\psi}_m + b_{r3}\right)\right), \tag{8}$$

$$\widetilde{\psi}_r = O_r \odot G_r + \psi_r, \tag{9}$$

where W_{r1},b_{r1},W_{r2},b_{r2},W_{r3},b_{r3} are learnable parameters, $O_r \in \mathbb{R}^{d\times N}$ is the selected complementary information from textual modality, $G_r \in \mathbb{R}^{d\times N}$ makes a gate role using bilinear interaction between two modalities to choose which complementary information should be encouraged and which should be suppressed. $\widetilde{\psi}_r \in \mathbb{R}^{d\times N}$ is the output of the fusion operation which elementwisely adds complementary features to original image features, and $\odot$ denotes the elementwise multiplication.

As for the operation of fusing reference image into text piece, it is completely symmetric with the text-to-image operation introduced above, we use the following formula to summarize this process:

$$\widetilde{\psi}_m = MCF(\psi_m, \psi_r), \tag{10}$$

where $\widetilde{\psi}_m \in \mathbb{R}^{d\times L}$ is the fused text feature.

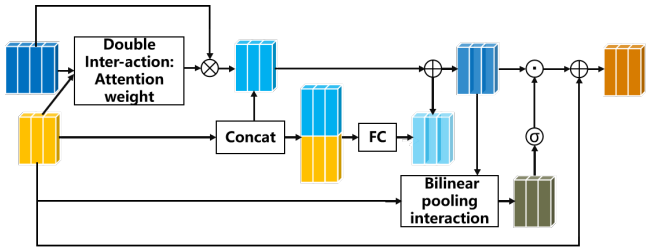

Figure 2: Multi-modal Complementary Fusion (MCF) module. MCF mines complementary information through multiple interactions between different modalities and then fuses the complementary features with original features.

Cross-modal Guided Pooling (CGP): Inspired by [60, 61], having obtained both the local fused image features and word-level fused text features, we need to summarize these two stream information to get a global embedding for the convenience of similarity measure with target embedding in embedding space. Note that it is also important to consider the interaction between these streams for achieving a more reasonable blending effect. So CGP adopts an attention pooling strategy not only considering the relation within

one modality but also exploiting the relevance between the current modality and another modality. Specifically, for the pooling of image features, text features also participate in it. We first compute two interaction maps (one is denoted as the intra-relation and another is denoted as the inter-relation) as follows:

$$A_{rm} = \widetilde{\psi}_r^T K_{rm} \widetilde{\psi}_m, \tag{11}$$

$$A_r = \widetilde{\psi}_r^T K_r \widetilde{\psi}_r, \tag{12}$$

where $K_{rm}, K_r \in \mathbb{R}^{d \times d}$ is learnable parameters, $A_{rm} \in \mathbb{R}^{N \times L}$ is similar with the first interaction map in MCF module but taking $\widetilde{\psi}_r$ and $\widetilde{\psi}_m$ as input, and $A_r \in \mathbb{R}^{N \times N}$ interacts each local image feature with all image features including itself. By establishing the two relation maps, we attempt to project those features from two modalities into one sharing space where the importance of different image features will be explored. The concrete procedure is written as:

$$X_r = tanh\left(W_1(A_{rm}\widetilde{\psi}_m^T)^T + b_1\right) \odot tanh\left(W_2(A_r\widetilde{\psi}_r^T)^T + b_2\right), \tag{13}$$

$$Z_r = softplus\left(W_3 X_r + b_3\right), \tag{14}$$

where $W_1, W_2 \in \mathbb{R}^{d \times d}$, $b_1, b_2, W_3 \in \mathbb{R}^d$, $b_3 \in \mathbb{R}$ are learnable parameters, $X_r \in \mathbb{R}^{d \times N}$ dynamically captures the salient information in the sharing space by comprehensively summarizing two sources (One is reflecting the intra-relation within local image features and another is considering inter-relation between local image features and word features). The interaction method between them also adopts the bilinear pooling. Finally, an weight distribution over the whole image features $Z_r \in \mathbb{R}^N$ can be the output.

Given the formulae about how to obtain pooling weights of image features, it is relatively easy to deduce the pooling method of text features which uses a completely symmetrical way to obtain the weights of text features. Here shows the summary formulae:

$$Z_m = CGP(\widetilde{\psi}_m, \widetilde{\psi}_r), \tag{15}$$

where $Z_m \in \mathbb{R}^N$ gives each word the weight.

Finally, we obtain the global embedding vector by linearly combining image embedding and text embedding which already have been solved above as follows:

$$eq = \gamma_r \widetilde{\psi}_r Z_r + \gamma_m \widetilde{\psi}_m Z_m, \tag{16}$$

where $\gamma_r, \gamma_m \in \mathbb{R}$ are learnable parameters to balance these two embeddings. eq is the query embedding which will be multiplied by a learned scale parameter as mentioned in MAAF [9] to get the final query embedding.

