Scene Graph-Aware Hierarchical Fusion Network for Remote Sensing Image Retrieval With Text Feedback

Fei Wang, Xianzhang Zhu, Xiaojian Liu, Yongjun Zhang, *Member, IEEE*, and Yansheng Li, *Senior Member, IEEE*

***Abstract*— In the realm of image retrieval with text feedback, existing studies have predominantly concentrated on the intrinsic attribute of target objects, neglecting extrinsic information essential for remote sensing (RS) images, such as spatial relationships. This research addresses this gap by incorporating RS image scene graphs as side information, given their capacity to encapsulate internal object attributes, external structural features between objects, and the relationships among images. To fully leverage the features from the reference RS image, scene graph, and modifier sentence, we propose a scene graph-aware hierarchical fusion network (SHF), which optimally integrates the multimodal features in a two-stage fusion process. Initially, image and scene graph features are fused hierarchically, followed by transforming content information with a proposed multimodal global content (MGC) block, ultimately transforming style information. To validate the superiority of SHF, we constructed three datasets with images from several popular RS datasets, named Airplane (3461 image + text–image pairs), Tennis (1924 image + text–image pairs), and WHIRT (3344 image + text–image pairs). Extensive experiments conducted on these datasets show that SHF significantly outperforms state-of-the-art methods.**

***Index Terms*— Image retrieval with text feedback, multimodal features, remote sensing (RS) image retrieval, scene graph.**

Manuscript received 28 February 2024; revised 26 April 2024; accepted 16 May 2024. Date of publication 23 May 2024; date of current version 10 June 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 42030102, Grant 42371321, and Grant 42192581; in part by the Fund for Innovative Research Groups of the Hubei Natural Science Foundation under Grant 2020CFA003; in part by the Major Special Project of Guizhou under Grant [2022]001; and in part by the Special Fund of Hubei Luojia Laboratory under Grant 220100032. *(Corresponding authors: Yongjun Zhang; Yansheng Li.)*

Fei Wang is with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, Hubei 430072, China, also with Changjiang Spatial Information Technology Engineering Company Ltd., Wuhan 430010, China, and also with the Water Resources Information Perception and Big Data Engineering Research Center of Hubei Province, Wuhan 430010, China (e-mail: flyking@whu.edu.cn).

Xianzhang Zhu is with the School of Municipal and Geomatics Engineering, Hunan City University, Yiyang, Hunan 413002, China (e-mail: zhuxianzhang1314@hotmail.com).

Xiaojian Liu is with the Tianjin Research Institute for Water Transport Engineering, M.O.T., Tianjin 300456, China (e-mail: xiaojianliu2018@whu.edu.cn).

Yongjun Zhang is with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, Hubei 430072, China (e-mail: zhangyj@whu.edu.cn).

Yansheng Li is with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, Hubei 430072, China, and also with the Hubei Luojia Laboratory, Wuhan, Hubei 430079, China (e-mail: yansheng.li@whu.edu.cn).

Digital Object Identifier 10.1109/TGRS.2024.3404605

I. Introduction

IN THE era of remote sensing (RS) big data, RS image retrieval has become an essential technology for acquiring RS images and plays a cornerstone role in a wide range of human-related applications such as disaster rescue and ecological prediction [1], [2], [3]. The predominant retrieval mode takes a reference image as an input query to seek images that bear similarity to the reference, known as image–image matching [4], [5]. However, the results often prioritize categories, overlooking the granular details of concrete content. Another widely-used retrieval paradigm is image–text matching [6], [7], wherein a natural language description of the target image's content serves as the input. However, it is laborious to accurately reflect the user's target concept with just a sentence. As a result, these two retrieval paradigms can only provide rough results that are insufficient to satisfy the user. In addition, both paradigms are infeasible for further refining results that fail to match the user's intent precisely.

Image retrieval with text feedback refines reference images through specific modifications that articulate changes from the reference to the target, as illustrated in Fig. 1(a). For instance, the user might envision changing the airplane color from white to purple while preserving other attributes like the parking place. By leveraging visual-linguistic information, users are allowed to express their intent more flexibly and precisely. The core principle of this task is to combine the representation of the reference image and the modifier sentence into an integrated representation that is as similar as possible to the representation of the target image. To achieve the goal, the main efforts of researchers have focused on designing a compositor capable of selectively preserving and transforming visual features according to modification from coarse-grained to fine-grained and from local to global [8], [9], [10]. Vo et al. [8] adopted a gated connection and a residual connection to determine what to change and how to change, respectively. Although the core idea is intuitive, it is suitable for concrete modifications (e.g., changing the airplane color from white to purple) but fails to deal with abstract ones (e.g., changing the airplane color darker). Chen et al. [9] harnessed multimodal nonlocal blocks (MNLs) to integrate the text feature with the image feature at varying depths. Lee et al. [10] disentangled the reference image into content and style, extending MNL to disentangled MNL (DMNL)

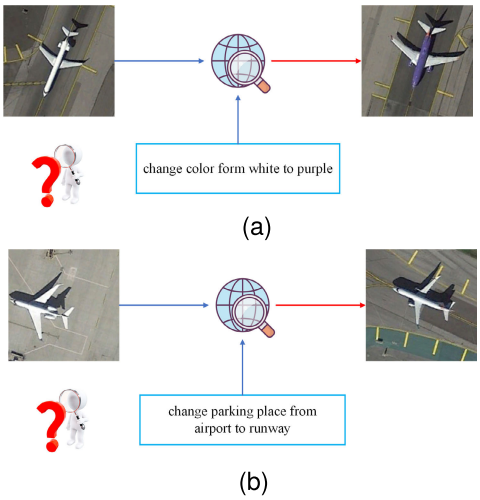

Fig. 1. Examples of RS image retrieval with text feedback. Given a reference RS image and a sentence as input, we consider searching for a new image from the database that is similar to the reference while resembling the user's feedback described by the text. The text generally states the visual content in the reference image that demands refinements, such as (a) intrinsic modification or (b) extrinsic modification.

for capturing the content modification. Although accommodating sentences of diverse granularity, these methods entail additional computational overhead due to intricate matrix computations. Primarily, existing methods focus on intrinsic modifications, altering properties of target objects themselves, which are suited for simple scenes with one or a few distinct targets. However, in complex RS scenarios, images often feature numerous nontarget objects. Moreover, beyond intrinsic properties, the environment of target objects should also be regarded as extrinsic properties as they play a crucial role in facilitating RS scene understanding [11]. The changes of extrinsic properties, as shown in Fig. 1(b), are considered as extrinsic modifications, and they generally entail changes in nontarget object properties and spatial relationships induced by modifier sentences in retrieval. While these algorithms showcase potential in computer vision, their performance is notably constrained in RS.

To address the aforementioned limitation in complex RS image retrieval with text feedback, we introduce scene graphs as side information for two key reasons: 1) a scene graph constitutes a structured representation of an image, typically comprising objects and the spatial relationship between them [12], [13], [14], therefore enhances our capacity to comprehend the image scene; and 2) shared edges among scene graphs signify consistency, aligning with the modifier's intent, while divergent edges encapsulate differences between images. In essence, our overarching objective pivots toward learning a unified visual-semantic-structural representation of the reference image, the modifier, and the reference scene graph, and converges it closer to an integrated representation of the target image and its scene graph in the embedding space. To summarize, our research unfolds with two pivotal objectives: 1) designing a comprehensive framework seamlessly integrating RS image features, scene graphs, and modifier sentences; and 2) validating the effectiveness of scene graphs and the envisioned framework.

To achieve this goal, we propose a new scene graph-aware hierarchical fusion network (short for SHF) that applies a two-stage multimodal information fusion strategy. We first perform a hierarchical fusion of multilevel graph–image features to generate the scene feature and then use a scene–text compositor to fuse the scene feature with the modifier feature. Specifically, in the first stage, considering that the lower level image and graph features capture local information while the higher level features capture global information, we fuse visual and structural representations of varying depths in order to understand the image scenes better as well as to explicitly inject signals reflecting the similarities and differences of the images contained in the structural representations into the scene representation. Drawing inspiration from [10], the second stage employs a content modulator and a style modulator to convey the content and style changes, respectively. The content modulator first extracts content information of scenes and sentences, leveraging a multimodal global content (MGC) block to transform the content information of scenes without the complex matrix computation of DMNL. The style modulator reintroduces style information into the transformed scene representation based on the style information of the sentence.

To substantiate the effectiveness of our proposed SHF, we construct three datasets, Airplane, Tennis, and WHIRT, comprising images from renowned RS datasets. The empirical findings on these datasets reveal the remarkable performance of SHF, showcasing a substantial improvement over the advanced model. Specifically, SHF outperforms Cosmo by 41.24%, 25.65%, and 2.76% points on Recall@1 for the Airplane, Tennis, and WHIRT datasets, respectively. Overall, our contributions are three-folds as follows.

1) For a better understanding of complex RS scenarios, we introduce scene graphs as side information. Then, we propose a new SHF to seamlessly integrate the features of scene graphs, images, and texts with a hierarchical multimodal information fusion strategy.
2) We propose a compositor MGC block, which reduces the computational complexity associated with DMNL, ensuring a streamlined yet effective modification of content information.
3) To rigorously validate our SHF, we meticulously crafted three datasets using publicly available RS image datasets: Airplane, Tennis, and WHIRT, which will be made publicly available along with this article. To the best of our knowledge, this represents the first dataset for RS image retrieval with text feedback.

The remainder of this article is organized as follows. Section II introduces literatures that most related to our work. Section III describes our methods in detail. Section IV presents our experimental result and analysis. Our conclusion and future work are discussed in Section V.

II. Related Work

In this section, we briefly summarize the previous literature that is most relevant to our work, including RS image retrieval and image retrieval with text feedback.

A. Image Retrieval

The primary objective of image retrieval lies in the quest for images that align with the conceptualization in the user's mind within an extensive collection of candidates, which can be divided into unimodal image retrieval and cross-modal image retrieval.

In unimodal image retrieval, where the query and target share the same modality, researchers engage in similarity matching based on the visual content features of images. In the early stages, these features were predominantly handcrafted descriptors, encompassing aspects such as color [15], textural [16], or a combination thereof [17], [18]. While these methods rely on high-quality low-level features, the advent of convolutional neural networks (CNNs) offers new opportunities as they empower researchers to adaptively extract fundamental features within an image without the need for complex hand-crafted features, proving their effectiveness in image retrieval [19], [20]. To address challenges like high storage requirements and slow retrieval associated with the high-dimensional feature computation, the deep hash neural network (DHNN) is utilized as encoder [21], [22], [23], [24], showcasing its capacity to map high-dimensional features into low-dimensional binary representations.

