Geometry Sensitive Cross-Modal Reasoning for Composed Query Based Image Retrieval

Feifei Zhang, Mingliang Xu, *Member, IEEE*, and Changsheng Xu, *Fellow, IEEE*

***Abstract*— Composed Query Based Image Retrieval (*CQBIR*) aims at retrieving images relevant to a composed query containing a reference image with a requested modification expressed via a textual sentence. Compared with the conventional image retrieval which takes one modality as query to retrieve relevant data of another modality, CQBIR poses great challenge over the semantic gap between the reference image and modification text in the composed query. To solve the challenge, previous methods either resort to feature composition that cannot model interactions in the query or explore inter-modal attention while ignoring the spatial structure and visual-semantic relationship. In this paper, we propose a geometry sensitive cross-modal reasoning network for *CQBIR* by jointly modeling the geometric information of the image and the visual-semantic relationship between the reference image and modification text in the query. Specifically, it contains two key components: a geometry sensitive inter-modal attention module (GS-IMA) and a text-guided visual reasoning module (TG-VR). The GS-IMA introduces the spatial structure into the inter-modal attention in both implicit and explicit manners. The TG-VR models the unequal semantics not included in the reference image to guide further visual reasoning. As a result, our method can learn effective feature for the composed query which does not exhibit literal alignment. Comprehensive experimental results on three standard benchmarks demonstrate that the proposed model performs favorably against state-of-the-art methods.**

***Index Terms*— Composed query based image retrieval, semantic gap, spatial structure, inter-modal attention, text-guided visual reasoning.**

I. Introduction

IMAGE retrieval, which enables similarity search across heterogeneous data, has emerged as a prominent research problem in the multimedia community and computer vision. It relates to numerous applications including person re-identification [1]–[3], fashion search [4], [5], face recognition [6]–[8], and product recommendation [9], [10], and thus garners wide interests.

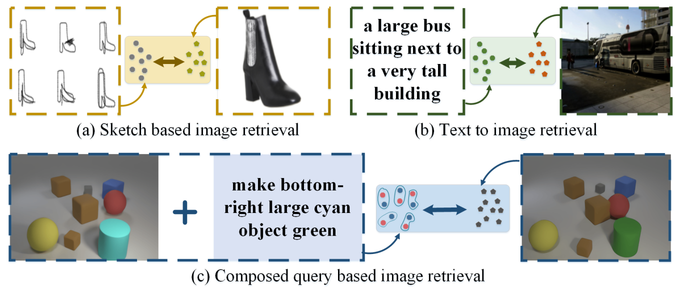

Fig. 1. Illustration of the difference between the *CQBIR* task (c) and the conventional image retrieval task [e.g., (a) sketch based image retrieval and (b) text to image retrieval].

In the general setting of the image retrieval, as shown in Figure 1 (a) and (b), a user searches for semantically relevant images in response to a single query item such as sketch and text. However, in many practical scenarios, the general image retrieval may not entirely fit the user's search intent. The users may have a 'concept' in mind, and they want to change the available query to match their specific requirements better. In particular, as shown in Figure 1 (c), given a query image, users may aspire to obtain a similar one but with some specific modifications. Therefore, in this paper, we devote to composed query based image retrieval (*CQBIR*), which aims at retrieving images relevant to a composed query. Specifically, as shown in Figure 1 (c), the query in *CQBIR* is composed of a reference image with a requested modification expressed via a sentence. This setting gives the user flexibility to express their intention in a more natural and meaningful way. Although the general image retrieval has been well studied [11]–[14], the *CQBIR* is a new direction and largely unexplored.

However, it is not easy to perform the *CQBIR*. The main challenge here is the semantic gap between the reference image and the modification text in the composed query. The problem becomes more severe when the image and text do not exhibit literal alignment. In detail, the image and text in the query usually have huge visual-semantic discrepancy, which calls for a comprehensive understanding of heterogeneous data. As shown in Figure 2, the text describes desired modifications to the image, which consequentially contains unequal semantic contents that do not embodied in the

Manuscript received March 16, 2021; revised August 19, 2021; accepted December 9, 2021. Date of publication December 31, 2021; date of current version January 10, 2022. This work was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0102200; in part by the National Natural Science Foundation of China under Grant 62036012, Grant 61720106006, Grant 62002355, Grant 61721004, Grant 61832002, Grant 62072455, Grant 62102415, Grant U1705262, and Grant U1836220; in part by the Key Research Program of Frontier Sciences of CAS under Grant QYZDJ-SSW-JSC039; and in part by the Beijing Natural Science Foundation under Grant L201001. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Zhu Li. *(Corresponding author: Changsheng Xu.)*

Feifei Zhang is with the School of Computer Science and Engineering, Tianjin University of Technology, Tianjin 300000, China, and also with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: feifei.zhang@ia.ac.cn).

Mingliang Xu is with the School of Information Engineering, Zhengzhou University, Zhengzhou 450000, China (e-mail: iexumingliang@zzu.edu.cn).

Changsheng Xu is with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China, and also with the Peng Cheng Laboratory, Shenzhen 518066, China (e-mail: csxu@nlpr.ia.ac.cn).

Digital Object Identifier 10.1109/TIP.2021.3138302

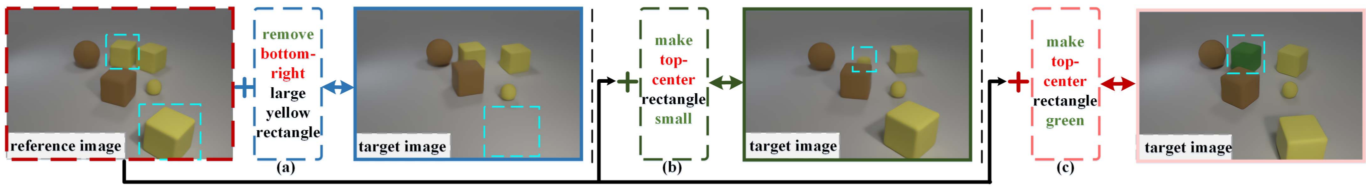

Fig. 2. Given a reference image and three different modification texts [15] [i.e., (a), (b), and (c)], the goal is to retrieve target images that are similar to the reference image but differ according to the requested modifications. To achieve this goal, the model needs to accurately find the image regions relevant to the given text and then correlates the regions with corresponding instructions in the text.

image (e.g., '*remove yellow rectangle*', '*make rectangle small*', and '*make rectangle green*'), and only some specific details (e.g., the size or color of one entity) in the reference image are required to change to retrieve the target. To deal with above issues, the predominant existing approaches [16]–[18] follow a two-stream strategy. First, they learn the visual and textual features by deep neural networks and well-known word embedding strategies, respectively. For the visual representation learning, several deep backbone models have been developed including ResNet [19], VGG [20], and GoogleNet [20], which have demonstrated early successes on a wide range of vision tasks. For language understanding, recurrent neural networks (RNNs) are impressively powerful in modeling the complex text data. After that, different feature composition methods, such as element-wise product, feature concatenation, and multi-modal residual networks, are exploited to combine the cross-modal feature. Although simple and intuitive, the visual and textual features in these methods are learned independently, making the approaches hard to model interactions between the text and image in the query. As a result, most of the existing methods might learn suboptimal representations for the composed query.

Recent studies on cross-modal learning [11], [13] suggest that incorporating interactions among vision-language elements would benefit in learning fine-grained representation for the heterogeneous data, such as the image and text. To achieve this goal, various inter-modal attention based methods have been proposed [12], [13]. To be specific, these methods devote to stressing possible interactions between image regions and sentence words. As shown in Figure 2, if the entity in the sentence (e.g., '*large yellow rectangle*', '*top-center rectangle*') shares inter-modality representation with their corresponding image regions, it would be helpful to capture the relevance among these two heterogeneous data. Although the attention-based methods have shown promising results in a wide range of visual tasks, such as image-text matching [12], referring expression comprehension [21], and video question answering [22], there are two issues with the existing methods. First, they usually treat the input image regions as 'bag-of-features' while neglecting the spatial structure of the regions. Nevertheless, the geometric information generally plays a vital role in the *CQBIR* task. For example, in Figure 2, the modification text describes not only the entity (e.g., '*rectangle*') but also the position of it (e.g., '*bottom-right*', '*top-center*'). Therefore, the spatial structure of the reference image is naturally critical to accurately localize the referent, i.e., the entity referred by the text. This inspires us to incorporate the geometric information into the inter-modal attention module. Second, the existing methods usually assume that the image and text contain equal semantics. However, as shown in Figure 2, the text in *CQBIR* describes desired modifications to the image, which consequentially contains semantics not embodied in the image. Particularly, as shown in Figure 2, given two different modification texts ('*make top-center rectangle small*' (b), '*make top-center rectangle green*' (c)), the same referent ('*top-center rectangle*') would be highlighted in the reference image, whereas the texts contain different modifications ('*make small*', '*make green*'). In cases like this, focusing only on the referent is not enough to learn effective representation for the composed query. We need to model the unequal semantics and conduct further reasoning about the referent with the unequally semantic contents.

