This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2021.3124252, IEEE Transactions on Geoscience and Remote Sensing

A Lightweight Multi-scale Crossmodal Text-Image Retrieval Method In Remote Sensing

Zhiqiang Yuan, *Student Member, IEEE,* Wenkai Zhang, *Member, IEEE,* Xuee Rong, Xuan Li, Jialiang Chen, Hongqi Wang *Member, IEEE,* Kun Fu, *Member, IEEE,* and Xian Sun, *Senior Member, IEEE*

***Abstract*—Remote sensing (RS) crossmodal text-image retrieval has become a research hotspot in recent years for its application in semantic localization. However, since multiple inferences on slices are demanded in semantic localization, designing a crossmodal retrieval model with less computation but well performance becomes an emergent and challenging task. In this paper, considering the characteristics of multi-scale and target redundancy in RS, a concise but effective crossmodal retrieval model (LW-MCR) is designed. The proposed model incorporates multi-scale information and dynamically filters out redundant features when encoding RS image while text features are obtained via lightweight group convolution. To improve the retrieval performance of LW-MCR, we come up with a novel hidden supervised optimization method based on knowledge distillation. This method enables the proposed model to acquire dark knowledge of the multi-level layers and representation layers in the teacher network, which significantly improves the accuracy of our lightweight model. Finally, on the basis of contrast learning, we present a method employing unlabeled data to boost the performance of RS retrieval model further. The experiment results on four RS image-text datasets demonstrate the efficiency of LW-MCR in RS crossmodal retrieval tasks.**

***Index Terms*—Cross-modal remote sensing text-image retrieval, knowledge distillation, contrast learning, lightweight retrieval, semantic localization.**

I. Introduction

In recent years, increasing remote sensing data has provided people with more opportunities to observe and explore the earth [1][2][3]. In order to adapt to the query of multiple data sources, RS crossmodal retrieval (RSCR) comes into being and has achieved remarkable success[4][5]. As a branch of RSCR, utilizing text to quickly and efficiently retrieve serviceable knowledge from plenty of RS images is also a meaningful and valuable research topic. However, although the existing methods of text-image retrieval can achieve satisfactory results, due to the limitations of storage space and computing resources, it is still a huge challenge to deploy these methods on small computing platforms. Automatically, designing a lightweight RS text-image retrieval model has attracted the interests of more and more scholars [6][7].

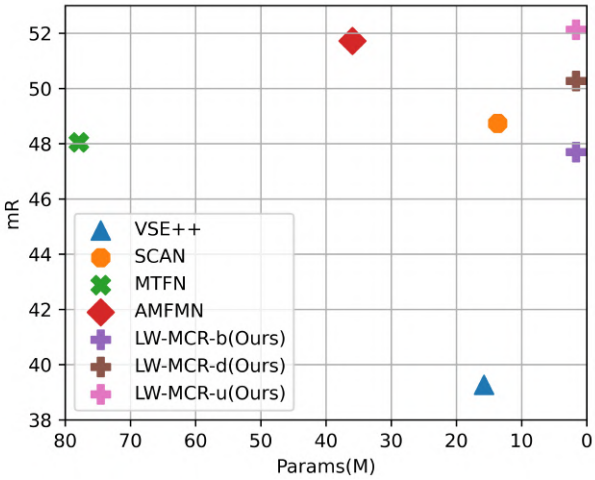


Fig. 1. Comparison of various methods on retrieval accuracy (mR) and the number of parameters (M) on the Sydney dataset. The proposed method (LW-MCR-b), which has fewer parameters than others, achieves high retrieval performance using only the traditional triplet loss for optimization. After using the hidden supervision loss to optimize the performance, LW-MCR-d reaches the top2 performance. Finally, we utilize the unlabeled data to boost the performance of the model and LW-MCR-u achieves the best performance while ensuring the number of parameters.

In general, the methods of implementing RS crossmodal text-image retrieval can be divided into two families: caption-based method and embedding-based method. The caption-based retrieval method first transforms the RS image into a sentence and then uses the input query to match the generated sentences in the retrieval stage. In order to provide data support for such tasks, researchers [8] construct a large-scale RS image text-image dataset and provide a comprehensive review of the proposed dataset. For the end-to-end generation of RS captions, authors establish a RS image caption framework using full convolution network in [9]. But the embedding-based retrieval method refers to map RS image and query sentence to the same high-dimensional space through feature extraction procedure, and follow by performing similarity calculation in the same feature space. To generate more robust features, authors build a triplet network to optimize the visual

This work was supported by the National Science Fund for Distinguished Young Scholars under Grant 67125105. *(Corresponding author: Xian Sun.)*

Z. Yuan, X. Rong, and X. Li are with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China, the Key Laboratory of Network Information System Technology (NIST), Institute of Electronics, Chinese Academy of Sciences, Beijing 100190, China, University of Chinese Academy of Sciences, Beijing 100190, China and the School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100190, China (e-mail: yuanzhiqiang19@mails.ucas.ac.cn; rongxuee19@mails.ucas.ac.cn; lixuan173@mails.ucas.ac.cn.

W. Zhang, J. Chen, H. Wang, K. Fu, and X. Sun are with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China and the Key Laboratory of Network Information System Technology (NIST), Institute of Electronics, Chinese Academy of Sciences, Beijing 100190, China (e-mail: zhangwk@aircas.ac.cn; chenjl@aircas.ac.cn; wiecas@sina.com; kunfuiecas@gmail.com; sunxian@mail.ie.ac.cn).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2021.3124252, IEEE Transactions on Geoscience and Remote Sensing

and text feature extractors in [10]. Further focus on the problem of fine-grained retrieval, researchers [11] present a fine-grained RS image-text dataset and propose an asymmetric multi-scale feature matching network to compute the distance of multimodal embedded vectors. Compared with the caption-based retrieval method, the embedding-based retrieval method calculates the similarity between the multi-modalities in a single stage thereby reducing information loss greatly. Due to the effectiveness of the latter, RS embedding-based retrieval methods has received unprecedented attention lately.

Although the current embedding-based method has achieved good retrieval accuracy, there are still some problems. The task of semantic localization requires the retrieval model to perform multiple similarity calculations between RS slices and query sentence, which results in heavy inference time consumption. In addition, semantic localization task is often realized by using small computing platforms [12], such as ARM platforms, thus bringing much pressure to the retrieval model on data transmission and processing. Consequently, lightweight retrieval models need to be devised to satisfy the requirements of inference time and model deployment [13][14]. In order to obtain lightweight model, some research directions have been explored. On the one hand, smaller and more efficient networks are proposed, such as MobileNet [15], SuqueezeNet [16], ShuffleNet [17], Xception [18]. Although these backbones can efficiently extract visual features in natural scenes, they may not be sufficient to characterize the multi-scale information of RS images. On the other hand, researchers [19][20][21] have explored the possibility of using knowledge distillation, which distills the student with the dark knowledge from the teacher model, to enhance the performance of lightweight models. It contributes to retaining the performance of large models on lightweight models. Even though researchers have made some attempts, using knowledge distillation methods to improve the performance of crossmodal retrieval still lacks of research. Within our knowledge, we are the first to introduce knowledge distillation to the field of RS crossmodal retrieval and prove that it helps improve the performance of the model.

In the past few years, unsupervised learning has attracted widespread attention because it can achieve comparable performance to supervised learning even without accessing to labeled data, which alleviates the algorithms dependence on annotations. Using unlabeled data to improve algorithm performance in RS has become a task worth studying. As an instance of unsupervised performance improvement of RS, researchers [22] attempt to utilize unsupervised-restricted deconvolution neural network to reduce the problems of overfitting and undertraining. In terms of unsupervised feature representation [23], researchers propose a probabilistic latent semantic hashing model to convert RS image into visual features. But to the best of our knowledge, there are few research focus on utilizing unlabeled samples to improve the performance of RS crossmodal retrieval methods. Based on the recent development of contrast learning [24][25], whether the method can be used to improve the performance of retrieval models by using unlabeled data is an area to be explored.

As mentioned above, even though some cross-modal retrieval models have been applied to RS, there are still many problems to be solved. **Firstly**, current RS crossmodal retrieval models suffer from the problems of high time-consuming and complex structures, which means that these models cannot be deployed to small computing platforms for semantic localization and crossmodal retrieval tasks. For this purpose, we construct a lightweight crossmodal retrieval model, which considers the multi-scale information of RS images and dynamically filters the redundant information from both channel and spatial levels. LW-MCR has achieved comparative retrieval accuracy with only about one-tenth of the parameters and one-sixth of the floating-point operations (FLOPS) of traditional retrieval algorithms.

Secondly, a lightweight crossmodal model, which may imply the lack of representing ability of the unimodal data, inevitably reduces the retrieval accuracy. In order to alleviate this problem, we set up a multi-level and multi-modal distillation loss, which can promote the lightweight model to learn the dark knowledge from the teacher network. The proposed approach aligns the multi-level feature distribution and unimodal representation of LW-MCR with the teacher network, which improves the performance of the lightweight network to the level of popular retrieval algorithms.

Thirdly, different with natural scenes, the scarcity of RS annotation makes it challenging to mine semantic information in retrieval systems. To cope with severe dependence on labeled data, a crossmodal semi-supervised optimization method is designed to enhance retrieval performance. Specifically, we first apply the contrast learning approach to generate negative sample mask, and then optimize the retrieval model parameters by analyzing the relative relationship between the labeled data and unlabeled data. The designed semi-supervised optimization algorithm further enhances the performance of LW-MCR.

The main contributions of our work are as follows:

- To reduce the occupancy and overhead of the retrieval algorithm, we come up with a novel lightweight RS crossmodal multi-scale retrieval model. LW-MCR has achieved comparative retrieval accuracy with only about one-tenth of the parameters and one-sixth of the FLOPS of traditional retrieval algorithms.
- In order to maintain the accuracy of lightweight retrieval model, a hidden supervision optimization method based on the knowledge distillation approach is proposed. The method improves the accuracy of LW-MCR to the same level as the popular retrieval model.
- We construct a semi-supervised optimization method based on contrast learning and manage to employ unlabeled RS images to further boost the retrieval precision of the lightweight network. The proposed LW-MCR model achieves better retrieval results after semi-supervised performance optimization.

We perform a large number of experiments and thoroughly analyze the sakes for the efficacy of LW-MCR and the proposed optimization method. The rest of the paper is organized as follows: Section II briefly summarizes some related works which are interwoven with our work. In Section III, LW-MCR and the proposed optimization algorithm are introduced

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2021.3124252, IEEE Transactions on Geoscience and Remote Sensing

in detail. Furthermore, Section IV provides the results and analysis of qualitative and quantitative experiments. At last, the conclusions of our work are given in Section V.

II. Related Work

This section first presents the development of crossmodal text-image retrieval in remote sensing, and then reviews the method of knowledge distillation. Finally, the evolution of contrast learning in recent years is introduced.

A. Remote Sensing Crossmodal Text-Image Retrieval

Remote sensing crossmodal text-image retrieval refers to retrieve RS images by text, which can be divided into caption-based methods and embedding-based methods. The caption-based method first converts RS images into sentences and performs matching between query and generated sentences subsequently. Considering the correlation between ground elements at different scales, a caption generation framework was proposed by Shi $et\ al.$ [9] to generate robust RS captions. In order to provide a favorable baseline, Lu $et\ al.$ [8] constructed a multi-scale RS image-text dataset and evaluated a series of caption generation models. For the problem of overfitting in RS caption generation caused by cross-entropy loss, Li $et\ al.$ [26] proposed truncated cross-entropy loss to mitigate the problem. The embedding-based method refers to embed RS images and query text into the same space and then performing the nearest neighbor search. To get robust feature representation, Abdullah $et\ al.$ [10] proposed a bidirectional triplet network and used an average fusion strategy to fuse the features of multiple sentences. To fully exploit the potential correspondence between images and text, Cheng $et\ al.$ [27] designed a semantic alignment module and used attention and gate mechanisms to filter and optimize features. Further focus on the problem of fine-grained retrieval, Yuan $et\ al.$ [11] proposed a multi-level RS retrieval framework and a fine-grained image-text dataset to advance the development of RS crossmodal retrieval tasks. Although there are many RS crossmodal retrieval models, existing methods may not satisfy the needs of the real-time retrieval and lightweight deployment due to complex structures and time-consuming calculations.

B. Knowledge Distillation

In recent years, to make the performance of lightweight networks comparable to large networks, knowledge distillation has become an emerging research field. Hinton $et\ al.$ [28] first proposed that using soft labels would enable student network to obtain better results than merely having hard labels. Next, Romero $et\ al.$ [29] mentioned that compared with using the soft labels of multi-model integration, aligning the corresponding hidden layer of the small network to the large network will achieve better results. To transfer attention from large networks to small networks, Zagoruyko $et\ al.$ [30] proposed several methods to force student network to mimic the attention maps of a powerful teacher network. Compared with the method of using the fusion model as a guide, Jin $et\ al.$ [31] demonstrated that employing some anchors in the parameter space for supervision will reduce the lower bound of the consistency loss. Different from the previous distillation method that minimizes the KL divergence, Tian $et\ al.$ [32] added contrast learning to make the student network be enable to obtain more information from the teacher network. Most current distillation methods focus on classification or detection domains but less on retrieval tasks. To the best of our knowledge, we are the first to apply knowledge distillation to lightweight retrieval in the field of remote sensing.

C. Contrast Learning

Lately, researchers have explored to apply contrast learning and fine-tuning in downstream tasks to alleviate the problems brought by the scarcity of annotations. To obtain variable negative samples and improve the learning quality, He $et\ al.$ [33] established a dynamic dictionary with a queue and a moving-averaged encoder. Further, Chen $et\ al.$ [34] greatly improved the representation quality of contrast learning by introducing a learnable non-linear transformation between feature representation and contrast loss. By integrating the above ideas, Chen $et\ al.$ [35] established a baseline that outperforms the former one without requiring a vast training batch. In order to alleviate the application limitations of the contrast learning method, Chen $et\ al.$ [36] used unlabeled data distillation to convert dark knowledge into the task-specific knowledge after unsupervised pre-training and supervised fine-tuning. Even though self-supervised learning paradigms are gaining popularity in natural scenarios, it is rare to employ such methods to enhance subtask performance in remote sensing. Using unlabeled data and contrast learning methods in RS to enhance the performance of specific tasks is an enormous challenge.