Cross-modal image retrieval, on the other hand, involves queries and targets from disparate modalities, necessitating the bridging of semantic gaps between them. A foundational solution paradigm involves using distinct networks to independently extract features for queries and targets, optimizing the networks with tricks to maximize feature similarity. In scenarios like cross-source image retrieval, where query and target images originate from two distinct data sources, Li et al. [25] leveraged two different DHNNs for feature extraction and devised optimization constraints for stable training. For text–image retrieval, where queries are in the form of natural language descriptions, a recurrent neural network (RNN)-like module is employed to extract content information. Abdullah et al. [26] utilized a long–short-term memory (LSTM) network with an average fusion strategy as a text encoder, Qin et al. [27] and Zhao et al. [28] employed Bi-BRU and Bert, respectively. Furthermore, taking sketches [29] and sound [30] as queries has captured the attention of researchers. Although these methods retrieve similar RS images based on available data, they fall short when the data inadequately reflects the user's intent. Thus, the refinement of results based on user feedback becomes imperative.

B. Image Retrieval With Text Feedback

Many studies have taken various forms of user feedback, encompassing aspects like relevance [31], concrete attributes [32], or modifier sentences [8], [33], all aimed at enhancing retrieval outcomes that may have content gaps misaligned with the true user intent. In this work, our focus is on leveraging modifier sentences as feedback, recognizing natural language as the quintessential medium for human-system interaction. To address this task, a prevalent paradigm is to adopt a multimodal compositor adept at efficiently amalgamating image–text features [8], [9], [10], [34], [35], [36]. Vo et al. [8] devised a gated residual connection to modify the image feature in the image embedding space, which was intuitive but difficult to cope with abstract properties. To comprehend semantic information from concrete to abstract, Chen et al. [9] and Jandial et al. [37] employed multiple compositors to integrate linguistic features with visual features at multiple levels of the image CNN. However, their approach involved complexities with numerous compositors or off-the-shelf models. Lee et al. [10] decomposed images into content and style, which were consistent with concrete and abstract linguistic granularity, and modified them based on accompanying text. Despite achieving stable and effective training, their compositor's matrix computation posed challenges with large-scale features. Specifically, these methods primarily focus on intrinsic attribute changes, overlooking extrinsic spatial relationships, rendering them more suitable for simple natural images rather than complex RS images. With the advancement of large-scale models, the utilization of pretrained models to enhance image and text features has garnered increasing attention from scholars, alongside the development of image–text composition modules. For instance, Tian et al. [38] employed the Swin transformer [39] and DistilBERT [40] to extract image and text features, respectively, subsequently integrating them into a transformer-based additive attention composition module. Similarly, Han et al. [41] and Baldrati et al. [42] initialized their language encoders and visual encoders with the pretrained clip model [43]. Then, in their second stage, Han et al. adopted cross-attention adapters and task-specific adapters for visual-linguistic representation generation, and Baldrati et al. merged the multimodal features with a simple nonlinear compositor. Another work line involves the design of regularization terms. For example, Chen et al. [44] extended Cosmo [10] by introducing an uncertainty regularization. In this article, we endeavor to design a network that also focuses on the extrinsic spatial relationships for better RS scene modification.

III. Methodology

Fig. 2 illustrates the overall framework of our proposed SHF. Given a reference RS image, its scene graph, and a modifier sentence as inputs, the ultimate aim of SHF is to learn an integrated representation that well-aligns with the joint representation of the target RS image and its scene graph. Our SHF mainly consists of five components: Fig. 2(a) an image encoder, Fig. 2(b) a graph encoder, Fig. 2(c) a text encoder for vision, structural, and linguistic representation learning, respectively; Fig. 2(d) multiple image–graph compositors that generate the scene representation by injecting structural features at different hops to visual features at varying layers; and Fig. 2(e) a scene–text compositor modifying the scene representation based on the text. All the components are optimized by minimizing the objective function in an end-to-end manner. In Section III-A, we overview the three basic encoders. In Section III-B, we introduce our image–graph compositor. In Section III-C, we elaborate on our scene–text compositor. And the model optimizer is introduced in Section III-D.

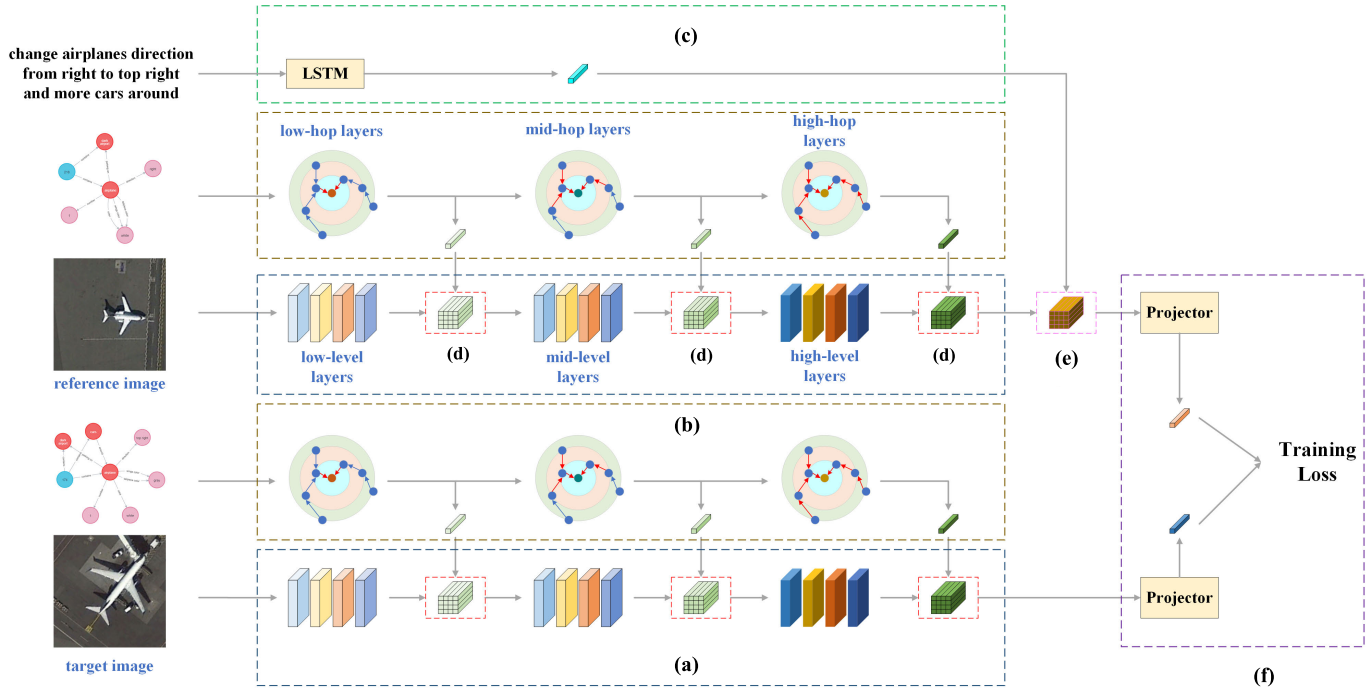

Fig. 2. Overall architecture of SHF framework. Given a triplet of the reference RS image, its corresponding scene graph, and the user feedback sentence as input, our goal is to learn a unified representation that is exclusively aligned with the joint representation of the target image and its scene graph. SHF is mainly composed of five components: (a) image encoder, (b) graph encoder, (c) text encoder (see Section III-A), (d) image–graph compositors that compose visual and structural features at varying layers (see Section III-B), and (e) scene–text compositor that integrates the scene and text feature (see Section III-C). All components are co-optimized by (f) optimizer learning (optimizer).

A. Basic Representation Learning

1) Image Representation: In order to encode visual information into discriminable representations, we employ a standard CNN as the image encoder for image representation learning. The convolutional layer extracts features by sliding the convolutional kernel over the image, which essentially extracts local features, so the lower convolutional layer captures information from different location regions, while the higher convolutional layer can extract global information by combining the information obtained from the lower convolutional layer [45], [46]. Considering the specificity of the features extracted from the low to high convolutional layers from local to global, we construct a feature pyramid [47] for the feature maps from different layers to generate a more powerful scene representation. Specifically, the feature pyramid Φ constructed by three different levels of feature maps from the image encoder f_{img}

$$\begin{aligned}\Phi_r &= \left\{X_r^L, X_r^M, X_r^H\right\} = f_{\text{img}}(I_r)\\ \Phi_t &= \left\{X_t^L, X_t^M, X_t^H\right\} = f_{\text{img}}(I_t).\end{aligned} \tag{1}$$

Here, Φ_r and Φ_t are the feature pyramids of the reference image I_r and the target image I_t, respectively, and each pyramid contains three feature maps X^L, X^M, and X^H, which are obtained from the low, middle, and high layers of f_{img}.

2) Graph Representation: Inspired by the significant progress of graph convolutional networks (GCNs) on graph task [48], [49], we employ a GCN-based network for scene graph representation learning. The standard GCN generates new node representations by stacking multiple graph convolutional layers to recursively aggregate the features of neighbor nodes. However, it does not distinguish between the types of edges (i.e., relationships) that are not negligible for the scene graph. Therefore, we use the relational GCN [50] as our image encoder, which assigns a specific weight matrix to each type of relationship during the feature aggregation process. Considering that the GCN inherently learns increasingly comprehensive structural information in a hierarchical manner, we also construct a feature pyramid Ψ for the different hops of graph feature maps of the graph encoder f_{graph}

$$\begin{aligned}\Psi_r &= \left\{Z_r^L, Z_r^M, Z_r^H\right\} = f_{\text{graph}}(G_r)\\ \Psi_t &= \left\{Z_t^L, Z_t^M, Z_t^H\right\} = f_{\text{graph}}(G_t).\end{aligned} \tag{2}$$

Here, G_r and G_t refer to the reference scene graph and the target scene graph, respectively; Φ_r and Φ_t are their feature pyramids, respectively, and each pyramid also consist of three feature maps Z^L, Z^M, and Z^H, which are generated by the low, middle, and high layers of f_{graph}.

3) Text Representation: To capture the linguistic information of modifications, we represent the user feedback with an RNN, which has been proven to be powerful in encoding natural language [51], [52]. Specifically, we implement the text encoder f_{text} as an LSTM [53], which outputs the text representation $T \in \mathbb{R}^D$, where D is the feature dimension.

B. Image–Graph Compositor

The common characteristic of visual features from CNN and structural features from GCN is that the lower level features are local and the higher level features are global, however, CNN is more concerned with pixel-level relationships

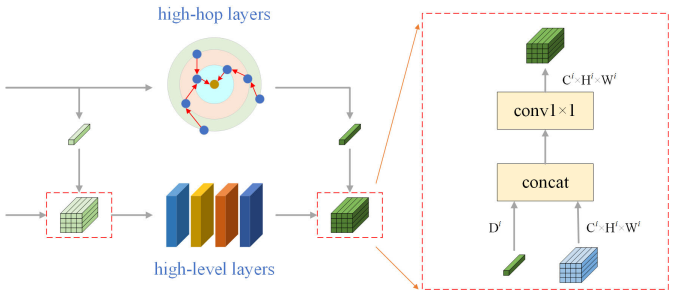

Fig. 3. Overall architecture of image–graph compositor. The image representation is concatenated with the graph representation broadcast along width and height, and then the matrix is multiplied with the feature transformation matrix (implemented as an 1×1 convolution) to generate the scene representation at the current level. The input image feature and output scene feature are kept as 3-D vectors.

while GCN is more concerned with object-level relationships, thereby visual and structural features are consistent and complementary. To generate scene representations with multigranularity features, we fuse visual representation Φ_r and structural representation Ψ_r, as shown in Fig. 3. For the visual feature X_r^i ($i \in \{L, M, H\}$ is the level in the feature pyramid), we concatenate it with the structural feature Z_r^i at the same level, followed by a transform function F_t to learn the scene representation

$$X_{\text{vs}}^i = F_t^i\left(X_r^i \| Z_r^i\right). \tag{3}$$

Here, $\cdot\|\cdot$ is the concatenation operation, which broadcasts the graph feature $Z_r^i \in \mathbb{R}^{D^i}$ spatially along the dimensions of height and width so that its shape is matched with the image feature $X_r^i \in \mathbb{R}^{C^i \times H^i \times W^i}$, in which D^i is the dimension, H^i is the feature height, W^i is the feature width, and C^i is the number of feature channel; F_t^i is a learnable weight matrix and is implemented as a 1×1 convolution.

