Multilanguage Transformer for Improved Text to Remote Sensing Image Retrieval

Mohamad M. Al Rahhal, *Senior Member, IEEE*, Yakoub Bazi, *Senior Member, IEEE*, Norah A. Alsharif, Laila Bashmal, *Graduate Student Member, IEEE*, Naif Alajlan, *Senior Member, IEEE*, and Farid Melgani, *Fellow, IEEE*

***Abstract*—Cross-modal text-image retrieval in remote sensing (RS) provides a flexible retrieval experience for mining useful information from RS repositories. However, existing methods are designed to accept queries formulated in the English language only, which may restrict accessibility to useful information for non-English speakers. Allowing multilanguage queries can enhance the communication with the retrieval system and broaden access to the RS information. To address this limitation, this article proposes a multilanguage framework based on transformers. Specifically, our framework is composed of two transformer encoders for learning modality-specific representations, the first is a language encoder for generating language representation features from the textual description, while the second is a vision encoder for extracting visual features from the corresponding image. The two encoders are trained jointly on image and text pairs by minimizing a bidirectional contrastive loss. To enable the model to understand queries in multiple languages, we trained it on descriptions from four different languages, namely, English, Arabic, French, and Italian. The experimental results on three benchmark datasets (i.e., RSITMD, RSICD, and UCM) demonstrate that the proposed model improves significantly the retrieval performances in terms of recall compared to the existing state-of-the-art RS retrieval methods.**

***Index Terms*—Contrastive loss, cross-modal retrieval, language transformer, remote sensing, vision transformer.**

I. Introduction

REMOTE sensing (RS) data play a substantial role in analyzing geographic phenomena and forecasting the future state of the earth's surface. In the last several years, RS technology has advanced at a breakneck pace [1]. This combined with the increasing number of the launched earth observation satellites that are constantly monitoring the earth has led to a significant growth in the RS image archive.

To make full use of such big data, the development of appropriate and efficient retrieval methods that can deal with RS archives in a manageable way is becoming urgently needed. The task of image retrieval, which aims to study how to extract a specific image out of a massive amount of data, has received a great deal of attention recently. The core idea is to narrow the search for the targeted image and retrieve the image that matches a particular query. This task has important value in many practical applications including deforestation detection, visual navigation, and urban planning.

The common approach in RS retrieval is the single-modal retrieval [2], which accepts an image as a query to match its content against all the images in the archive. This process involves extracting representative features from the set of images and then, applying a certain measure to quantify the similarity between the query image and the images in the archive to retrieve a list of candidate images. Early single-modal methods have adopted hand-crafted features to represent the visual content of images [3], [4]. However, these manually designed features are inefficient at describing the rich semantic information contained in RS images. On the contrary, the developments of deep-learning models such as convolutional neural networks (CNNs), have brought crucial achievements in boosting the accuracy of retrieval systems [5] duo to its ability to automatically learn high-level features from complex RS scenes.

Although single-modal retrieval has been extensively studied in the RS domain, it still suffers from a fundamental problem in terms of usability. Single-modal retrieval requires the user to formulate the query using a preexisting image. This constraint, in many cases, can be problematic and impractical as the availability of an exemplar query is not always guaranteed. Allowing the user to formulate a spoken, written, or even drawn query can give the user more flexibility to describe the content of the targeted image. Hence, developing cross-modal retrieval models has become increasingly important for enhancing the retrieval experience.

Cross-modal retrieval basically aims to let the user search for data in one modality by a query in another modality. Today, as we are witnessing the era of big data, data from various sources (e.g., optical, radar, or laser) and a growing number of domains (e.g., image, text, and sound) have become available. As a result, a new category of multimodal applications has emerged, and

Manuscript received 17 May 2022; revised 10 August 2022 and 10 September 2022; accepted 11 October 2022. Date of publication 20 October 2022; date of current version 1 November 2022. This work was supported by the Researchers Supporting Project under Grant RSP-2021/69, King Saud University, Riyadh, Saudi Arabia. *(Corresponding author: Yakoub Bazi.)*

Mohamad M. Al Rahhal is with the Applied Computer Science Department, College of Applied Computer Science, King Saud University, Riyadh 4545, Saudi Arabia (e-mail: mmalrahhal@ksu.edu.sa).

Yakoub Bazi, Norah A. Alsharif, Laila Bashmal, and Naif Alajlan are with the Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 4545, Saudi Arabia (e-mail: ybazi@ksu.edu.sa; 441202939@student.ksu.edu.sa; lailabashmal@outlook.com; najlan@ksu.edu.sa).