III. Method

In this paper, a lightweight crossmodal RS retrieval model is designed. We will introduce our work from five aspects: formulation, architecture, hidden supervision optimization, performance boost by unlabeled data and training algorithm. All the notations and their meanings have been summarized for better clarity and ease of understanding in Appendix A.

A. Formulation

Firstly, given an image I and a sentence S, we need to map both to a high-dimensional embedded space for similarity calculation. For image I, the weight matrix W_v is constructed to map I to the embedded feature space $\mathbb{R}$, and the process can be expressed as:

$$f_v = W_v I, f_v \in \mathbb{R}^{d_f} \tag{1}$$

where d_f represents the dimensionality of the embedded feature space, and f_v is the representation of the image I in the embedded feature space. Similarly, for the query sentence S, W_t is constructed to project S to the embedded feature space, and the process is denoted as:

$$f_t = W_t S, f_t \in \mathbb{R}^{d_f} \tag{2}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2021.3124252, IEEE Transactions on Geoscience and Remote Sensing

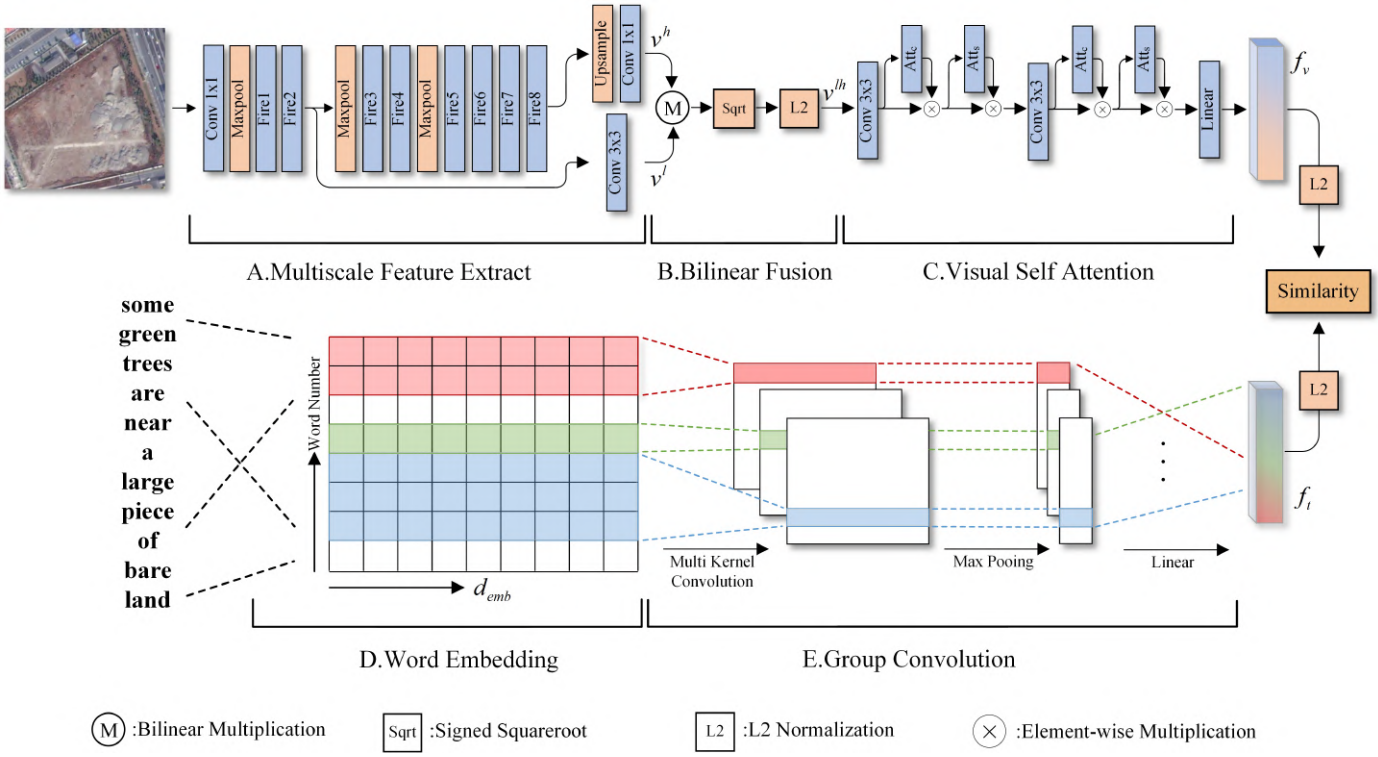


Fig. 2. Framework of lightweight multi-scale crossmodal retrieval model. The proposed method fuses multi-scale features of RS images by bilinear pooling and filters redundant information by visual self attention. After encoding text features by group convolution, the cosine distance is used to calculate the similarity between the visual and text features after $l2$ normalization.

f_t is the representation of the query sentence S in the embedded feature space. For the obtained embedded vectors, the similarity is calculated to measure their distance:

$$Sim = \mathfrak{S}(f_v, f_t) \tag{3}$$

among them, $\mathfrak{S}$ denotes calculating the similarity of the vectors x and y, which is implemented by cosine similarity in this paper. Sim is the crossmodal similarity.

On the one hand, to achieve concise and fast crossmodal similarity computation, it is necessary to construct lightweight embedded weights W_v and W_t. Meanwhile, in order to reduce the time consumption on retrieval, it is also essential to make the model less dependent on computing power. On the other hand, a pertinent embedded matrix of different modalities is still indispensable to ensure retrieval accuracy. Inspired by [11], when constructing the visual embedded weights W_v, the model should consider the multi-scale characteristic of RS images and pay attention to the salient features adaptively.

In the next part, we establish a lightweight multi-scale crossmodal retrieval model LW-MCR. Subsequently, a hidden supervised optimization method based on knowledge distillation (kd) is proposed, which dramatically improves the retrieval performance of LW-MCR. Further, we present a method of leveraging unlabeled data to boost the performance of the crossmodal retrieval system. Finally, the detailed training procedures of LW-MCR are introduced.

B. LW-MCR

Fig. 2 shows the overall architecture of the proposed lightweight multiscale crossmodal retrieval model. We will present our model in two parts: visual representation and text representation.

*1) **Visual Representation**:* To extract salient features of RS images, a multi-scale lightweight image feature extractor is devised as shown at the top of Fig. 2. The extractor comprehensively considers the multi-scale characteristics of RS and applies the visual self attention mechanism to generate a focus on the salient features after multi-scale fusion.

Firstly, to ensure that the network has competitive accuracy with fewer parameters, we choose the SqueezeNet [16] as the backbone of our network. 1×1 filters is leveraged to instead of 3×3 filters, while significantly reducing the number of 3×3 filters to some extent. The multi-scale characteristics of RS images mean that the pooling layer could filter out some small targets as the depth of the network increases. Automatically, fine-grained target information may not appear in high-level semantic information [11].

Accordingly, the visual extractor needs to extract different levels of semantic information. Precisely, we extract the output from the Fire2 layer and the Fire8 layer of SqueezeNet architecture as low- and high- level semantic features. In the mean time, considering that multi-level information needs to be embedded in the same dimension, convolution matrices of different kernel sizes are applied for feature re-transformation.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2021.3124252, IEEE Transactions on Geoscience and Remote Sensing

The above process can be expressed as follows:

$$v^l = \text{PReLU}(conv_{3\times3}(squeeze_{Fire2}(I))) \tag{4}$$

$$v^h = \text{PReLU}(conv_{1\times1}(squeeze_{Fire8}(I))) \tag{5}$$

among them, $conv(x)$ denotes the convolution transformation of the matrix x, and $\text{PReLU}(x)$ means activating matrix x using the PReLU activation function. v^l and v^h are regarded as the low- and high- level semantic features extracted from the backbone network.

Furthermore, a convolution bilinear pool with fewer parameters is utilized to fuse multi-scale information. Specifically, we use bilinear multiplication to accomplish the above operations. For the two features $v^l(p) \in \mathbb{R}^{N\times M}$ and $v^h(p) \in \mathbb{R}^{N\times M}$ of dimension $N \times M$ at position p, this process can be expressed as:

$$b(p, v^l, v^h) = (v^l(p))^T v^h(p) \in \mathbb{R}^{M\times M} \tag{6}$$

where $b(p, v^l, v^h)$ is the bilinear vector at position p. Automatically, the mean pooling function aggregates all features to obtain the average fusion result b_{ave}:

$$b_{ave} = \frac{1}{M^2}\sum_p b(p, v^l, v^h) \in \mathbb{R}^{M\times M} \tag{7}$$

Enlightened by Perronnin *et al.* [37], signed squareroot step ($y \leftarrow sign(b_{ave}\sqrt{|b_{ave}|})$ and $l2$ normalization ($v^{lh} \leftarrow y/||y||^2$) are used to constrain the bilinear fusion results, where $|x|$ represents the absolute value of x, $||x||^2$ represents the l2-norm of x, and v^{lh} is the multi-scale fusion feature following bilinear pooling.

RS images often contain numerous targets. How to filter redundant targets to obtain salient features is still a significant challenge. After fusing the multi-level information, the visual extractor needs to generate self-attention to suppress the noise in the high-dimensional embedded features. Inspired by Woo *et al.* [38], the attention mechanism is affiliated in both channel and space to make the model dynamically focus on the salient features. For the feature map $f \in \mathbb{R}^{B\times M\times M}$, where B represents the dimensionality of the channel, we decompose it to average component f_{ave} and impulse component f_{max}:

$$f_{ave} = Ave(f) \in \mathbb{R}^{M\times M} \tag{8}$$

$$f_{max} = Max(f) \in \mathbb{R}^{M\times M} \tag{9}$$

where $Ave(x)$ and $Max(x)$ denote the channel averaging and max pooling of the feature map x, respectively. Next, two components are utilized to generate the channel attention $Att_c(f)$ and spatial attention $Att_s(f)$ of the feature map f:

$$Att_c(f) = \sigma(conv_{1\times1}(f_{ave}) + conv_{1\times1}(f_{max})) \tag{10}$$

$$Att_s(f) = \sigma(conv_{1\times1}(Cat(f_{ave}, f_{max}))) \tag{11}$$

where σ is the activation function, $Cat(a, b)$ denotes the channel-wised concatenation. Further, we define the visual self attention (VSA) block as:

$$f^{'} = Att_c(f) \otimes f \tag{12}$$

$$VSA(f) = Att_s(f^{'}) \otimes f^{'} \tag{13}$$

where $\otimes$ is element-wise multiplication. Therefore, the final image embedded vector f_v represented in the high-dimensional embedded feature space $\mathbb{R}$ can be expressed as:

$$f_v = Linear(\underbrace{VSA(...VSA}_{n}(v^{lh}))) \tag{14}$$

where n is the number of VSA blocks, and $Linear(x)$ represents the linear transformation of the vector x.

*2) **Text Representation**:* When representing text, recurrent neural networks (RNN) are ordinarily the first choice of the traditional methods. Although RNN can provide a better representation of temporal information, the high computational complexity of the network makes it unsuitable for lightweight models. In addition, retrieval sentences tend to be shorter statements, which are ideal for extracting features by variable-length windows. Automatically, inspired by [39], CNN with less computation and parameters is constructed for the text representation.

For the query sentence S which contains N words $\{w_n\}_{n=1}^N$, the word embedding is performed:

$$e_n = W_e\, w_n(n \in [1, N]) \in \mathbb{R}^{d_{emb}} \tag{15}$$

where W_e is the word embedding matrix, d_{emb} is the dimension of the word embedding, and e_n is the embedded vector of n^{th} word. After acquiring the representation of each word, we concatenate them to get the sentence embedded matrix $e_{1:N}$:

$$e_{1:N} = Cat(e_1, e_2, ..., e_N) \tag{16}$$

$e_{i:j}$ denotes the embedded matrix of the window between the i^{th} word and the j^{th} word.

Moreover, for a window that contains h words, a convolution transformation with a kernel size of $h \times d_{emb}$ is applied to obtain the semantic information $c_{i,h}$:

$$c_{i,h} = conv_{h\times d_{emb}}(e_{i:i+h}) \tag{17}$$

the above operation is used for every starting point i to generate a set of feature maps. Then the max-pooling is employed to obtain the salient information $\overline{c_h}$ of the window h:

$$\overline{c_h} = Max((Cat(c_{1,h}, c_{2,h}....c_{N-h+1,h})) \tag{18}$$

Finally, for different windows h, the corresponding $\overline{c_h}$ are spliced, and a linear transformation is exploited to obtain the embedded text vector f_t:

$$f_t = Linear(Cat(\overline{c_{h_1}}, \overline{c_{h_2}}, ..., \overline{c_{h_m}})) \tag{19}$$

among them, m is the number of windows.