Note that the lower level scene representations are fed into the image encoder to generate the higher level visual representation. After obtaining the final scene representation $X_s = X_{\text{vs}}^H \in \mathbb{R}^{C \times H \times W}$, we feed it into the scene–text compositor for the preservation and transformation of content and style.

C. Scene–Text Compositor

Fig. 4 illustrates the overall pipeline of our compositor to generate the scene–text joint representation X_{ss}, which consists of two main components: 1) a content modulator that modifies the content of scenes based on the content of the text and 2) a style modulator which modifies the style of scenes corresponding to the style of the text and reintroduces the style information into scene representations.

Before feeding the scene representation into the content modulator, we remove its potential style information $(\mu_{X_s}, \sigma_{X_s})$ by applying an instance normalization that has been shown to normalize the style of each image to the target style, allowing the network to focus solely on the content while ignoring its initial style information [54], which is formulated as

$$X_{\text{si}} = \text{IN}(X_s) = \frac{X_s - \mu_{X_s}}{\sigma_{X_s}} \tag{4}$$

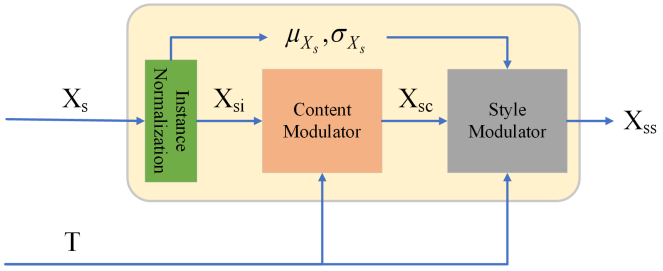

Fig. 4. Overall pipeline of our scene–text compositor.

where μ_{X_s} and σ_{X_s} are the channel-wise mean and variance of X_s.

1) Content Modulator:

a) Disentangled multimodal nonlocal (DMNL) block: The basic DMNL [10], as shown in Fig. 5, aims at capturing the long-range dependencies between two locations of multimodal input features. DMNL takes image and text representations as input, and here we replace the image representation with a scene representation, which is formulated as

$$y_s^i = \sum_{j=1}^{N_p} f\left(X_{\text{si}}^i, X_{\text{si}}^j, T\right) g\left(X_{\text{si}}^j \| T\right) \tag{5}$$

where i is the query position, $N_p = H \cdot W$ is the number of positions in the scene feature map, and j enumerates all the possible positions. The value function $g(\cdot)$ outputs a new representation of scene representation at position j under the constraint of text representation T, and is implemented as an 1×1 convolution. A triplet function $f(\cdot, \cdot, \cdot)$ computes a scalar representing the relationship between i and all j under T, which can be decomposed into a pixel-wise self-attention and a cross-attention, as

$$f\left(X_{\text{si}}^i, X_{\text{si}}^j, T\right) = \frac{1}{C(X_{\text{si}})}\left(\psi\left(\left(W_{q_i} X_{\text{si}}^i\right)^t \left(W_{k_i} X_{\text{si}}^j\right)\right) + \psi\left(\left(W_{q_g} T\right)^t \left(W_g X_{\text{si}}^j\right)\right)\right) \tag{6}$$

where $C(X_{\text{si}})$ is a normalization factor, ψ is softmax function, t is transpose operation, and W_q, W_k, and W_g are trainable transformation matrices. After obtaining the transformed scene–text feature y_s, the normalized scene feature X_{si} is content modified based on it, which is formulated as

$$X_{\text{sc}} = \text{con}_{1\times1}(y_s) + X_{\text{si}} \tag{7}$$

where $\text{con}_{1\times1}$ is a 1×1 convolution. $+X_{\text{si}}$ can be regarded as a residual connection, that allows content modifications to occur. Similar to the multihead self-attention block [55], DMNL can be implemented with multiple heads and being stacked multiple times for performance improvements.

DMNL block pursues efficient and stable training through the synergy of two independent modules—pixel-wise self-attention and cross-attention. Moreover, it can be regarded as a global multimodal context modeling block that aggregates query-specific global context features under specific constraints (another modality, such as text) to each query position. Therefore, the time and space complexity of the

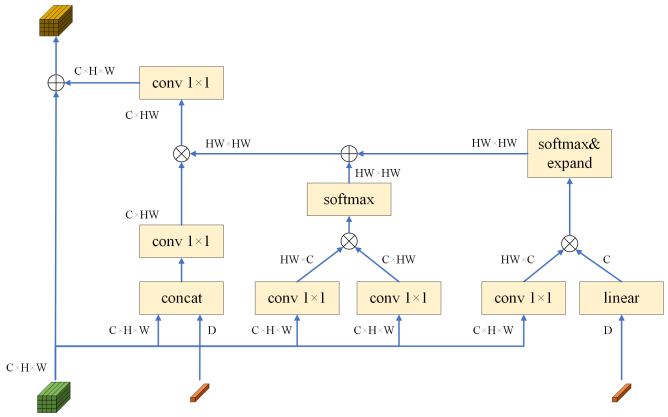

Fig. 5. Architecture of DMNL. The features are expressed in simplified terms as their dimensions, e.g., $C \times H \times W$ and D. $\oplus$ denotes element-wise addition, $\otimes$ refers to matrix multiplication.

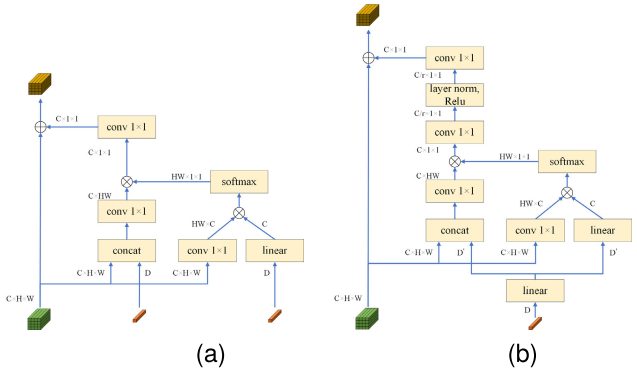

Fig. 6. Architectures of (a) SDMNL block and (b) our MGC block.

DMNL block are both quadratic to the position number Np, which is obviously not friendly to large feature maps.

b) Simplying DMNL: Considering that numerous matrix computation is contained in the self-attention module, we simplify DMNL by omitting the pixel-wise self-attention module but just keeping the cross-attention module. Our simplified DMNL (SDMNL) block is formulated as

$$f\left(X_{\mathrm{si}}^{i}, X_{\mathrm{si}}^{j}, T\right) = \frac{1}{C(X_{\mathrm{si}})}\psi\left(\left(W_{q_g}T\right)^{t}\left(W_g X_{\mathrm{si}}^{j}\right)\right) \tag{8}$$

where W_{q_g} and W_g are trainable linear transformation matrices. The architecture of SDMNL is shown in Fig. 6(a). We evaluate the importance of self-attention under our framework in Table IV, and the result shows that our model can achieve comparable results with or without it.

Although SDMNL greatly reduces the complexity, it retains the complex $\mathrm{con}_{1\times1}$ computation in 7, which is computationally intensive when the number of feature channels is large. In this article, the scene feature is obtained from layer 4 of ResNet-50 [56], whose number of channels is 2048, so the parameter number of this $\mathrm{con}_{1\times1}$ convolution is $C{\cdot}C = 2048 \times 2048$, which greatly increases the parameters of this block.

c) MGC block: Squeeze-excitation (SE) block [57] employs a specially designed bottleneck containing a 1×1 convolution, a nonlinear activation layer Relu, another 1×1 convolution, and a nonlinear activation layer Sigmoid to flexibly learn nonlinear interrelationships between channels, where a scaling ratio r is designed to decrease the parameters by scaling the number of channels. As the channel is C, the parameter is $2 \cdot C \cdot C/r$. Benefiting from the lightweight computation of SDMNL and SE block, we further simplify SDMNL and propose a new block named MGC. As shown in Fig. 6(b), in addition to simplifying the base DMNL, we introduce an additional content extractor F_{tc} for content representation generation of text, as we conjecture that the content modifications should be determined by the content of modifiers, which is given by

$$T_c = F_{\mathrm{tc}}(T). \tag{9}$$

Here, $F_{\mathrm{tc}}{:}\mathbb{R}^D \rightarrow \mathbb{R}^{D\prime}$ is implemented as a fully connected layer followed by a normalization layer. Thereafter, the output of our content modulator can be reformulated as

$$X_{\mathrm{sc}} = F_c(y_s) + X_{\mathrm{si}} \tag{10}$$

where F_c is a bottleneck containing two transformation matrices with scaling ratio r and a nonlinear activation Relu with layer normalization. We set the ratio r to 16, at which point the number of parameters of the module is 1/8 of the original, and we show the effect of different ratios on the retrieval results in Table VI. Note that the layer normalization is inserted before Relu because it effectively alleviates the optimization difficulty caused by the two transformation layers in the bottleneck [58]. Similarly, MGC can be implemented with multiheads as well as being stacked multiple times.

2) Style Modulator: To update the style of the scene representation, an affine transformation is applied to the individual channel of the output of the content modulator X_{sc} to modulate its channel-wise statistics, which is formulated as

$$X_{\mathrm{ss}} = \alpha X_{\mathrm{sc}} + \beta \tag{11}$$

where $\alpha \in \mathbb{R}^C$ and $\beta \in \mathbb{R}^C$ are affine parameters, and are defined as

$$\begin{aligned} \alpha &= \sigma(\varphi_\alpha(T_s)) \cdot \sigma_{X_s} + f_\alpha(T_s) \\ \beta &= \sigma\left(\varphi_\beta(T_s)\right) \cdot \mu_{X_s} + f_\beta(T_s) \end{aligned} \tag{12}$$

where $\sigma(\cdot)$ is the sigmoid function, $\varphi(\cdot){:}\mathbb{R}^{D''} \rightarrow \mathbb{R}^C$ is a gating unit that determines the certain information of the original scene feature (i.e., X^s) to be retained while discarding others, $f(\cdot){:}\mathbb{R}^{D''} \rightarrow \mathbb{R}^C$ is a linear transformation function that injects new information into the scene features (i.e., X_{sc}) based on its input signal, and T_s is the style of text, which is formulated as

$$T_s = F_{\mathrm{ts}}(T) \tag{13}$$

where $F_{\mathrm{ts}}{:}\mathbb{R}^D \rightarrow \mathbb{R}^{D''}$ is a solely fully connected layer.