To address the aforementioned issues, we propose a geometry sensitive cross-modal reasoning network for *CQBIR* by jointly modeling the spatial structure of images and the visual-semantic relationship. The relationship here denotes the equal or unequal semantic relation between the reference image and the modification text. Our proposed method consists of two crucial modules: a Geometry Sensitive Inter-Modal Attention module (GS-IMA) and a Text-Guided Visual Reasoning module (TG-VR). Specifically, the GS-IMA takes the visual, textual, and geometry feature as inputs, and stresses the correspondence between the image regions and the text words in a geometry-aware way to learn aligned visual and textual features. It improves the vanilla cross-modal attention by considering the geometric information in both implicit and explicit manner. The former (implicit manner) integrates the image regions and their corresponding spatial coordinate features into a unified representation, while the latter (explicit manner) directly adopts a key-query attention mechanism between the text and the geometric information. In this way, we can explicitly assign different weights to the image regions according to the similarities between the textual and the geometry feature. The TG-VR aims to manipulate the aligned visual feature with the unequal semantics. First, it constructs the unequally semantic contents in the modification text through modeling the discrepancy between the aligned and the original textual features. Then it exploits a visual reasoning network for further visual manipulation. By combining the GS-IMA and TG-VR, our proposed method can learn rectified geometry sensitive cross-modal representation for the composed query, and it does not require the heterogeneous data to exhibit literal alignment. We demonstrate the effectiveness of our approach on three popular *CQBIR* benchmarks.

The major contributions of this work can be summarized as follows.

1). We propose a geometry sensitive cross-modal reasoning network for composed query based image retrieval by jointly modeling the spatial structure and the visual-semantic relationship.
2). The geometric information of the image is novelly incorporated into inter-modal attention in both implicit and explicit manners, and the unequal semantics in the text are constructed for further visual reasoning. As a result, our method can learn effective cross-modal representation for unequal visual-semantic pairs in a geometry aware way, and it does not require the image and text to exhibit literal alignment.
3). We advance the state-of-the-art for the composed query based image retrieval on three benchmarks including CSS [15], Fashion200K [23], and MIT States [24], which powerfully demonstrates the effectiveness of our proposed method.

II. Related Work

A. Image Retrieval

Conventional image retrieval aims at searching semantically relevant images in response to a single query item, such as the text to image retrieval (TIR) [25], [26], content based image retrieval (CBIR) [10], [27], [28], and sketch based image retrieval (SBIR) [29], [30]. In particular, the TIR tries to bridge the domain gap between the image and text containing equally semantic contents. For example, Liu *et al.* [25] propose a graph structured matching network to learn fine-grained correspondence for the TIR. Wang *et al.* [26] investigate two-branch neural networks for learning the similarity between images and sentences. CBIR is the problem in which the query is in the form of a single image, which has been extensively explored for the tasks of face recognition [27], person re-identification [31], and product search [10]. For example, Choi *et al.* [31] introduce an ID-preserving person image generation network and a hierarchical feature learning module for visible-infrared person re-identification. The SBIR searches images associated with a sketch. For example, Bhunia *et al.* [29] introduce an on-the-fly sketch to image retrieval model to retrieve photos using an incomplete sketch.

Another line of work focuses on interactive retrieval. To be specific, the interactive retrieval aims at incorporating user feedback, such as the modification text [15], [32], [33], attribute [18], [34], [35], and spatial layout [36], [37], into an image retrieval system to refine or modify the image retrieval results tailored to user's expectations. Therefore, compared with conventional image retrieval, the interactive retrieval can better fit user's search intent. Our task (*CQBIR*) belongs to the interactive retrieval, and it takes modification text as the feedback because the text can naturally serve as an effective modality to express fine-grained intentions of users. The basic idea of *CQBIR* is to retrieve semantically related images for a query composed of a reference image with a modification text. The key to this task is to develop a cross-modal feature learning method that can effectively integrate the reference image and the modification text in the composed query. Guo *et al.* [15] propose the first work on the *CQBIR*, and they exploit a gated residual module to connect the image and the text. To learn image-text embeddings, Chen and Bazzani [38] propose a unified joint visual semantic matching approach with several compositional losses. Although intuitive and straightforward, these methods process the reference image and modification text independently, which fail to model the interactions in the query. Different from them, we use a variation of inter-modal attention to connect the image and text in a geometry aware way and adopt a text-guided visual reasoning network for further visual manipulation.

B. Attention Mechanism

Recently, the attention mechanism has gained popularity and been widely applied to various visual and textual applications, including image classification [39]–[41], image captioning [42], [43], visual question answering [44], [45], and object detection [46], [47]. For example, Huynh and Elhamifar [39] propose a shared multi-attention model for multi-label zero-shot learning, which improves the state of the art by a clear margin in two datasets. Pan *et al.* [42] introduce a unified attention block that fully employs bilinear pooling to selectively capitalize on visual information or perform multi-modal reasoning. Kim *et al.* [44] design a modality shifting attention network for multimodal video question answering, which can predict answers using an attention mechanism on both the visual and textual modalities. Chandran *et al.* [46] propose an end-to-end attention-driven fully convolutional architecture for facial landmark detection, which achieves superior performance over holistic state of the art convolutional architectures across different image resolutions from 256 to 4K. Benefiting from the great power of the attention mechanism, it has also been applied to the *CQBIR* task to fuse the visual and textual features in the composed query [32], [33]. For instance, to implement the *CQBIR*, Chen *et al.* [33] first concatenate the reference image with modification text and then utilize multiple transformers to selectively preserve and transform multi-level visual features conditioned on the semantics. Furthermore, some work has shown that the geometric information is a useful clue in the visual tasks. For example, Hosseinzadeh and Wang [32] first represent the image using a set of local areas and utilize a cross-modal attention module to fuse the image and text in the composed query. Then, this method averages the visual and positional features for each region to leverage the image's geometric information and uses a linear layer to integrate them to facilitate the matching task. Unlike the existing methods, first, the geometry feature is novelly incorporated into the inter-modal attention model, which makes our proposed method be able to model the geometric structure of the image in both implicit and explicit manners. Besides, the method in [32] directly performs cross-modal attention between the visual and textual feature, while ignores the unequal semantics. In contrast, our approach could model the unequally semantic contents, and precisely modify the visual features with the modification text.

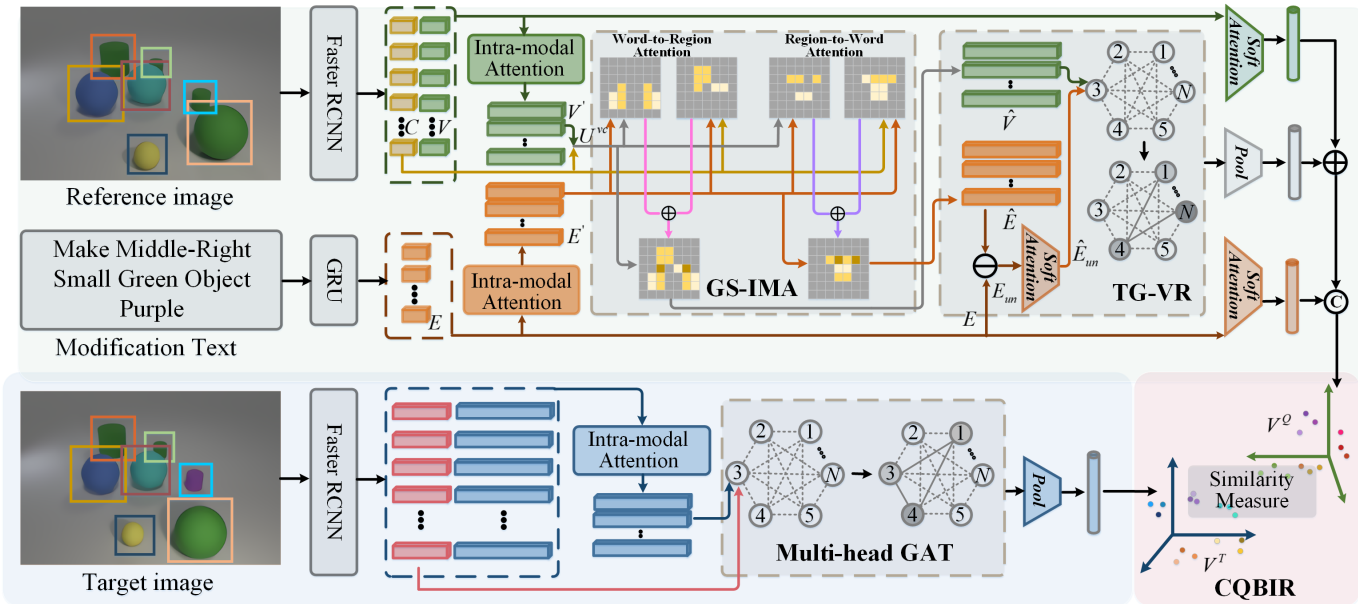

Fig. 3. The overall architecture of our proposed method, which contains two key components: GS-IMA and TG-VR. By considering the geometric structure and the unequal semantics, the proposed model can first learn cross-modal representation for the composed query in a geometry-aware way and then rectify the visual feature with the guidance of the text for *CQBIR*. In the figure, V, E, and C represent the preliminary visual, textual, and geometry feature, V' and E' are obtained by performing intra-modal attention on the V and E, U^{vc} is the visual feature that contains geometry information, $\hat{V}$ and $\hat{E}$ represent the visual and textual feature after our GS-IMA module, the E_{un} and $\hat{E}_{un}$ are the unequal semantics, and the V^Q and V^T are the final features for the composed query and target.

