Cross-Modal Joint Prediction and Alignment for Composed Query Image Retrieval

Yuchen Yang[1], Min Wang[2]*, Wengang Zhou[1,2]*, Houqiang Li[1,2]
[1]CAS Key Laboratory of Technology in GIPAS, EEIS Department,
University of Science and Technology of China
[2]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
yyc123@mail.ustc.edu.cn,wangmin@iai.ustc.edu.cn,zhwg@ustc.edu.cn,lihq@ustc.edu.cn

ABSTRACT

In this paper, we focus on the composed query image retrieval task, namely retrieving the target images that are similar to a composed query, in which a modification text is combined with a query image to describe a user’s accurate search intention. Previous methods usually focus on learning the joint image-text representations, but rarely consider the intrinsic relationship among the query image, the target image and the modification text. To address this problem, we propose a new cross-modal joint prediction and alignment framework for composed query image retrieval. In our framework, the modification text is regarded as an implicit transformation between the query image and the target image. Motivated by that, not only the combination of the query image and modification text should be similar to the target image, but also the modification text should be predicted according to the query image and the target image. We devote to aligning this relationship by a novel Joint Prediction Module (JPM). Our proposed framework can seamlessly incorporate the JPM into the existing methods to effectively improve the discrimination and robustness of visual and textual representations. The experiments on three public datasets demonstrate the effectiveness of our proposed framework, proving that our proposed JPM can be simply incorporated with the existing methods while effectively improving the performance.

CCS CONCEPTS

• **Computing methodologies → Machine translation**; • **Information systems → Image search**.

KEYWORDS

Image Retrieval, Multimedia Search, Image-Text Representation

ACM Reference Format:
Yuchen Yang, Min Wang, Wengang Zhou, Houqiang Li. 2021. Cross-Modal Joint Prediction and Alignment for Composed Query Image Retrieval. In *Proceedings of the 29th ACM Int’l Conference on Multimedia(MM ’21), Oct. 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3475483

1 INTRODUCTION

With the surge of visual data on the Internet, the image retrieval technologies become more and more important for many downstream tasks, such as product search [18] [11] and landmark retrieval [4] [29]. The common setting in image retrieval is taking an image as a query, searching for the most semantically relevant target image in the database. Generally speaking, there are two fundamental challenges in image retrieval, namely semantic gap and intention gap. Despite the fact that image retrieval community has made great breakthroughs in solving the semantic gap during the past decades [22] [28] [35] [20], the intention gap receives much less attention. This is because it’s hard for users to accurately express search intentions with a single image.

To alleviate the above problem, TIRG [32] makes the first attempt by introducing a novel framework namely composed query image retrieval. In this novel setting, users compose a single image with a modification text which describes the modification on the attributes or layout in the query image, and use the composed query to retrieve the target image in the database. In this manner, the query image can be selected more flexibly and does not need to accurately correspond to user’s search intention. The user’s input modification text can mitigate the gap between the query image and the target image.

On the composed query image retrieval task, most of the previous works [32] [14] [38] [7] focus on designing an appropriate network structure to learn joint image-text representations. Generally, they use text to implicitly guide the combination of visual and textual representations with various attention mechanisms. These attention mechanisms jointly express the semantics of the visual and textual representations. Specifically, TIRG [32] composes visual and textual representations by learning a gated residual connection. VAL [7] composes the visual representations and textual representation at CNN layers of different scales. It combines self-attention, channel attention and spatial attention in a composite transformer to obtain visiolinguistic representations. After obtaining the composited feature, these methods mentioned above generally align these two heterogeneous data (composited feature and target image’s feature) by exploiting a cross-modal matching loss such as tripet loss [13] and batch-based classification loss [32] [19].

In general, most of the previous work regards the composed query image retrieval task as a matching problem between heterogeneous modalities (image plus text vs image). They mainly focus on learning composited features and then directly optimizing the matching between the two heterogeneous modalities, namely

*Corresponding authors.

MM ’21, October 20–24, 2021, Virtual Event, China.
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475483

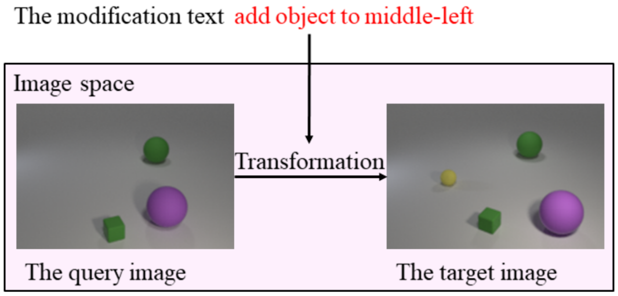

Figure 1: The modification text works as a kind of transformation in image space. The query image and the target image are viewed as a pair of transformed images. Our goal is to project the textual representation into the image space and make it focus on meaningful areas.

composed modality and image modality. Considering the diversity of text description and image content, directly optimizing the matching between the two modalities is difficult to learn optimal image-text representations, which limits the final retrieval performance. To address this problem, we propose to improve the feature fusion between the query image and modification text by mining the intrinsic relationship in different modalities.

To capture the intrinsic relationship in the composed query, the reserved region and modified region in the query image should be discriminated against according to the modification text. This requires the alignment of text semantics and image semantics. Actually, according to the difference between the query image and the target image, we can naturally obtain the prior knowledge of the image area that should be focused on by the modification text.

Based on the above analysis, we propose a novel composed query image retrieval framework, which directly aligns visual and textual representations by fully exploiting the relationship while inheriting the advantages of the previous methods. Given the query image and the target image, the modification text represents a kind of transformation and is able to be predicted directly. As shown in Fig. 1, the modification text “add object to middle-left” can be regarded as a sort of transformation in image space, which describes the difference in attributes between the query and target image. Our proposed framework is motivated by AET [39], which also learns image representations by viewing the query image and the target image as a pair of transformed images. Different from AET, we regard the modification text as an implicit transformation, which is more general than the geometric transformation used in AET. With the semantic discrimination introduced by the modification text, our method could be utilized in the composed image retrieval tasks while AET is usually used to learn image representations.