D. Optimizer Learning

After obtaining the unified representation of the reference input $X_{\mathrm{ss}} \in \mathbb{R}^{C\times H\times W}$ and the joint image–graph representation of the target input $X_{\mathrm{ts}} \in \mathbb{R}^{C\times H\times W}$, we project them into the same embedding place with a projector layer F_p, which is formulated as

$$\begin{aligned} X_{\mathrm{ref}} &= F_p(X_{\mathrm{ss}}) \\ X_{\mathrm{tar}} &= F_p(X_{\mathrm{ts}}). \end{aligned} \tag{14}$$

Here, F_p:$\mathbb{R}^{C\times H\times W} \rightarrow \mathbb{R}^L$ is implemented as a global average pooling followed by a linear transformation layer and a l_2 normalization layer, in which L is the dimension of final features.

Toward the purpose that the reference representation (i.e., X_{ref}) is exclusively aligned with the target representation (i.e., X_{tar}), we adopt the batch-based classification loss (BBCL) [8] for parameter optimization, which encourages positive sample pairs to be closer while negative pairs to be further away, and is proven to be more discriminative and can converge faster. BBCL is formulated as

$$L = \frac{1}{B}\sum_{i=1}^{B} -\log \frac{\exp\left(k\left(X_{\text{ref}}^{i}, X_{\text{tar}}^{i}\right)\right)}{\sum_{j=1}^{B}\exp\left(k\left(X_{\text{ref}}^{i}, X_{\text{tar}}^{j}\right)\right)} \tag{15}$$

where B is the number of sample pairs in each training minibatch, $k(\cdot,\cdot)$ is an arbitrary distance function and is implemented as the cosine similarity due to its simplicity, and X_{ref} and X_{tar} with the same superscript are positive sample pairs, otherwise they are negative sample pairs.

IV. Experiments

In this section, we conduct extensive experiments and analyses to evaluate our proposed SHF. Section IV-A introduces the experimental settings. Section IV-B shows the overall comparison of SHF and baselines. Section IV-C evaluates how the quality of the scene graph affects the retrieval results. Section IV-D conducts extensive ablation studies.

A. Experimental Settings

In Section IV-A1, we introduce our constructed datasets. In Section IV-A2, we summarize the baselines to be compared. In Section IV-A3, we describe the implementation details. The evaluation metric for quantitative evaluation is introduced in Section IV-A4.

1) Experimental Datasets: To demonstrate the effectiveness of our proposed SHF, we created three datasets named Airplane, Tennis, and WHIRT, respectively. The datasets are organized in terms of quintets, consisting of a reference RS image and its scene graph, a target RS image and its scene graph, and a pair of modifier sentences, examples are shown in Fig. 7. For quality assurance, ten domain experts are involved in the annotation and calibration of the dataset. When training the model, two texts are combined by "and" into one as input.

a) Airplane dataset: The dataset contains 1600 RS images with the category airplane and 3461 pairs of modifier sentences. Among them, images are collected from three public and widely used RS datasets, including UCM [59], PatternNet [19], and NWPU-RESISC45 [60]. The scene graph mainly contains the attributes (e.g., color, number, and orientation) of the target object (i.e., airplane) and the spatial relationships (e.g., one side, two sides, and round) between the target object and other nontarget objects, while it does not contain the spatial relationships between nontarget objects. The modifier sentences focus on the description of the differences in attributes of the target object and the differences in spatial relationships between the target object and other objects.

b) Tennis dataset: This dataset focuses on further refining the reference image, i.e., identifying target images that are similar to the reference image and meeting the modifier sentences from the candidate images with the category tennis court. The spatial relationship between target and nontarget objects and the properties of the target object is emphasized in the scene graph of each image, while the relationship between nontarget objects is ignored. The Tennis dataset contains 1200 tennis court images from UCM, PatternNet, and NWPU-RESISC45 and 1924 manually annotated pairs of modifier sentences.

c) WHIRT dataset: The dataset contains 4940 RS images from WHDLD [61] and 3344 manually generated ⟨reference image, target image⟩ pairs. The scene graphs are constructed based on the attributes of all objects in the image and the spatial relationships between two objects. The modifier sentences reflect all the differences between the reference and target images.

To train the models, we split the datasets. The training set of the Airplane contains 1280 RS images, of which 2667 pairs are used for training, and the test set of 794 queries for evaluation. As for the Tennis, the training set contains 960 RS images, and the test set contains 458 queries. Among the 3952 images in the training set of the WHIRT, there are 2203 image pairs.

2) Baselines: We compare our SHF with the following baselines.

1) *Image Only:* It takes the image representation extracted from the final layer of a CNN as the composed representation.
2) *Text Only:* It takes the text representation extracted from an RNN as the composed representation.
3) *Concatenation:* It concatenates the image and text representations and then feeds the integrated representation to a two-layer MLP with Relu to obtain the composed representation.
4) *TIRG [8]:* It combines the image and text representations by concatenation and then obtains the composed representation by learning a gated residual connection.
5) *ComposeAE [35]:* It projects the image and text representations into a complex embedding space and learns the combination of the projected representations through a deep metric learning method.
6) *Cosmo [10]:* It uses a content modulator and a style modulator to update the image representation to the composed representation according to the text representation.
7) *CLIP4Cir [42]:* It encodes the image and text into visual and linguistic representations using clip and then integrates the multimodal representations through a combiner composed of multiple linear layers, nonlinear activation functions, and dropout layers.
8) *AACL [38]:* It utilizes the Swin transformer [39] as the image encoder and DistilBERT [40] as the text encoder, with the additive self-attention layer similar to FastFormer [62] serving as the image–text compositor.
9) *UncerRe [44]:* It augments Cosmo by adding Gaussian Noise to the target features and incorporating an uncertainty regularization term to the original BBCL loss.

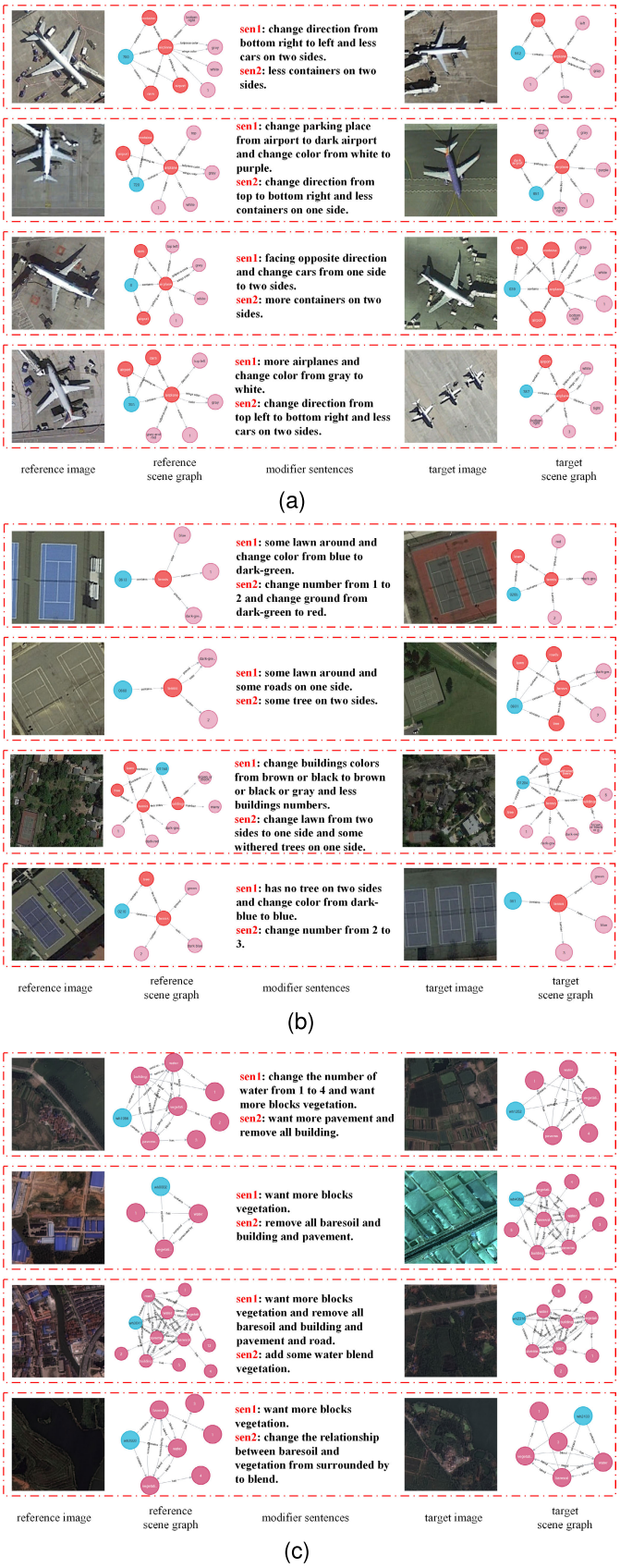

Fig. 7. Example of (a) Airplane, (b) Tennis, and (c) WHIRT datasets.

For a fair comparison, all baselines are implemented using the same image encoder and text encoder. However, we defer the attempts of pretrained transformer-based encoders to future work.

3) Implementation Details: Our image encoder, f_{img}, is implemented as ResNet-50 pretrained on ImageNet [63] without the final fully connected layer, whose output feature map has 2048 channels. Both target and reference images share the same image encoder. In the feature aggregation of graph encoder, f_{graph}, we set the number of relational GCN layers to 2, the hidden dimension and output dimension to 512, and each layer is followed by a batch normalization layer. In addition, the graph encoder is shared by both the target and reference scene graphs. Our text encoder is composed of a one-layer LSTM with 512 hidden units, which outputs the final text feature $T \in \mathbb{R}^{512}$. The text content extractor F_{tc} and the style extractor F_{ts} map 512-D features into a new 512-D vector, respectively. We stack the content modulator twice and set the number of attention heads to 8. The final projector, F_p, consists of a global average pooling layer that compresses the 3-D feature map into 1-D, followed by a linear layer that projects the 1-D feature onto a 512-dimension vector to obtain the final representation X^{ref}, $X^{\text{tar}} \in \mathbb{R}^{512}$ for retrieval.

A rectified Adam [64] optimizer is applied to optimize our model whose initial learning rate is set to $2 \times e^{-4}$ and decays once after 30 epochs with a factor of 10. The training epoch is set to 80. The batch size is set to 32 for the Airplane and Tennis datasets and 16 for the WHIRT dataset. The parameters of the baselines are set consistent with SHF or adjusted according to the original paper. All the experiments are conducted in Pytorch [65].