C. Text-Guided Visual Reasoning

The text-guided visual reasoning aims at learning visual representation with the guidance of the given text, which usually exists in the image manipulation [48]–[50], visual question answering [22], [51], [52], and referring expression grounding [25], [53]–[55]. For example, Li *et al.* [48] incorporate a text-image affine combination module into a generative adversarial network, which can select image regions relevant to the given text and manipulate the regions with corresponding semantic words. With the help of the powerful context reasoning ability of the graph convolution network (GCN) [56], Vo *et al.* [22] propose a deep heterogeneous graph alignment network for video question answering and get the state of the art result. Li *et al.* [52] exploit graph attention networks to address the visual question answering. Wang *et al.* [53] propose a language-guided graph attention network to address the referring expression comprehension task. Recently, there has been a surge of interest in using graph convolution algorithms for learning enhanced embeddings in a collaborative way [57], [58]. In this paper, we utilize a multi-head graph attention network (GAT) [59] to implement our TG-VR module. After the feature reasoning in the TG-VR, the visual representation of the reference image can be manipulated with the guidance of the modification text.

III. Proposed Method

Figure 3 shows a detailed illustration of our method. We aim to train a geometry sensitive cross-modal reasoning network to learn effective representation for the composed query (a reference image and a modification text) and compute the similarity between the query and target. As the representation of the target image is learned similarly with the query by omitting the cross-modal attention module and the textual inputs, we use the query as an example to present our model in detail in the following. As shown in Figure 3, we first feed the image into a faster R-CNN [60] to extract n image region feature $V \in \mathbb{R}^{n \times d_v}$ and n spatial coordinate feature $C \in \mathbb{R}^{n \times d_c}$. Meanwhile, we adopt a bidirectional GRU to generate text embedding $E \in \mathbb{R}^{m \times d_e}$, where m represents the number of words in the modification text. Then, we perform intra-modal attention on the feature V and E to stress the relations within each modality, and denote the updated feature as $V' \in \mathbb{R}^{n \times d_v}$ and $E' \in \mathbb{R}^{m \times d_e}$, respectively. Specifically, the two intra-modal attention modules have an identical architecture, but they are independent of each other and do not share the weights. Take the image region feature V as an example, we first employ three fully-connected layers to transform the V into query, key, and value features following [61], and calculate the within-modality importance between image regions through performing dot products between the transformed query and key features. Then, we pass information flows according to the learned importance weights and aggregate features to update the image region feature, which can be formulated as

$$V' = (softmax(\frac{VW^QV^TW^K}{\sqrt{d_v}}) + I)VW^V, \tag{1}$$

where W^Q, W^K, and W^V are the learned weight matrices, and I is the identity matrix used to preserve the original information. Via the message passing between visual regions (or text words), we can better capture the relations within each modality and extract global context information. Thereafter, we learn the integrated feature for the query with two crucial modules: 1) a Geometry Sensitive Inter-Modal Attention module (GS-IMA), which takes the V', E', and C as inputs to learn aligned visual feature $\hat{V} \in \mathbb{R}^{n\times d_v}$ and textual feature $\hat{E} \in \mathbb{R}^{m\times d_e}$ in a geometry aware way. 2) A Text-Guided Visual Reasoning module (TG-VR), which first constructs unequal semantics $E_{un} \in \mathbb{R}^{m\times d_e}$ by the aligned and original textual feature $\hat{E}$ and E, and then rectify the feature $\hat{V}$ to $\hat{V}' \in \mathbb{R}^{n\times d_v}$ under the guidance of the E_{un}. Finally, we combine the variation of the V, $\hat{V}'$, and E to represent the composed query and train our model end-to-end via a matching objective.

A. Geometry Sensitive Inter-Modal Attention

As shown in Figure 3, the modification text usually contains positional words (e.g., 'middle-right'). Therefore, the spatial structure is beneficial to cross-modal feature learning. To this end, we propose a geometry aware inter-modal attention network, which improves the vanilla cross-modal attention by considering the image region's geometric information in both implicit and explicit manners.

First, we integrate the visual feature V' and coordinate feature C into a unified feature $U^{vc} \in \mathbb{R}^{n\times d_v}$ as follows:

$$u_i = F_c([v_i', c_i]), \tag{2}$$

where $u_i \in U^{vc}$, v_i' and c_i represent the ith region feature and coordinate feature in V' and C, $[\cdot]$ means concatenation, and F_c is a linear layer. Furthermore, c_i is defined as $c_i = [c_i^x, c_i^y, w_i, h_i, w_i h_i]$, where (c_i^x, c_i^y) denotes the normalized center coordinate of the ith region, and w_i, h_i correspond to the normalized width and height of the region.

Then, taking the textual feature E', visual feature U^{vc}, and the coordinate feature C as inputs, the word-to-region attention A^{tv} is computed as follows:

$$A^{tv} = softmax(\frac{Q_t K_t^T}{\sqrt{d_e}}) + \alpha * softmax(\frac{Q_t' K_t'^T}{\sqrt{d_e}}) \tag{3}$$

$$\begin{aligned} Q_t &= E' W^{Q_t^0}, \quad K_t = U^{vc} W^{K_t^0}, \quad Q_t' = E'' W^{Q_t^1}, \\ K_t' &= C' W^{K_t^1}. \end{aligned} \tag{4}$$

In the above formulation, K_t, K_t' are the visual and geometric keys, and Q_t, Q_t' are the text queries corresponding to the K_t and K_t'. W with superscript represents the learned weight matrix, α is the trade-off hyperparameter, $C' = Relu(fc(C))$, and $E'' = Dropout(fc(E'))$.

The attention matrix $A^{tv} \in \mathbb{R}^{m\times n}$ in Eq. (3) represents the geometry aware similarities of word-region pairs. In this setting, the implicit manner in our GS-IMA is implemented by the Eq. (2), and the first item in Eq. (3). Specifically, since U^{vc} in Eq. (2) contains coordinate features, the word-region similarity in the first item in Eq. (3) is calculated conditioned on the geometric information. Besides, the explicit manner is realized by the second item in Eq. (3). In this manner, the coordinate feature C is directly converted to a key vector K_t', and the textual feature is taken as the query vector Q_t'. Through the key-query attention between the K_t' and Q_t', we can explicitly assign different weights to the image regions according to the similarities of the word-coordinate pairs. Overall, the attention matrix A^{tv} between the word and image regions is calculated through comprehensively considering the coordinate feature of each image region, which therefore demonstrates geometry sensitive. Likewise, the region-to-word attention $A^{vt} \in \mathbb{R}^{n\times m}$ is calculated as follows:

$$A^{vt} = softmax(\frac{Q_v K_v^T}{\sqrt{d_v}}) + \beta * softmax(\frac{Q_v' K_v'^T}{\sqrt{d_v}}), \tag{5}$$

$$\begin{aligned} Q_v &= U^{vc} W^{Q_v^0}, \quad K_v = E' W^{K_v^0}, \quad Q_v' = C' W^{Q_v^1}, \\ K_v' &= E'' W^{K_v^1}, \end{aligned} \tag{6}$$

where Q_v, Q_v' are the region and coordinate queries, and K_v, K_v' are the corresponding word keys. W with superscript is the learned weight matrix, and β is the trade-off hyperparameter. Finally, based on Eq. (3) and (5), we can reconstruct features for one modality by all samples in another modality, that is

$$\hat{E} = A^{tv} U^{vc} W^e, \quad \hat{V} = A^{vt} E' W^v, \tag{7}$$

where $\hat{E}$, $\hat{V}$ represent the aligned textual and visual feature, and W^e, W^v are the learned weight matrices. To facilitate model optimization, residual connection followed by layer normalization is applied to the $\hat{E}$ and $\hat{V}$ in our method.