Farid Melgani is with the Department of Information Engineering and Computer Science, University of Trento, 38123 Trento, Italy (e-mail: melgani@disi.unitn.it).

Digital Object Identifier 10.1109/JSTARS.2022.3215803

researchers have started to pay more attention to the interpretation tasks that involve multimodality interactions, such as image captioning [6], [7], [8] and visual question answering [9], [10].

Motivated by this, RS image retrieval has been explored using queries of different modalities such as images from different sources or sensors [11], [12], [13], [14], sketches [15], [16], speech [17], [18], [19], [20], [21], and text [22], [23], [24], [25], [26], [27]. Among these modalities, textual descriptions represent the most intuitive way of communicating with machines. It can exhaustively describe the vast and complex content of the scene with a few concise words. Therefore, using textual information for querying RS images can enhance the retrieving experience and bring more flexibility in terms of the query description, but at the same time, expressing RS scene using natural language introduces new challenges due to the visual-semantic discrepancy between language and vision worlds.

In the literature, only few works have been developed for text-image retrieval [22], [23], [24], [25], [26]. All of these works have been designed to allow English as the primary language of the query. However, the use of English as the sole language may create a barrier for those who are non-English speakers. Thus, developing a retrieval model that can cross language boundaries and process queries in the user's mother tongue is highly desirable.

Moreover, although the prior methods of RS text-image retrieval have achieved promising results, most of the proposed models rely on CNN for encoding the visual data and recurrent models for encoding the textual data. Yet, these models have shown some limitations in capturing global dependencies within the input and have been gradually replaced by a self-attention-based model known as a transformer. The transformer is now considered the state-of-the-art model in natural language processing (NLP), due to its ability to handle long-range interrelationships within the data and hence, providing better representations. In addition, transformer-based retrieval models have shown promising results in the context of computer vision [28], [29].

With these considerations in mind, in this article, we propose a multilanguage text-based retrieval model for RS images where the textual query can be of any of the following languages—English, Arabic, French, and Italian. To the best of our knowledge, no work has addressed the use of multilanguage text for querying RS images. To fully extract representative features from the vision and language domains, the model employs two transformers, one for visual features and the second for textual features. The model aims at exploiting the semantic relation between the image and the corresponding textual description to learn their representation in a joint embedding space.

The main contributions of this article are summarized as follows.

1) To the best of our knowledge, this is the first study in RS community that incorporates multilanguage queries to perform cross-modal text-image retrieval.
2) This article proposes a dual transformer-based model for better learning of the visual and linguistic features from the image and the corresponding captions, respectively. The model is trained with bidirectional contrastive loss to encourage the model to correctly align the features of the image and the corresponding text into the same cross-modal space, and thereby improving the performance of the retrieval task.
3) The proposed model was extensively evaluated on three RS text-image datasets using single-language and multi-language training settings. The results show that it can achieve better performance compared to state-of-the-art methods based on recurrent networks.

The rest of this article is organized as follows. Section II presents an overview of the state-of-the-art works in the cross-modal RS retrieval. Section III describes the methodology of the multilanguage text-image retrieval model. The experimental results on three benchmark datasets are presented in Section IV. Finally, Section V concludes this article.

II. Related Work

Many efforts have been dedicated by the RS community to automate information retrieval from large repositories. The existing techniques can be roughly divided into two main streams: single-modal and cross-modal retrieval methods depending on the type of the query. The single-modal retrieval methods take a query image as an input and retrieve a list of images that are mostly similar to the content of the query image. The other approach is cross-modal retrieval, in which the query can be of any type of data (e.g., speech or a descriptive sentence). In this case, the semantic concept of the query is extracted and matched against the visual content of all images in the archive to find the most relevant ones. In the following, we present a review of existing works related to cross-modal image retrieval.

A. Cross-Modal Image-Image Retrieval

As a result of the rapid development of RS technologies, the number of RS images captured by different types of sensors has increased. Consequently, cross-model image-image retrieval has received widespread attention in recent years.

Cross-modal image-image retrieval allows retrieving the target RS imagery by using a query image from different sources [11], different sensors [12], [30], or even a sketch image [15], [16].

For example, Li et al. [11] introduced a cross-source image retrieval approach, which utilizes deep hashing CNNs to perform image retrieval. Xiong et al. [13] proposed a cross-source discriminative distillation network to elevate the effect of data drift. In [14], a model based on cycle-GAN is proposed to translate the image from one source to the other.

Chaudhuri et al. [12] proposed a method for performing cross-modal retrieval between panchromatic (PAN) and multispectral imagery. Xiong et al. [30] proposed a cross-modal hashing network for retrieving optical images from synthetic aperture radar (SAR) images. The method transforms the optical image into a single channel image and pairs it with the corresponding SAR image to train the model.