4) Evaluation Metric: To quantitatively compare our model with baselines, a standard evaluation metric in retrieval, recall at rank k (i.e., Recall@k, short for $R@k$) is adopted, defined as the percentage of evaluation queries where the target image is within the top k retrieved images. A larger $R@k$ score indicates a better performance. We report performance with $k \in \{1, 5, 10, 20\}$ on each dataset.

B. Comparison With Baselines

1) Overall Comparison on Airplane: Table I reports our results on the Airplane dataset. It can be seen that our method significantly outperforms the baselines. For instance, SHF surpasses the best scores of baselines by approximately 41% and 32% points on $R@1$ and $R@10$, respectively, showcasing the effectiveness of introducing scene graphs in RS image retrieval with text feedback. Comparing the results of image only and text only reveals that solely relying on visual features may not suffice, particularly when substantial differences exist between the contents of reference and target images. In comparison to text only, CLIP4Cir demonstrates an improvement of over 3% scores for both $R@1$ and $R@10$, indicating that combining visual and linguistic features using an appropriate compositor leads to enhanced performance. In addition, UncerRe outperforms Cosmo, underscoring the effectiveness of introducing uncertainty into the model. The retrieval results for the first ranking of the different methods are visualized in Fig. 8, and the results show that SHF significantly outperforms the baselines. In the third example with better retrieval results

TABLE I
QUANTITATIVE RESULTS ON AIRPLANE DATASET. THE BEST RESULTS ARE MARKED IN BOLD AND THE SECOND BEST RESULTS ARE UNDERLINED

	R@1	R@5	R@10	R@20
Image Only	0.25	2.14	5.29	8.06
Text Only	17.82	45.21	59.51	71.60
Concatenation	14.42	39.04	54.03	66.18
TIRG [8]	15.05	37.28	52.20	66.18
ComposeAE [35]	16.18	45.34	61.40	75.76
Cosmo [10]	20.91	49.31	62.66	74.87
CLIP4Cir [42]	20.53	51.19	63.85	76.95
AACL [38]	14.29	45.53	61.08	72.73
UncerRe [44]	19.14	51.13	66.25	77.14
SHF	**62.15**	**96.22**	**98.43**	**99.37**

TABLE II
QUANTITATIVE RESULTS ON TENNIS DATASET

	R@1	R@5	R@10	R@20
Image Only	1.31	5.68	10.26	16.16
Text Only	14.84	38.43	52.95	64.63
Concatenation	7.86	24.67	39.85	53.93
TIRG [8]	5.68	22.16	35.04	50.55
ComposeAE [35]	5.90	23.58	37.55	54.80
Cosmo [10]	11.68	39.63	55.02	68.89
CLIP4Cir [42]	17.58	46.51	61.68	72.82
AACL [38]	7.42	28.71	45.41	59.50
UncerRe [44]	13.97	43.78	60.59	73.58
SHF	**43.23**	**84.93**	**93.89**	**96.83**

TABLE III
QUANTITATIVE RESULTS ON WHIRT DATASET

	R@1	R@5	R@10	R@20
Image Only	0.35	1.14	1.93	2.98
Text Only	2.32	8.50	14.07	23.53
Concatenation	1.05	2.19	4.56	8.19
TIRG [8]	0.83	4.18	4.56	8.19
ComposeAE [35]	1.62	5.17	9.58	14.86
Cosmo [10]	1.71	7.23	13.15	20.73
CLIP4Cir [42]	2.10	9.86	17.62	26.95
AACL [38]	2.28	7.93	14.64	22.57
UncerRe [44]	1.84	9.16	17.66	28.31
SHF	**5.08**	**21.60**	**36.33**	**54.12**

for the baselines, the modifier sentences only modify the intrinsic properties (color and orientation) of the target object, which indicates that the baseline methods perform better in understanding the changes in the intrinsic properties of the target object. The qualitative results of SHF on the Airplane dataset shown in Fig. 9 indicate that even the erroneous images retrieved by SHF satisfy the modifier sentences to change the intrinsic properties of the target object. Combining Figs. 8 and 9, it can be seen that SHF can not only change the intrinsic properties of the target object (e.g., color and orientation) and the external spatial relationship with other objects according to the conditions of the text feedback but also accurately respond to the change of the environment in which the target object is located (i.e., parking place).

2) Overall Comparison on Tennis: The quantitative comparison results between SHF and various baseline methods on the Tennis dataset are presented in Table II. These results indicate that SHF significantly outperforms the baseline method across all evaluation metrics, akin to the findings on the Airplane dataset, albeit with a slightly smaller margin. Compared to the best scores of the baseline methods, SHF achieved performance gains of about 25% and 32% points on $R@1$ and $R@10$, respectively. In contrast to the results observed on the Airplane dataset, CLIP4Cir consistently outperforms UncerRe. This observation suggests that in certain cases, a simpler compositor may prove to be more effective. The retrieval results in Fig. 10 reveal that the baselines can make better modifications to the intrinsic properties of the target object based on the modifier sentences. For example, Cosmo accurately retrieved the correct images with corresponding changes to the number of tennis courts in all three examples but exhibited limitations in understanding the global changes of images, such as variations in the numbers of nontarget objects and alterations in spatial relationships between objects. Furthermore, the qualitative results in Fig. 11 further demonstrate that SHF excels in jointly comprehending the intrinsic properties and extrinsic spatial relationships of the target object.

3) Overall Comparison on WHIRT: The quantitative evaluation results of SHF and various baselines on the WHIRT dataset are shown in Table III. Unlike the Airplane and Tennis datasets, the WHIRT dataset lacks a clearly defined target object category, features more spatial relationships, and comprises more complex image scenes. In addition, the similarity between the reference image and the target image is lower. Consequently, retrieving images poses a greater difficulty, resulting in relatively lower evaluation metric scores. Overall, the evaluation metric scores of SHF are still significantly ahead of other baseline methods, including 11% and 18% points higher than the suboptimal scores on $R@5$ and $R@10$, respectively, once again proving the superiority of SHF in complex RS image retrieval with text feedback. Furthermore, the baseline method text only consistently outperforms most other methods, probably because the modifier sentences focus more on alterations to the spatial relationships of objects, again underscoring the challenge of accurately perceiving changes in the extrinsic properties with the existing methods. The comparison results in Fig. 12 prove the superiority of SHF, and even the retrieved error images can also meet the sentence content well. The qualitative results in Fig. 13 show that SHF can perform better even if the content of the target image has changed significantly, proving that SHF can cope with complex content changes in RS image retrieval.

C. Impact of Scene Graph Quality

SHF significantly improves the retrieval accuracy of RS images with text feedback by introducing scene graphs to generate better scene representations. To explore the effect of scene graph quality on retrieval results, we randomly

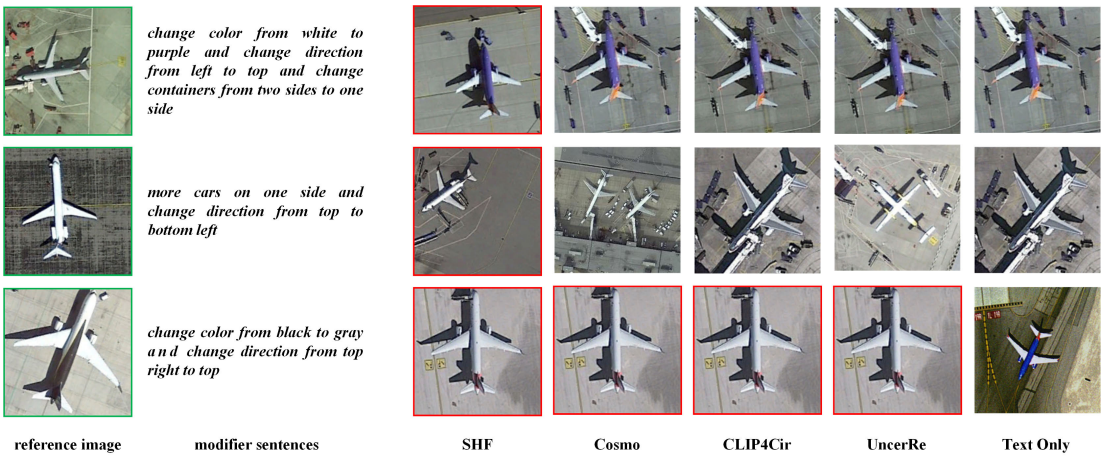

Fig. 8. Visualization of the first ranked retrieval results on the Airplane dataset. The correct retrieval results are marked in red.

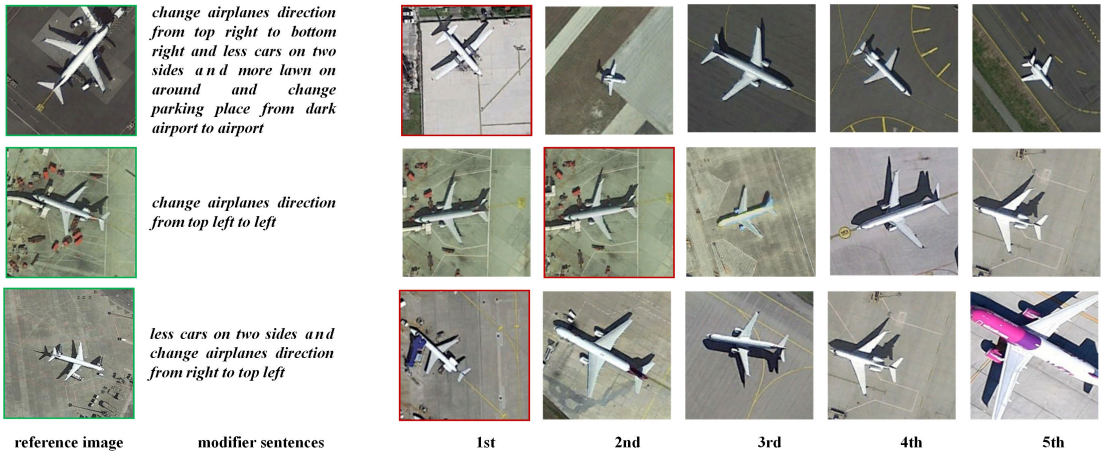

Fig. 9. Qualitative results of SHF on Airplane dataset. The top-5 retrieval results are reported. Green/red boxes: reference/target RS images.