B. Text-Guided Visual Reasoning

The GS-IMA in Section III-A aims at learning aligned visual and textual features, while the TG-VR devotes to further visual reasoning with the guidance of the unequal semantics. To be specific, given the original and the aligned textual feature E and $\hat{E}$, the unequal semantics E_{un} is formulated as the subtraction of them:

$$E_{un} = E - \hat{E}. \tag{8}$$

Then we apply soft attention on the sequence of E_{un} to generate the unequally semantic embedding $\hat{E}_{un} \in \mathbb{R}^{d_e}$ as:

$$\alpha_i = softmax(W^{e_2}\sigma(W^{e_1} e_{un}^i + b^{e_1}) + b^{e_2}), \tag{9}$$

$$\hat{e}_{un}^i = \sum_{i=1}^{m} \alpha_i e_{un}^i \tag{10}$$

where W and b with superscript are the learnable parameters, σ is a tanh function, e_{un}^i is the ith word in E_{un}, α_i is the attention vector for word i, and $\hat{E}_{un} = \{\hat{e}_{un}^1, \hat{e}_{un}^2, \ldots, \hat{e}_{un}^m\}$. After that, the $\hat{E}_{un}$ acts as an instruction to guide the visual feature reasoning, which is capable of coping with the aligned visual feature $\hat{V}$ in any neural network. In this section, we utilize a multi-head GAT [59] to implement the text-guided feature reasoning. Specifically, we first concatenate the feature $\hat{E}_{un}$ to each region feature in $\hat{V}$ to construct a heterogeneous input matrix S. Then we build a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, S)$ where $\mathcal{V}$ and $\mathcal{E}$ are the sets of vertexes and edges and S is the set of vertex features. Finally, the rectified visual feature $\hat{V}' \in \mathbb{R}^{n\times d_v}$ is calculated as follows:

$$\begin{aligned} \hat{v}_i' &= \big\Vert_{h=1}^{H} \sigma(\sum_{j\in n_i} \alpha_{ij}^h W^{h_1} s_j), \\ \alpha_{ij}^h &= \frac{\exp((W^{h_2} s_i)^T \cdot (W^{h_3} s_j))}{\sum_{j\in n_i} \exp((W^{h_2} s_i)^T \cdot (W^{h_3} s_j))} \end{aligned} \tag{11}$$

where $\hat{v}'_i$ is the rectified feature for the ith region in $\hat{V}'$, H is the number of the independent attention module, s_i means the ith feature in S, n_i denotes the neighborhood of region i, α_{ij}^h indicates the attention coefficient in the hth attention module to measure the importance of feature i for a neighbor j, σ is a nonlinear function such as ReLU, and W with superscript is the learnable parameter. Through Eq. (11), we can obtain the $\hat{V}' = \{\hat{v}'_1, \hat{v}'_2, \ldots, \hat{v}'_n\}$ with the guidance of the unequally semantic contents in the modification text.

C. Query and Target Representation

As shown in Figure 3, the final representation V^Q for the composed query is a combination of: 1) a visual feature $V^f \in \mathbb{R}^{d_v}$ from the visual feature V encoded by a soft attention layer, which serves as the visual content preservation purpose. 2) A visual representation $\hat{V}'^f \in \mathbb{R}^{d_v}$ by conducting mean pooling over the $\hat{V}'$. 3) A textual feature $E^f \in \mathbb{R}^{d_e}$ from the E encoded by a soft attention layer. To be specific, as shown in Figure 3, we first perform element-wise addition operation on the V^f and $\hat{V}'^f$, which are then concatenated with the E^f. Finally, we feed them to a fully-connected layer with batch normalization and dropout to learn the final feature $V^Q \in \mathbb{R}^{d_f}$ for the composed query.

The representation $V^T \in \mathbb{R}^{d_f}$ for the target image is learned similarly with the query, and the implementation details for it are shown in the bottom part in Figure 3. Specifically, taking a target image as input, we also use a Faster RCNN [60] to extract the visual feature and spatial coordinate feature, and then perform self-attention on the visual feature to stress the relations among the regions. After that, taking the enhanced visual feature and the coordinate feature as inputs, a multi-head GAT [59] followed by meaning pooling is used to produce the V^T for the target.

D. Objective Function

Given the representation of the query (V^Q) and target (V^T), we aim to introduce a loss function that can enforce the matched query-target pairs to be clustered and the unmatched ones to be separated in the embedding spaces. To accomplish this task, we employ a hinge-based triplet ranking loss [12], [13] to train our model end-to-end. Instead of comparing with all negatives, we focus on optimizing hard negative samples that produce the highest loss. Formally, our objective function is defined as follows:

$$\begin{aligned}\mathcal{L} = &\max[0, \gamma - F(V^Q, V^T) + F(V^Q, \tilde{V}^T)] \\ &+ \max[0, \gamma - F(V^Q, V^T) + F(\tilde{V}^Q, V^T)], \quad (12)\end{aligned}$$

where γ is a margin value, and $F(\cdot)$ denotes the semantic similarity function. We use the usual inner product as the $F(\cdot)$ in our experiment. $\tilde{V}^T$ and $\tilde{V}^Q$ represent the hard negatives for the positive pair (V^Q, V^T), i.e., $\tilde{V}^T = \arg\max_{x \neq V^T} F(V^Q, x)$ and $\tilde{V}^Q = \arg\max_{y \neq V^Q} F(y, V^T)$. In practice, instead of summing over all the negative samples, we only use the hard negatives in a mini-batch, which has proved to be effective for the retrieval performance [62].

IV. Experimental Results

In this section, we first introduce the datasets used in our method and the implementation details of our approach. Then, to demonstrate the effectiveness of our proposed method, we carry out extensive experiments on these three benchmarks. Finally, we conduct ablation studies and display some qualitative results to thoroughly investigate our model.

A. Datasets and Evaluation Metric

To evaluate the effectiveness of our proposed method for the *CTBIR*, we conduct thorough experiments on three benchmarks: Fashion200k [23], MIT-States [24], and CSS [15]. Details are as follows.

1) CSS: [15] is one of the most popular datasets for the *CQBIR* task. It contains 18,012 training queries and 18,057 test queries generated based on the CLEVR toolkit [63]. In this dataset, each query consists of a reference image (2D or 3D) and a modification text. The objects in the image are of different colors, shapes, and sizes, and the modification text has three different types (adding, removing, and changing). Following [15], [32], we train and test our model by all the training and test queries.

2) Fashion200K: [23] is a dataset for fashion search, which consists of about 200k person images with fashion clothes. Following [15], [32], we use 172,049 images and 33,480 test queries for training and evaluation. During training, pairs of images with one word difference in their descriptions are selected as the query image and target image, and the modification text is comprised of the different words.

3) MIT States: Ref. [24] contains about 60k images, and each image is annotated with a noun and an adjective. In total, it has 245 nouns and 115 adjectives. Following [15], [32], we use 80 nouns for training and the rest for testing. Pairs of images with different adjectives but the same nouns are selected as the query image and target image, and the different adjectives are taken as the modification text.

4) Evaluation Metric: Conventionally, we take the Recall at K (R@K) as the evaluation metric, i.e., the fraction of queries for which the correct item is retrieved among the top K results. The experiment on each dataset is repeated 5 times, and both the mean and variance are reported.

B. Implementation Details

Our network is constructed according to Figure 3. We adopt a Faster RCNN [60] with ResNet-101 [19] backbone to extract 36 region features and coordinate features for each image. The modifications in the text are first tokenized and each word is embedded using a 300-dimensional textual feature. After that, we feed them into a bi-directional GRU and perform intra-modal attention on the sequence of the GRU hidden states to generate text embedding, which is then updated in our geometry sensitive inter-modal attention module and the text-guided visual reasoning module to learn text-adaptive representation for the query image. The d_v, d_e and d_f are set to 2048, and the d_c is 5. The dimensionality of the coordinate feature is mapped to 2048 in Eq. (4) and (6). The number of heads H is set to 16,

TABLE I

COMPARISON OF OUR METHOD WITH OTHER STATE-OF-THE-ART METHODS ON CSS. THE BEST RESULT IS HIGHLIGHTED IN BOLD

Method	3D-to-3D		2D-to-3D
	R@1	R@5	R@1
Text only [22]	0.1	-	0.1
Image only [22]	6.3	-	6.3
Concatenation [22]	$60.6^{\pm 0.8}$	-	27.3
Show and Tell [65]	$33.0^{\pm 3.2}$	-	6.0
Param Hashing [66]	$60.5^{\pm 1.9}$	-	31.4
Relationship [64]	$62.1^{\pm 1.2}$	-	30.6
MRN [67]	$60.1^{\pm 2.7}$	-	26.8
FiLM [68]	$65.6^{\pm 0.5}$	-	43.7
TIRG [22]	$73.7^{\pm 1.0}$	-	46.6
JAMMA [69]	$76.07^{\pm 0.4}$	-	48.85
LBF [32]	$79.2^{\pm 1.2}$	94.08	$55.69^{\pm 0.9}$
Ours	**81.81**$^{\pm 0.6}$	**99.11**	**58.74**$^{\pm 0.5}$