C. Hidden Supervision Optimization

In this subsection, we first introduce the traditional triplet loss in the retrieval field. Then, we propose the embedded kd loss and multi-scale information supervision loss to do hidden supervision optimization of LW-MCR.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2021.3124252, IEEE Transactions on Geoscience and Remote Sensing

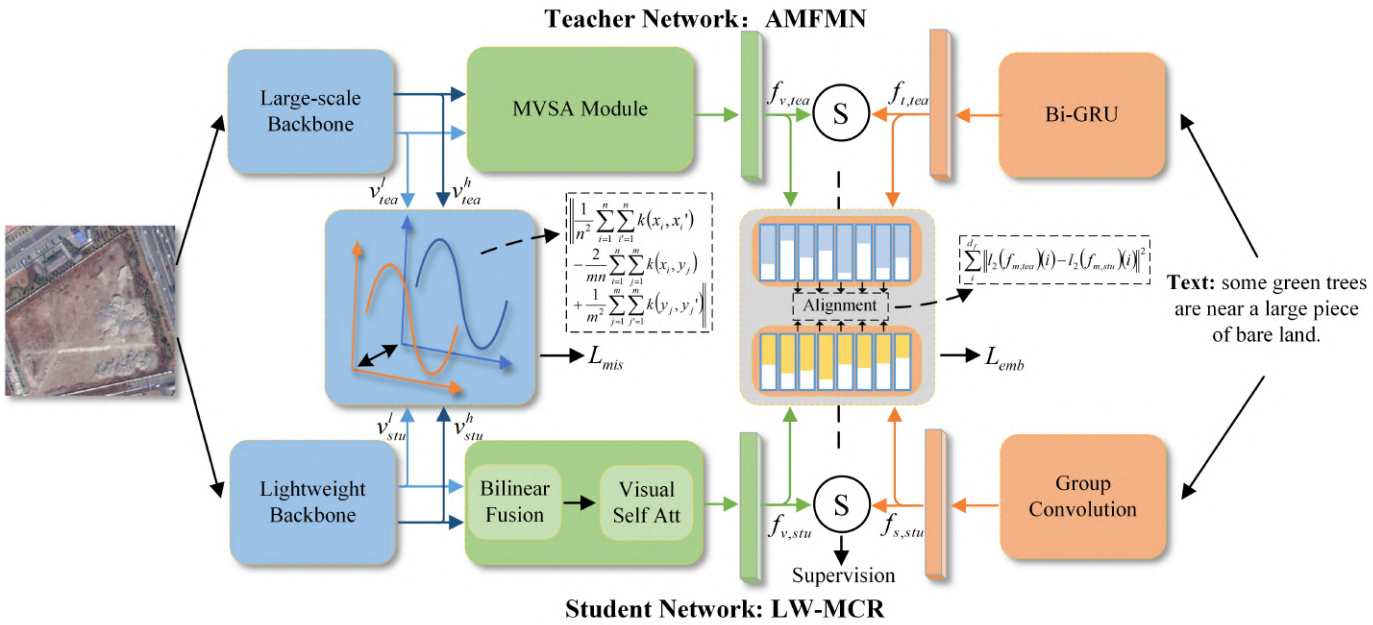


Fig. 3. The proposed hidden supervised optimization method. The method regards LW-MCR as the student network and uses AMFMN with SoTA performance as the teacher network. The approach of knowledge distillation enables LW-MCR to learn the dark knowledge from the teacher network, aiming to make the model learn the multi-level information and embedded feature representations in the large model.

1) ***Triplet Loss*** L_{tpt}*:* In multimodal feature alignment tasks, the triplet loss is usually one of the classical loss functions. The triplet function aims to shorten the distance between the intra-class samples as close as possible and make the distance between the inter-class samples as far as possible. Faghri $et\ al.$ [40] is the first to apply triplet loss to the crossmodal retrieval task:

$$L_{tpt}(I,S) = \sum\nolimits_{\widehat{S}}[\varepsilon - \mathfrak{S}(I,S) + \mathfrak{S}(I,\widehat{S})]_+ + \sum\nolimits_{\widehat{I}}[\varepsilon - \mathfrak{S}(I,S) + \mathfrak{S}(\widehat{I},S)]_+ \tag{20}$$

where ε represents the minimum margin, $[x]_+ \equiv max(x,0)$. $\mathfrak{S}(I,S)$ denotes the similarity between image I and sentence S from the same sample pair. $\mathfrak{S}(I,\widehat{S})$ and $\mathfrak{S}(\widehat{I},S)$ denote the similarity of image I and sentence S from different sample pairs. L_{tpt} considers all samples where the intra-class distance is greater than the inter-class distance, and uses the difference between the two distances as the final loss. The optimization goal is to make the multimodal features with the same semantics as close as possible in the high-dimensional feature space.

Even if the triplet loss directly optimizes the model from the retrieval accuracy, it does not constrain the representation of individual modality. However, due to the small number of network parameters, some representations may not be captured by lightweight models.

2) ***Embedding Kd Loss*** L_{emb}*:* To obtain excellent unimodal representations, we let the lightweight model (student network) learn the hidden representation of the deep model (teacher network) directly by utilizing the knowledge distillation approach. On the one hand, this approach enables dark knowledge to be transferred from the teacher network to the student network. On the other hand, it also allows the student network to imitate the teacher and optimize the network parameters to the optimum. Since AMFMN [11] is one of the superior models in the field of RS crossmodal retrieval, we regard AMFMN as the teacher network for LW-MCR.

As shown in Fig. 3, for the unimodal embedded features in the student network $f_{m,stu} \in \mathbb{R}^{d_f}$, it is reasonable to be aligned with that of teacher network $f_{m,tea} \in \mathbb{R}^{d_f}$. Such an approach allows the student network to be optimized directly towards the teacher network, thereby obtaining a more intuitive embedded expression. After performing $l2$ normalization of $f_{m,stu}$ and $f_{m,tea}$, the Sum Mean Square Error is applied to calculate the loss $L_{emb,m}$ of the two embedded features under modality m:

$$L_{emb,m}(f_{m,tea}, f_{m,stu}) = \sum_{i}^{d_f} ||l_2(f_{m,tea})(i) - l_2(f_{m,stu})(i)||^2 \tag{21}$$

where $l_2(x)$ stands for $l2$ normalization of vector x.

3) ***Multi-scale Information Supervision Loss*** L_{mis}*:* Even if the objectives of network optimization are clear, because of the difference in the structure and parameters between the two networks, the student network may not be sufficient enough to learn the dark knowledge in the teacher network. In order to transfer the multi-scale knowledge learned from the teacher network to the representation of different but related student networks, the multi-scale information supervision loss L_{mis} is designed. Firstly, channel-averaging is exploited to transform the single-scale feature map into a set of one-dimensional feature vectors. Next, we aim to find a transformation function $\phi(\cdot)$ to minimize the distance between the two feature

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2021.3124252, IEEE Transactions on Geoscience and Remote Sensing

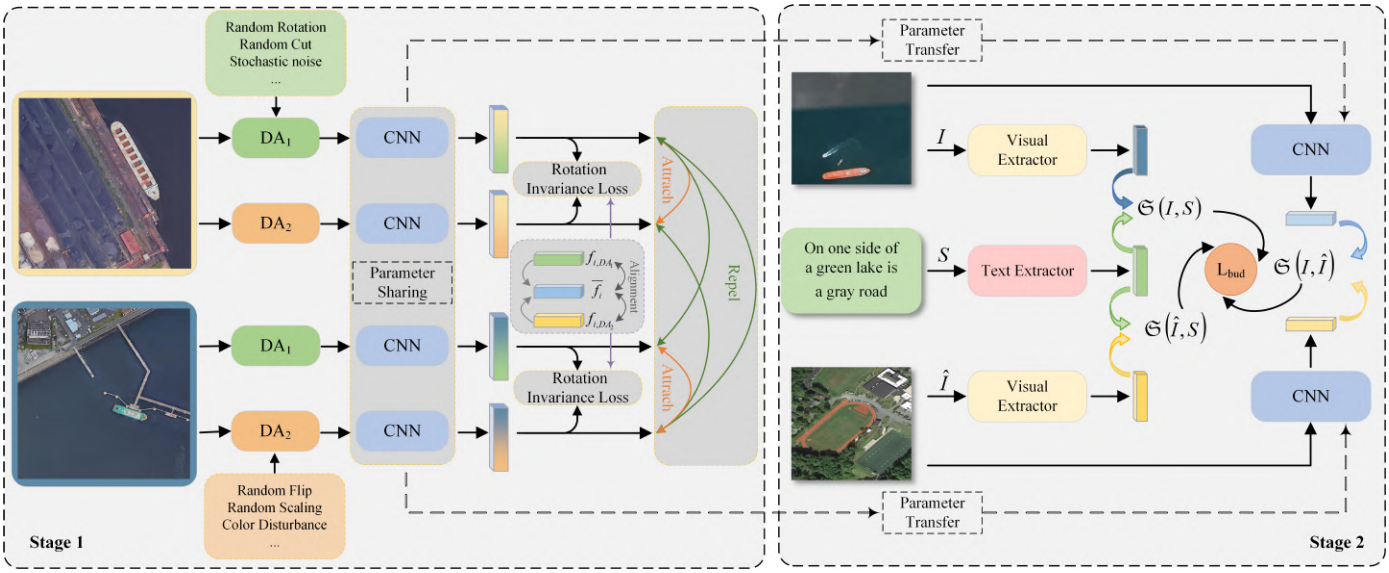


Fig. 4. Boost retrieval model performance by unlabeled data. In the first stage, we use unlabeled data to establish an image semantic similarity calculation model based on contrast learning. In the second stage, we utilize the model in the first stage to generate the negative sample mask, which adds more negative samples during training thus increasing the discrimination difficulty to the retrieval model.

distributions, and the process is defined as:

$$L_{mis}(x,y) = ||\frac{1}{n}\sum_{i=1}^{n}\phi(x_i) - \frac{1}{m}\sum_{j=1}^{m}\phi(y_j)||^2 \tag{22}$$

where n and m are the dimensions of the multi-scale feature in the teacher network and the student network, x and y are the multi-scale features of both respectively. Then, the equation (22) is expanded as:

$$\begin{aligned} L_{mis}(x,y) = ||&\frac{1}{n^2}\sum_{i=1}^{n}\sum_{i'=1}^{n}\phi(x_i)\phi(x_{i'}) \\ &- \frac{2}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}\phi(x_i)\phi(y_j) \\ &+ \frac{1}{m^2}\sum_{j=1}^{m}\sum_{j'=1}^{m}\phi(y_j)\phi(y_{j'})|| \end{aligned} \tag{23}$$

Inspired by [41], we use the kernel function $k(\cdot)$ to skip the calculation of the function $\phi(\cdot)$ directly:

$$\begin{aligned} L_{mis}(x,y) = ||&\frac{1}{n^2}\sum_{i=1}^{n}\sum_{i'=1}^{n}k(x_i,x_{i'}) \\ &- \frac{2}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}k(x_i,y_j) \\ &+ \frac{1}{m^2}\sum_{j=1}^{m}\sum_{j'=1}^{m}k(y_j,y_{j'})|| \end{aligned} \tag{24}$$

Since the Gaussian kernel function can map the original data to an infinite dimensional space, we choose $k(u,v) = e^{\frac{-||u-v||^2}{\sigma}}$ as the kernel function when calculating multi-scale information supervision loss L_{mis}.

*4) **Integration of the loss**:* After obtaining the above loss, loss integration with different weights is performed to obtain the final distillation loss L_{dis}:

$$\begin{aligned} L_{dis}(I,S) = &\alpha L_{tpt}(I,S) \\ &+ \beta(L_{emb,v}(f_{v,tea},f_{v,stu}) + L_{emb,t}(f_{t,tea},f_{t,stu})) \\ &+ \gamma(L_{mis}(V^h_{tea},V^h_{stu}) + L_{mis}(V^l_{tea},V^l_{stu})) \end{aligned} \tag{25}$$

Among them α, β, and γ are three trade-off parameters. The purpose of setting different weights is to keep the various losses at the same magnitude and avoid prominent inverse gradients for a single loss.

D. Performance Boost by Unlabeled Data

There are a vast amount of unlabeled data in RS, and utilizing these data to further improve the performance of RS retrieval models is a valuable task. Our initial idea is to add numerous negative samples during training to enhance the model's ability to judge positive samples, inevitably making the model obtains a set of reasonably robust parameters. However, since the image semantics of these unlabeled samples are not clear, using these samples that are very similar to ground truth (GT) samples may cause positive sample ambiguity [11] unavoidably. For this reason, when performing negative sample expansion, we first calculate the semantic similarity between the unlabeled sample and the labeled sample. Furthermore, a negative sample mask is designed to obtain beneficial unlabeled samples so as to improve retrieval performance.

*1) **Semantic Similarity Calculation**:* In the first stage, for RS images without any labels, the contrast learning method is used to establish the model for image semantic similarity computation as shown in the left of Fig. 4. Specifically, for the two unlabeled RS images I_1 and I_2, two different data augmentation methods $DA_1(\cdot)$ and $DA_2(\cdot)$ are utilized to argument their samples respectively. Next, the convolution

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2021.3124252, IEEE Transactions on Geoscience and Remote Sensing

transformation $\varphi(\cdot)$ with the shared parameter matrix θ_c is applied to transform the argumented samples. The above process can be expressed as:

$$f_{m,DA_n} = \varphi(DA_n(I_m); \theta_c), m \in \{1,2\}, n \in \{1,2\} \quad (26)$$

where f_{m,DA_n} represents the feature of image I_m following DA_n.