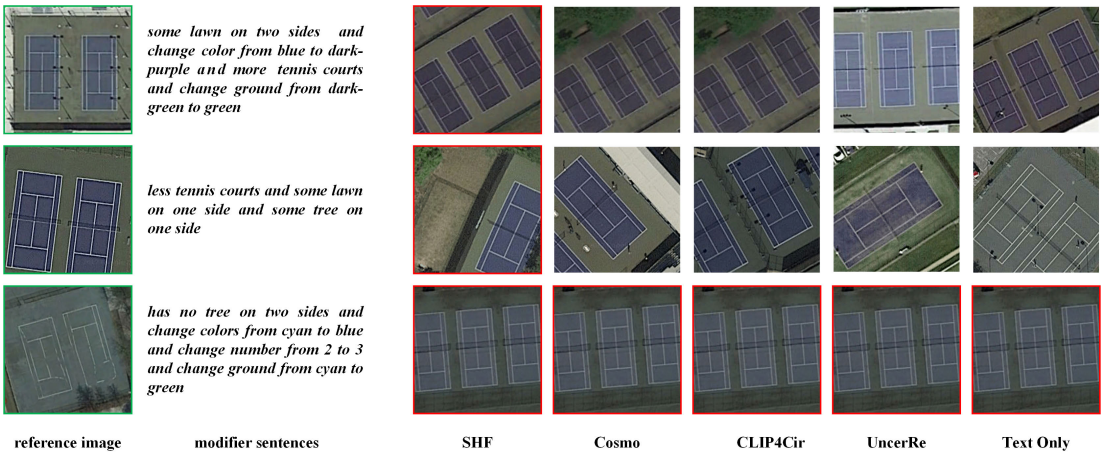

Fig. 10. Visualization of the first ranked retrieval results on Tennis dataset.

masked a certain ratio of edges of the scene graph and varied the masking ratio within $\{0, 0.1, 0.2, 0.3, 0.4\}$. Fig. 14 shows the results on the Airplane and Tennis datasets. We observe that as the masking ratio increases, the scores of evaluation metrics decrease dramatically, indicating that the scene graph quality has a significant impact on the model performance. Specifically, as the proportion of edge masking reaches 0.4, the $R@10$ of SHF on the Airplane and Tennis datasets decreases by 40% and 38% points, respectively, and is even lower than the performance of the baseline model Cosmo, indicating that when the quality of the scene graph is relatively low, it will have a negative impact on the retrieval. However, when the

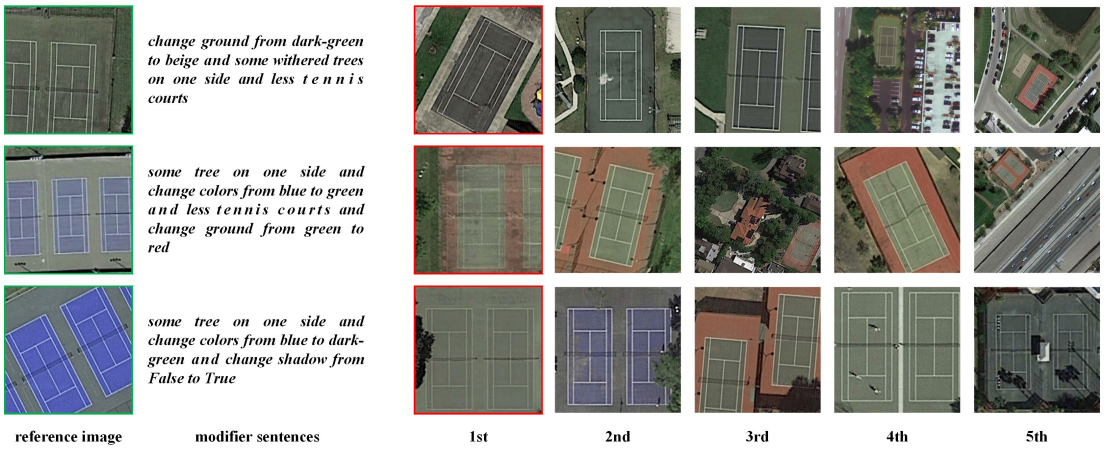

Fig. 11. Qualitative results of SHF on Tennis dataset.

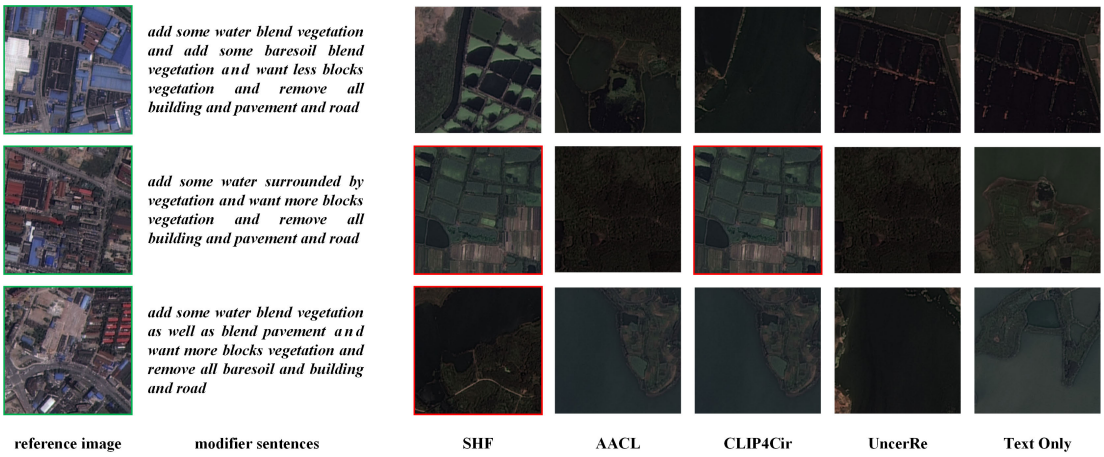

Fig. 12. Visualization of the first ranked retrieval results on the WHIRT dataset.

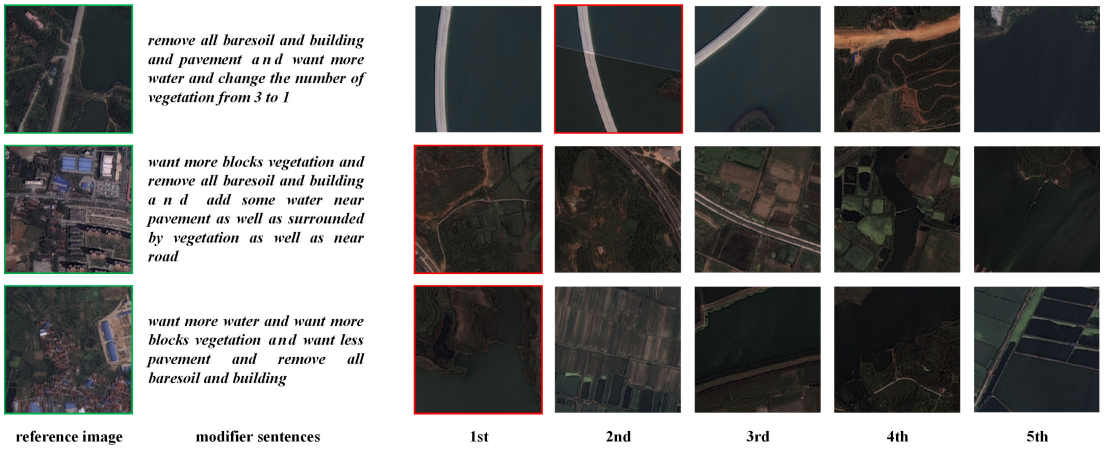

Fig. 13. Qualitative results of SHF on WHIRT dataset.

masking ratio is lower than 0.4, SHF still performs better than the baseline method, which reflects the superiority and competitiveness of SHF.

D. Ablation Studies

1) Impact of Proposed Components: SHF is a model with the following improvements over Cosmo: 1) the scene graph is introduced as auxiliary information; 2) the text content and style extractors are used to control the change of scene features; 3) the self-attention module in DMNL is discarded; and 4) the 1×1 convolution is replaced by an excitation module with a scaling factor r when the content is modified. To verify the effectiveness of the improvements, we conducted ablation experiments, and Table IV shows the experimental results on

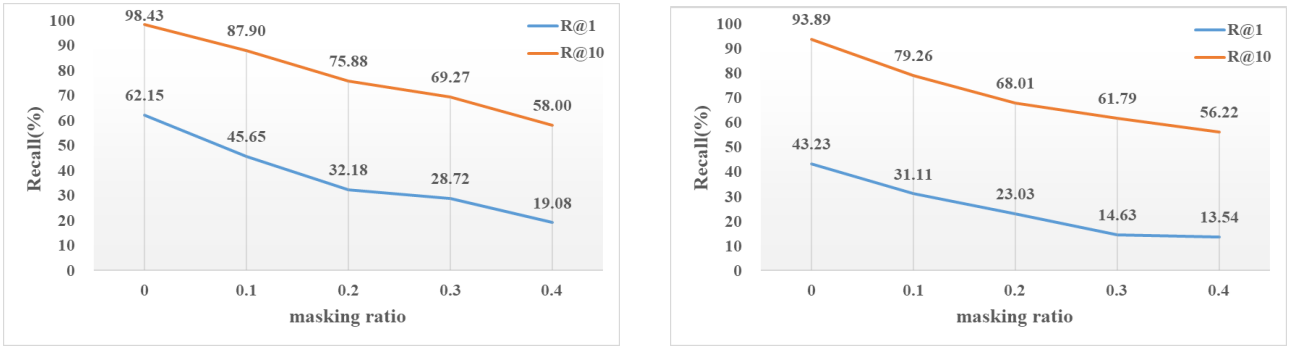

Fig. 14. Performance variation under different masking ratios for scene graphs on (a) Airplane and (b) Tennis datasets.

TABLE IV
ABLATION STUDY OF COMPONENTS

	Airplane		Tennis		WHIRT	
	R@1	R@10	R@1	R@10	R@1	R@10
Cosmo	20.91	62.66	11.68	55.02	1.71	13.15
add te	19.40	63.04	14.30	56.77	1.67	11.83
add te&sg	44.65	93.58	41.70	93.34	4.34	33.74
add te&sg w/o ss	60.08	98.80	41.81	93.01	5.30	36.06
SHF	62.15	98.43	43.23	93.89	5.08	36.33

[1] te is the text content & style extractors.
[2] sg is the scene graph.
[3] ss is the self-attention module .
[4] w/o means remove operation.

TABLE V
PERFORMANCE VARIATION UNDER DIFFERENT INPUTS

	Airplane		Tennis		WHIRT	
	R@1	R@10	R@1	R@10	R@1	R@10
image	0.25	5.29	1.31	10.26	0.35	1.91
graph	0.44	5.46	0.63	5.42	0.00	2.89
image + text	23.87	64.04	9.17	47.27	1.62	10.78
text + graph	0.13	5.29	0.55	6.77	0.26	1.71
image + graph	1.01	6.68	0.44	10.26	0.35	1.75
image + text + graph	62.15	98.43	43.23	93.89	5.08	36.33

the three datasets. It can be seen that each of the improvements contributed to the retrieval performance. Comparing the third and fourth rows shows that the text extractor can significantly improve the performance of the model on the Tennis dataset; comparing the fourth and fifth rows shows that the introduction of scene graphs has significantly improved the retrieval results, and the model has improved the $R@10$ on the three datasets by 30%, 36%, and 22% points, respectively, verifying the conjecture that scene graphs can help further understand the RS image scenes and modify the text; comparing the fifth and sixth rows shows that the self-attention module has even a negative effect on retrieval when removing this module, the model improves $R@1$ on the Airplane dataset by 15% points, indicating that changes to the scene should be mainly determined by the text (cross-attention module); the sixth and seventh rows demonstrate the boosting effect of using the bottleneck module F_C to replace the 1×1 convolution on retrieval results.