TABLE II

COMPARISON OF OUR METHOD WITH OTHER STATE-OF-THE-ART METHODS ON FASHION200K. THE BEST RESULT IS HIGHLIGHTED IN BOLD

Method	R@1	R@10	R@50
Text only [22]	1.0	12.3	21.8
Image only [22]	3.5	22.7	43.7
Concatenation [22]	$11.9^{\pm 1.0}$	$39.7^{\pm 1.0}$	$62.6^{\pm 0.7}$
Show and Tell [65]	$12.3^{\pm 1.1}$	$40.2^{\pm 1.7}$	$61.8^{\pm 0.9}$
Param Hashing [66]	$12.2^{\pm 1.1}$	$40.0^{\pm 1.1}$	$61.7^{\pm 0.8}$
Relationship [64]	$13.0^{\pm 0.6}$	$40.5^{\pm 0.7}$	$62.4^{\pm 0.6}$
MRN [67]	$13.4^{\pm 0.4}$	$40.0^{\pm 0.8}$	$61.9^{\pm 0.6}$
FiLM [68]	$12.9^{\pm 0.7}$	$39.5^{\pm 2.1}$	$61.9^{\pm 1.9}$
TIRG [22]	$14.1^{\pm 0.6}$	$42.5^{\pm 0.7}$	$63.8^{\pm 0.8}$
JVSM [38]	15.6	44.0	63.7
JVSM* [38]	19.0	52.1	70.0
JAMMA [69]	$17.34^{\pm 0.6}$	$45.28^{\pm 0.9}$	$65.65^{\pm 0.8}$
LBF [32]	$17.78^{\pm 0.5}$	$48.35^{\pm 0.6}$	$68.5^{\pm 0.5}$
Ours	**21.57**$^{\pm 0.7}$	**52.84**$^{\pm 0.5}$	**70.12**$^{\pm 0.9}$

TABLE III

COMPARISON OF OUR METHOD WITH OTHER STATE-OF-THE-ART METHODS ON MIT STATES. THE BEST RESULT IS HIGHLIGHTED IN BOLD

Method	R@1	R@10	R@50
Text only [22]	7.4	21.5	32.7
Image only [22]	3.3	12.8	20.9
Concatenation [22]	$11.8^{\pm 0.2}$	$30.8^{\pm 0.2}$	$42.1^{\pm 0.3}$
Show and Tell [65]	$11.9^{\pm 0.1}$	$31.0^{\pm 0.5}$	$42.0^{\pm 0.8}$
Param Hashing [66]	$8.8^{\pm 0.1}$	$27.3^{\pm 0.3}$	$39.1^{\pm 0.3}$
Relationship [64]	$12.3^{\pm 0.5}$	$31.9^{\pm 0.7}$	$42.9^{\pm 0.9}$
MRN [67]	$11.9^{\pm 0.6}$	$30.5^{\pm 0.3}$	$41.0^{\pm 0.2}$
FiLM [68]	$10.1^{\pm 0.3}$	$27.7^{\pm 0.7}$	$38.3^{\pm 0.7}$
TIRG [22]	$12.2^{\pm 0.4}$	$31.9^{\pm 0.3}$	$43.1^{\pm 0.3}$
JAMMA [69]	$14.27^{\pm 0.5}$	$33.21^{\pm 0.5}$	$45.34^{\pm 0.6}$
LBF [32]	$14.72^{\pm 0.6}$	$35.30^{\pm 0.7}$	$46.56^{\pm 0.5}$
Ours	**17.28**$^{\pm 0.4}$	**36.45**$^{\pm 0.4}$	**47.04**$^{\pm 0.5}$

and the margin γ is set to 0.2. We implement our framework with PyTorch. The algorithm is trained for 150 epochs on CSS and 400 epochs on Fashion200K and MIT States. Adam [64] is used as the training optimizer, with a batch size of 128 and an initial learning rate of 0.0002. The learning rate is decayed by $3/5$ every 40 epochs on the CSS and 60 epochs on the Fashion200K and MIT States. The setting of hyperparameter α in Eq. (3), β in Eq. (5), the number of heads in multi-head GAT in Eq. (10) (H), and the margin γ in Eq. (11) are investigated in Section IV-D. Based on the experimental setting described above, the memory usage of our proposed method is 9.76G. Besides, based on a platform with a GTX 1080TI GPU and IntelXeon CPU E5-2687W (3.0GHZ), the inference time is 70.59ms per query on the CSS dataset.

C. Comparisons With State-of-the-Arts

1) Baselines: We evaluate our method by comparing its performance with several state-of-the-art methods on the CSS, Fashion200K, and MIT States in Table I, Table II, and Table III, respectively. The following is a brief introduction to the compared methods: Text only [15] and Image only [15] just use the modification text or the reference image for *CQBIR*. Concatenation [15] fuses the image and text in the query by MLP. Show and Tell [65] utilizes LSTM to encode the query. Param Hashing [66], Relationship [64], MRN [67], and FiLM [68] are originally used for the VQA task. These methods mainly leverage different composition methods to integrate the composed query, such as transformation matrix, element-wise multiplication, and feature-wise affine transformation. TIRG [15], JVSM, JVSM* [38], LBF [32], and JAMMA [69] are the state-of-the-art methods designed for the *CQBIR*. *TIRG* [15] fuses the visual and textual features with a gated residual connection network. JVSM and JVSM* [38] learn composition features for the query by jointly associating visual and textual modalities in a shared feature space. '*' indicates that the model is trained with privileged information, such as the original text description for the reference image and target image. LBF [32] adopts intra-modal attention and inter-modal attention for *CQBIR*, and JAMMA [69] utilizes graph attention networks for the task.

2) Results on CSS: The results on CSS are shown in Table I. We use the 3D and 2D versions of the dataset in our experiments. Clearly, our proposed method outperforms other methods in both the 3D-to-3D and 2D-to-3D tasks. Specifically, among all the methods, Text only [15] achieves the lowest retrieval results. Compared with this method, although the Concatenation [15] just combines the image and text by a simple two-layer MLP, it achieves a remarkable improvement on all metrics, which demonstrates that the modification text contains huge unequally semantic contents compared with the target image. Besides, as shown in the middle part in Table I, our proposed method significantly outperforms the methods that originally used for image caption (Show and Tell) and VQA (Param Hashing, Relationship, MRN, and FiLM), which achieves 16.21% to 48.81% and 15.04% to 52.74% performance gains in terms of R@1 in 3D-to-3D and 2D-to-3D tasks respectively. The results indicate the superiority to consider the interactions in the composed query during the cross-modal feature learning. In addition, compared with the state-of-the-art method LBF, our approach can also obtain clear improvements of 2.61% (R@1 in 3D-to-3D), 5.03% (R@5 in 3D-to-3D), and 3.05% (R@1 in 2D-to-3D). Unlike the LBF that directly embeds position information into visual features, our approach further adopts an explicit key-query attention between the

TABLE IV

ABLATION STUDIES ON ALL DATASETS IN TERMS OF R@1. IMP: IMPLICIT; EXP: EXPLICIT; SUB: SUBTRACTION; ADD: ADDITION; CON: CONCATENATION

Method	imp	exp	sub	add	con	mlp	gcn	gat	CSS	Fashion200k	MIT-states
Baseline	-	-	-	-	-	✓	-	-	75.34	15.64	13.81
Baseline+Imp.	✓	-	-	-	-	✓	-	-	75.92	17.07	15.26
Baseline+Exp.	-	✓	-	-	-	✓	-	-	76.45	17.11	14.64
Baseline+GS-IMA	✓	✓	-	-	-	✓	-	-	77.14	17.82	15.35
GS-IMA+TG-VR$_{mlp+sub}$	✓	✓	✓	-	-	✓	-	-	78.24	18.91	15.89
GS-IMA+TG-VR$_{gcn+sub}$	✓	✓	✓	-	-	-	✓	-	79.15	20.05	16.53
GS-IMA+TG-VR$_{gat+sub}$	✓	✓	✓	-	-	-	-	✓	**81.81**	**21.57**	**17.28**
GS-IMA+TG-VR$_{gat+add}$	✓	✓	-	✓	-	-	-	✓	79.51	19.98	16.82
GS-IMA+TG-VR$_{gat+con}$	✓	✓	-	-	✓	-	-	✓	80.06	20.21	16.65

words and the coordinate features. The improvements suggest that explicitly modeling the spatial structure can improve the retrieval performance.

3) Results on Fashion200K: Table II illustrates the results of each method on Fashion200K. As can be seen, the proposed method outperforms other methods on all metrics. Among all the compared methods, LBF and JAMMA are more relevant to our approach. In particular, the LBF adopts inter-modal attention between the reference image and the modification text to learn representation for the composed query. The JAMMA utilizes a multi-level graph attention network to progressively modify the visual feature of the reference image conditioned on the holistic textual feature. Instead of adopting the whole textual representation for visual feature reasoning, our method models the unequally semantic contents to rectify the visual feature in a fine-grained manner. With the help of this change, our method surpasses the JAMMA and LBF by 4.23% and 3.79% on R@1 accuracy, indicating the effectiveness of our TG-VR module. Note that although JVSM* leverages privileged information (text description corresponds to the image), our method can also surpass it by 2.57% on R@1 accuracy. Compared with its variant JVSM, which does not use the privileged data, our method outperforms it with clear margins on all the metrics. The gains over these methods may benefit from the joint modeling of the geometric information and the relationship between the reference image and modification text.

4) Results on MIT States: The quantitative results on a more complicated dataset MIT States are shown in Table III. From the table, we can observe that the proposed model also performs best compared with other methods. In particular, it outperforms the state-of-the-art method LBF with a 2.56% improvement in terms of R@1. These results well demonstrate that the proposed approach exhibits effectiveness for the *CQBIR*.