In order to obtain a visual representation network without any labels, we use self-supervised information to optimize the network parameters θ_c. On the one hand, the distance between two features of the same image after different data augmentation methods should be as small as possible, while on the other hand, the distance between the features of different images should be greater than the distance of the same image with different data augmentation methods. Thus inspired by [42], the enhanced features f_{i,DA_k} and f_{i,DA_l} for the same sample with different argumentation methods in one batch have the following loss:

$$\begin{aligned} \urcorner_{nmr}(f_{i,DA_k}, f_{i,DA_l}) = \\ \Upsilon_{k \neq l} exp(\frac{\mathfrak{S}(f_{i,DA_k}, f_{i,DA_l})}{\tau}) \end{aligned} \quad (27)$$

$$\begin{aligned} \urcorner_{dnm}(f_{i,DA_k}, f_{i,DA_l}) = \\ \sum_{i=1}^{B} \sum_{k=0}^{1} \sum_{l=0}^{1} \Upsilon_{i \neq j} exp(\frac{\mathfrak{S}(f_{i,DA_k}, f_{j,DA_l})}{\tau}) \end{aligned} \quad (28)$$

$$L(f_{i,DA_k}, f_{i,DA_l}) = -log(\urcorner_{nmr} / \urcorner_{dnm}) \quad (29)$$

where τ represents the temperature parameter, B is the sample number in one batch, and $\Upsilon_{k \neq l}$ is an indicator function assigned to 1 if $k \neq l$. $\urcorner_{nmr}$ calculates the similarity of the anchor sample with the different data argumentation methods, and $\urcorner_{dnm}$ aims to calculate the sum of the similarity of the anchor and all different samples. The optimization direction of $L_{(f_{i,DA_k}, f_{i,DA_l})}$ aims to maximize $\urcorner_{nmr}$. Further, self-supervised information loss L_{nce} is calculated as:

$$L_{nce}(f_{i,DA_k}, f_{i,DA_l}) = \frac{1}{B} \sum_{i=1}^{B} L_{(f_{i,DA_k}, f_{i,DA_l})} \quad (30)$$

Taking the rotation invariance of RS images into account and inspired by [43], rotation invariant loss is added to the self-supervised network to improve the quality of visual representation. In particular, for two features f_{i,DA_k} and f_{i,DA_l} $(k \neq l)$ of the same sample with different data argumentation methods, we first calculate the mean values $\overline{f_i}$:

$$\overline{f_i} = \frac{1}{2}(f_{i,DA_k} + f_{i,DA_l}) \quad (31)$$

Subsequently, in order to align the output feature of each sample consistent with the average value of the rotation feature, the rotation invariant loss L_{inv} is defined as:

$$L_{inv}(f_{i,DA_k}, f_{i,DA_l}) = \frac{1}{2B} \sum_{i=1}^{B} \sum_{k=0}^{1} \|f_{i,DA_k} - \overline{f_i}\|_2^2 \quad (32)$$

Ultimately, L_{nce} and L_{inv} are jointly used to optimize the model to calculate image semantic similarity.

*2) **Semi-supervised Loss Construction**:* In the second stage, we construct a negative sample mask through the image semantic similarity calculation network, and later use the filtered negative samples to boost the retrieval performance.

On the one hand, for a sample pair (I, S) and a suitable negative sample images $\widehat{I}$ in one batch, the similarity has the following relationship:

$$\mathfrak{S}(I, S) > \mathfrak{S}(\widehat{I}, S) \quad (33)$$

In order to make the positive and negative sample similarities judged by the model as far as possible, we set a distance ε between the above two similarities:

$$\mathfrak{S}(I, S) > \mathfrak{S}(\widehat{I}, S) + \varepsilon \quad (34)$$

On the other hand, subsequent to obtain the semantic similarity model, the algorithm in stage 1 is used to calculate the similarity $\mathfrak{S}(I, \widehat{I}) \in (0, 1)$ between the supervised image I and the unsupervised image $\widehat{I}$. It should be noted that during the second stage of training, the parameters of the semantic similarity calculation network are fixed, which means that it does not participate in the gradient calculation. When I and $\widehat{I}$ have a high degree of semantic similarity, the unlabeled sample is a soft positive sample. At this time, due to the positive sample ambiguity [11], it is not suitable for loss calculation as a negative sample. Considering the above reasons, the negative sample masks $Mask_{neg}$ is designed as follows:

$$Mask_{neg}(I, \widehat{I}) = (1 + tanh(\kappa(0.5 - \mathfrak{S}(I, \widehat{I}))))/2 \quad (35)$$

where κ is the edge softening coefficient. The function aims to establish a suitable negative sample selection mechanism through image similarity between unlabeled and anchor samples. This mechanism controls the proportion that each unlabeled image with different semantics contributes to the total loss, which in turn enables adaptive selection of negative sample images.

Therefore, the semi-supervised loss L_{bud} is proposed automatically as follows:

$$\begin{aligned} L_{bud}(I, S, \widehat{I}) = \\ [Mask_{neg}(I, \widehat{I})(\mathfrak{S}(\widehat{I}, S) + \varepsilon - \mathfrak{S}(I, S))]_+ \end{aligned} \quad (36)$$

The loss is implemented in every batch. L_{bud} aims to enable the model to have a well understanding of crossmodal similarity comparison, so as to obtain a set of robust model parameters.

E. Training Algorithm

The training procedure of the proposed LW-MCR is shown in Algorithm 1.

A labeled RS image-text dataset $\mathbf{D_l}$ and an unlabeled RS image dataset $\mathbf{D_u}$ are needed. We use pre-trained SqueezeNet to initialize the visual representation branch of the LW-MCR. Parameters α, β, γ, and λ need to be taken to balance the gradient of each loss to the optimization.

The auxiliary model is utilized to improve the retrieval accuracy of the proposed model. $\mathbf{D_l}$ is leveraged to train a

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2021.3124252, IEEE Transactions on Geoscience and Remote Sensing

Algorithm 1 Training Procedure of the Proposed LW-MCR.

Require:

Labeled dataset $\mathbf{D_l} = \{\{x_1, y_1\}, \{x_2, y_2\}, ...\}$ (x is labeled image, y is corresponding sentence),

Unlabeled dataset $\mathbf{D_u} = \{z_1, z_2, ...\}$ (z is unlabeled image),

LW-MCR model $f_{v,stu}, f_{t,stu}, v^l_{stu}, v^h_{stu} = m_1(x, y; \theta_1)$ with SqueezeNet pre-training paramters,

Parameter α, β, γ, λ

Auxiliary Model Construction:

Train the AMFMN model $f_{v,tea}, f_{t,tea}, v^l_{tea}, v^h_{tea} = m_2(x, y; \theta_2)$ as teacher network,

Utilize contrast learning method train the image semantic similarity calculation network $s_z = \varphi(z; \theta_c)$

Repeat until convergence:

for each batch $\mathbf{X}, \mathbf{Y} \in \mathbf{D_l}$ **do**

Calculate the triplet loss

$F_{v,stu}, F_{t,stu}, V^l_{stu}, V^h_{stu} = m_1(\mathbf{X}, \mathbf{Y}; \theta_1)$

$l_{tpt} = L_{tpt}(F_{v,stu}, F_{t,stu})$

Calculate the embedded kd loss

$F_{v,tea}, F_{t,tea}, V^l_{tea}, V^h_{tea} = m_2(\mathbf{X}, \mathbf{Y}; \theta_2)$

$l_{emb} = L_{emb,v}(F_{v,tea}, F_{v,stu}) + L_{emb,t}(F_{t,tea}, F_{t,stu})$

$l_{mis} = L_{mis}(V^h_{tea}, V^h_{stu}) + L_{mis}(V^l_{tea}, V^l_{stu})$

Calculate the semi-supervised loss

Random fetch one batch $\mathbf{Z} \in \mathbf{D_u}$

$F_{v,uld} = m_1(\mathbf{Z}; \theta_1)$

$l_{bud} = L_{bud}(F_{v,stu}, F_{t,stu}, F_{v,uld})$

Calculate the total loss

$l_{total} = \alpha l_{tpt} + \beta l_{emb} + \gamma l_{mis} + \lambda l_{bud}$

Update θ_1 **by** l_{total}

end for

return $m_1(x, y; \theta_1)$

teacher network AMFMN to provide dark knowledge for LW-MCR. Hereafter, in order to boost performance by unlabeled data, we leverage $\mathbf{D_u}$ to train an image semantic similarity calculation network $\varphi(\cdot)$ to generate negative sample masks.

During training stage, for a batch of data $\mathbf{X}, \mathbf{Y} \in \mathbf{D_l}$, we calculate l_{tpt}, l_{emb}, l_{mis}, and l_{bud} respectively. l_{tpt} is calculated using only the visual features $F_{v,stu}$ and text features $F_{t,stu}$ which are extracted by LW-MCR. For embedded kd loss, we first align the visual and text features extracted by AMFMN with $F_{v,stu}$ and $F_{t,stu}$ to obtain l_{emb}. Furthermore, the high-level features V^h and low-level features V^l of AMFMN and LW-MCR are aligned to get the loss l_{mis}. In order to use unlabeled data to enhance performance, we randomly fetch one batch of samples $\mathbf{Z}$ from $\mathbf{D_u}$. The image semantic similarity calculation network is then utilized to calculate the similarity between $\mathbf{Z}$ and $\mathbf{X}$, and a negative sample mask is generated by this similarity. Mask is used to filter for beneficial negative samples and combine $F_{v,stu}$, $F_{t,stu}$, and $F_{v,uld}$ to obtain l_{bud}. Finally, the input parameters α, β, γ, and λ are leveraged to weigh the above losses to obtain the total loss l_{total} and then update the parameters θ_1 of LW-MCR.

IV. Experiments Results And Analysis

To analyze the performance of the LW-MCR method, we conduct an abundant of experiments on several remote sensing image-text datasets. This section shows and analyzes the experiment results, which verify the efficacy of the proposed method.

A. Dataset and Evaluation Metrics

Experiments are carried out on four RS image-text datasets: Sydney, UCM, RSITMD, and RSICD. As shown in Fig. 5, for each sample pair in these datasets, a RS image and five corresponding sentences are included. The Sydney and UCM datasets are the earliest RS image-text datasets proposed by [44], and these two datasets contain 555 and 2100 sample pairs, respectively. These two smaller datasets extensively test the generalization performance of the retrieval model. The RSITMD dataset [11] is a more fine-grained image-text dataset proposed by Yuan *et al.*, which contains a total of 4743 sample pairs from 24 categories. This dataset will be more challenging for fine-grained retrieval tasks. The RSICD dataset [8] proposed by Lu *et al.* is a large-scale multi-scale RS image dataset containing a total of 10,921 sample pairs. Since RSICD has a large sample size, the retrieval accuracy is more dependent on the parameter size of the model. We regard 80% of the samples from each dataset as the training set, 10% as the validation set, and the remaining 10% as the test set. For performance boost by unlabeled data, 29,251 unlabeled images of 500x500 acquired from Google Earth are used.

Cap1: There is a baseball field beside the green amusement park around the red track.

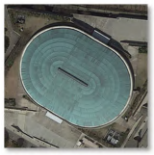
Cap2: A green baseball field adjacent to the playground and red square.

Cap3: There is a long path in the field next to the red playground.

Cap4: The green playground around the red runway is a baseball field.

Cap5: The green baseball field is adjacent to the playground and the red playground.

(a) RSITMD Sample



Cap1: a building with light blue roof in the middle .

Cap2: a oval building in the middle while with some intensive plants in side .

Cap3: a building in the middle while with light gray ground around .

Cap4: a oval building with light blue roof and white edge in the middle .

Cap5: an almost circle building is next to roads .

(b) RSICD Sample

Fig. 5. For each sample in remote sensing image-text datasets, an RS image and five corresponding sentences are included. Samples from RSITMD and RSICD are shown in this figure respectively.

Two evaluation criteria mR [45] and $R@K(K = 1, 5, 10)$, are used to evaluate the various methods. $R@K$ represents the proportion of ground truth in the top$K$ results. mR represents the average of each R@K result, and this indicator can evaluate the overall performance of the model more reasonably.

B. Implementation Details

All experiments are performed on a single NVIDIA GTX 2080ti GPU. The input RS images are cropped to 256×256 following a series of data augmentation methods. The word

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2021.3124252, IEEE Transactions on Geoscience and Remote Sensing

TABLE I
COMPARISONS OF CROSS-MODAL RETRIEVAL RESULTS ON FOUR RS TEXT-IMAGE TEST SET.

Approach	RSICD dataset							RSITMD dataset						
	Sentence Retrieval			Image Retrieval			mR	Sentence Retrieval			Image Retrieval			mR
	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
VSE++	3.38	9.51	17.46	2.82	11.32	18.10	10.43	10.38	27.65	39.60	7.79	24.87	38.67	24.83
SCAN	5.85	12.89	19.84	3.71	16.40	26.73	14.23	**11.06**	25.88	39.38	9.82	29.38	42.12	26.28
MTFN	5.02	12.52	19.74	**4.90**	17.17	29.49	14.81	10.40	27.65	36.28	9.96	31.37	45.84	26.92
AMFMN	**5.39**	**15.08**	**23.40**	**4.90**	18.28	31.44	**16.42**	10.63	24.78	**41.81**	**11.51**	**34.69**	**54.87**	**29.72**
LW-MCR-b (ours)	4.57	13.71	20.11	4.02	16.47	28.23	14.52	9.07	22.79	38.05	6.11	27.74	49.56	25.55
LW-MCR-d (ours)	3.29	12.52	19.93	4.66	17.51	30.02	14.66	10.18	**28.98**	39.82	7.79	30.18	49.78	27.79
LW-MCR-u (ours)	4.39	13.35	20.29	4.30	**18.85**	**32.34**	15.59	9.73	26.77	37.61	9.25	34.07	54.03	28.58
	UCM dataset							Sydney dataset						
Approach	Sentence Retrieval			Image Retrieval			mR	Sentence Retrieval			Image Retrieval			mR
	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
VSE++	12.38	44.76	65.71	10.10	31.80	56.85	36.93	24.14	53.45	67.24	6.21	33.56	51.03	39.27
SCAN	12.85	47.14	**69.52**	12.48	46.86	71.71	43.43	20.69	55.17	67.24	**15.52**	57.59	76.21	48.74
MTFN	10.47	47.62	64.29	**14.19**	52.38	78.95	44.65	20.69	51.72	68.97	13.79	55.51	77.59	48.05
AMFMN	**16.67**	45.71	68.57	12.86	**53.24**	79.43	**46.08**	**29.31**	58.62	67.24	13.45	**60.00**	**81.72**	51.72
LW-MCR-b (ours)	12.38	43.81	59.52	12.00	46.38	72.48	41.10	17.24	48.28	72.41	14.13	56.90	77.24	47.70
LW-MCR-d (ours)	15.24	**51.90**	62.86	11.90	50.95	75.24	44.68	18.97	58.63	75.86	13.45	57.59	78.97	50.57
LW-MCR-u (ours)	18.10	47.14	63.81	13.14	50.38	**79.52**	45.35	20.69	**60.34**	**77.59**	**15.52**	58.28	80.34	**52.13**

TABLE II
COMPARISON OF PARAMETERS AND FLOPS OF VARIOUS APPROACHES ON THE SYDNEY DATASET

Approach	VSE++	SCAN	MTFN	AMFMN	LW-MCR (ours)
Pramaters(M)	15.78	13.68	77.90	35.94	**1.65**
Flops(G)	2.44	2.42	2.80	2.75	**0.46**

embedded dimension is set to 300, and the embedded dimension of unimodal is 512. The visual feature extraction network of LW-MCR uses 2 VSA blocks, and the text encoding network uses windows of sizes 3, 4, and 5, respectively. The minimum margin of triplet loss is set to 0.25. The multi-scale loss weights γ and embedded kd loss weights β are restricted to 0.02 and 0.7 for the knowledge distillation step, the triplet loss weight α is set to 1. In the semi-supervised loss construction, λ is set to 0.003, the distance ε is set to 0.2, and the edge softening coefficient κ is set to 20. ResNet-18 [48], which has no final classification layer, is regarded as image semantic similarity calculation network φ to extract the 512-dimensional features of the RS image. The Adam optimizer is used to train the network with 120 epochs, and the batchsize is set to 70. During training, the learning rate is initialized to 1e-4 and decreases by 0.7 after every 20 epochs.