In sum, we split the composed query image retrieval task into three sub-tasks:

- Modeling the implicit transformation according to the input modification text.
- Learning visual representations that contain enough information to predict the transformation defined by the modification text.
- Learning joint image-text representations of the query image and the modification text to retrieve the target image.

The previous work usually devotes to the third sub-task and achieves good results. In this work, on the basis of the existing work, we jointly optimize these three sub-tasks by training our framework in an end-to-end manner. Our framework mainly includes four parts: image backbone to extract primitive visual representations, text backbone to extract textual representations, the Feature Fusion Module to transfer appropriate attributes from text to image and the Joint Prediction Module (JPM) to directly align the vision-language modalities. The Feature Fusion Module in our framework can be adopted from existing method, and is improved by optimizing JPM simultaneously. By training our framework end-to-end, we optimize the discrimination of visual and textual representations and promote the alignment among the three heterogeneous modalities (image vs text; image plus text vs image).

To summarize, our contributions are three-fold as follows::

- We propose a new auxiliary module CCDC for combined query image retrieval tasks, which can be incorporated into any existing framework to significantly alleviate the overfitting problem and improve performance.
- We propose a Joint Prediction Module to align the difference between the query image and the target image with the modification text, through which the whole model can capture the intrinsic relationship among them more accurately.
- We conduct experiments on three benchmarks. Results show that by incorporating the JPM, our proposed framework can improve the discrimination of visual and textual features and promote the learning of the Feature Fusion Module. We achieve a significant improvement of Recall @ K compared to the baseline methods on three benchmarks.

2 RELATED WORK

2.1 Image Retrieval

Although traditional content-based image retrieval has made great progress in the past decades, they still suffer two main difficulties, namely semantic gap and intention gap. The semantic gap refers to the gap between low-level features of an image and semantic meanings that people recognize from the image. Nowadays, many image retrieval methods [22] [20] [24] [8] [27] use CNN to learn the discriminative representations of images, which can mitigate the semantic gap to a certain extent. However, the intention gap has been pending for a long time. The intention gap means that a single image is difficult to accurately convey the user’s search intention.

Therefore, some people use text or other modality queries to express the search intention and the task of cross-modal retrieval (especially using text to retrieve images) has emerged. Cross-modal retrieval focuses on mapping different modalities to the same semantic space, and uses supervision information to guide the alignment of images and texts [6] [17] [26] [33] [15] [34] [37] . However, the information conveyed by the text is very abstract and sparse, which makes cross-modal retrieval very difficult and the application scenarios are not extensive.

In order to combine the advantages of image and text query, namely the rich semantic information of images and flexibility of texts, [32] first proposes the composed query image retrieval task. Following this setting that input query is specified in the form of an image with a modification text that describes desired modifications to the query image, many researchers devote to learning the joint expression of vision-language. [14] represents the input image as a set of local regions by applying a pre-trained region proposal network [23]. Then it uses self-attention [30] and cross-modal attention to learn a bidirectional correlation between the words in the modification text and local areas in the image. [38] uses Graph Convolutional Networks (GCNs) [16] to model the visual and textual features. Then it leverages a Jumping Graph Attention Network to inject semantic information from the text into the visual representations. Spatial attention and channel attention also help to fuse the features of images and texts. VAL [7] makes good use of different attention mechanisms, extracts CNN features of different scales to fuse them with textual representations, which make composited feature contain stronger discrimination. In this work, on the basis of previous methods, we aim to further explore the intrinsic relationship among the query image, target image and the modification text, which is often ignored by previous work. To be concrete, we not only consider the matching between composed query and target image but also focus on the alignment between image and text.

2.2 Visiolinguistic Representation Learning

Learning the joint representation of image and text is a fundamental work for many multimedia tasks, such as VQA [2] [36] [3] [25] and image captioning [31] [9]. Attention mechanism plays an important role in visiolinguistic representation learning. Various attention mechanisms can capture meaningful information explicitly, such as self-attention [30] mines richer semantic information and cross-modal attention captures the mutual information between heterogeneous modalitis. [40] exploits intra-modal and inter-modal attention to aggregate the context information in heterogeneous modality. [5] leverages an iterative matching scheme and a memory distillation unit to explore the correspondence between vision and language. Different from the descriptive text in many other tasks, the modification text in the composed query image retrieval task expresses the editorial intent of the image. Hence, in other tasks, the image contents that do not conform to the text may be considered as background noise, while in this task these image contents may be the important preserved parts for reflecting the accurate users' search intention. In this work, to explicitly formulate the connection between text and image, we exploit the spatial attention to exploring the difference between the query image and target image, which provides the necessary information for predicting the modification text.

2.3 Auto-Encoding Transformation

[39] presents a novel Auto-Encoding Transformation (AET) paradigm for unsupervised training mechanism. In the structure of AET, with a pair of transformed images and a random parametric transformation, they seek to train an auto-encoder, which can directly reconstruct the parameters of the transformation by leveraging the learned feature representations of original and transformed images. AET uses this unsupervised training method to improve the discrimination and robustness to various transformations of visual features.

In this work, we view the modification text as an implicit transformation in the image space and then reconstruct the modification text by joint prediction using visual information. Unlike AET with preset transformation parameters, we jointly optimize the transformation (the textual representation) and the visual representation in an end-to-end manner. This allows us to learn better joint image-text representations that accurately capture the corresponding intrinsic relationship and reflect the true search intention.