D. Model Analysis

1) Network Components: We perform ablation studies to testify the effectiveness of each proposed module, including the implicit and explicit geometric information construction in the GS-IMA module, different feature reasoning networks, and different unequally semantics construction methods in the TG-VR module. First, we provide the *Baseline* model, which learns aligned visual feature and textual feature with

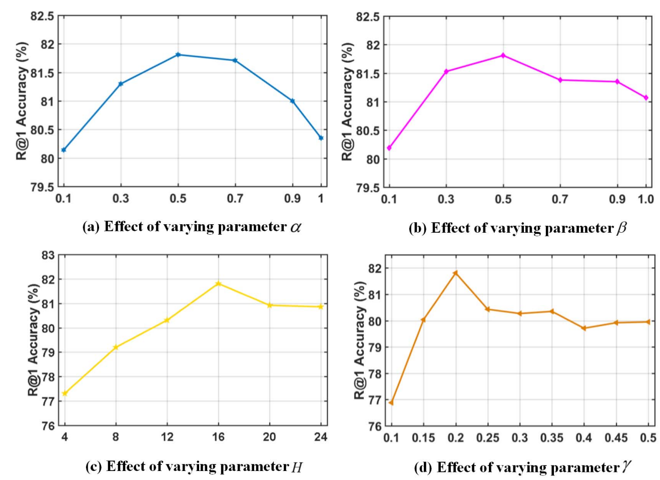

Fig. 4. The performance variation with respect to different parameters α and β on CSS dataset.

the vanilla inter-modal attention [70] and fuses them by a two-layer MLP. After that, the representation for the query is calculated according to Section III-C. In all variants, the feature for the target image is generated in the same way as the reference image by omitting the inter-modal attention and the textual inputs. Table IV shows the ablation study results. From the table, we can observe that the *Baseline* achieves a fairly competitive result compared with the method which can construct interactions in the composed query, such as the JAMMA shown in Table I. It shows the reasonability to focus on the vision-language interactions rather than treating them independently.

2) Impact of Geometric Information: We separately incorporate the coordinate feature into the *Baseline* with implicit, explicit, and both implicit and explicit (i.e., the GS-IMA) manner, and denote the variants as *Baseline+imp*, *Baseline+exp*, and *Baseline+GS-IMA* in Table IV. Specifically, taking the visual, textual, and geometry feature as inputs, in the implicit experiment (i.e., *Baseline+imp*), we first integrate the visual and geometry feature into a unified representation (U^{vc}) by Eq. (2), and then perform inter-modal attention between U^{vc} and the textual feature just by the first item in Eq. (3) and Eq. (5) to implicitly incorporate the geometric information into the textual feature. The other parts are the same as the *Baseline* method. The explicit experiment (i.e., *Baseline+exp*) directly

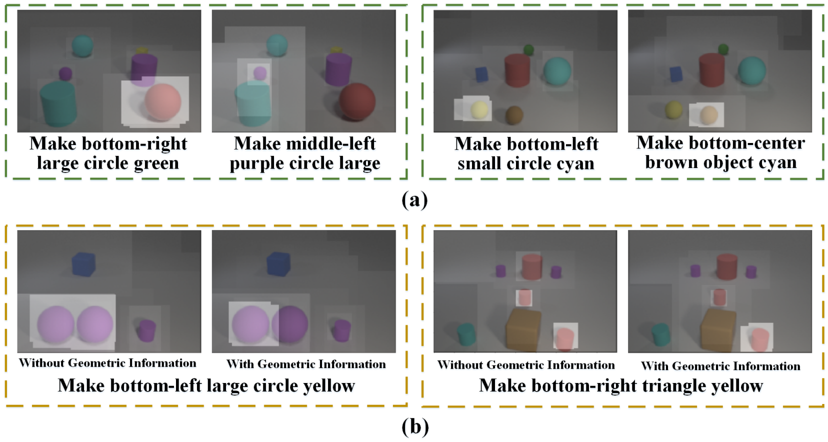

Fig. 5. Visualization of the attention weight for each image region with respect to the corresponding modification text. Sub-figure (a) shows the qualitative examples for the same image with different modification texts. Sub-figure (b) compares the attention results from models with or without geometric information.

adopts key-query attention between the textual and geometry feature by the second item in Eq. (3) and Eq. (5). Moreover, similar to the *Baseline* method, it also adopts an inter-modal attention between the visual feature and textual feature. The *Baseline+GS-IMA* combines the implicit and explicit manner in *Baseline+imp* and *Baseline+exp*. Compared with the *Baseline*, all the three variants can achieve better performance, demonstrating the effectiveness of considering the geometric information of the image regions during the inter-modal attention. Besides, compared with *Baseline+imp* and *Baseline+exp*, the *Baseline+GS-IMA* could further improve the retrieval performance on all the datasets, suggesting that the two manners (implicit and explicit) can complement each other in our approach.

3) Impact of Different Network Structures in TG-VR: We incorporate our TG-VR module into the *Baseline+GS-IMA* with three different feature learning networks, including MLP, GCN [56], and multi-head GAT [59]. The variants are denoted as *GS-IMA+TG-VR*$_{mlp+sub}$, *GS-IMA+TG-VR*$_{gcn+sub}$, and *GS-IMA+TG-VR*$_{gat+sub}$ (i.e., our method) in Table IV. From the table, we can observe that variants with the TG-VR consistently outperform the methods without it, indicating that modeling the unequal semantics plays a critical role in our task. Moreover, the *GS-IMA+TG-VR*$_{gat+sub}$ outperforms another two variants implemented by MLP and GCN, suggesting that the multi-head graph attention is a better solution to conduct the unequally semantics guided visual feature reasoning.

4) Impact of Different Unequal Semantics Construction Methods: The aggregation function in Eq. (8) is essential for our TG-VR module. We replace the subtraction operation in *GS-IMA+TG-VR*$_{gat+sub}$ with addition and concatenation, and denote the variants as *GS-IMA+TG-VR*$_{gat+add}$ and *GS-IMA+TG-VR*$_{gat+con}$, respectively. From Table IV, one can observe that the subtraction operation conduces to obtain better retrieval results.

5) Parameter Choice: In this section, we conduct quantitative experiments on evaluating the *CQBIR* performance on CSS dataset when four main parameters involved in our method change, i.e., the tradeoff parameter α in Eq. (3), β in Eq. (5), the number of heads in multi-head GAT in Eq. (10) (H), and the margin γ in Eq. (11). Specifically, α and β are set as $\{0.1, 0.3, 0.5, 0.7, 0.9, 1.0\}$, γ is ranging from 0.1 to 0.5 with a 0.05 step size, and H is set as $\{4, 8, 12, 16, 20, 24\}$. Each time when we evaluate a parameter, we fix another one. The results are depicted in Figure 4, from which we can observe that $\alpha = 0.5$ and $\beta = 0.5$ perform the best in our experiment. Moreover, from Figure 4 (c), we can observe that with the increase of the number of heads ($4 \rightarrow 16$), the performance raises accordingly. When even more heads are added ($20, 24$), the final performances show slight drops. The results in Figure 4 (d) suggest that $\gamma = 0.2$ performs the best in our experiment. Therefore, we set $\alpha = \beta = 0.5$, $H = 16$ and $\gamma = 0.2$ in our experiments.

E. Qualitative Analysis

1) Attention Results: To better understand the effectiveness of our proposed method, we visualize the attention results from our GS-IMA module in Figure 5. To be specific, from the qualitative examples shown in Figure 5 (a), we can observe that given the same reference image, our approach can accurately assign different attention weights with respect to the corresponding modification text. For example, as shown in the left rectangle in Figure 5 (a), the region 'bottom-right large circle' receives more attention in the left reference image while the region 'middle-left purple circle' is the focus in the right image. It indicates that our model could infer inter-modal alignments based on our GS-IMA module. Besides, we compare the attention results from models with or without geometric information in Figure 5 (b). From the figure, we can observe that when the model considers geometric information (the second column in each rectangle in Figure 5(b)), the corresponding region can be assigned more attention, suggesting that the geometric structure is beneficial to learn fine-grained cross-modal features.