C. Comparisons with the SoTA Methods

We have compared the proposed LW-MCR with the following methods on four RS image-text datasets.

- **VSE++ [40]:** VSE++ directly embeds image and text through a deep learning framework, and then exploits cosine similarity to calculate the distance between them.
- **SCAN [46]:** Contrast to VSE++, the SCAN method attempts to align the image and text features at the instance level and thus to judge the similarity of the two modalities.
- **MTFN [47]:** MTFN utilizes a multi-modal fusion method to predict the similarity of crossmodal features end-to-end.
- **AMFMN [11]:** AMFMN takes advantage of the MVSA module to extract salient features from RS images, and uses the features to guide the text representation to obtain higher retrieval accuracy.

In the above experiments, following the configuration of the literature [11], ResNet-18 is utilized as the backbone of each network. We list three LW-MCR methods to compare with the above models:

- **LW-MCR-b:** The LW-MCR method only performs triplet loss optimization.
- **LW-MCR-d:** Optimized LW-MCR-b method by using AMFMN as a teacher network.
- **LW-MCR-u:** Quadratic optimization of LW-MCR-d by using semi-supervised loss L_{bud}.

We run several experiments to record the mean values of these experiment results, and the test results of the various methods on the four datasets are shown in Table I.

- On the RSICD dataset, the LW-MCR-b method achieves similar retrieval accuracy comparable to other large models while ensuring the smaller number of parameters. Following optimization using the knowledge distillation method, the retrieval accuracy of the model is slightly improved. We employ unlabeled data to enhance its performance and obtain an accuracy that exceeds other methods except for AMFMN. LW-MCR-u gets the first

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2021.3124252, IEEE Transactions on Geoscience and Remote Sensing

TABLE III
RETRIEVAL RESULTS OF LW-MCR WITH VARIOUS CONFIGURATIONS ON RSITMD TEST SET.

Ablation Model	Triplet Loss	KD Loss		Semi-supervised Loss	Sentence Retrieval			Image Retrieval			mR
	L_{tpt}	L_{emb}	L_{mis}	L_{bud}	R@1	R@5	R@10	R@1	R@5	R@10	
s1	✓				5.08	20.13	31.64	5.97	27.52	44.25	22.43
s2	✓				5.97	22.34	32.96	6.59	27.03	43.50	23.07
s3	✓				7.96	22.78	33.41	6.59	27.83	48.10	24.45
m1	✓				9.07	22.79	38.05	6.11	27.74	49.56	25.55
m2		✓			8.41	24.34	34.96	5.62	27.70	48.05	24.85
m3	✓	✓			8.63	26.33	36.95	7.88	30.40	49.56	26.62
m4	✓		✓		**10.40**	28.76	**40.27**	7.65	26.37	42.65	26.02
m5	✓	✓	✓		10.18	**28.98**	39.82	7.79	30.18	49.78	27.79
m6	✓	✓	✓	✓	9.73	26.77	37.61	**9.25**	**34.07**	**54.03**	**28.58**

place in the $R@5$ and $R@10$ metrics of the image retrieval, which verifies the performance of the proposed method.

- The RSITMD dataset is a relatively fine-grained dataset, which validates the model's ability to distinguish detailed information. On this dataset, the performance of the LW-MCR-b method with fewer parameters has outperformed VSE++ with an mR metric of 25.55. After optimizing the model using the hidden supervised optimization strategy, the mR indicator of LW-MCR-d reaches 27.79, which verifies the effectiveness of the proposed knowledge distillation method in retrieval tasks. Further, we use L_{bud} to enhance the performance of the student network and the mR indicator rises again by 0.79 points, reaching the top2 compared to other large models.
- On the UCM dataset, the performance gain which brings by the hidden supervised optimization is particularly prominent. Compared with LW-MCR-b, the mR of LW-MCR-d is increased by 3.58 points, which proves the importance of dark knowledge to improve retrieval performance. Subsequent to utilize unlabeled data to enhance its performance, the mR indicator of LW-MCR-u reaches 45.35. Simultaneously, the $R10$ indicator of LW-MCR-u on image retrieval reaches the best with 79.52.
- The Sydney dataset has s small size and it could better validate the robustness of the models. On this dataset, the performance of LW-MCR-b is basically the same as that of large model retrieval indicators except for AMFMN. After knowledge distillation, the mR indicator of LW-MCR-d reaches top2 with 50.57. Due to the small size of the Sydney dataset, the unlabeled data boost method shows superiority at this time. With the addition of L_{bud}, the mR indicator of the model improved by 1.56, which strongly validate the importance of unlabeled data for enhancing the performance of the retrieval system.

At the same time, the proposed method has significant advantages in the number of parameters and FLOPS. As shown in Table II, LW-MCR has only 1.65 million parameters, which is about one-tenth of VSE++ and one-48th of MTFN method. Compared with other methods, the proposed model immensely reduces the number of parameters and perfectly meets the deployment requirements of the lightweight platform. In terms of FLOPS, LW-MCR is far ahead with 0.46G FLOPS compared with other models with about 2.4G FLOPS. This makes it possible to realize more efficient and faster RS image retrieval when performing the task of semantic localization mentioned in [11]. The comparison with other models in terms of number of parameters and FLOPS greatly verifies the feasibility of the LW-MCR method in lightweight deployment and small platform applications.

D. Ablation Studies

In this subsection, a detailed ablation experiment is performed to analyze the LW-MCR method systematically. In order to study the effectiveness of the proposed method in alleviating the multi-scale and target redundancy of RS images, we first conduct ablation experiments on the model structure. At the same time, in order to analyze the impact of each optimization method on the results, we conduct the ablation experiment for the loss function.

We list a series of configurations as shown in Table III. s(1-3) are the comparative experiment of model structure, and L_{tpt} is used to optimize the model. m(1-6) are the ablation experiment of optimization method, which aims to verify the effectiveness of various optimization methods.

- **s1(L_{tpt}):** only use SqueezeNet to extract RS image features;
- **s2(L_{tpt}):** only use SqueezeNet and multi-scale fusion module to extract RS image features;
- **s3(L_{tpt}):** only use SqueezeNet and visual self-attention module to extract RS image features;
- **m1(L_{tpt}):** only the traditional triplet loss function is utilized for optimization;
- **m2(L_{emb}):** only embedded kd loss is used to optimize the feature layer;
- **m3($L_{tpt}+L_{emb}$):** joint optimization using triplet loss and embedded kd loss;
- **m4($L_{tpt}+L_{mis}$):** joint optimization using triplet loss and multi-scale supervision loss;
- **m5($L_{tpt}+L_{emb}+L_{mis}$):** joint optimization using triplet loss and loss of knowledge distillation;
- **m6($L_{tpt}+L_{emb}+L_{mis}+L_{bud}$):** unlabeled data is applied to enhance the performance of the m5 model.

In these models, s(1-3) verify the performance improvement of each module in the visual representation. m(1-2) provide

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2021.3124252, IEEE Transactions on Geoscience and Remote Sensing

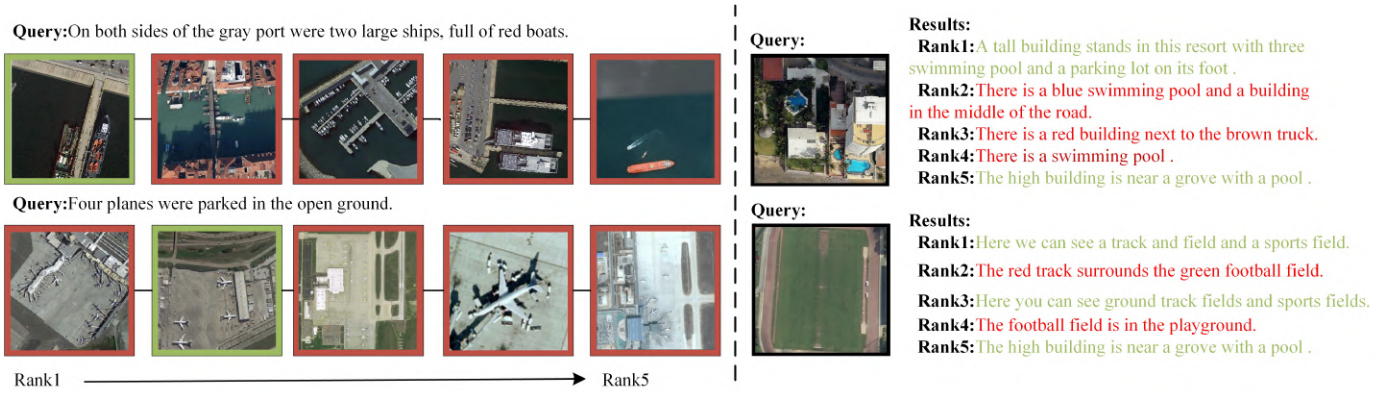


Fig. 6. Visualization of retrieval results. The left of the figure shows two results of image retrieval, where rank1 to rank5 are shown from left to right. The right of the figure shows the two results of sentence retrieval. The green result in the figure is the ground truth.

the results achieved by a single optimization method. m(3-5) aim to analyze the impact of hidden supervised optimization on retrieval performance. m6 explore the impact of unlabeled data on boosting model performance.

Table III shows the experiment results of various variants of LW-MCR.

- Compared with s1, the s2 model with multi-scale fusion achieves higher retrieval accuracy, and the $R@1$ of both sentence retrieval and image retrieval has been improved.
- Similarly, compared with s1, the s3 model has added the visual self-attention module, and the mR indicator has been increased to 24.45.
- Compared with s(1-3), the m1 model combines multi-scale fusion and visual self attention modules, and the mR indicator is improved by 1.10%.
- Compared with m1 and m2, it can be found that only using embedded supervision and using ground truth optimization can achieve almost the same results, which undoubtedly verifies the effectiveness of L_{emb} loss.
- The m3 model integrates the L_{tpt} and L_{emb} optimization methods and obtains better performance relative to both methods, with an mR metric of 26.62.
- In addition, we attempt to use multi-scale information supervised loss L_{mis} to assist the L_{tpt} during training, and the performance is slightly improved compared with m1.
- Then, both triplet loss and multiple hidden supervised losses are used to optimize the model, which obtains better performance. In this optimization method, the model has achieved retrieval performance similar to other large models on RSITMD dataset, with the mR indicator of 27.79.
- In the end, unlabeled data is used to boost the performance of the retrieval model further. Compared with m5, the m6 model improve 0.79 in the mR indicator.

E. Analysis of Time Consumption

When performing the semantic localization task [11] in large scenes, since a large number of crossmodal sample similarities need to be calculated, the inference time of the model is particularly important. In this part, the proposed LW-MCR is compared with other models in terms of retrieval time. The hardware used in this experiment is an Intel Xeon(R) Gold 6226 CPU @ 2.70GHz and a single NVIDIA GeForce GTX 2080ti graphics card. The time consumption of the model is contrasted in two indicators: inference time and evaluation time. In this experiment, the inference time refers to the time for one complete calculation of the crossmodal similarity. Evaluation time refers to the time spent on calculating the text-image similarity in the test set. In order to compare each model reasonably, we run the experiment several times with no other load on the device and record its average time.

TABLE IV
TIME-CONSUMING COMPARISON OF DIFFERENT METHODS.

Approach	Evalution Time(s) RSICD	RSITMD	UCM	Sydney	Inference Time(ms)
VSE++	23.46	4.53	1.71	0.71	3.36
SCAN	66.23	11.94	4.69	1.56	5.02
MTFN	47.36	8.71	3.00	1.31	4.76
AMFMN	40.39	8.89	2.91	1.23	4.69
LW-MCR	**18.46**	**4.43**	**1.64**	**0.68**	**2.93**

Table IV shows the results of five models using the above twon evaluation indicators on different test sets. In terms of inference time, LW-MCR achieves first place with an inference speed of 2.93ms. While in evaluation time, LW-MCR has achieved a leading position on different test sets. The comparison of the retrieval times powerfully illustrates the superiority of LW-MCR in performing multimodal retrieval.

F. Remote Sensing Image Recall Analysis

Some qualitative results are as shown in Fig. 6. The left of Fig. 6 shows two examples of image retrieval, and the right shows two examples of text retrieval.