In [15], a multiscale model consisting of CNNs and fully connected layers is proposed for retrieving RS images using a coarse sketch. Another work [16] proposed a sketch-based retrieval model that learns domain-invariant representations by adversarial training.

B. Cross-Modal Sound-Image Retrieval

The goal of cross-modal sound-image retrieval is to leverage sound to retrieve relevant RS images. Indeed, using speech as a query can be efficient in practice as it has a broad range of application scenarios. Some notable works have been published very recently for sound-image retrieval. For example, the authors in [18] proposed a triplet network containing a branch for the image, another one for the positive sound, and a branch for the negative sound. Then, the model parameters were learned by optimizing a triplet loss. The model integrated hash code learning to reduce storage costs. In another work [19], a CNN with an inception dilated convolution [31], [32] module layer was introduced to extract multiscale contextual information from both images and voices. The model also learns hash code for faster retrieval and lower storage. The authors in [21] proposed a model that extracts high-level features from the images using a CNN and extracts high-level voice features using a 1-D dilated convolutional model. To find the similarity between the two different modalities, the consistency loss and the classification loss are minimized jointly to learn the representations needed for the RS image–voice retrieval. Guo et al. [20], proposed an RS speech-image retrieval model which uses a 1-D convolutional network to extract high-level semantic features of the spoken query, and a CNN to extract high-level visual features from each RS image. A multimodal fusion layer was added on the top for fusing both modalities.

C. Cross-Modal Text-Image Retrieval

Over the last two years, only a few studies have been proposed for RS image retrieval using a textual query. This is mainly due to the challenging nature of the problem and the special characteristics of RS images. The first work in text-based RS image retrieval [22] proposed a deep bidirectional triplet network to match natural language descriptions to images. The triplet network is composed of a long short term memory (LSTM) and a pretrained CNN. On top of this architecture, an average fusion strategy was used to fuse the features pertaining to different sentences. Hoxha et al. [23] developed a text-based image retrieval system that combines a CNN with a recurrent neural network. The textual query can be directly given or generated by captioning the query image. Rahhal et al. [24] proposed an unsupervised learning method for text-image retrieval. The model used a CNN for encoding the image and a bidirectional LSTM for encoding the text description. The authors in [25] introduced a semantic alignment module in order to discover the semantic relationships between image and text in the joint embedding space. The module employed attention and gate mechanisms to extract discriminative visual and textual feature representations. Specifically, the attention mechanism was utilized to optimize the corresponding relationships between visual and textual features, and the gate function was employed to filter the unnecessary information to obtain the discriminative features. Yuan et al. [26] designed an asymmetric network to solve the target redundancy and multiscale scarcity problems in RS retrieval tasks. This method filters redundant features and adapt to multiscale feature inputs by using a multiscale visual self-attention module. The authors also addressed the problem of high intraclass similarity in RS images by designing a triplet loss function to train the model. In [27], the authors proposed a text-image retrieval model and applied two methods to improve retrieval performance—a knowledge distillation-based method and a semisupervised optimization method based on contrastive learning.

III. Methodology

In this section, we introduce the multilanguage transformer method that we propose in detail. Since the transformer is a central model in our method, we first describe the general architecture of the language transformer encoder, and then we describe the vision transformer encoder.

First, we assume having a set of image-text pairs denoted as $D = \{X_{\mathrm{i}}, t_i\}_{i=1,}^{N}$ where X_{i} represents an image, and t_{i} represents the corresponding sentence in one of the following languages—English, Arabic, French, and Italian. In the text-to-image retrieval task, given a text query, the goal is to search for the most relevant image X_{i} to the given text query. Similarly, in the image-to-text retrieval task, the goal is to retrieve the most similar sentence t_{i} to the query image. To achieve that, we adopt two transformers one for image encoding and the other for text encoding. The vision encoder accepts a mini-batch of b images, while the language encoder accepts a set of tokens of b sentences. The outputs of each encoder are injected into the global average pooling (GAP) layer to obtain a global feature representation for each modality. The output features are then normalized with L2-normlization to obtain the visual features $\{f_{vi}\}_{i=1}^{b}$ and the textual features $\{f_{ti}\}_{i=1}^{b}$. Afterward, a similarity matrix of size $b \times b$ is constructed between all the text-image pairs in the mini-batch. The model learns the weights by optimizing text-image and image-text contrastive classification loss using a stochastic gradient descent (SGD) optimizer.