More qualitative attention results from our geometry sensitive inter-modal attention module are shown in Figure 6. Visually, our proposed model could successfully attend the referent with respect to the modification text. The results further suggest that our model could learn effective aligned visual and textual features for the composed query, which

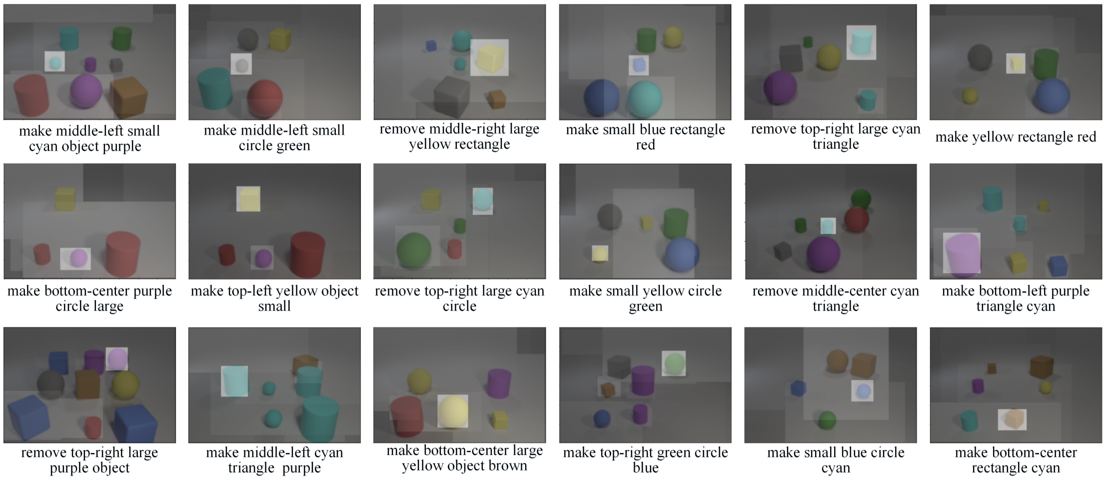

Fig. 6. Visualization of the attention weights for each image region with respect to the corresponding modification text.

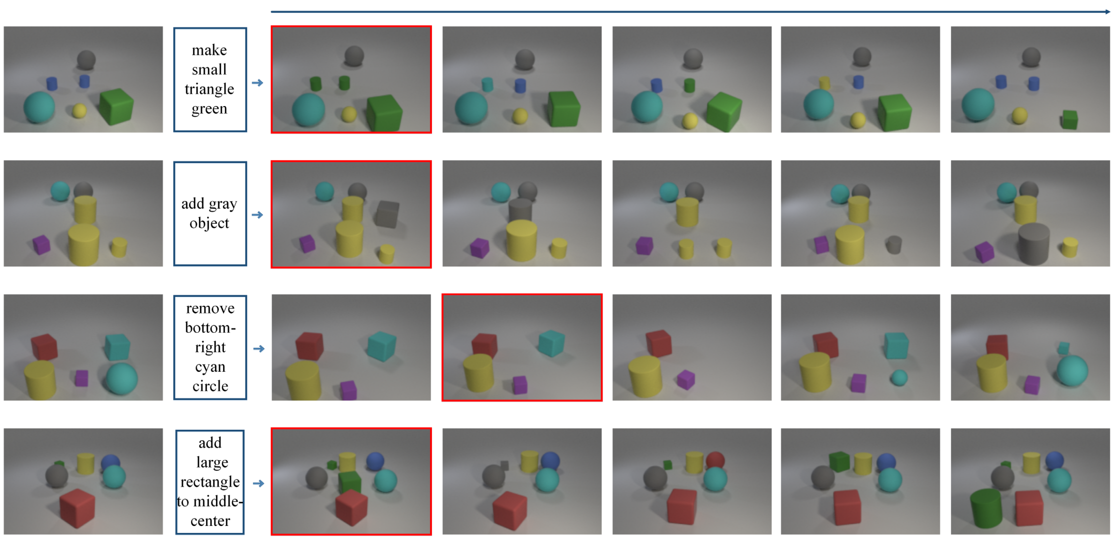

Fig. 7. Visualization of image retrieval on CSS dataset. For each composed query (a reference image and a modification text shown in the left), we show top-5 ranked images from left to right. The images with red boundaries indicate correct retrievals.

are then used in our text-guided visual reasoning module for further visual manipulation.

2) Image Retrieval Results: In this section, we show some qualitative retrieval results of our model, and the results are shown in Figure 7. From the figure we can observe that our approach always retrieves ground truth with a high rank. In addition, our approach is able to learn fine-grained correspondence between the composed query and the target image. For example, the images retrieved in case of 'make small triangle green' in Figure 7 (the first row), our proposed model could locate all the 'small triangles' and then make them green. The result suggest that our model could learn discriminative features for the composed query and the target image, and it can capture effective semantic information across the heterogeneous modalities.

V. Conclusion

In this paper, we propose a geometry sensitive cross-modal reasoning network for *CQBIR*. To capture the spatial structure and exploit the unequally visual-semantic relations, we propose a geometry sensitive inter-modal attention module (GS-IMA) and a text-guided visual reasoning module (TG-VR). The GS-IMA utilizes the geometric information of image regions to stress the alignment between the visual and textual feature and learn cross-modal feature for the composed query in a geometry aware way. Based on the aligned textual feature, the TG-VR constructs unequally semantic contents to further conduct visual reasoning for *CQBIR*. Extensive studies on three publicly available datasets demonstrate the effectiveness of our method. In the future, we plan to explore more compact unequal semantics construction methods and evaluate the effectiveness of it on more vision-language tasks.

References

[1] H. Feng, M. Chen, J. Hu, D. Shen, H. Liu, and D. Cai, "Complementary pseudo labels for unsupervised domain adaptation on person re-identification," *IEEE Trans. Image Process.*, vol. 30, pp. 2898–2907, 2021.

[2] J. Sun, Y. Li, H. Chen, Y. Peng, and J. Zhu, "Unsupervised cross domain person re-identification by multi-loss optimization learning," *IEEE Trans. Image Process.*, vol. 30, pp. 2935–2946, 2021.

[3] W. Dong, Z. Zhang, C. Song, and T. Tan, "Bi-directional interaction network for person search," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2839–2848.

[4] Y. Lang, Y. He, F. Yang, J. Dong, and H. Xue, "Which is plagiarism: Fashion image retrieval based on regional representation for design protection," in *Proc. CVPR*, Jun. 2020, pp. 2595–2604.

[5] Y.-L. Lin, S. Tran, and L. S. Davis, "Fashion outfit complementary item retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3311–3319.

[6] Y. Zhong, W. Deng, J. Hu, D. Zhao, X. Li, and D. Wen, "SFace: Sigmoid-constrained hypersphere loss for robust face recognition," *IEEE Trans. Image Process.*, vol. 30, pp. 2587–2598, 2021.

[7] A. Sepas-Moghaddam, A. Etemad, F. Pereira, and P. L. Correia, "CapsField: Light field-based face and expression recognition in the wild using capsule routing," *IEEE Trans. Image Process.*, vol. 30, pp. 2627–2642, 2021.

[8] J. Chang, Z. Lan, C. Cheng, and Y. Wei, "Data uncertainty learning in face recognition," in *Proc. CVPR*, Jun. 2020, pp. 5710–5719.

[9] Z. Lu, Y. Hu, Y. Jiang, Y. Chen, and B. Zeng, "Learning binary code for personalized fashion recommendation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10562–10570.

[10] W.-C. Kang, E. Kim, J. Leskovec, C. Rosenberg, and J. McAuley, "Complete the look: Scene-based complementary product recommendation," in *Proc. CVPR*, Jun. 2019, pp. 10532–10541.

[11] H. Chen, G. Ding, X. Liu, Z. Lin, J. Liu, and J. Han, "IMRAM: Iterative matching with recurrent attention memory for cross-modal image-text retrieval," in *Proc. CVPR*, Jun. 2020, pp. 12655–12663.

[12] Q. Zhang, Z. Lei, Z. Zhang, and S. Z. Li, "Context-aware attention network for image-text retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3536–3545.

[13] X. Wei, T. Zhang, Y. Li, Y. Zhang, and F. Wu, "Multi-modality cross attention network for image and sentence matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10941–10950.

[14] M. Meng, H. Wang, J. Yu, H. Chen, and J. Wu, "Asymmetric supervised consistent and specific hashing for cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 30, pp. 986–1000, 2021.

[15] N. Vo *et al.*, "Composing text and image for image retrieval—An empirical Odyssey," in *Proc. CVPR*, Jun. 2019, pp. 6439–6448.

[16] X. Guo, H. Wu, Y. Cheng, S. Rennie, G. Tesauro, and R. Feris, "Dialog-based interactive image retrieval," in *Proc. NIPS*, 2018, pp. 678–688.

[17] I. Misra, A. Gupta, and M. Hebert, "From red wine to red tomato: Composition with context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1792–1801.

[18] B. Zhao, J. Feng, X. Wu, and S. Yan, "Memory-augmented attribute manipulation networks for interactive fashion search," in *CVPR*, Jul. 2017, pp. 1520–1528.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, Jun. 2016, pp. 770–778.

[20] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. CVPR*, Jun. 2015, pp. 1–9.

[21] S. Huang *et al.*, "Referring image segmentation via cross-modal progressive comprehension," in *Proc. CVPR*, Jun. 2020, pp. 10488–10497.

[22] P. Jiang and Y. Han, "Reasoning with heterogeneous graph alignment for video question answering," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 11109–11116.

[23] X. Han *et al.*, "Automatic spatially-aware fashion concept discovery," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1463–1471.

[24] P. Isola, J. J. Lim, and E. H. Adelson, "Discovering states and transformations in image collections," in *CVPR*, Jun. 2015, pp. 1383–1391.