3 METHODOLOGY

The composed query image retrieval aims to return the relevant target images with an image and a modification text as a composed query. Let (I, M, I_t) represents the query image, the modification text and a candidate target image, respectively. The composed query image retrieval usually maps the query image, the modification text and the target image to a common space in order to learn a joint representation $f(I, M)$ which is similar with the representation $f_{target}(I_t)$. Following [32], we extract the $2D$ spatial feature map of the query image and the target image using a CNN, (*e.g.* ResNet18), obtaining $\phi_q \in R^{W \times H \times C} = f_{query}(I)$, $\phi_t \in R^{W \times H \times C} = f_{target}(I_t)$, where W is the width, H is the height , and C is the number of feature channels. Let $g_t \in R^C$ be the target image's average-pooled feature over the $2D$ spatial feature map ϕ_t. We use an LSTM to extract textual representations $\phi_m \in R^D$. We start with an overview of our method in Sec. 3.1 and introduce the details of the proposed Joint Prediction Module and the training and inference procedure in Sec. 3.2 and Sec. 3.3, respectively.

3.1 Overview

Fig. 2 shows an overview of our framework. We divide our framework into four parts:

- **Text Backbone.** We deliver the modification text to a standard LSTM, followed by a max-pooling and linear projection layer, obtaining $\phi_m \in R^D$, where $D = 512$ is the hidden size of the last linear projection layer.
- **Image Backbone.** We use a CNN backbone (*e.g.* ResNet18) to extract primitive visual representations $\phi_q \in R^{W \times H \times C}$ for query image, $\phi_t \in R^{W \times H \times C}$ for target image.
- **Feature Fusion Module.** We achieve the composited feature ϕ_{com} as the joint embedding $f(I, M)$ by feeding (ϕ_q, ϕ_m) into the Feature Fusion Module. This part can be adopted from any existing method. We adopt the two representative architecture of TIRG and VAL in the experiments. By optimizing the whole framework, the referred method used in this module is obviously improved and learns better joint feature representations, which is proved in the experiments.
- **Joint Prediction Module.** Considering that the ultimate goal of our framework is to fuse the visual and textual features ϕ_q and ϕ_m to obtain the composited feature ϕ_{com}, which should be similar to the target image's feature g_t. To achieve this goal, not only the feature fusion of ϕ_q and ϕ_m should accurately capture the semantic information of heterogeneous modality but also the expression of textual

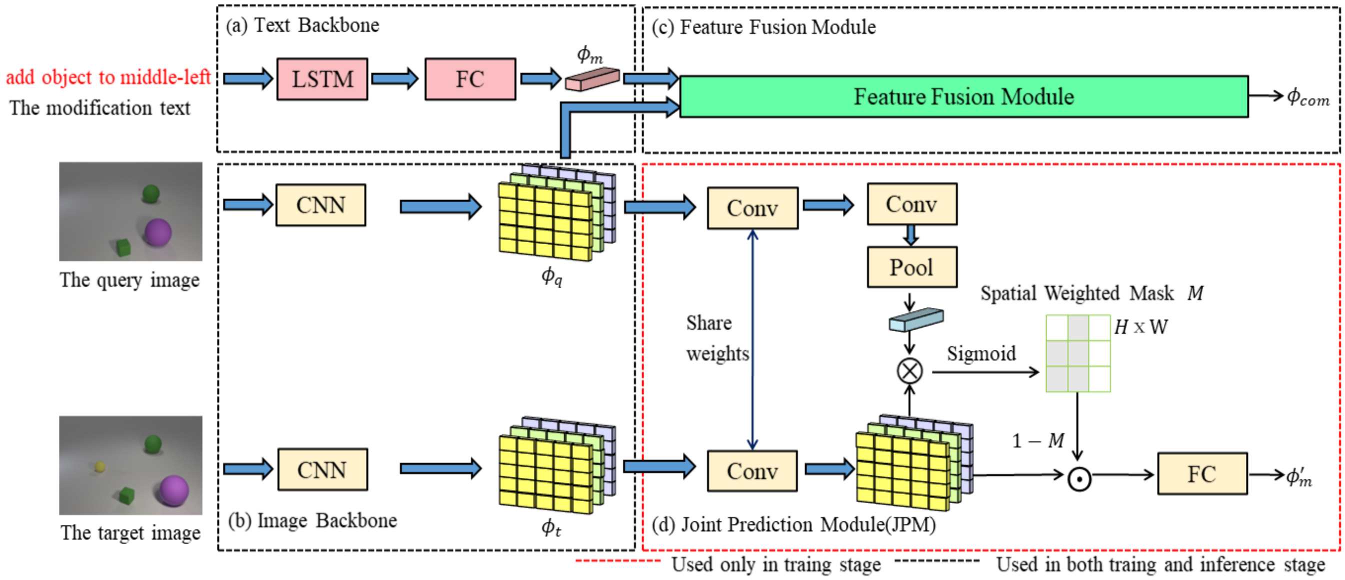

Figure 2: An overview of our framework. We divide the whole framework into four parts: (a) Text backbone to extract textual representation. (b) Image backbone to extract visual representation. (c) The Feature Fusion Module (can be adopted from any existing method) takes visual and textual representations from (a) and (b) as input and learns the composited feature. (d) Our proposed the Joint Prediction Module(JPM). The JPM takes visual representations from (b) and explores their difference to calculate a difference map. The difference map is used to highlight the corresponding semantic with the modification text in the target image's feature. Then we pass the weighted-sum feature through a linear projection layer to predict the modification text's feature. Symbols $\otimes$, $\odot$ stand for the element-wise dot product and weighted-sum pooling, respectively.

semantic and visual semantics should be robust and discriminative enough. However, visual features ϕ_q and ϕ_t and textual features ϕ_m are extracted from different pre-trained backbones (*e.g.* Imagenet pre-trained and Glove pre-trained). They are usually in different latent spaces and the semantic information they focus on is not necessarily consistent, which may hinder the subsequent feature fusion. To handle these problems, we propose a new Joint Prediction Module (JPM) to capture the difference between the query image and the target image for the prediction of transformation. We feed ϕ_q, ϕ_t into the Joint Prediction Module to predict the modification text $\phi_m^{'}$ and explore the intrinsic relationship among ϕ_q, ϕ_t and ϕ_m.

As shown in Fig. 2, during the training stage, the Feature Fusion Module and the Joint Prediction Module are trained together end-to-end. During the inference stage, the Joint Prediction Module is discarded, while the refined CNN backbone and LSTM are used for the feature extraction and feature fusion.