2) Impact of Different Inputs: An important improvement of SHF is the introduction of RS image scene graphs as part of the input to enhance the understanding of both scene and modifier text. We explored the impact of scene graphs on the results by varying the query and target inputs. The search results are reported in Table V, where the scene representation generation of the query image is consistent with that of the target image, e.g., "graph" and "text + graph" indicate that the scene representations of both the query and the target are derived from the scene graph representation only. Comparing the third and fourth rows, it can be found that the retrieval results on the Tennis dataset drop significantly when using only scene graphs for retrieval compared to using only images as input, probably because the overall similarity of images in

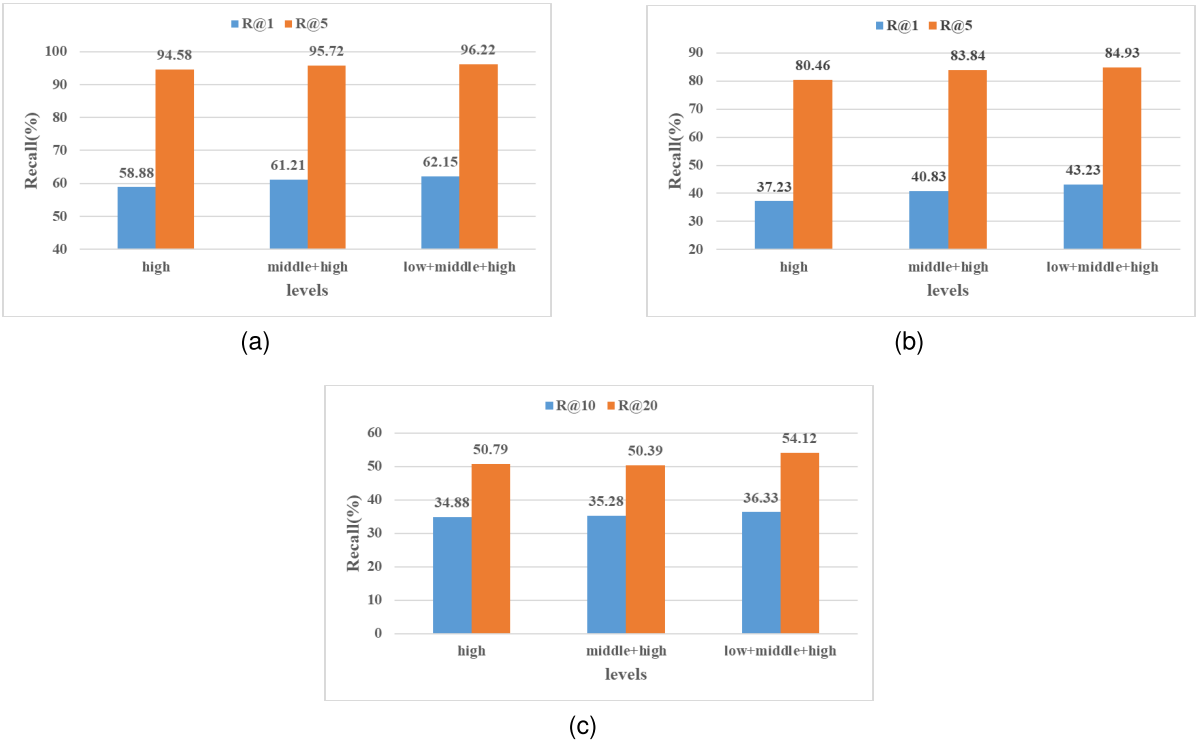

Fig. 15. Performance variation under different feature levels. (a) Performance on Airplane dataset. (b) Performance on Tennis dataset. (c) Performance on WHIRT dataset.

the Tennis dataset is relatively high, which further narrows the differences between images when using scene graphs to represent images, making retrieval difficult; comparing the fifth and sixth rows, the retrieval accuracy decreases significantly when the scene graph representation is directly converted using modifier sentences, which may be mainly due to the fact that the scene graph is a global structural representation of RS images, while it lacks the most basic fine-grained features such as spectra and textures; comparing the third, fourth, and seventh rows, the retrieval accuracy is higher when using a hybrid image-scene graph representation on the Airplane dataset than when using only a single representation, which indicates that scene graphs can help increase the variability of images to a certain extent; the fifth and eighth rows show that introducing scene graphs in the original retrieval structure can effectively improve the retrieval accuracy, indicating that scene graphs can help better understand the text.

3) Impact of Multilevel Composition: In the image–graph compositor, SHF fuses low-, middle-, and high-level features of image and scene graphs, respectively. We explore how the combination of features at different levels can help scene representation learning by comparing SHF (low + middle + high) with two baselines: 1) middle + high and 2) high. Fig. 15 shows that except in the WHIRT dataset, where the $R@20$ of SHF (middle + high) is slightly lower than that of SHF (high), the evaluation metric scores improve significantly with more fusion levels, indicating that the combination of features at multiple levels helps improve the overall performance, which validates the efficacy of using feature compositors of different depths to capture multigranularity information, consistent with the fact that CNNs and GCNs continuously learn global features from low to high levels.

TABLE VI
PERFORMANCE VARIATION UNDER DIFFERENT SCALING RATIO

	Airplane		Tennis		WHIRT	
	R@1	R@10	R@1	R@10	R@1	R@10
w/o r	60.08	98.80	41.81	93.01	5.30	36.06
4	63.60	98.30	**44.87**	93.23	**6.18**	**39.13**
8	**64.48**	**98.87**	37.33	92.58	5.30	35.32
16	62.15	98.43	43.23	**93.89**	5.08	36.33
32	61.71	98.74	37.77	92.36	5.48	35.89

4) Impact of Scaling Ratio: The bottleneck module F_C is designed to reduce the redundancy of parameters and provide a good tradeoff between performance and parameters. The effect of the scaling ratio on the results is shown in Table VI, where w/o r indicates the use of the original 1×1 convolution as the bottleneck module. The experimental results show that an appropriate scaling ratio can improve the performance of the model. When r is 8, 4, and 4, the model achieves the best performance on the Airplane, Tennis, and WHIRT datasets, respectively, indicating that the setting of r should be appropriately selected according to the dataset. Overall, the larger the number of parameters, the better the performance of the model, indicating that SHF has a good balance between the number of parameters and the model performance.

5) Impact of the Number of MGC: Fig. 16 shows that as the number of stacks of the MGC module increases, the performance of SHF on all three datasets tends to increase and

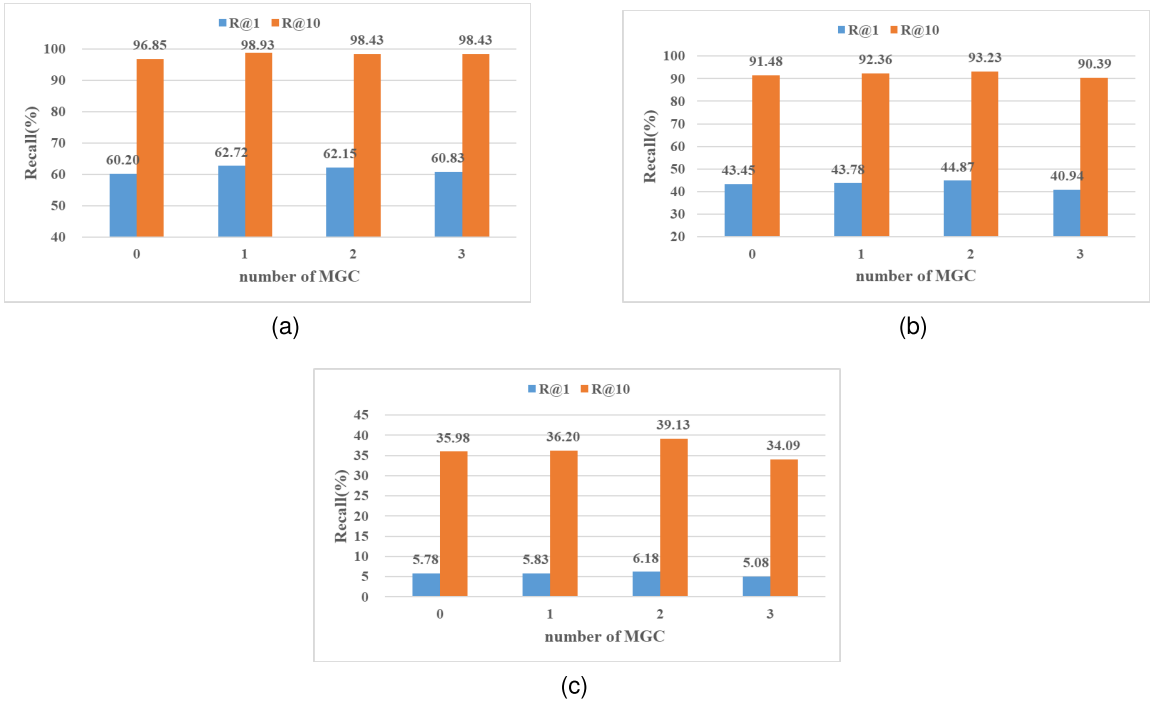

Fig. 16. Performance variation under different MGC numbers. (a) Performance on Airplane dataset. (b) Performance on Tennis dataset. (c) Performance on WHIRT dataset.

then decrease. The optimal performance is reached when the number of stacks is 1, 2, and 2, respectively, probably because multiple stacks help the model understand the content information of the text and further modify the content information of the scene representation based on this information; when the number of stacks exceeds 2, the performance decreases due to the overload of parameters and over-modification of the content.

V. Conclusion

In this article, we propose a new SHF for RS image retrieval with text feedback. The core idea is to leverage scene graphs as side information to enrich image features and to elucidate the consistency and difference between images. Specifically, we first fuse multilevel visuals and structural features to generate scene representations with multiple granularity, ranging from local to global. We then employ a content modulator to modify the content of the scene based on the modifier text content and a style modulator to modify the style of the scene based on the text style, aligning the scene representations of the target image. Extensive comparative experiments conducted on three datasets demonstrate that SHF outperforms the state-of-the-art methods significantly.

The field of RS image retrieval with text feedback remains relatively unexplored, and our work represents just the beginning. In the future, we aim to design a compositor capable of simultaneously fusing features from the image, scene graph, and modifier sentence. In addition, we believe that joint learning of scene graph generation and image retrieval with text feedback holds promise within our framework.

References

[1] Y. Li, J. Ma, and Y. Zhang, "Image retrieval from remote sensing big data: A survey," *Inf. Fusion*, vol. 67, pp. 94–115, Mar. 2021.

[2] W. Zhou, H. Guan, Z. Li, Z. Shao, and M. R. Delavar, "Remote sensing image retrieval in the past decade: Achievements, challenges, and future directions," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 1447–1473, 2023.

[3] Y. Li, J. Luo, Y. Zhang, Y. Tan, J.-G. Yu, and S. Bai, "Learning to holistically detect bridges from large-size VHR remote sensing imagery," 2023, *arXiv:2312.02481*.

[4] P. Staszewski, M. Jaworski, J. Cao, and L. Rutkowski, "A new approach to descriptors generation for image retrieval by analyzing activations of deep neural network layers," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 7913–7920, Dec. 2022.