[25] C. Liu, Z. Mao, T. Zhang, H. Xie, B. Wang, and Y. Zhang, "Graph structured network for image-text matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10921–10930.

[26] L. Wang, Y. Li, J. Huang, and S. Lazebnik, "Learning two-branch neural networks for image-text matching tasks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 394–407, Feb. 2019.

[27] Y. Kim, W. Park, M. Roh, and J. Shin, "Groupface: Learning latent groups and constructing group-based representations for face recognition," in *Proc. CVPR*, Jun. 2020, pp. 5620–5629.

[28] W. Dong, Z. Zhang, C. Song, and T. Tan, "Instance guided proposal network for person search," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2582–2591.

[29] A. K. Bhunia, Y. Yang, T. M. Hospedales, T. Xiang, and Y.-Z. Song, "Sketch less for more: On-the-fly fine-grained sketch-based image retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9779–9788.

[30] K. Pang, Y. Yang, T. M. Hospedales, T. Xiang, and Y.-Z. Song, "Solving mixed-modal jigsaw puzzle for fine-grained sketch-based image retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10344–10352.

[31] S. Choi, S. Lee, Y. Kim, T. Kim, and C. Kim, "Hi-CMD: Hierarchical cross-modality disentanglement for visible-infrared person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10257–10266.

[32] M. Hosseinzadeh and Y. Wang, "Composed query image retrieval using locally bounded features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3596–3605.

[33] Y. Chen, S. Gong, and L. Bazzani, "Image search with text feedback by visiolinguistic attention learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2998–3008.

[34] Z. Ma *et al.*, "Fine-grained fashion similarity learning by attribute-specific embedding network," *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, pp. 11741–11748, Apr. 2020.

[35] K. E. Ak, A. A. Kassim, J. Lim, and J. Y. Tham, "Learning attribute representations with localization for flexible fashion search," in *Proc. CVPR*, Jun. 2018, pp. 7708–7717.

[36] J. Ma, S. Pang, B. Yang, J. Zhu, and Y. Li, "Spatial-content image search in complex scenes," in *IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, 2020, pp. 2492–2500.

[37] L. Mai, H. Jin, Z. L. Lin, C. Fang, J. Brandt, and F. Liu, "Spatial-semantic image search by visual feature synthesis," in *Proc. CVPR*, Jul. 2017, pp. 1121–1130.

[38] Y. Chen and L. Bazzani, "Learning joint visual semantic matching embeddings for language-guided retrieval," in *Proc. ECCV*, 2020, pp. 1–17.

[39] D. Huynh and E. Elhamifar, "A shared multi-attention framework for multi-label zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8773–8783.

[40] L. Zhu and Y. Yang, "Inflated episodic memory with region self-attention for long-tailed visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4343–4352.

[41] D. Zoran, M. Chrzanowski, P.-S. Huang, S. Gowal, A. Mott, and P. Kohli, "Towards robust image classification using sequential attention models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9480–9489.

[42] Y. Pan, T. Yao, Y. Li, and T. Mei, "X-linear attention networks for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10968–10977.

[43] L. Guo, J. Liu, X. Zhu, P. Yao, S. chen Lu, and H. Lu, "Normalized and geometry-aware self-attention network for image captioning," in *CVPR*, Jun. 2020, pp. 10324–10333.

[44] J. Kim, M. Ma, T. Pham, K. Kim, and C. Yoo, "Modality shifting attention network for multi-modal video question answering," in *CVPR*, Jun. 2020, pp. 10103–10112.

[45] C. Fan, X. Zhang, S. Zhang, W. Wang, C. Zhang, and H. Huang, "Heterogeneous memory enhanced multimodal attention model for video question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1999–2007.

[46] P. Chandran, D. Bradley, M. Gross, and T. Beeler, "Attention-driven cropping for very high resolution facial landmark detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5860–5869.

[47] Q. Fan, W. Zhuo, C.-K. Tang, and Y.-W. Tai, "Few-shot object detection with attention-RPN and multi-relation detector," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4012–4021.

[48] B. Li, X. Qi, T. Lukasiewicz, and P. H. S. Torr, "Manigan: Text-guided image manipulation," in *Proc. CVPR*, Jun. 2020, pp. 7877–7886.

[49] Y. Cheng, Z. Gan, Y. Li, J. Liu, and J. Gao, "Sequential attention GAN for interactive image editing," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 4383–4391.

[50] A. El-Nouby *et al.*, "Tell, draw, and repeat: Generating and modifying images based on continual linguistic instruction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10303–10311.

[51] D. Guo, H. Wang, H. Zhang, Z.-J. Zha, and M. Wang, "Iterative context-aware graph inference for visual dialog," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10052–10061.

[52] L. Li, Z. Gan, Y. Cheng, and J. Liu, "Relation-aware graph attention network for visual question answering," in *Proc. ICCV*, Oct. 2019, pp. 10312–10321.

[53] P. Wang, Q. Wu, J. Cao, C. Shen, L. Gao, and A. V. D. Hengel, "Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks," in *Proc. CVPR*, Jun. 2019, pp. 1960–1968.

[54] X. Rong, C. Yi, and Y. Tian, "Unambiguous scene text segmentation with referring expression comprehension," *IEEE Trans. Image Process.*, vol. 29, pp. 591–601, 2020.

[55] J. Liu, W. Wang, L. Wang, and M.-H. Yang, "Attribute-guided attention for referring expression generation and comprehension," *IEEE Trans. Image Process.*, vol. 29, pp. 5244–5258, 2020.

[56] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. ICLR*, Sep. 2017, pp. 1–14.

[57] S. Huang *et al.*, "Referring image segmentation via cross-modal progressive comprehension," in *Proc. CVPR*, Jun. 2020, pp. 10485–10494.

[58] L. Mi and Z. Chen, "Hierarchical graph attention network for visual relationship detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13883–13892.

[59] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *Proc. ICLR*, 2018, pp. 1–12.

[60] P. Anderson *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6077–6086.

[61] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[62] F. Faghri, D. J. Fleet, J. Kiros, and S. Fidler, "VSE++: Improving visual-semantic embeddings with hard negatives," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2018, pp. 1–14.

[63] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick, "CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning," in *Proc. CVPR*, Jul. 2017, pp. 2901–2910.

[64] A. Santoro *et al.*, "A simple neural network module for relational reasoning," in *Proc. NIPS*, 2017, pp. 4967–4976.

[65] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3156–3164.

[66] H. Noh, P. H. Seo, and B. Han, "Image question answering using convolutional neural network with dynamic parameter prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 30–38.

[67] J.-H. Kim *et al.*, "Multimodal residual learning for visual QA," in *Proc. NIPS*, 2016, pp. 361–369.

[68] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 3942–3951.

[69] F. Zhang, M. Xu, Q. Mao, and C. Xu, "Joint attribute manipulation and modality alignment learning for composing text and image to image retrieval," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 3367–3376.

[70] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, "Deep modular co-attention networks for visual question answering," in *Proc. CVPR*, Jun. 2019, pp. 6274–6283.

Feifei Zhang received the Ph.D. degree from Jiangsu Univerisity, Zhenjiang, Jiangsu, China, in 2019. Then, she worked as an Assistant Professor with the National Laboratory of Pattern Recognition, Institute of Automation Chinese Academy of Sciences. She is currently a Professor with the School of Computer Science and Engineering, Tianjin University of Technology, Tianjin, China. Her current research interests include multimedia analysis, computer vision, deep learning, especially multimedia computing, facial expression recognition, and cross-modal image retrieval.

Mingliang Xu (Member, IEEE) received the B.S. and M.S. degrees from the Department of Computer Science, Zhengzhou University, Zhengzhou, China, and the Ph.D. degree in computer science and technology from the State Key Lab of CAD&CG, Zhejiang University, Hangzhou, China. He is currently a Full Professor with the School of Information Engineering, Zhengzhou University, China, and the Director of the Center for Interdisciplinary Information Science Research (CIISR). He is the Vice General Secretary of ACM SIGAI China. His current research interests include computer graphics, multimedia, and artificial intelligence. He has authored more than 60 journal and conference papers in these areas, including the *ACM Transactions on Graphics*, the *ACM Transactions on Intelligent Systems and Technology*, the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CYBERNETICS, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, ACM SIGGRAPH (Asia), ACM MM, and ICCV.

Changsheng Xu (Fellow, IEEE) is currently a Distinguished Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His research interests include multimedia content analysis/indexing/retrieval, pattern recognition, and computer vision. He holds 40 granted/pending patents and published over 300 refereed research papers in these areas. He is a Fellow of IAPR and a Distinguished Scientist of ACM. He has served as an associate editor, a guest editor, the general chair, the program chair, the area/track chair, a special session organizer, the session chair, and a TPC member of over 20 IEEE and ACM prestigious multimedia journals, conferences, and workshops, including IEEE TRANSACTIONS ON MULTIMEDIA, *ACM Transactions on Multimedia Computing, Communications and Applications*, and ACM Multimedia conference.