3.2 Joint Prediction Module

In this section, we detail the Joint Prediction Module that captures the difference between visual features ϕ_q and ϕ_t to predict the transformation defined by modification text M. To capture the specific characteristics of different datasets and improve the retrieval performance, visual representations should be discriminative to different attributes for different datasets. Actually, the modification text partially represents the characteristics of the datasets, (*e.g.* in the CSS dataset, the modification text tends to describe the color, size and position of different geometric objects, while in Fashion datasets, the modification text tends to describe the style, material and color of various clothes). Therefore, the visual representations should pay more attention to the characteristics described in the modification text. We achieve these goals by leveraging the JPM to exxplore difference information of visual representations to jointly predict the modification text.

The detailed architecture of the proposed Joint Prediction Module is illustrated in Fig. 2. ϕ_q and ϕ_t are input into two convolutional layers (denoted by Θ_x) with shared weights to extract the discriminative information A_q and A_t which contain the characteristics of the dataset:

$$A_q = \Theta_x(\phi_q), \tag{1}$$

$$A_t = \Theta_x(\phi_t). \tag{2}$$

Then, we characterize the difference between A_q and A_t by calculating a difference map. We transform A_q into a latent space and pool on the spatial dimension to express the desired semantics:

$$A_{se} = Pool(\Theta_y(A_q)), \tag{3}$$

where Θ_y is implemented as a 3×3 convolutional layer and *Pool* refers to average pooling. To calculate the difference map, we convolve A_t with semantic representation A_{se} to obtain a similarity map $M_{sim} \in R^{H^{'} \times W^{'}}$:

$$M_{sim} = A_t \otimes A_{se}. \tag{4}$$

The desired difference map M_{dif} is computed to describe the difference between A_q and A_t:

$$M_{dif} = 1 - sigmoid(M_{sim}). \tag{5}$$

We then use the obtained difference map M_{dif} as attention to highlight the dissimilar part from A_q in A_t by sum-pooling each channel of A_t:

$$\phi_{qt} = \sum_{h=1}^{H'}\sum_{w=1}^{W'} M_{dif}(h,w) * A_t(h,w), \tag{6}$$

where $\phi_{qt} \in R^C$ is the differential expression of query image I and target image I_t. The highlighted part in A_t reflects the modification applied to the query image I by the modification text M. Therefore ϕ_{qt} should have enough ability to predict the modification text's feature ϕ_m. We then map the ϕ_{qt} from the visual space to the semantic space by a linear projection layer f_{map} to predict ϕ_m:

$$\phi_m^{'} = f_{map}(\phi_{qt}), \tag{7}$$

where the $\phi_m^{'} \in R^D$ is the predicted text feature. Through the training of the JPM, visual and textual features can gradually grasp the corresponding intrinsic relationship, which is beneficial for the feature fusion between visual and textual features.

3.3 Training and Inference Procedure

In this section, we will introduce the training stage and inference stage of the proposed framework.

1)**Training.** During the training stage, the Feature Fusion Module and the Joint Prediction Module are trained jointly based on the triplets (I, M, I_t) of query image, modification text and target image. To be precise, a training minibatch B consists of K triplets and each triplet consists of (I_i, M_i, I_{t_i}), which represents the ith query image, modification text and target image, respectively. Our ultimate training objective is to close the gap between the learned composited representation ϕ_{com_i} and g_{t_i}. For ease of expression, we use $\phi_i^+ = g_{t_i}$ to represent the positive sample of the combination query (I_i, M_i), which should be similar with ϕ_{com_i}. $\phi_i^- = g_{t_j}, i \neq j$ (j is randomly sampled within the minibatch) represents the negtive sample of the ϕ_{com_i}. In the same way, we use $\psi_i^+ = \phi_{m_i}$ to represent the positive sample of pair (I_i, I_{t_i}), and $\psi_i^- = \phi_{m_j}, i \neq j$ to represent the negative sample of pair (I_i, I_{t_i}). Specially, we repeat the process of selecting negative samples H times to obtain $\phi_{i,h}^-$, which represents the hth negative sample of ϕ_{com_i}. We repeat the process of selecting negative samples N times to obtain $\psi_{i,n}^-$, which represents the nth negative sample of $\phi_{m_i}^{'}$. With these denotations, we next introduce the loss functions used to train our framework.

Composited Feature Learning. Following TIRG, we consider the soft triplet loss and the batch classification loss for learning composited feature. The soft triplet loss aims to close the distance between query and positive sample meanwhile extend the distance between the query and negative sample:

$$L_{Tri} = \frac{1}{HK}\sum_{i=1}^{K}\sum_{h=1}^{H} log\{1 + exp[\kappa(\phi_{com_i}, \phi_i^+) - \kappa(\phi_{com_i}, \phi_{i,h}^-)]\}, \tag{8}$$

where κ is an arbitrary similarity kernel function. We implement κ as the negative l_2 distance and set H to 2 same to TIRG.

The batch classification loss aligns the pair (I_i, M_i) with the target image I_{t_i} through a batch-based classification, which assigns an independent label to each target image:

$$L_{BC} = \frac{1}{K}\sum_{i=1}^{K} -log\{\frac{\kappa(\phi_{com_i}, \phi_i^+)}{\sum_t^K \kappa(\phi_{com_i}, \phi_t^+)}\}, \tag{9}$$

where κ is an arbitrary similarity kernel function. We implement κ as the dot product similar to TIRG. To summary, both the soft triplet loss and the batch classification loss aim to align the composited representation ϕ_{com_i} with the target image's representation g_{t_i}.