When using text to retrieve images of the two ships next to the harbor, LW-MCR successfully retrieves the ground truth (GT) images. Although all the retrieved images are related to the GT, the rank1 image matches the best with the query. When retrieving the two aircrafts in the open field, the GT result is incorrect to rank2. Since the rank1 image and the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2021.3124252, IEEE Transactions on Geoscience and Remote Sensing

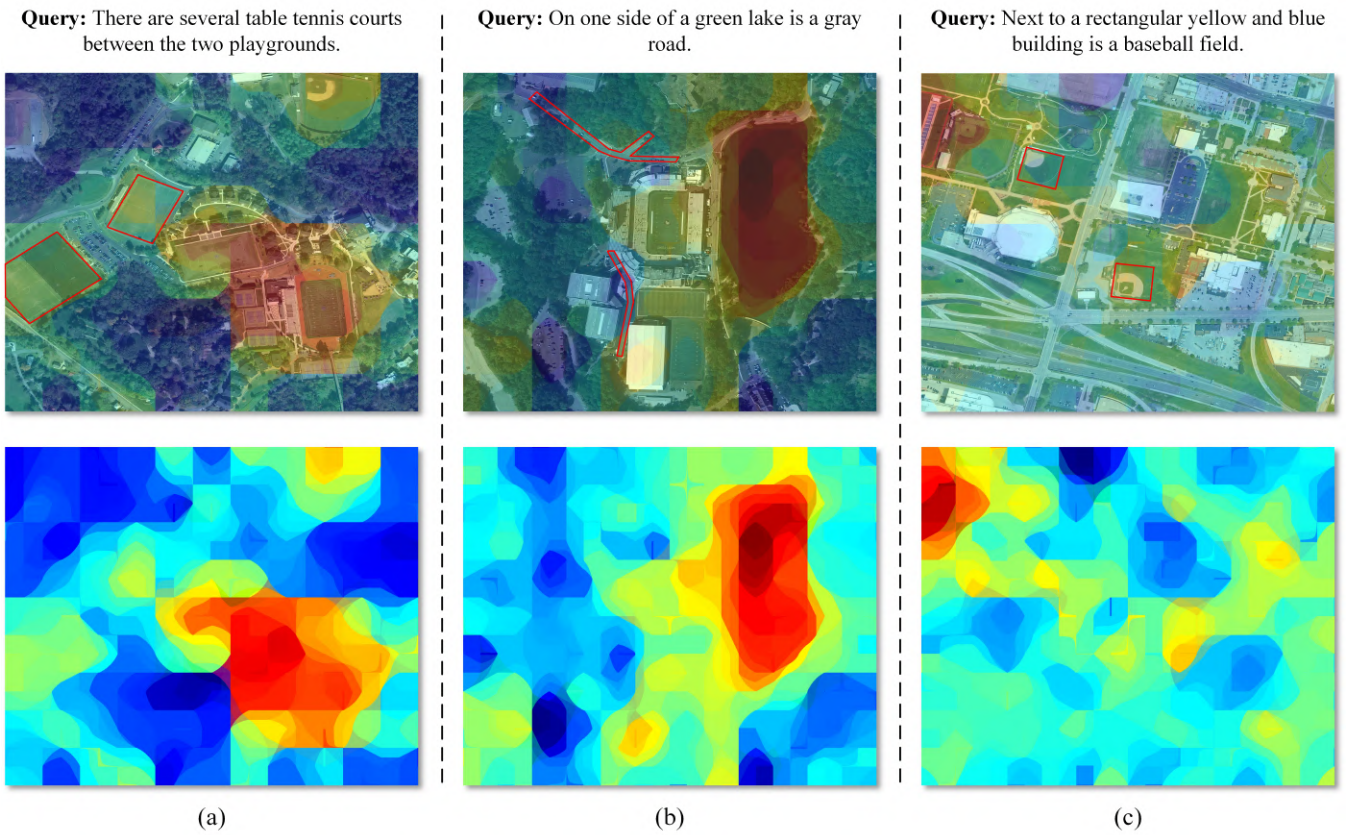


Fig. 7. Visualization of semantic localization results. (a)-(c) respectively show three semantic localization results. Each result from top to bottom is the query text, the located image, and the corresponding probability map. The red boxes are used to draw the objects that are present in the query sentence but not in the GT area.

scene of the GT image are incredibly similar, the model cannot obtain a more fine-grained retrieval semantics, thus causing the confusion of the model.

Two samples when using images to retrieve text are shown in the right of Fig. 6. When sentence retrieval is performed on images with swimming pools, almost all the retrieved sentences have mentioned "pool", which indicates that the model has successfully perceived the swimming pool target in the RS images. Even if the retrieved rank2-5 sentences are not GT results, they contain the same semantics. When retrieving on images containing playgrounds, the top5 sentences contain 3 GT results and the wrong sentences can also describe the images.

The above experiments qualitatively analyze the effectiveness of the LW-MCR model in performing sentence and image retrieval, and show that the proposed model can perceive important objects in images and sentences. However, due to the problem of fine-grained scarcity, it will inevitably retrieve some error samples which contain the similar semantics as the GT.

G. Examples of Semantic Localization

In this subsection, we verify the feasibility of the LW-MCR method on the semantic localization task. The semantic localization task refers to the model locates the region that best matches the text in a large scene. We clean up the implementation of semantic localization and made it open to access[1]. Following Yuan *et al.* [11], after cutting the large scene image with a multi-scale sliding window, the probability distribution between the text and each slice is calculated. Then the obtained probability distributions are merged and the median filter is used to remove the impact noise in the probability map. Three examples of semantic localization are shown in Fig. 7. And more examples of semantic localization have been presented in Appendix B. The upper, middle, and lower areas of each example respectively display the query text, the located image, and the corresponding probability map. We have used the red box to draw the objects that are present in the query sentence but not in the GT area.

In Fig.7(a), we attempt to locate several table tennis courts in the middle of two playgrounds. It can be seen that in the probability map, the GT region has gained attention. Although the playground is still present in some other areas as shown in the red box in the Fig. 7(a), the model still does not place attention on it. This example shows that the model can already understand the semantic information in the query sentence and can analyze the spatial relationship among objects in RS image.

In the second sample, the input query is "On one side of a green lake is a gray road". The area of the lake is almost completely localized in the obtained probability map.

[1]https://github.com/xiaoyuan1996/retrievalSystem

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2021.3124252, IEEE Transactions on Geoscience and Remote Sensing

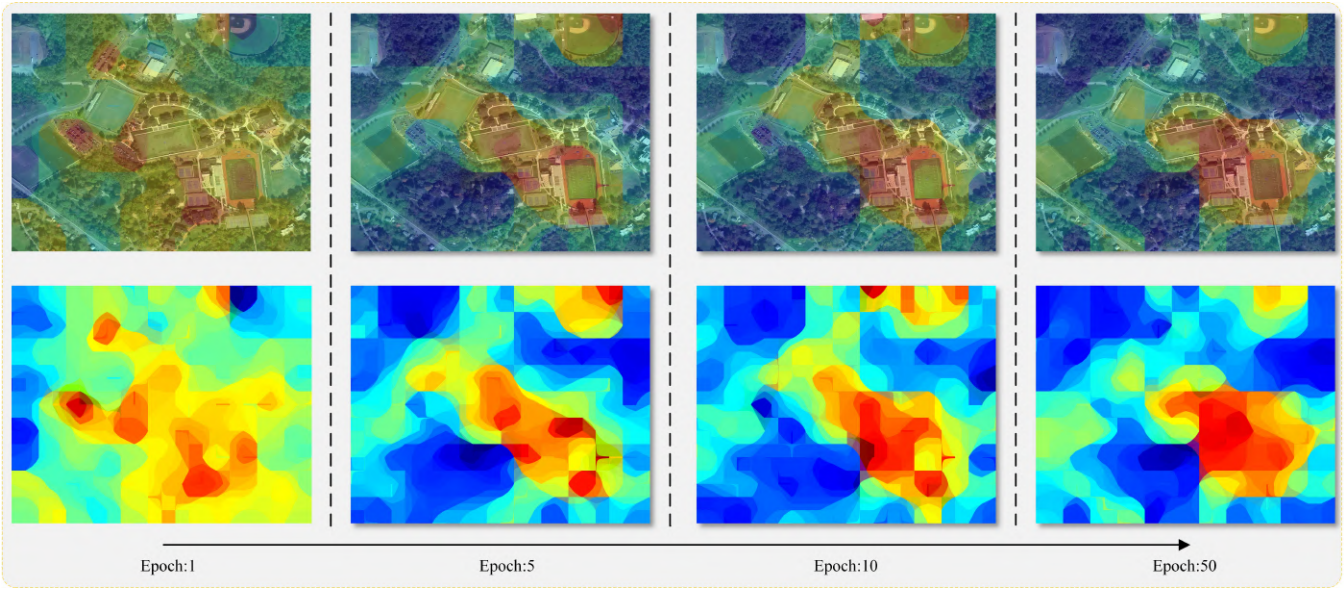


Fig. 8. The visualization of RoI varies with the number of epochs. The figure shows how the RoI changes when the number of epochs increases from 1 to 50. The top of each sample is the located image, and the bottom is the corresponding probability map. Query: "There are several table tennis courts between the two playgrounds."

Although the object "road" in the sentence is still present in this image, attention is not focused on these positions. Although we only teach the model whether the lake exists during training, the proposed model is still able to locate the area of the lake. This sample demonstrates the feasibility of crossmodal retrieval in performing unsupervised segmentation tasks.

In Fig. 7(c), we locate a more detailed scene and try to retrieve the baseball field next to a specific building. As with the previous two examples, it can be seen that the object "baseball field" in the query is present elsewhere in the image, but attention is only present in the upper left corner of the GT area. This example strongly illustrates that the model provides a good understanding of the query and is able to understand the semantic relationships in the image.

H. Analysis of RoI Change

In this subsection, we analyze the impact caused by epochs on probability map generation. The probability maps with epochs are shown in Fig. 8 for the same query sentences and located images. The query is "There are several table tennis courts between the two playgrounds.", and the number of epochs is 1, 5, 10, and 50, respectively.

In epoch 1, there are abundant errors in the region of interest (RoI). Even though the model can focus on table tennis, a great deal of attention has still been given to the other area. In epoch 5, the RoI gradually shifts to the playground and the table tennis field. Nevertheless, the RoI is still not block-shaped, which is single-level target features at this time. In epoch 10, while the model can already pay attention to the table tennis field, some attention is still incorrectly focused on the baseball field at the top of the image. After training with 50 epochs, the model has corrected the error and put almost all the RoI on the described area. This experiment demonstrates the effectiveness of the proposed optimization mechanism through the change of RoI with epochs.

I. Qualitative Comparison of Retrieval Performance on Different Datasets

In [11], the author proposes a fine-grained RS image-text dataset for carrying out crossmodal RS retrieval tasks. Compared with the traditional Sydney, UCM, and RSICD, the dataset is more variable in sentences and contains more categories and words. However, it is a pity that the author does not compare the performance of the models which are trained using these datasets. On this basis, we utilize these datasets to train LW-MCR respectively, and carry out semantic localization to investigate the performance of these datasets in RS retrieval tasks. Fig. 9 shows two comparative examples. In each set of examples, we use the same query sentence and the localized image, but employ LW-MCR trained on different datasets. The upper part of each sample is the localized image, and the lower part is the probability map.

In Fig. 9(a), the query is "There are several table tennis courts between the two playgrounds.". When using the Sydney dataset, we can see that the RoI is incorrectly focused on trees. Although a small part of the attention is on the table tennis court, most of the RoI is in the wrong area. Even if the table tennis court receives most of the attention when using the UCM dataset, a portion of the RoI falls in other areas. At this time, there is attention falling on the playground on the left, which indicates that the model still does not fully understand the semantic information of the sentence. When using the RSITMD dataset, it can be seen that the localization effect is better. Almost all of the RoI focus on the described GT area at this time. When using the RSICD dataset, while essentially locating the area described, a portion of the RoI still shifts to the upper right side of the baseball field.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2021.3124252, IEEE Transactions on Geoscience and Remote Sensing

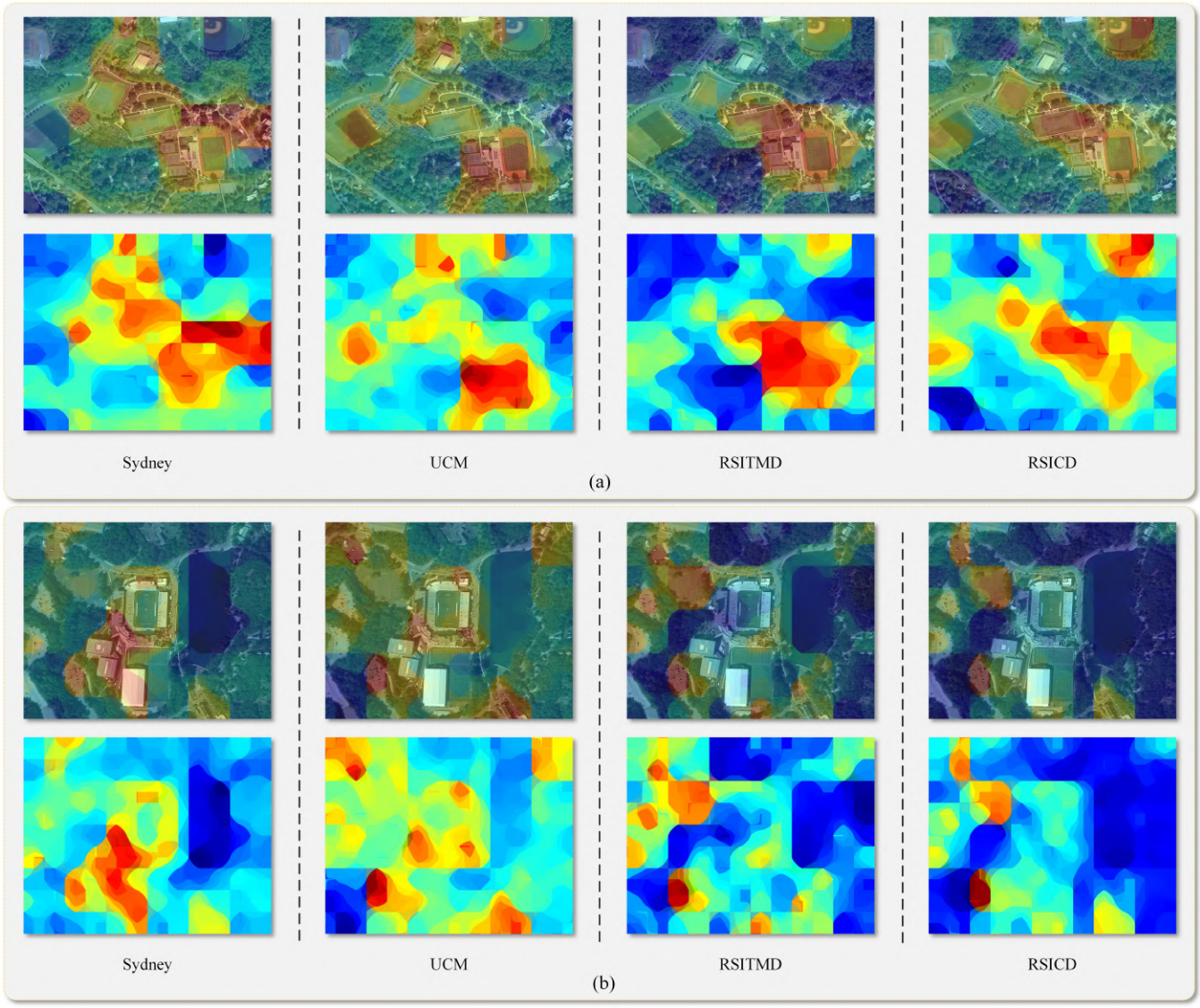


Fig. 9. Visual comparison of retrieval results on different datasets. (a)(b) show two examples respectively. The top of each example is the located image, and the bottom is the corresponding probability map. The query of (a): "There are several table tennis courts between the two playgrounds." The query of (b): "There are some cars parked in the parking lot."