For the obtaining of target embedding, the MCF module is unnecessary obviously and the CGP module can be slightly modified to adapt to handling the pooling operation for the target image. we only consider the intra-relation within target image features, and the solution of target embedding et is listed as follows:

$$A_t = \psi_t^T K_t \psi_t, \tag{17}$$

$$X_t = tanh\left(W_4(A_t\psi_t^T)^T + b_4\right), \tag{18}$$

$$Z_t = softplus\left(W_5 X_t + b_5\right), \tag{19}$$

$$et = \psi_t Z_t, \tag{20}$$

where $K_t, W_4, W_5 \in \mathbb{R}^{d \times d}$, $b_4 \in \mathbb{R}^d$, $b_5 \in \mathbb{R}$ are learnable parameters. Same as the query embedding eq, et will multiply the same learned scale parameter mentioned above to get the final target embedding.

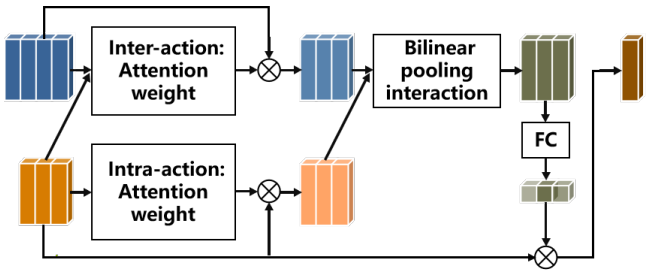

Figure 3: Cross-modal Guided Pooling (CGP) module. CGP comprehensively considers the relation between modalities and the relation within a modality, and then dynamically weights different local features of one modality based on this.

Relative Caption-aware Consistency (RCC): The module MCF and CGP mentioned above are used to resolve the issue of heterogeneous feature integration between the reference image and text piece. However, in this image retrieval task, it is quite important to align the cross-modal semantics in the final embedding space. The previous methods usually utilize extra information such as image labels to constrain the model to learn a high-level semantic space where the same semantics of different modalities will be aligned, but it needs the cost of manual annotations. Based on our composed image retrieval tasks, we propose a relative image caption strategy to capture such high-level semantic information without any extra manual labeling. Relative image caption generator in [10, 15, 24] is used for generating a description reflecting the difference between two image, but in our method it is used for aligning the semantics of different modalities.

After passing the weight-sharing Transformer block, both target image features and query "blended" features transform to hidden state output, and after this process, all these features are inserted into a sharing space. The space can become more perfect (the matching pairs become closer) as the training loss gets smaller. In order to make the features in the space possess higher-level semantic so that different modality features can fuse better and matching pairs can become more similar, we take the hidden state output of target image features ψ_t and the hidden state output of only reference image features without text piece features (the output named ψ_r^*) as input, and then feed them into a modified LSTM structure by which we attempt to generate the text piece T word by word. The T can be encoded as a sequence of one hot vector as follows:

$$T = \{w_1, ..., w_L\}, w \in \mathbb{R}^K, \tag{21}$$

where K is the size of the vocabulary constructed from words of all the text pieces and L is the length of the current text piece. And the LSTM structure exhibits as follows:

$$\begin{aligned}&(i_s\ f_s\ o_s\ g_s)^T \\ &= (\sigma\ \sigma\ \sigma\ tanh)^T A_{d+d+2*d,d}\left(Ew_{s-1}\ h_{s-1}\ \left[z_{t,s}, z_{r,s}\right]\right)^T,\end{aligned} \tag{22}$$

$$c_s = f_s \odot c_{s-1} + i_t \odot g_s, \tag{23}$$

$$h_s = o_s \odot tanh\left(c_s\right), \tag{24}$$

where $i, f, c, o, h \in \mathbb{R}^d$ are the input, forget, memory, output and hidden state of the LSTM, s and $s-1$ mean current step and last step of the inference procedure. c_0, h_0 are initialized randomly. $A_{a,b} : \mathbb{R}^a \rightarrow \mathbb{R}^b$ is an affine transformation with learnable parameters. $E \in \mathbb{R}^{d \times K}$ is a completely new embedding matrix which is different from the one used for encoding text, σ is the logistic sigmoid activation. $z_{t,s}, z_{r,s}$ are context vectors dynamically capturing the visual information within different regions of target image and reference image which are obtained by the same attention operation used in [55] briefly written as:

$$z_{t,s} = f_{att,t}\left(\psi_t, h_{s-1}\right), \tag{25}$$

$$z_{r,s} = f_{att,r}\left(\psi_r^*, h_{s-1}\right), \tag{26}$$

where $z_{t,s}, z_{t,s} \in \mathbb{R}^d$. Similar to [55] we compute the output word probability by using the LSTM hidden state, the previous word, the context vector from the reference image, and the context vector from the reference image as follows:

$$\begin{aligned} &p\left(w_s|\psi_t, \psi_r^*, w_{s-1}\right) \\ &\propto exp\left(W_o\left(Ew_{s-1} + W_h h_s + W_z\left[z_{t,s}, z_{r,s}\right]\right)\right), \end{aligned} \tag{27}$$

where $W_o \in \mathbb{R}^{K \times d}$, $W_h \in \mathbb{R}^{d \times d}$, $W_z \in \mathbb{R}^{d \times 2d}$. Given the text piece, the objective of relative image captioning is to make the following cross entropy loss as small as possible:

$$L_{rel} = \frac{1}{L}\sum_{s=1}^{L} -log\left(p\left(w_s|w_1, ..., w_{s-1}\right)\right). \tag{28}$$

3.4 Loss Function

The whole network contains two components including the embedding network for embedding both the query (reference image plus text piece) and the target (target image) and the text generator for predicting the text piece from the reference image and target image. We jointly train these two components including the loss relative image captioning mentioned in Sec. 3.2 and original batch-based classification loss mentioned in Sec. 3.3, the final loss function is composed of two parts as follows:

$$L_{fin} = \alpha L_{bat} + \beta L_{rel}, \tag{29}$$

where α, β are hyperparameters to balance the weights of these two losses.

4 EXPERIMENT

4.1 Datasets and Evaluation Metric

There are four benchmark datasets used in our experiments: Fashion200K, FashionIQ, Shoes, MIT-States.

(1) Fashion200K dataset [16] has ~ 200k images attached with attribute-like product descriptions (such as black pilot jacket) of fashion products, in which the reference image and target image having one-word difference in their descriptions become a pair and the text piece is constructed based on this difference.

(2) FashionIQ dataset [15] has ~ 60k real product images from *amazon.com*. The text piece just like the real-world user feedback is used to retrieval what the user wants based on a reference product image. The dataset has three categories of product: Dresses, Tops/Tees, and Shirts, and the distributions of them are even. Finally, the dataset is divided into ~ 18k queries for training, ~ 6K queries for validation, and ~ 6K queries for test following the splitting method used in [9].

(3) Shoes dataset [5] is crawled from *like.com*. Its text piece is natural language, which is similar to FashionIQ from this point, but the difference is that the dataset only has 10K images for training and 4658 images for test. The splitting method is following [6].

(4) MIT-States dataset [23] has ~ 60K images, in which each image is tagged with an adjective or state label and a noun or object label (such as "dry bathroom" or "cooked beef"). There are 115 adjectives and 245 nouns. Following the setting of [50], the goal is to retrieve the target image which has the same object as the query image with different adjective or state label. To inspect the ability of dealing with unseen objects of the model, 49 nouns are used for the test and the rest is for training.

Evaluation Metric: we adopt the same evaluation metrics Recall at K (R@K) in all datasets as [9] to compare our proposed method with other methods. Recall at K (R@K) computes the percentage of test queries whose labeled target image appears in the top K retrieved images.

4.2 Implementation Details

We use PyTorch in our experiments. We use Resnet-50 [17, 18] pretrained on ImageNet as our backbone for image encoder (output feature size = 512) and the LSTM [20] of random initial weights as our text encoder (hidden size is 512). By default, The model is trained by SGD optimizer at an initial learning rate of 0.01 which is decreased by a factor of 10 every 25k iterations and training is stopped at 75k iterations. We use a batch size of 32 for all experiments. The hyperparameter α and β of the final loss function are set as 0.7 and 0.3 by default. Each experiment is repeated 4 times to obtain a stable retrieval performance, and both mean and standard deviation are reported.

4.3 Quantitative Results

We compare our proposed method with the baseline model and published state-of-the-art models in the four benchmark datasets. We take the best-reported results from corresponding papers when they are available. The method with an asterisk is our reproduction of the original method.

Quantitative Results on FashionIQ: We test our complete model (including module MCF, module CGP, and Constraint RCC) on FashionIQ, a medium-sized dataset. Table 2 shows the comparison to other methods, and the experiment results of them all come from [9]. Based on the baseline model, MAAF [9], which is also the SOTA method, our method has a performance improvement of 2.1%.