Vision-Language Alignment. The JPM is used to improve the discrimination of visual and textual features while mapping them to the same latent space. We consider the soft triplet loss and the MSE loss for aligning the output of JPM $\phi_m^{'}$ with the textual feature ϕ_m. The soft triplet loss is similar to the above-mentioned, with the anchor replaced by $\phi_m^{'}$ and the positive sample and negative sample replaced by ψ_i^+, ψ_i^-, respectively:

$$L_{mTri} = \frac{1}{NK}\sum_{i=1}^{K}\sum_{n=1}^{N} log\{1+exp[\kappa(\phi_{m_i}^{'}, \psi_i^+)-\kappa(\phi_{m_i}^{'}, \psi_{i,n}^-)]\}, \tag{10}$$

where κ is an arbitrary similarity kernel function. We implement κ as the negative l_2 distance and set N to 3. However, when dealing with datasets with concise modification texts, the MSE loss performs better than the soft triplet loss:

$$L_{MSE} = \frac{1}{K}\sum_{i=1}^{K}(\phi_{m_i}^{'} - \phi_{m_i})^2. \tag{11}$$

It is worth mentioning that, we find in the experiment that the LSTM network, Joint Prediction Module and CNN backbone possess different parameter quantities, which may lead to unbalanced parameter updates. When training the Joint Prediction Module, it is better to freeze other layers of CNN backbone and train only the last linear projection layer with the whole LSTM. This means that the gradient returned from the Joint Prediction Module is truncated at the last linear projection layer of CNN backbone. This stop gradient strategy does not affect the training of Feature Fusion Module.

2)**Inference.** During the inference stage, the Joint Prediction Module is discarded, while the refined CNN backbone and LSTM are used for the feature extraction and feature fusion. This is because the discrimination of visual and textual features has been improved through the training of Joint Prediction Module.

4 EXPERIMENTS

In order to verify the effectiveness of our proposed framework over different baseline methods, we conduct experiments on three benchmarks including CSS3D [32], Fashion200k [12] and FashionIQ [10]. In this section, we introduce the compared methods, the implementation details and the experimental results on different datasets in Sec. 4.1, Sec. 4.2 and Sec. 4.3.

4.1 Compared Methods

- **Image Only.** It uses query image's feature only as composited feature for pair (I, M) ,*i.e.*, $\phi_{com} = g_q \in R^C$, where g_q is the average-pooled feature of ϕ_q.
- **Text Only.** It uses modification text's feature only as composited feature for pair (I, M) ,*i.e.*, $\phi_{com} = g_m \in R^C$, where

g_m is obtained by passing the $\phi_m \in R^D$ through a linear projection layer whose output size is C.

- **Concat Only.** It directly concatenates the query image’s feature and modification text’s feature as composited feature for pair (I, M) ,*i.e.*, $\phi_{com} = Linear(\phi_q \bigoplus \phi_m) \in R^C$, where $\bigoplus$ represents the concatenation and *Linear* represents a linear projection layer.
- **TIRG [32].** It composes visual and textual representations by learning a gating connection and a residual connection to obtain composited feature ϕ_{com}.
- **[38].** It models the visual and textual features by building relation graphs. Then it leverages a Jumping Graph Attention Network to inject semantic information from the text into the visual representations.
- **VAL [7].** It composes visual and textual representations at multiple CNN layers using a composite transformer to obtain composited feature ϕ_{com}, which is able to better capture detail information of different image scales.
- **[14].** It tries to learn a many-to-many matching relationship between the words in the modification text and local areas in the image by leveraging self-attention and cross-modal attention.

4.2 Implementation Details

Our framework is a general boosting method for existing composed query image retrieval method such as TIRG and VAL by introducing JPM to align more accurate intrinsic relationship. For fair comparison with the referred methods, we implement our framework based on the released source code of these referred methods. When built upon TIRG, we conduct the experiments in Pytorch [21]. To be concrete, on the CSS dataset, we train the Feature Fusion Module using Triplet loss, we implement the Θ_x in our JPM as a 1×1 convolutional layer with stride = 1 and implement the Θ_y as a 3×3 convolutional layer with stride = 1, both followed by a ReLU. We use Adam optimizer with an initial learning rate = 0.0001. On the Fashion200k dataset, we train the Feature Fusion Module using Batch Classification loss, we implement the Θ_x and the Θ_y as 3×3 convolutional layers with stride = 1, both followed by a ReLU. We use SGD optimizer with an initial learning rate = 0.01, momentum = 0.9 and a weight decay of 10^{-6}. When built upon VAL, we conduct the experiments in Tensorflow [1]. To be concrete, on the FashionIQ dataset, we train the Feature Fusion Module using Triplet loss, we implement the Θ_x and the Θ_y as 3×3 convolutional layers with stride = 1, both followed by a ReLU. We use Adam optimizer with an initial learning rate = 0.0002. In order to evaluate performance fairly, we use recall at rank k(R@K) as evaluation metric, which is the same as previous work.

4.3 Experimental Results

CSS3D Dataset. The CSS3D dataset is proposed by TIRG, which describes the layout of geometric objects with different sizes, shapes, and colors. It contains about 19K training images and 18K testing images. The CSS3D dataset has complicated modification text which contains spatial position information such as “add object to middle-left”. The CSS3D dataset divides modification text into three categories: adding, removing, or changing object attributes.

Method	CSS3D	Fashion200k		
	R@1	R@1	R@10	R@50
Image Only	6. 3	3. 5	22. 7	43. 7
Text Only	0. 1	1. 0	12. 3	21. 8
Concat	61. 4	11. 9	39. 7	62. 6
TIRG [32]	76. 0	14. 1	42. 5	63. 8
[38]	76. 1	17. 3	45. 2	65. 7
[14]	79. 2	17. 8	48. 4	68. 5
TIRG* [32]	78. 5	14. 8	43. 7	64. 1
TIRG+JPM(MSE)	**83. 8**	**19. 8**	**46. 5**	**66. 6**
TIRG+JPM(Tri)	**83. 2**	**17. 7**	**44. 7**	**64. 5**

Table 1: Results on the CSS3D and Fashion200K dataset. The * means our implementation based on the source code. The “MSE” and “Tri” refer to training JPM by the Mean Squared Error loss or the Triplet loss.