In Fig. 9(b), the query of "There are some cars parked in the parking lot." is used to perform semantic localization. When using the Sydney dataset, it is evident that the model focuses on the wrong regions, which indicates that the model does not have a better perception of the categories. When using the UCM dataset, while some correct results are retrieved, there are also many missed detections and false detections. When the RSITMD dataset is used, good retrieval results are obtained. Almost all regions matching with the description are obtained. When using the RSICD dataset, whilst the model retrieved some results, the degree of attention is not as significant as the case of RSITMD.

With these two representative examples, we can already draw the following basic conclusions:

- The Sydney dataset is too small to achieve the desired results when performing the semantic localization task, even if good results can be achieved in the metrics.
- The models trained on Sydney and UCM datasets could not be capable to accept a wide range of query input, due to their small query language base, which will lead to a large amount of missed detections and false detections.
- When the RSICD dataset is used for the retrieval task, even though relatively good results can be obtained due to the large size of the dataset, a small portion of the region is still missed. This may be caused by the fact that the dataset contains fewer words and more solidified sentences.
- Compared with the other datasets, the performance of the retrieval model obtained by training with the RSITMD dataset is optimal. This is caused by the fact that RSITMD has more categories, more words, and more fine-grained semantic representations. Therefore, RSITMD may be a better choice when performing semantic localization tasks.

V. Conclusion

This paper first proposes a lightweight multi-scale cross-modal retrieval framework. Furthermore, we apply knowledge distillation and semi-supervised optimization to improve the performance of the retrieval model further. The results on four RS image-text datasets indicate the strength of the proposed framework and optimization method. Qualitative and quantitative analyses illustrate that lightweight RS retrieval with less calculation and parameters may be the next research hotspot.

References

[1] Y. Ma et al., "Remote sensing big data computing: Challenges and opportunities," Future Gener. Comput. Syst., vol. 51, pp. 47-60, Oct. 2015

[2] D. Mandal, K. N. Chaudhury, and S. Biswas, "Generalized semantic preserving hashing for N-label cross-modal retrieval," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 4076-4084

[3] M. Chi, A. Plaza, J. A. Benediktsson, Z. Sun, J. Shen, and Y . Zhu, "Big data for remote sensing: Challenges and opportunities," Proc. IEEE, vol. 104, no. 11, pp. 2207-2219, Nov. 2016.

[4] Li, Y., Ma, J., & Zhang, Y. (2021). Image retrieval from remote sensing big data: A survey. Information Fusion, 67, 94-115.

[5] Y. Li, Y. Zhang, X. Huang and J. Ma, "Learning Source-Invariant Deep Hashing Convolutional Neural Networks for Cross-Source Remote Sensing Image Retrieval," in IEEE Transactions on Geoscience and Remote Sensing, vol. 56, no. 11, pp. 6521-6536, Nov. 2018, doi: 10.1109/TGRS.2018.2839705.

[6] Z. Huang, W. Li, X. -G. Xia, H. Wang, F. Jie and R. Tao, "LO-Det: Lightweight Oriented Object Detection in Remote Sensing Images," in IEEE Transactions on Geoscience and Remote Sensing, doi: 10.1109/TGRS.2021.3067470.

[7] Q. Ran, Q. Wang, B. Zhao, Y. Wu, S. Pu and Z. Li, "Lightweight Oriented Object Detection using Multi-scale Context and Enhanced Channel Attention in Remote Sensing Images," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, doi: 10.1109/JSTARS.2021.3079968.

[8] Lu, X., Wang, B., Zheng, X., & Li, X. (2017). Exploring models and data for remote sensing image caption generation. IEEE Transactions on Geoscience and Remote Sensing, 56(4), 2183-2195.

[9] Shi, Z., & Zou, Z. (2017). Can a machine generate humanlike language descriptions for a remote sensing image?. IEEE Transactions on Geoscience and Remote Sensing, 55(6), 3623-3634.

[10] Abdullah, T., Bazi, Y., Al Rahhal, M. M., Mekhalfi, M. L., Rangarajan, L., & Zuair, M. (2020). TextRS: Deep bidirectional triplet network for matching text to remote sensing images. Remote Sensing, 12(3), 405.

[11] Z. Yuan et al., "Exploring a Fine-Grained Multiscale Method for Cross-Modal Remote Sensing Image Retrieval," in IEEE Transactions on Geoscience and Remote Sensing, doi: 10.1109/TGRS.2021.3078451.

[12] X. Wu, D. Hong, P. Ghamisi, W. Li and R. Tao, "LW-ODF: A Light-Weight Object Detection Framework for Optical Remote Sensing Imagery," IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium, 2019, pp. 1462-1465, doi: 10.1109/IGARSS.2019.8898673.

[13] Li, Y., Zhang, Y., Huang, X., Zhu, H., & Ma, J. (2017). Large-scale remote sensing image retrieval by deep hashing neural networks. IEEE Transactions on Geoscience and Remote Sensing, 56(2), 950-965.

[14] P. Li et al., "Hashing Nets for Hashing: A Quantized Deep Learning to Hash Framework for Remote Sensing Image Retrieval," in IEEE Transactions on Geoscience and Remote Sensing, vol. 58, no. 10, pp. 7331-7345, Oct. 2020, doi: 10.1109/TGRS.2020.2981997

[15] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

[16] Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., & Keutzer, K. (2016). SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and ¡0.5 MB model size. arXiv preprint arXiv:1602.07360.

[17] Zhang, X., Zhou, X., Lin, M., & Sun, J. (2018). Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6848-6856).

[18] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1251-1258).

[19] Yin, H., Molchanov, P., Alvarez, J. M., Li, Z., Mallya, A., Hoiem, D., ... & Kautz, J. (2020). Dreaming to distill: Data-free knowledge transfer via deepinversion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8715-8724).

[20] Pilzer, A., Lathuiliere, S., Sebe, N., & Ricci, E. (2019). Refine and distill: Exploiting cycle-inconsistency and knowledge distillation for unsupervised monocular depth estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9768-9777).

[21] Ji, M., Heo, B., & Park, S. (2021, February). Show, Attend and Distill: Knowledge Distillation via Attention-based Feature Matching. In Proceedings of the AAAI Conference on Artificial Intelligence.

[22] Y. Tao, M. Xu, F. Zhang, B. Du and L. Zhang, "Unsupervised-Restricted Deconvolutional Neural Network for Very High Resolution Remote-Sensing Image Classification," in IEEE Transactions on Geoscience and Remote Sensing, vol. 55, no. 12, pp. 6805-6823, Dec. 2017, doi: 10.1109/TGRS.2017.2734697.

[23] R. Fernandez-Beltran, B. Demir, F. Pla and A. Plaza, "Unsupervised Remote Sensing Image Retrieval Using Probabilistic Latent Semantic Hashing," in IEEE Geoscience and Remote Sensing Letters, vol. 18, no. 2, pp. 256-260, Feb. 2021, doi: 10.1109/LGRS.2020.2969491.

[24] Cai, J., Gu, S., & Zhang, L. (2018). Learning a deep single image contrast enhancer from multi-exposure images. IEEE Transactions on Image Processing, 27(4), 2049-2062.

[25] Lee, J., Koh, J., & Yoon, S. (2020). Momentum Contrast Speaker Representation Learning. arXiv preprint arXiv:2010.11457.

[26] Li, X., Zhang, X., Huang, W., & Wang, Q. (2020). Truncation Cross Entropy Loss for Remote Sensing Image Captioning. IEEE Transactions on Geoscience and Remote Sensing.

[27] Q. Cheng, Y. Zhou, P. Fu, Y. Xu and L. Zhang, "A Deep Semantic Alignment Network for the Cross-Modal Image-Text Retrieval in Remote Sensing," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 14, pp. 4284-4297, 2021, doi: 10.1109/JSTARS.2021.3070872.

[28] Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

[29] Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., & Bengio, Y. (2014). Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550.

[30] Zagoruyko, S., & Komodakis, N. (2016). Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv preprint arXiv:1612.03928.

[31] Jin, X., Peng, B., Wu, Y., Liu, Y., Liu, J., Liang, D., ... & Hu, X. (2019). Knowledge distillation via route constrained optimization. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1345-1354).

[32] Tian, Y., Krishnan, D., & Isola, P. (2019). Contrastive representation distillation. arXiv preprint arXiv:1910.10699.

[33] He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9729-9738).

[34] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020, November). A simple framework for contrastive learning of visual representations. In International conference on machine learning (pp. 1597-1607). PMLR.

[35] Chen, X., Fan, H., Girshick, R., & He, K. (2020). Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297.

[36] Chen, T., Kornblith, S., Swersky, K., Norouzi, M., & Hinton, G. (2020). Big self-supervised models are strong semi-supervised learners. arXiv preprint arXiv:2006.10029.

[37] Perronnin, F., Snchez, J., & Mensink, T. (2010, September). Improving the fisher kernel for large-scale image classification. In European conference on computer vision (pp. 143-156). Springer, Berlin, Heidelberg.

[38] Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). Cbam: Convolutional block attention module. In Proceedings of the European conference on computer vision (ECCV) (pp. 3-19).

[39] Zhang, Y., & Wallace, B. (2015). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. arXiv preprint arXiv:1510.03820.

[40] Faghri, F., Fleet, D. J., Kiros, J. R., & Fidler, S. (2017). Vse++: Improving visual-semantic embeddings with hard negatives. arXiv preprint arXiv:1707.05612.

[41] Borgwardt, K. M., Gretton, A., Rasch, M. J., Kriegel, H. P., Scholkopf, B., & Smola, A. J. (2006). Integrating structured biological data by kernel maximum mean discrepancy. Bioinformatics, 22(14), e49-e57.

[42] Oord, A. V. D., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748.

[43] Cheng, G., Zhou, P., & Han, J. (2016). Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 54(12), 7405-7415.

[44] Qu, B., Li, X., Tao, D., & Lu, X. (2016, July). Deep semantic understanding of high resolution remote sensing image. In 2016 International conference on computer, information and telecommunication systems (Cits) (pp. 1-5). IEEE.

[45] Huang, Y., Wu, Q., Song, C., & Wang, L. (2018). Learning semantic concepts and order for image and sentence matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6163-6171).

[46] Lee, K. H., Chen, X., Hua, G., Hu, H., & He, X. (2018). Stacked cross attention for image-text matching. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 201-216).

[47] Wang, T., Xu, X., Yang, Y., Hanjalic, A., Shen, H. T., & Song, J. (2019, October). Matching images and text with multi-modal tensor fusion and re-ranking. In Proceedings of the 27th ACM international conference on multimedia (pp. 12-20).

[48] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[49] Guo, M., Zhou, C., & Liu, J. (2019). Jointly Learning of Visual and Auditory: A New Approach for RS Image and Audio Cross-Modal Retrieval. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 12(11), 4644-4654.

[50] Chen, Y., Lu, X., & Wang, S. (2020). Deep Cross-Modal Image-Voice Retrieval in Remote Sensing. IEEE Transactions on Geoscience and Remote Sensing, 58(10), 7049-7061.

[51] Lin, T. Y., RoyChowdhury, A., & Maji, S. (2015). Bilinear cnn models for fine-grained visual recognition. In Proceedings of the IEEE international conference on computer vision (pp. 1449-1457).

[52] Kim, Y. (2014). Convolutional neural networks for sentence classification. EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 1746-1751. doi:10.3115/v1/d14-1181

Zhiqiang Yuan received the B.Sc. degree from the Harbin Engineering University, Harbin, China, in 2019. He is currently pursuing the Ph.D. degree with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China.

His research interests include multi-modal remote sensing image interpretation and multi-modal signal processing. (yuanzhiqiang19@mails.ucas.ac.cn)

Wenkai Zhang recevied the B.Sc. degree from China University Of Petroleum, TsingTao, China, in 2013 and the Ph.D. from Institute of Electronics, Chinese Academy of Sciences in 2018.

He is currently an assistant professor in the Aerospace Information Research Institute, Chinese Academy of Sciences, China. His research interests include multi-modal signal processing, image segmentation, and pattern recognition. (zhangwk@aircas.ac.cn)

Xuee Rong received the B.Sc. degree from the Minzu University of China, Beijing, China, in 2019. She is currently pursuing the Ph.D. degree with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China.