Quantitative Results on Fashion200K: We test our complete model on Fashion200K. Because of the relatively large scale of Fashion200K, specifically, we set the learning rate of training is decreased by a factor of 10 every 50k iterations and the training process is stopped at 150k iterations. We set an initial learning rate of 0.002 for the image model parameters to prevent overfitting. The experiment results of other methods are taken from [21] and [6], and due to the lack of detailed results of the baseline model, we reproduce MAAF based on the code provided by the author following the same parameter set. Table 3 shows the comparison of

Table 1: Ablation on the FashionIQ validation set.

Method	MCF	CGP	RCC	Dresses R10 (R50)	Shirts R10 (R50)	Tops/Tees R10 (R50)	(R@10 + R@50)/2
Baseline	-	-	-	23.8 ± 0.6 (48.6 ± 1.0)	21.3 ± 0.7 (44.2 ± 0.3)	27.9 ± 0.8 (53.6 ± 0.6)	36.6 ± 0.4
Ours	-	-	✓	25.1 ± 1.3 (50.1 ± 0.9)	21.8 ± 0.7 (45.5 ± 0.6)	28.5 ± 0.6 (55.7 ± 0.7)	37.8 ± 0.4
Ours	-	✓	-	26.1 ± 0.3 (50.4 ± 0.4)	21.5 ± 0.6 (44.8 ± 0.5)	29.4 ± 0.8 (55.1 ± 0.9)	37.8 ± 0.4
Ours	-	✓	✓	26.7 ± 0.4 (50.3 ± 0.7)	22.2 ± 0.2 (45.6 ± 0.8)	29.3 ± 0.6 (56.2 ± 0.3)	38.4 ± 0.3
Ours	✓	-	-	25.6 ± 0.4 (50.7 ± 0.3)	21.0 ± 0.8 (44.8 ± 0.6)	28.5 ± 0.2 (54.6 ± 0.2)	37.5 ± 0.2
Ours	✓	-	✓	26.4 ± 0.5 (51.6 ± 0.8)	21.2 ± 0.5 (44.8 ± 0.4)	29.0 ± 1.2 (55.8 ± 0.5)	38.2 ± 0.2
Ours	✓	✓	-	26.7 ± 0.9 (51.2 ± 0.5)	21.4 ± 0.5 (45.1 ± 0.5)	28.9 ± 0.3 (55.3 ± 0.4)	38.1 ± 0.2
Ours	✓	✓	✓	26.2 ± 0.3 (51.2 ± 0.6)	22.4 ± 0.3 (46.0 ± 0.5)	29.7 ± 0.5 (56.4 ± 0.4)	38.7 ± 0.2

Table 2: Results on the FashionIQ validation set.

Method	(R@10 + R@50)/2
TIRG[50]	31.20
VAL[6]	35.4
MAAF[9]	36.6 ± 0.4
Ours	**38.7 ± 0.2**

our model and other methods. Compared with the baseline method, our method also has certain boosts in performance. At R@10 and R@50, The improvements are 1.89%, and 1.46% respectively. VAL[6] uses a hierarchical matching method to comprehensively match low, medium, and high-level features, so better results can be obtained, but it also triples the computational complexity in the inference process.

Quantitative Results on Shoes: We test our complete model on Shoes. Due to the relatively small scale of Shoes, we set that the learning rate of training is decreased by a factor of 10 every 15k iterations, and the training process is stopped at 45k iterations. Meanwhile, we adjust the hyperparameter α and β to 0.8 and 0.2. The experiment results of other methods come from [6]. We reproduce the baseline results by only keeping one Transformer block. The parameter set is the same as the set of our model. Table 4 shows the comparison of our method and other methods. Our method surpasses the baseline model with margins of 1.40%, 1.00%, and 0.88% at R@1, R@5, and R@10.

Quantitative Results on MIT-States: Note that the text pieces in MIT-States are some label words describing the target image as has introduced in Sec. 4.1. In the experiment, We find that the use of batch-based classification loss can cause bad performance, so we replace it with its variation (soft triplet based loss used in [19, 36]) and set that the learning rate is decreased by a factor of 10 every 50k iterations and training process is stopped at 150k iterations. Meanwhile, we adjust the hyperparameter α and β to 1 and 0.01. We reproduce the MAAF with only one Transformer block and find that the Transformer can't do very well in handling word-liked text piece. Actually, we test our model without a Transformer block on the MIT-States dataset. Results of other methods are from Locally Bounded [21]. As shown in Table 5, our model still has good performance when the Transformer block is absent.

Table 3: Results on the Fashion200K test set.