As shown in Table 1, we conduct our experiments following the setting of TIRG, and the Recall@1 is improved by about 5% when JPM is incorporated. It outperforms the SOTA [14] by about 4% on the Recall@1. We use the two different losses (mean squared error and triplet loss) to train our JPM as mentioned in Sec. 3.3, and both achieve good results. Fig. 3 shows our qualitative results on CSS3D dataset. We observe that our method can handle attributes including spatial position, shape, and size well, and at the same time clearly understand the intention conveyed by the modification text. These results demonstrate that our proposed JPM could better capture the intrinsic relationship among the images and modification text, and improve the discrimination of the composed query.

Fashion200K Dataset. Fashion200K is a diverse dataset consisting of about 200k clothes images of various styles. Each image is equipped with some tags describing attributes. When generating training triplets, if the tags of two fashion images differ one word, we choose them as query image and target image, and the modification text is formulated as “replace A with B”. For an instance, an image of tags “black sequin dress" and an image of tags “black mesh dress" form a training triplet with the modification text “replace sequin with mesh". We incorporate our JPM into TIRG and use the same training split of around 172k images for training and the testset of 33,480 test queries for evaluation. It should be noted that the training triplets are generated during the training stage in the manner described above.

As shown in the Table 1, we have gained an improvement of about 5% in Recall@1 compared to the TIRG method. It is worth mentioning that, despite the huge performance gap between baseline methods TIRG and VAL, we have achieved a close performance when incorporating our JPM with the TIRG. We also use the two different losses to train our JPM, and the MSE loss performs better than the Triplet loss on Fashion200k dataset. We believe that this is due to the modification text in Fashion200k is very concise and easy to be predicted, while the intra-class difference between clothes is too diverse to construct effective enough triplets. Fig. 4 shows our qualitative results on Fashion 200k. We observe that our method captures various complex attributes. Through alignment with visual and textual representations, not only concise and direct

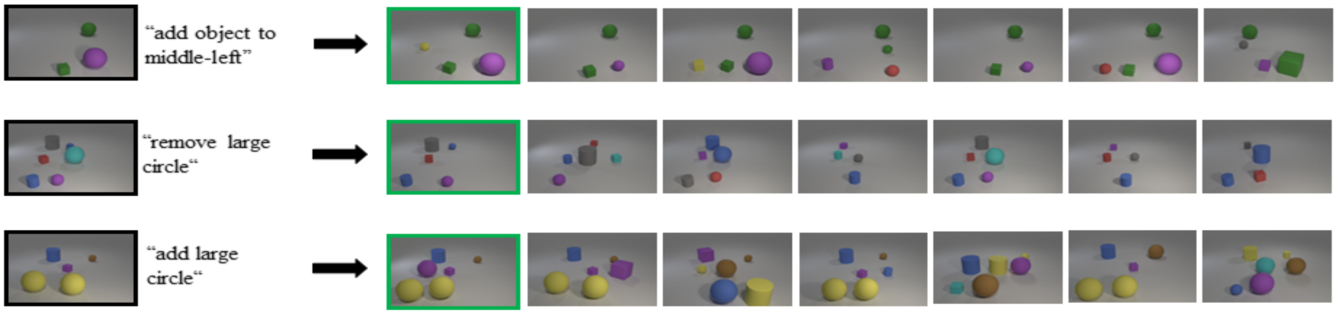

Figure 3: Qualitative results of our methond on CSS3D. Black box refers to the query image and green box refers to the target image.

(*e.g.* green, pink, sequin) but also abstract and subtle (*e.g.* side-zip, mesh, straight-leg) semantics are well understood.

FashionIQ Dataset. FashionIQ [10] is a natural language-based interactive fashion product retrieval dataset. Different from the Fashion200K dataset, in which the modification text is automatically generated with fixed length. The modification text in FashionIQ is more natural and complicated with an average length of 10.69 words. It contains 77,684 images, covering three categories: Dress, Toptee and Shirt. There are 18,000 image pairs in the 46,609 training images. Each pair is accompanied with around two natural language sentences as modification text.

Table 2 shows our results on FashionIQ. To verify the effectiveness of our proposed framework over different baseline methods, we conduct experiment following the setting of VAL. When incorporating with VAL, our proposed JPM outperforms other approaches in all the metrics with a 5.5% performance improvement in terms of the Recall@10 metric. We also notice that when training our JPM with MSE loss on FashionIQ, the effect of JPM is not obvious. We believe that the length and variety of modification text in FashionIQ lead to high complexity, and is hard to directly predict. Hence, training the JPM using the triplet loss performs better on this dataset. Fig. 5 shows our qualitative results on Fashion IQ. It is obvious that the complexity of text expression is higher than the other two datasets. Nevertheless, our method can retrieve the desired target image more accurately regardless of the difference in details and advanced semantics.

5 ABLATION STUDIES

In this section, we conduct ablation studies to analyze the influence of various network architecture and stop gradient training strategy of our proposed JPM. Particularly, we conduct experiments on Fashion200k and use the same evaluation metric as before.

Effect of Spatial Weighted Mask. In JPM, we calculate a spatial weighted mask to highlight the differences between feature maps A_q and A_t, and then catch their difference to predict text transformation ϕ_m. A natural baseline is that since A_q and A_t contain all the required information, the network learns the intrinsic relationship between them automatically without calculating the spatial weighted mask. Concretely, we concatenate them directly and feed the concatenation feature map of A_q and A_t into an averaged-pooling layer followed by a linear projection layer to predict text transformation ϕ_m.