[5] Ş. Öztürk, E. Çelik, and T. Çukur, "Content-based medical image retrieval with opponent class adaptive margin loss," *Inf. Sci.*, vol. 637, Aug. 2023, Art. no. 118938.

[6] Z. Yuan et al., "A lightweight multi-scale crossmodal text-image retrieval method in remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5612819.

[7] Z. Yuan et al., "Remote sensing cross-modal text-image retrieval based on global and local information," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5620616.

[8] N. Vo et al., "Composing text and image for image retrieval—An empirical Odyssey," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6439–6448.

[9] 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. 3001–3011.

[10] S. Lee, D. Kim, and B. Han, "CoSMo: Content-style modulation for image retrieval with text feedback," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 802–812.

[11] W. Cui et al., "Multi-scale semantic segmentation and spatial relationship recognition of remote sensing images based on an attention model," *Remote Sens.*, vol. 11, no. 9, p. 1044, May 2019.

[12] J. Johnson et al., "Image retrieval using scene graphs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3668–3678.

[13] Y. Li, D. Kong, Y. Zhang, Y. Tan, and L. Chen, "Robust deep alignment network with remote sensing knowledge graph for zero-shot and generalized zero-shot remote sensing image scene classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 179, pp. 145–158, Sep. 2021.

[14] F. Wang, X. Zhu, X. Cheng, Y. Zhang, and Y. Li, "MMKDGAT: Multimodal knowledge graph-aware deep graph attention network for remote sensing image recommendation," *Expert Syst. Appl.*, vol. 235, Jan. 2024, Art. no. 121278.

[15] J. Wang and X.-S. Hua, "Interactive image search by color map," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 1, pp. 1–23, Oct. 2011.

[16] S. Newsam, L. Wang, S. Bhagavathy, and B. S. Manjunath, "Using texture to analyze and manage large collections of remote sensed image and video data," *Appl. Opt.*, vol. 43, no. 2, pp. 210–217, 2004.

[17] M. Wang and T. Song, "Remote sensing image retrieval by scene semantic matching," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 5, pp. 2874–2886, May 2013.

[18] P. J. Bosco and S. Janakiraman, "Content-based image retrieval (CBIR): Using combined color and texture features (TriCLR and HistLBP)," *Int. J. Image Graph.*, Sep. 2023, Art. no. 2550021.

[19] W. Zhou, S. Newsam, C. Li, and Z. Shao, "PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 197–209, Nov. 2018.

[20] W. Ma, T. Zhou, J. Qin, X. Xiang, Y. Tan, and Z. Cai, "Adaptive multi-feature fusion via cross-entropy normalization for effective image retrieval," *Inf. Process. Manage.*, vol. 60, no. 1, Jan. 2023, Art. no. 103119.

[21] Y. Li, Y. Zhang, X. Huang, H. Zhu, and J. Ma, "Large-scale remote sensing image retrieval by deep hashing neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 950–965, Feb. 2018.

[22] P. Li et al., "Hashing nets for hashing: A quantized deep learning to hash framework for remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7331–7345, Oct. 2020.

[23] W. Song, Z. Gao, R. Dian, P. Ghamisi, Y. Zhang, and J. A. Benediktsson, "Asymmetric hash code learning for remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5617514.

[24] Y. Li, M. Hao, R. Liu, Z. Zhang, H. Zhu, and Y. Zhang, "Semantic-aware attack and defense on deep hashing networks for remote-sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5627214.

[25] Y. Li, Y. Zhang, X. Huang, and J. Ma, "Learning source-invariant deep hashing convolutional neural networks for cross-source remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6521–6536, Nov. 2018.

[26] T. Abdullah, Y. Bazi, M. M. Al Rahhal, M. L. Mekhalfi, L. Rangarajan, and M. Zuair, "TextRS: Deep bidirectional triplet network for matching text to remote sensing images," *Remote Sens.*, vol. 12, no. 3, p. 405, Jan. 2020.

[27] X. Qin, L. Li, F. Hao, M. Ge, and G. Pang, "Multi-level knowledge-driven feature representation and triplet loss optimization network for image–text retrieval," *Inf. Process. Manage.*, vol. 61, no. 1, Jan. 2024, Art. no. 103575.

[28] G. Zhao, C. Zhang, H. Shang, Y. Wang, L. Zhu, and X. Qian, "Generative label fused network for image–text matching," *Knowl.-Based Syst.*, vol. 263, Mar. 2023, Art. no. 110280.

[29] A. K. Bhunia et al., "Adaptive fine-grained sketch-based image retrieval," in *Proc. Eur. Conf. Comput. Vis.* Tel Aviv-Yafo, Israel: Springer, 2022, pp. 163–181.

[30] M. Guo, Y. Yuan, and X. Lu, "Deep cross-modal retrieval for remote sensing image and audio," in *Proc. 10th IAPR Workshop Pattern Recognit. Remote Sens.*, Aug. 2018, pp. 1–7.

[31] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra, "Relevance feedback: A power tool for interactive content-based image retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, no. 5, pp. 644–655, Sep. 1998.

[32] B. Zhao, J. Feng, X. Wu, and S. Yan, "Memory-augmented attribute manipulation networks for interactive fashion search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1520–1528.

[33] H. Wu et al., "Fashion IQ: A new dataset towards retrieving images by natural language feedback," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11307–11317.

[34] M. Shin, Y. Cho, B. Ko, and G. Gu, "RTIC: Residual learning for text and image composition using graph convolutional network," 2021, *arXiv:2104.03015*.

[35] M. U. Anwaar, E. Labintcev, and M. Kleinsteuber, "Compositional learning of image-text query for image retrieval," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1140–1149.

[36] S. Goenka et al., "FashionVLP: Vision language transformer for fashion retrieval with feedback," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 14105–14115.

[37] S. Jandial, P. Badjatiya, P. Chawla, A. Chopra, M. Sarkar, and B. Krishnamurthy, "SAC: Semantic attention composition for text-conditioned image retrieval," 2020, *arXiv:2009.01485*.

[38] Y. Tian, S. Newsam, and K. Boakye, "Fashion image retrieval with text feedback by additive attention compositional learning," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 1011–1021.

[39] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[40] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," 2019, *arXiv:1910.01108*.

[41] X. Han, X. Zhu, L. Yu, L. Zhang, Y.-Z. Song, and T. Xiang, "FAME-ViL: Multi-tasking vision-language model for heterogeneous fashion tasks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 2669–2680.

[42] A. Baldrati, M. Bertini, T. Uricchio, and A. Del Bimbo, "Effective conditioned and composed image retrieval combining CLIP-based features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 21466–21474.

[43] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 8748–8763.

[44] Y. Chen, Z. Zheng, W. Ji, L. Qu, and T.-S. Chua, "Composed image retrieval with text feedback via multi-grained uncertainty regularization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2024.

[45] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.* Zürich, Switzerland: Springer, 2014, pp. 818–833.

[46] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[47] T. Y. Lin, P. Dollàr, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2117–2125.

[48] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*.

[49] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," 2017, *arXiv:1710.10903*.

[50] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. Van Den Berg, I. Titov, and M. Welling, "Modeling relational data with graph convolutional networks," in *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*. Springer, 2018, pp. 593–607.

[51] I. Sutskever, J. Martens, and G. E. Hinton, "Generating text with recurrent neural networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2011, pp. 1017–1024.

[52] W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent neural network regularization," 2014, *arXiv:1409.2329*.

[53] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[54] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6924–6932.

[55] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[56] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[57] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[58] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: Non-local networks meet squeeze-excitation networks and beyond," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1971–1980.

[59] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst.*, Nov. 2010, pp. 270–279.

[60] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.

[61] Z. Shao, K. Yang, and W. Zhou, "Performance evaluation of single-label and multi-label remote sensing image retrieval using a dense labeling dataset," *Remote Sens.*, vol. 10, no. 6, p. 964, Jun. 2018.

[62] C. Wu, F. Wu, T. Qi, Y. Huang, and X. Xie, "Fastformer: Additive attention can be all you need," 2021, *arXiv:2108.09084*.
[63] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[64] L. Liu et al., "On the variance of the adaptive learning rate and beyond," 2019, *arXiv:1908.03265*.
[65] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–12.

Fei Wang received the B.S. degree in geographic information systems and the M.S. degree in environmental science and engineering from the Wuhan University of Technology, Wuhan, China, in 2016 and 2019, respectively, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, in 2023.

He is currently an Engineer with Changjiang Spatial Information Technology Engineering Company Ltd., Wuhan. His research interests include remote sensing knowledge graphs, image retrieval, and recommender systems.

Xianzhang Zhu received the B.S. degree in geographic information systems from the China University of Geosciences, Wuhan, China, in 2015, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, in 2022.

He is currently a Lecturer with the School of Municipal and Geomatics Engineering, Hunan City University, Yiyang, China. His research interests include plane segmentation, 3-D reconstruction, image processing, and remote sensing image retrieval.

Xiaojian Liu was born in 1990. He received the B.S. degree in geomatics engineering from Shandong Agricultural University, Tai'an, China, in 2015, the M.S. degree in geomatics engineering from Shandong University of Science and Technology, Qingdao, China, in 2018, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2023.

He is currently working with the Tianjin Research Institute for Water Transport Engineering, M.O.T, Tianjin, China. His research interests include modeling and application of remote sensing knowledge graphs and semantic segmentation.

Yongjun Zhang (Member, IEEE) received the B.S. degree in geodesy, the M.S. degree in geodesy and surveying engineering, and the Ph.D. degree in geodesy and photography from Wuhan University, Wuhan, China, in 1997, 2000, and 2002, respectively.

He is currently the Dean with the School of Remote Sensing and Information Engineering, Wuhan University. He has published more than 180 research articles and three books. His research interests include aerospace and low-attitude photogrammetry, image matching, combined block adjustment with multisource datasets, object information extraction and modeling with artificial intelligence, integration of LiDAR point clouds and images, and 3-D city model reconstruction.

Dr. Zhang is the Coeditor-in-Chief of The Photogrammetric Record.

Yansheng Li (Senior Member, IEEE) received the B.S. degree in information and computing science from Shandong University, Weihai, China, in 2010, and the Ph.D. degree in pattern recognition and intelligent systems from the Huazhong University of Science and Technology, Wuhan, China, in 2015.

From 2017 to 2018, he was a Visiting Assistant Professor with the Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA. He is currently a Full Professor with the School of Remote Sensing and Information Engineering, Wuhan University (WHU), Wuhan. He has authored more than 100 peer-reviewed journal articles and conference papers. His research interests include multimodal remote sensing foundation model, remote sensing spatiotemporal knowledge graph, and their applications in remote sensing big data mining.

Dr. Li was awarded the Young Surveying and Mapping Science and Technology Innovation Talent Award of the Chinese Society for Geodesy, Photogrammetry, and Cartography in 2022. He received the recognition of the Best Reviewers of IEEE GEOSCIENCE AND REMOTE SENSING LETTERS in 2022. He is an Associate Editor of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and a Junior Editorial Member of The Photogrammetric Record.