Method	R@1	R@10	R@50
Han et al.[16]	6.3	19.9	38.3
Show and Tell[49]	12.3 ± 1.1	40.2 ± 1.7	61.8 ± 0.9
FiLM[40]	12.9 ± 0.7	39.5 ± 2.1	61.9 ± 1.9
Param hashing[38]	12.2 ±1.1	12.2 ±1.1	12.2 ±1.1
Relationship[44]	13.0 ± 0.6	40.5 ± 0.7	62.4 ± 0.6
TIRG[50]	14.1 ± 0.6	42.5 ± 0.7	63.8 ± 0.8
Locally Bounded[21]	17.8 ± 0.5	48.35 ± 0.6	68.5 ± 0.5
VAL[6]	**22.9**	**50.8**	**72.7**
MAAF *[9]	18.22 ± 0.3	47.52 ± 1.1	67.91 ± 0.5
Ours	18.24 ± 0.5	49.41 ± 0.4	69.37 ± 0.8

Table 4: Results on the Shoes test set.

Method	R@1	R@10	R@50
FiLM[40]	10.19	38.89	68.30
MRN[25]	11.74	41.70	67.01
Relationship[44]	12.31	45.10	71.4
TIRG[50]	12.60	45.45	69.39
VAL[6]	17.18	**51.52**	75.83
MAAF *[9]	16.45 ± 0.2	49.95 ± 0.3	76.36 ± 0.2
Ours	**17.85 ± 0.3**	50.95 ± 0.7	**77.24 ± 0.4**

Table 5: Results on the MIT-States test set.

Method	R@1	R@5	R@10
Show and Tell[49]	11.9 ± 0.1	31.0 ± 0.5	42.0 ± 0.8
Attribute Op.[36]	8.8 ± 0.1	27.3 ± 0.3	39.1 ± 0.3
Relationship[44]	12.3 ± 0.5	31.9 ± 0.7	42.9 ± 0.9
FiLM[40]	10.1 ± 0.3	27.7 ± 0.7	42.9 ± 0.9
TIRG[50]	12.2 ± 0.4	31.9 ± 0.3	41.3 ± 0.3
Locally Bounded[21]	**14.72 ± 0.6**	35.30 ± 0.7	46.56 ± 0.5
MAAF *[9]	11.22 ± 0.3	31.20 ± 0.4	42.26 ± 0.5
Ours (w/o Transformer)	14.30 ± 0.3	**35.36 ± 0.4**	**47.12 ± 0.2**

4.4 Ablation Studies

In this section, we also display the results of the ablation experiment on the FashionIQ dataset to study the role of each part (including

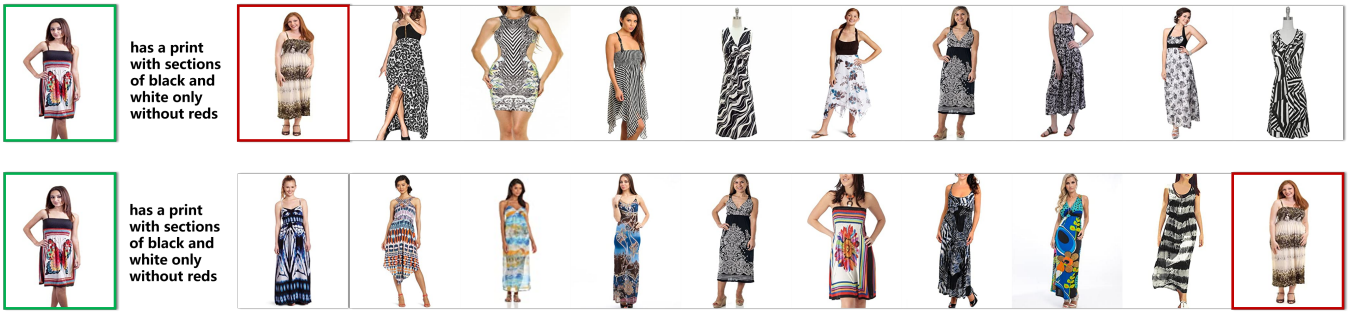

Figure 4: Qualitative Results of FashionIQ. The green and red boxes represent the reference image and target image respectively. The first row shows the results of our model and the second row shows the results of the baseline model.

Figure 5: Qualitative Results of Shoes. The reference image and target image are marked by green and red boxes respectively. The two-row exhibits the results of our model (top) and the baseline model (bottom).

module MCF, module CGP, and constraint RCC) plays in performance improvement. These results are shown in Table 1. **Effect of Multi-modal Complementary Fusion (MCF)**: As a supplement to the original Transformer module, MCF makes the fusion of modalities more sufficient through multiple interactions. Judging from the experimental results, a certain improvement can be achieved on the basis of the Transformer model. **Effect of Cross-modal Guided Pooling (CGP)**: The function of CGP is to dynamically assign weights to local features according to the relation within one modality and the relation between two modalities. Such function is obviously not available in the Transformer block or MCF module. From the results of the ablation experiment, CGP is a very useful module. Even if directly following the Transformer module, it can achieve considerable results. **Effect of Relative Caption-aware Consistency (RCC)**: RCC can enforce the output hidden features to learn a higher-level semantic space. Once the model has learned the high-level semantics, not only can the semantic gap between the modalities be eliminated, but the query embeddings and target embeddings can also be well measured. So whether this constraint is added to the original Transformer model or our improved model, a better improvement can be achieved.