Method	Dress		Toptee		Shirt		Avg	
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
Image Only	2. 92	10. 10	4. 53	11. 63	5. 34	14. 62	4. 26	12. 12
Text Only	8. 67	25. 08	9. 68	28. 25	8. 30	25. 02	8. 88	26. 11
Concat	9. 06	27. 27	10. 45	29. 83	9. 66	28. 06	9. 72	28. 33
TIRG [32]	14. 87	34. 66	19. 08	39. 62	18. 26	37. 89	17. 40	37. 39
VAL [7]	21. 12	42. 19	25. 64	49. 49	21. 03	43. 44	22. 60	45. 04
VAL* [7]	20. 39	43. 66	25. 47	49. 68	22. 41	43. 83	22. 75	45. 72
VAL+JPM(MSE)	21. 27	43. 12	25. 81	50. 27	21. 88	43. 3	22. 98	45. 59
VAL+JPM(Tri)	**21. 38**	**45. 15**	**27. 78**	**51. 70**	**22. 81**	**45. 18**	**23. 99**	**47. 34**

Table 2: Results on the FashionIQ dataset. The * means our implementation based on the source code. The "MSE" and "Tri" refer to training JPM by the Mean Squared Error loss or the Triplet loss. The "Avg" column refers to the average results on three categories.

Method	R@1	R@10	R50
TIRG*	14. 8	43. 7	64. 1
TIRG+JPM(MSE)	**19. 8**	**46. 5**	**66. 6**
TIRG+JPM(w/o SWM)(MSE)	18. 3	45. 8	66. 0
TIRG+JPM(Tri)	**17. 7**	**44. 7**	**64. 5**
TIRG+JPM(w/o SWM)(Tri)	16. 7	44. 7	64. 3

Table 3: Ablation study on effect of Spatial Weighted Mask(SWM).

Our experimental results are presented in Table 3. We can observe that without the Spatial Weighted Mask (SWM) design, our method still outperforms the baseline method by a gain of 3.5% on R@1. This further illustrates the motivation of our JPM is reasonable, textual feature and visual feature are well aligned by the JPM module. We also notice that without the SWM design, the performance degrades 1.5% on R@1 compared with our best results. This is also in line with our intuition that the modification text describes the difference between query and target images, and therefore the difference map represents the more accurate information.

Effect of Stop Gradient Training Strategy. Due to the great difference in parameter quantity of CNN backbone and LSTM, we freeze the parameter update for CNN backbone except the last linear projection layer when training the JPM. Therefore, we construct an ablation study on Fashion200k to verify the effect of this training

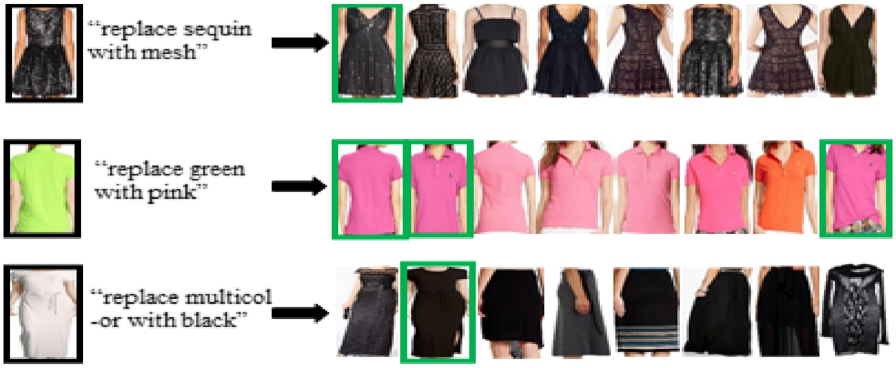
"replace sequin with mesh"
"replace green with pink"
"replace multicol -or with black"

Figure 4: Qualitative results of our methond on Fashion200K. Black box refers to the query image and green box refers to the target image.

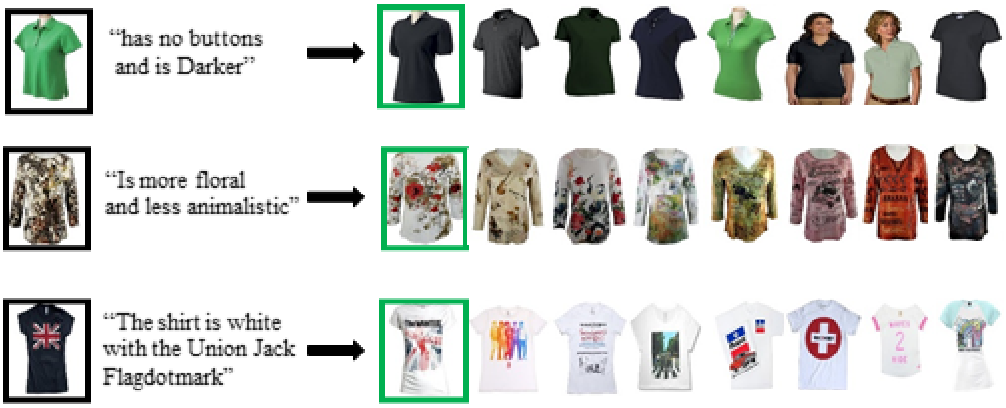
"has no buttons and is Darker"
"Is more floral and less animalistic"
"The shirt is white with the Union Jack Flagdotmark"

Figure 5: Qualitative results of our methond on FashionIQ. Black box refers to the query image and green box refers to the target image.

strategy. It can be seen from Table 4 that if the training strategy of freezing parameters is not applicable, the overall performance will drop obviously. This training strategy makes the training process update visual and textual features equally.

Method	R@1	R@10	R50
TIRG*	14. 8	43. 7	64. 1
TIRG+JPM(MSE)	**19. 8**	**46. 5**	**66. 6**
TIRG+JPM(w/o Stop)(MSE)	16. 9	44. 3	64. 6
TIRG+JPM(Tri)	**17. 7**	**44. 7**	**64. 5**
TIRG+JPM(w/o Stop)(Tri)	15. 1	43. 9	63. 9

Table 4: Ablation study on effect of Stop gradient training strategy.

6 CONCLUSIONS

We propose a new composed query image retrieval framework, which regards the modification text as an implicit transformation in image space while viewing the query image and the target image as a pair of transformed images. To explore the potential intrinsic relationship between them, we propose a Joint Prediction Module which can be incorporated into any existing method. The JPM leverages the difference between the query image and the target image to predict the modification text. With the flexible JPM, the proposed framework can be applied in many existing methods and improve their retrieval performance. Through extensive experiments, we demonstrate that our proposed method outperforms the baseline methods by a large margin.