Her research interests include computer vision, pattern recognition, and remote sensing image processing, especially on continual semantic segmentation. (rongxuee19@mails.ucas.ac.cn)

Xuan Li received the B.Sc. degree from the Jilin University, Jilin, China, in 2017. He is currently pursuing the Ph.D. degree with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China.

His research interests include multi-modal signal processing, image caption, and pattern recognition. (lixuan173@mails.ucas.ac.cn)

Jialiang Chen received the B.Sc. degree from the Zhengzhou University, Zhengzhou, China in 2012 and the M.Sc. degree from the Beijing Institute of Technology, Beijing, China in 2016.

He is currently an assistant professor in the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research interests include pattern recognition and remote sensing image processing. (chenjl@aircas.ac.cn)

Hongqi Wang received the B.Sc. degree from the Changchun University of Science and Technology, Changchun, China, in 1983, the M.Sc. degrees from the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China, in 1988, and the Ph.D. degree from the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, in 1994.

He is currently a Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences, China. His research interests include computer vision, and remote sensing image understanding.

Kun Fu received the B.Sc., M.Sc. and Ph.D. degrees from the National University of Defense Technology, Changsha, China, in 1995, 1999 and 2002, respectively.

He is currently a Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research interests include computer vision, remote sensing image understanding, geospatial data mining and visualization.

Xian Sun received the B.Sc. degree from the Beijing University of Aeronautics and Astronautics, Beijing, China, in 2004, and the M.Sc. and Ph.D. degrees from the Institute of Electronics, Chinese Academy of Sciences, China, in 2009.

He is currently a Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences, China. His research interests include computer vision, geospatial data mining, and remote sensing image understanding.

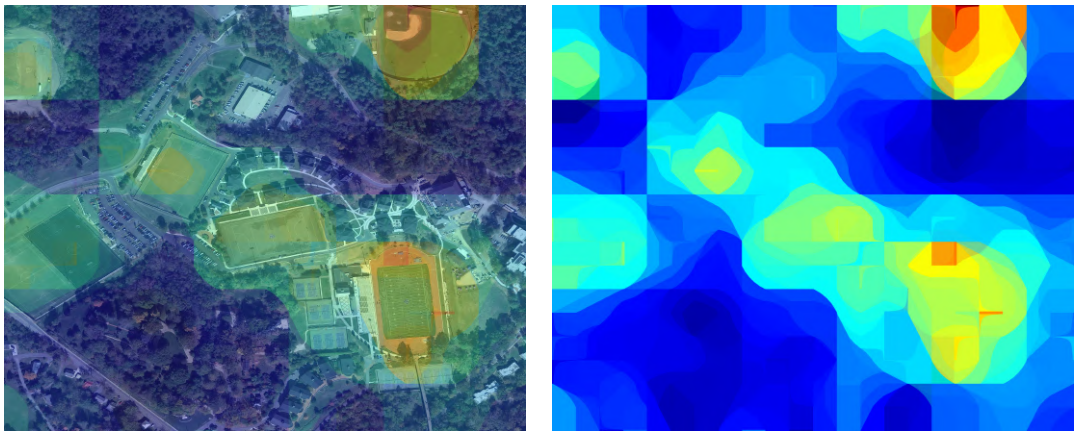
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2021.3124252, IEEE Transactions on Geoscience and Remote Sensing

APPENDIX A
NOTATIONS AND THEIR MEANINGS IN THE PROPOSED METHOD (IN ORDER OF OCCURRENCE)

Symbol	Meaning
$\mathbb{R}$	high-dimensional embedded feature space
d_f	dimensionality of the embedded feature space
I,S	image, sentence
Wv, Wt	weight matrix for visual and sentence
Sim	crossmodal similarity
$\mathfrak{S}$	cosine similarity
f_v,f_t	representation of the image and sentence in the embedded feature space
v^l, v^h, v^{lh}	low-level semantic features, high-level semantic feature and multi-scale fusion feature
PReLU	PReLU activation function
$conv$	convolution transformation
b_{ave}	average fusion result after bilinear multiplication
$l2$	L2 normalization
Ave	channel averaging
Max	max pooling
f_{ave},f_{max}	average component and impulse component of the multi-scale fusion feature following bilinear pooling
Cat	channel-wised concatenation
Att_c, Att_s	channel attention, spatial attention
$\otimes$	element-wise multiplication
VSA	visual self attention block
$Linear$	linear transformation
We	word embedding matrix
d_{emb}	dimension of the word embedding
e_n	embedded vector of n^{th} word
$e_{i:j}$	embedded matrix of the window between the i^{th} word and the j^{th} word
$c_{i,h}$	semantic information between the i^{th} word and the $(i+h)^{th}$ word
$\overline{c_h}$	salient information for the window h
L_{tpt}	triplet Loss
$\widehat{I}$,$\widehat{S}$	image which from different sample pairs with anchor S, sentence which from different sample pairs with anchor I
ε	minimum margin
$f_{m,tea}$, $f_{m,stu}$	unimodal embedded features f_m in the teacher network and student network
$L_{emb,m}$	embedded kd Loss for the unimodal m
ϕ	transformation function
L_{mis}	multi-scale information supervision loss
α,β,γ,λ	trade-off parameters to integrate total loss
DA_n	n^{th} data augmentation methods
φ	image semantic similarity computation network
f_{m,DA_n}	feature of image I_m following DA_n
τ	temperature parameter
$\Upsilon_{k\neq l}$	an indicator function assigned to 1 if $k \neq l$
L_{nce}	self supervised information loss
L_{inv}	rotation invariant loss
κ	edge softening coefficient
L_{bud}	semi-supervised loss
$Mask_{neg}$	negative sample masks
D_l	labeled dataset
D_u	unlabeled dataset
L_{total}	total loss

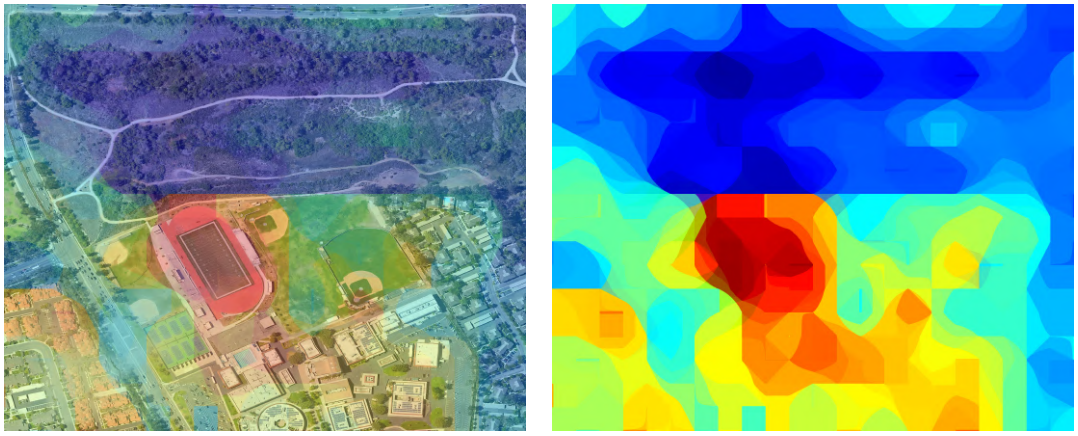
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2021.3124252, IEEE Transactions on Geoscience and Remote Sensing

APPENDIX B
MORE EXAMPLES OF SEMANTIC LOCALIZATION



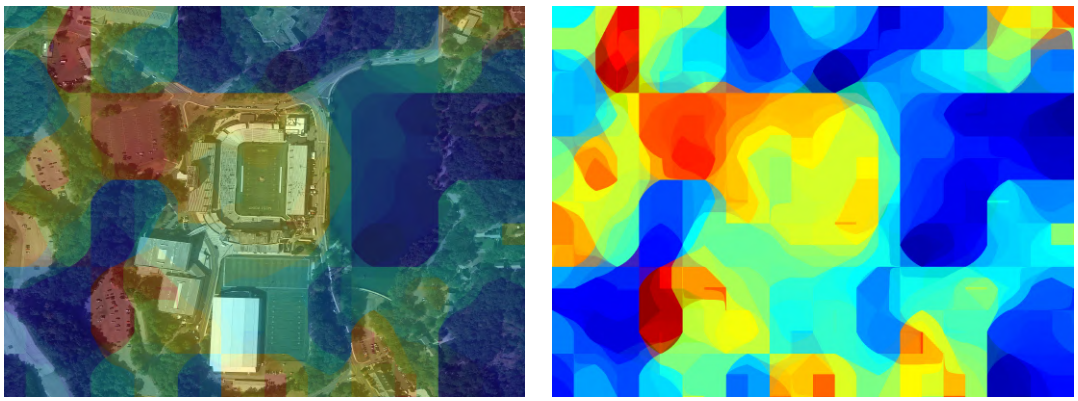
(a) Located image (b) Probability map

Fig. 10. Query: There is a green baseball field next to a small gray house.



(a) Located image (b) Probability map

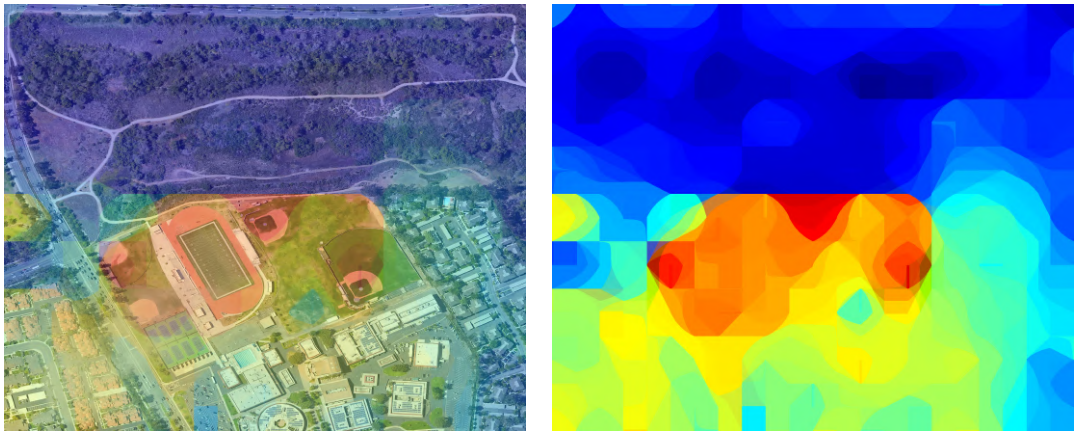
Fig. 11. Query: The red track surrounds the green playground.



(a) Located image (b) Probability map

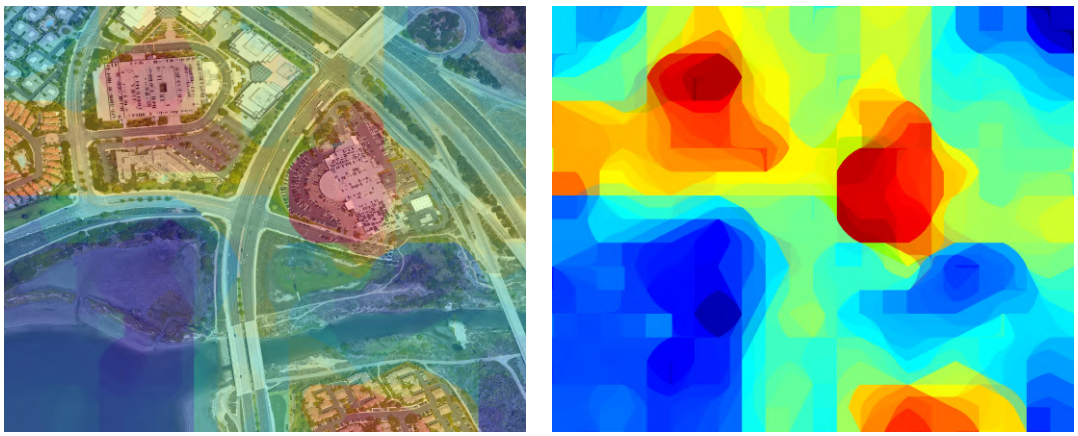
Fig. 12. Query: There are many cars parked in the parking lot.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2021.3124252, IEEE Transactions on Geoscience and Remote Sensing



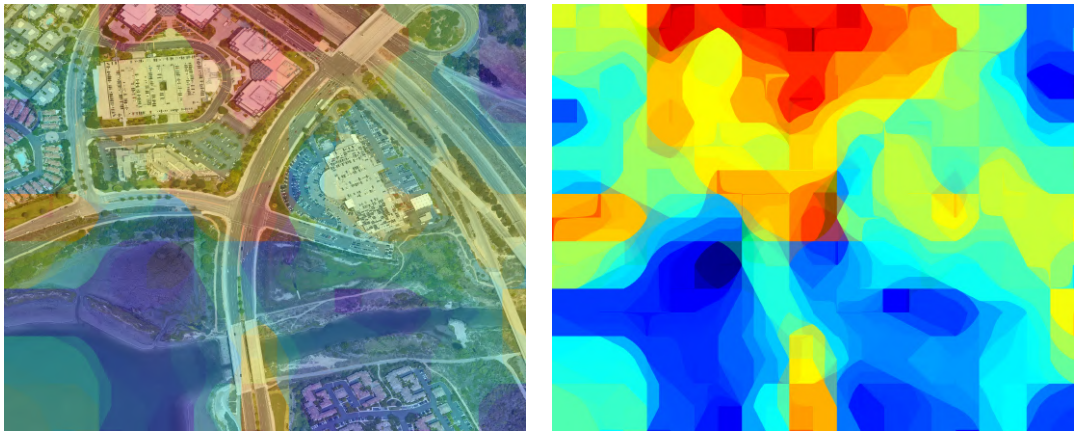
(a) Located image (b) Probability map

Fig. 13. Query: The two baseball fields are symmetrically next to each other.



(a) Located image (b) Probability map

Fig. 14. Query: The parking lot is packed with a large number of cars.



(a) Located image (b) Probability map

Fig. 15. Query: A white right-angled building surrounded by the gray road