4.5 Qualitative Results and Visualizations

Figure 4 and Figure 5 show the qualitative comparisons between our model and the baseline model, of which the samples of Figure 4 come from the validation set of Fashion IQ and samples of Figure 5 come from the test set of Shoes. Each graph includes a reference image, a text piece, and the top ten retrieved results corresponding to two different models. Compared to the baseline model, the improved model can capture some specific words better within the text piece, such as "sections" and "buckle" in the samples. This also shows our model's ability to grasp high-level semantic information.

5 CONCLUSION

In this paper, we propose a useful model including two modules (MCF and CGP) and a constraint (RCC) to handle the modality fusion problem and similarity matching problem in the task of composed image retrieval. Specifically, (1) our method uses multiple interactions between modalities to mine the complementary information that exists between modalities and fuse features from reference images and modified texts. Our method can further improve performance based on the original Transformer model. (2) Our method adopts a relative image caption strategy to bridge the semantics between visual and textual features and creates a more ideal common space for fusion operation and measure operation. We validate our models on four benchmark datasets. The results of the experiment on these datasets signify that our method can achieve the state-of-the-art performance.

ACKNOWLEDGMENTS

This work is supported by the National Key Research and Development of China (2017YFC1703503) and the National Natural Science Foundation of China (61972022, 1936212).

REFERENCES

[1] Kenan E. Ak, Ashraf A. Kassim, Joo Hwee Lim, and Jo Yew Tham. 2018. Learning Attribute Representations With Localization for Flexible Fashion Search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2017. Bottom-Up and Top-Down Attention for Image Captioning and VQA. *CoRR* abs/1707.07998 (2017). arXiv:1707.07998 http://arxiv.org/abs/1707.07998

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[4] John Arevalo, Thamar Solorio, Manuel Montes y Gómez, and Fabio A. González. 2017. Gated Multimodal Units for Information Fusion. arXiv:1702.01992 [stat.ML]

[5] Tamara L. Berg, Alexander C. Berg, and Jonathan Shih. 2010. Automatic Attribute Discovery and Characterization from Noisy Web Data. In *Computer Vision – ECCV 2010*, Kostas Daniilidis, Petros Maragos, and Nikos Paragios (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 663–676.

[6] 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 (CVPR)*.

[7] Yi-Wen Chen, Yi-Hsuan Tsai, Tiantian Wang, Yen-Yu Lin, and Ming-Hsuan Yang. 2019. Referring Expression Object Segmentation with Caption-Aware Consistency. *CoRR* abs/1910.04748 (2019). arXiv:1910.04748 http://arxiv.org/abs/1910.04748

[8] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. {UNITER}: Learning {UN}iversal Image-{TE}xt Representations. https://openreview.net/forum?id=S1eL4kBYwr

[9] Eric Dodds, Jack Culpepper, Simao Herdade, Yang Zhang, and Kofi Boakye. 2020. Modality-Agnostic Attention Fusion for visual search with text feedback. *arXiv preprint arXiv:2007.00145* (2020).

[10] Maxwell Forbes, Christine Kaeser-Chen, Piyush Sharma, and Serge J. Belongie. 2019. Neural Naturalist: Generating Fine-Grained Image Comparisons. *CoRR* abs/1909.04101 (2019). arXiv:1909.04101 http://arxiv.org/abs/1909.04101

[11] Yang Fu, Yunchao Wei, Yuqian Zhou, Honghui Shi, Gao Huang, Xinchao Wang, Zhiqiang Yao, and Thomas S. Huang. 2018. Horizontal Pyramid Matching for Person Re-identification. *CoRR* abs/1804.05275 (2018). arXiv:1804.05275 http://arxiv.org/abs/1804.05275

[12] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. *CoRR* abs/1606.01847 (2016). arXiv:1606.01847 http://arxiv.org/abs/1606.01847

[13] Peng Gao, Hongsheng Li, Haoxuan You, Zhengkai Jiang, Pan Lu, Steven C. H. Hoi, and Xiaogang Wang. 2018. Dynamic Fusion with Intra- and Inter- Modality Attention Flow for Visual Question Answering. *CoRR* abs/1812.05252 (2018). arXiv:1812.05252 http://arxiv.org/abs/1812.05252

[14] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. 2016. Deep Image Retrieval: Learning Global Representations for Image Search. In *Computer Vision – ECCV 2016*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer International Publishing, Cham, 241–257.