7 ACKNOWLEDGEMENTS

This work was supported in part by the National Key R&D Program of China under contract 2018YFB1402605, in part by the National Natural Science Foundation of China under Contract 61822208 and 62021001, and in part by the Youth Innovation Promotion Association CAS under Grant 2018497. It was also supported by the GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC.

REFERENCES

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*. 265–283.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 6077–6086.

[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.* 2425–2433.

[4] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. 2016. NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 5297–5307.

[5] Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. 2020. Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 12655–12663.

[6] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. 2020. Fine-grained video-text retrieval with hierarchical graph reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 10638–10647.

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

[8] Albert Gordo, Jon Almazan, Jerome Revaud, and Diane Larlus. 2017. End-to-end learning of deep visual representations for image retrieval. *International Journal of Computer Vision* (2017), 237–254.

[9] Longteng Guo, Jing Liu, Jinhui Tang, Jiangwei Li, Wei Luo, and Hanqing Lu. 2019. Aligning linguistic words and visual semantic units for image captioning. In *Proceedings of the 27th ACM International Conference on Multimedia.* 765–773.

[10] Xiaoxiao Guo, Hui Wu, Yupeng Gao, Steven Rennie, and Rogerio Feris. 2019. Fashion IQ: A New Dataset towards Retrieving Images by Natural Language Feedback. *arXiv preprint arXiv:1905.12794* (2019).

[11] Yangyang Guo, Zhiyong Cheng, Liqiang Nie, Yinglong Wang, Jun Ma, and Mohan Kankanhalli. 2019. Attentive long short-term preference modeling for personalized product search. *ACM Transactions on Information Systems* 37, 2 (2019), 1–27.

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

[13] Alexander Hermans, Lucas Beyer, and Bastian Leibe. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737* (2017).

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

[15] Mengmeng Jing, Jingjing Li, Lei Zhu, Ke Lu, Yang Yang, and Zi Huang. 2020. Incomplete Cross-modal Retrieval with Dual-Aligned Variational Autoencoders. In *Proceedings of the 28th ACM International Conference on Multimedia.* 3283–3291.

[16] Thomas N Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. (2017), 1–14.

[17] Zhanghui Kuang, Yiming Gao, Guanbin Li, Ping Luo, Yimin Chen, Liang Lin, and Wayne Zhang. 2019. Fashion retrieval via graph reasoning networks on a similarity pyramid. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 3066–3075.

[18] 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.* 1096–1104.

[19] Yair Movshovitz-Attias, Alexander Toshev, Thomas K Leung, Sergey Ioffe, and Saurabh Singh. 2017. No fuss distance metric learning using proxies. In *Proceedings of the IEEE International Conference on Computer Vision.* 360–368.

[20] Tony Ng, Vassileios Balntas, Yurun Tian, and Krystian Mikolajczyk. 2020. SOLAR: second-order loss and attention for image retrieval. In *European Conference on Computer Vision.* 253–270.

[21] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703* (2019).

[22] Filip Radenović, Giorgos Tolias, and Ondřej Chum. 2018. Fine-tuning CNN image retrieval with no human annotation. *IEEE transactions on pattern analysis and machine intelligence* 41, 7 (2018), 1655–1668.

[23] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Conference and Workshop on Neural Information Processing Systems.* 91–99.

[24] Jerome Revaud, Jon Almazán, Rafael S Rezende, and Cesar Roberto de Souza. 2019. Learning with average precision: Training image retrieval with a listwise loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 5107–5116.

[25] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 8317–8326.

[26] Shupeng Su, Zhisheng Zhong, and Chao Zhang. 2019. Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 3027–3035.

[27] Marvin Teichmann, Andre Araujo, Menglong Zhu, and Jack Sim. 2019. Detect-to-retrieve: Efficient regional aggregation for image search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 5109–5118.

[28] Yurun Tian, Xin Yu, Bin Fan, Fuchao Wu, Huub Heijnen, and Vassileios Balntas. 2019. SOSNet: Second Order Similarity Regularization for Local Descriptor Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

[29] Giorgos Tolias, Tomas Jenicek, and Ondřej Chum. 2020. Learning and aggregating deep local descriptors for instance-level recognition. In *Proceedings of the European Conference on Computer Vision.* 460–477.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. (2017), 5998–6008.

[31] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 3156–3164.

[32] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. 2019. Composing text and image for image retrieval-an empirical odyssey. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 6439–6448.

[33] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. 2017. Adversarial cross-modal retrieval. In *Proceedings of the 25th ACM international conference on Multimedia.* 154–162.

[34] Yongxin Wang, Xin Luo, and Xin-Shun Xu. 2020. Label Embedding Online Hashing for Cross-Modal Retrieval. In *Proceedings of the 28th ACM International Conference on Multimedia.* 871–879.

[35] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. 2020. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 2575–2584.

[36] Li-Ming Zhan, Bo Liu, Lu Fan, Jiaxin Chen, and Xiao-Ming Wu. 2020. Medical Visual Question Answering via Conditional Reasoning. In *Proceedings of the 28th ACM International Conference on Multimedia.* 2345–2354.

[37] Yu-Wei Zhan, Xin Luo, Yongxin Wang, and Xin-Shun Xu. 2020. Supervised Hierarchical Deep Hashing for Cross-Modal Retrieval. In *Proceedings of the 28th ACM International Conference on Multimedia.* 3386–3394.

[38] Feifei Zhang, Mingliang Xu, Qirong Mao, and Changsheng Xu. 2020. Joint Attribute Manipulation and Modality Alignment Learning for Composing Text and Image to Image Retrieval. In *Proceedings of the 28th ACM International Conference on Multimedia.* 3367–3376.

[39] Liheng Zhang, Guo-Jun Qi, Liqiang Wang, and Jiebo Luo. 2019. Aet vs. aed: Unsupervised representation learning by auto-encoding transformations rather than data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2547–2555.

[40] 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.* 3536–3545.