[15] Xiaoxiao Guo, Hui Wu, Yupeng Gao, Steven Rennie, and Rogério Schmidt Feris. 2019. The Fashion IQ Dataset: Retrieving Images by Combining Side Information and Relative Natural Language Feedback. *CoRR* abs/1905.12794 (2019). arXiv:1905.12794 http://arxiv.org/abs/1905.12794

[16] 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. *CoRR* abs/1708.01311 (2017). arXiv:1708.01311 http://arxiv.org/abs/1708.01311

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Identity Mappings in Deep Residual Networks. In *Computer Vision – ECCV 2016*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer International Publishing, Cham, 630–645.

[19] Alexander Hermans, Lucas Beyer, and Bastian Leibe. 2017. In Defense of the Triplet Loss for Person Re-Identification. *CoRR* abs/1703.07737 (2017). arXiv:1703.07737 http://arxiv.org/abs/1703.07737

[20] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[21] 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 (CVPR)*.

[22] Zhiwei Hu, Guang Feng, Jiayu Sun, Lihe Zhang, and Huchuan Lu. 2020. Bi-Directional Relationship Inferring Network for Referring Image Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[23] Phillip Isola, Joseph J. Lim, and E. Adelson. 2015. Discovering states and transformations in image collections. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), 1383–1391.

[24] Harsh Jhamtani and Taylor Berg-Kirkpatrick. 2018. Learning to Describe Differences Between Pairs of Similar Images. *CoRR* abs/1808.10584 (2018). arXiv:1808.10584 http://arxiv.org/abs/1808.10584

[25] Jin-Hwa Kim, Sang-Woo Lee, Dong-Hyun Kwak, Min-Oh Heo, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. 2016. Multimodal Residual Learning for Visual QA. *CoRR* abs/1606.01455 (2016). arXiv:1606.01455 http://arxiv.org/abs/1606.01455

[26] Jin-Hwa Kim, Kyoung Woon On, Woosang Lim, Jeonghee Kim, JungWoo Ha, and Byoung-Tak Zhang. 2016. Hadamard Product for Low-rank Bilinear Pooling. *CoRR* abs/1610.04325 (2016). arXiv:1610.04325 http://arxiv.org/abs/1610.04325

[27] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. *CoRR* abs/1411.2539 (2014). arXiv:1411.2539 http://arxiv.org/abs/1411.2539

[28] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked Cross Attention for Image-Text Matching. *CoRR* abs/1803.08024 (2018). arXiv:1803.08024 http://arxiv.org/abs/1803.08024

[29] Gen Li, Nan Duan, Yuejian Fang, Daxin Jiang, and Ming Zhou. 2019. Unicoder-VL: A Universal Encoder for Vision and Language by Cross-modal Pre-training. *CoRR* abs/1908.06066 (2019). arXiv:1908.06066 http://arxiv.org/abs/1908.06066

[30] Wei Li, Xiatian Zhu, and Shaogang Gong. 2018. Harmonious Attention Network for Person Re-Identification. *CoRR* abs/1802.08122 (2018). arXiv:1802.08122 http://arxiv.org/abs/1802.08122

[31] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. 2016. DeepFashion: Powering Robust Clothes Recognition and Retrieval With Rich Annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[32] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. *CoRR* abs/1908.02265 (2019). arXiv:1908.02265 http://arxiv.org/abs/1908.02265

[33] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical Question-Image Co-Attention for Visual Question Answering. *CoRR* abs/1606.00061 (2016). arXiv:1606.00061 http://arxiv.org/abs/1606.00061

[34] J. Ma, S. Pang, B. Yang, J. Zhu, and Y. Li. 2020. Spatial-Content Image Search in Complex Scenes. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2492–2500. https://doi.org/10.1109/WACV45572.2020.9093427

[35] L. Mai, H. Jin, Z. Lin, C. Fang, J. Brandt, and F. Liu. 2017. Spatial-Semantic Image Search by Visual Feature Synthesis. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1121–1130. https://doi.org/10.1109/CVPR.2017.125

[36] Tushar Nagarajan and Kristen Grauman. 2018. Attributes as Operators. *CoRR* abs/1803.09851 (2018). arXiv:1803.09851 http://arxiv.org/abs/1803.09851

[37] Duy-Kien Nguyen and Takayuki Okatani. 2018. Improved Fusion of Visual and Language Representations by Dense Symmetric Co-Attention for Visual Question Answering. *CoRR* abs/1804.00775 (2018). arXiv:1804.00775 http://arxiv.org/abs/1804.00775

[38] Hyeonwoo Noh, Paul Hongsuck Seo, and Bohyung Han. 2016. Image Question Answering Using Convolutional Neural Network With Dynamic Parameter Prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[39] OM Parkhi, A Vedaldi, and A Zisserman. 2015. Deep face recognition. 1–12.

[40] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. 2017. FiLM: Visual Reasoning with a General Conditioning Layer. *CoRR* abs/1709.07871 (2017). arXiv:1709.07871 http://arxiv.org/abs/1709.07871

[41] Filip Radenovic, Giorgos Tolias, and Ondrej Chum. 2016. CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples. *CoRR* abs/1604.02426 (2016). arXiv:1604.02426 http://arxiv.org/abs/1604.02426

[42] Filip Radenovic, Giorgos Tolias, and Ondrej Chum. 2018. Deep Shape Matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[43] Yong Rui, Thomas S. Huang, Michael Ortega, and Sharad Mehrotra. 1998. Relevance feedback: A power tool for interactive content-based image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology* 8, 5 (1998), 644–655. https://doi.org/10.1109/76.718510 Funding Information: Manuscript received October 31, 1997; revised May 30, 1998. This work was supported in part by NSF/DARPA/NASA DLI Program under Co-operative Agreement 94-11318, in part by ARL Cooperative Agreement DAAL01-96-2-0003, and in part by NSF CISE Research Infrastructure Grant CDA-9624396. The work of Y. Rui was also supported in part by a CSE Fellowship, the University of Illinois. The work of M. Ortega was also supported in part by CONACYT Grant 89061. This paper was recommended by Associate Editor S. Panchanathan.

[44] Adam Santoro, David Raposo, David G. T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter W. Battaglia, and Timothy P. Lillicrap. 2017. A simple neural network module for relational reasoning. *CoRR* abs/1706.01427 (2017). arXiv:1706.01427 http://arxiv.org/abs/1706.01427

[45] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A Unified Embedding for Face Recognition and Clustering. *CoRR* abs/1503.03832

(2015). arXiv:1503.03832 http://arxiv.org/abs/1503.03832

[46] Yaser Souri, Erfan Noury, and Ehsan Adeli-Mosabbeb. 2015. Deep Relative Attributes. *CoRR* abs/1512.04103 (2015). arXiv:1512.04103 http://arxiv.org/abs/1512.04103

[47] Yi Sun, Xiaogang Wang, and Xiaoou Tang. 2014. Deep Learning Face Representation from Predicting 10,000 Classes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *CoRR* abs/1706.03762 (2017). arXiv:1706.03762 http://arxiv.org/abs/1706.03762

[49] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2014. Show and Tell: A Neural Image Caption Generator. *CoRR* abs/1411.4555 (2014). arXiv:1411.4555 http://arxiv.org/abs/1411.4555

[50] 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 (CVPR)*.

[51] Liwei Wang, Yin Li, and Svetlana Lazebnik. 2015. Learning Deep Structure-Preserving Image-Text Embeddings. *CoRR* abs/1511.06078 (2015). arXiv:1511.06078 http://arxiv.org/abs/1511.06078

[52] Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. 2019. CAMP: Cross-Modal Adaptive Message Passing for Text-Image Retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[53] Huijuan Xu and Kate Saenko. 2016. Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering. In *Computer Vision – ECCV 2016*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer International Publishing, Cham, 451–466.

[54] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. 2018. Attention-Aware Compositional Network for Person Re-identification. *CoRR* abs/1805.03344 (2018). arXiv:1805.03344 http://arxiv.org/abs/1805.03344

[55] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *CoRR* abs/1502.03044 (2015). arXiv:1502.03044 http://arxiv.org/abs/1502.03044

[56] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. 2019. Cross-Modal Self-Attention Network for Referring Image Segmentation. *CoRR* abs/1904.04745 (2019). arXiv:1904.04745 http://arxiv.org/abs/1904.04745

[57] Sasi Kiran Yelamarthi, Shiva Krishna Reddy, Ashish Mishra, and Anurag Mittal. 2018. A Zero-Shot Framework for Sketch based Image Retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[58] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M. Hospedales, and Chen-Change Loy. 2016. Sketch Me That Shoe. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[59] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. 2017. Multi-modal Factorized Bilinear Pooling with Co-Attention Learning for Visual Question Answering. *CoRR* abs/1708.01471 (2017). arXiv:1708.01471 http://arxiv.org/abs/1708.01471

[60] Qi Zhang, Zhen Lei, Zhaoxiang Zhang, and Stan Z. Li. 2020. Context-Aware Attention Network for Image-Text Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[61] Chen Zhu, Yanpeng Zhao, Shuaiyi Huang, Kewei Tu, and Yi Ma. 2017. Structured Attentions for Visual Question Answering. *CoRR* abs/1708.02071 (2017). arXiv:1708.02071 http://arxiv.org/abs/1708.02071