In this article, we are particularly interested in investigating two learning paradigms as shown in Fig. 1. In the single-language learning paradigm, the text encoder is trained on each language independently so that it accepts a query from one language only. In the second paradigm, the text encoder is trained on sentences from multiple languages jointly, so the model can accept a query formulated in any of these languages. Detailed descriptions of the proposed model are provided in the following sections.

A. Language Transformer Encoder

The first step in processing the textual description for language transformer encoder [33] is sentence tokenization, in which the sentence is represented as word tokens $t_i = (w_1, w_2, \ldots w_m)$, where m is the length of the sentence. Then, this word vector is projected into an embedding space using a learnable embedding layer E_t, that converts these tokens into a sequence of textual features of dimension d_t.

Before feeding the sequence into the encoder, a learnable positional embedding is appended to supply the sequence with information about the order of each word. In addition, two special tokens CLS and SEP are added to the input tokens to

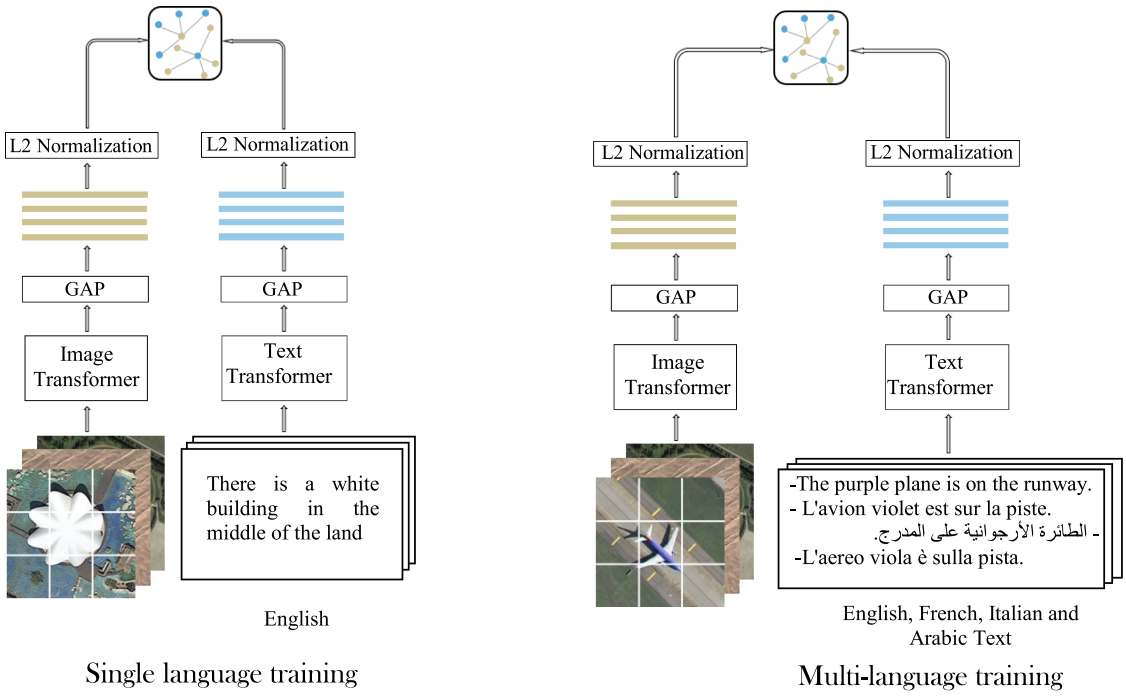


Fig. 1. Difference between the single-language (left) and multilanguage (right) retrieval models. In the single-language model, the query text can be formulated in only one language. In the multilanguage model, the query can be formulated in any of the four considered languages. For training the network, we sample a mini-batch of $\boldsymbol{b}$ images-text pairs and feed them to the vision and language transformer encoders to generate L2-normalized visual $\{\boldsymbol{f_{vi}}\}_{i=1}^{\boldsymbol{b}}$ and textual $\{\boldsymbol{f_{ti}}\}_{i=1}^{\boldsymbol{b}}$ features. Then, we generate a similarity matrix of size $\boldsymbol{b} \times \boldsymbol{b}$ by computing the similarity between all possible visual and textual pairs in the mini-batch. We learn the model weights by optimizing text-to-image and image-to-text contrastive classification loss $\mathcal{L}_{\boldsymbol{total}} = \boldsymbol{\lambda}_1 \mathcal{L}_{\boldsymbol{v} \rightarrow \boldsymbol{t}} + \boldsymbol{\lambda}_2 \mathcal{L}_{\boldsymbol{t} \rightarrow \boldsymbol{v}}$ using an SGD optimizer.

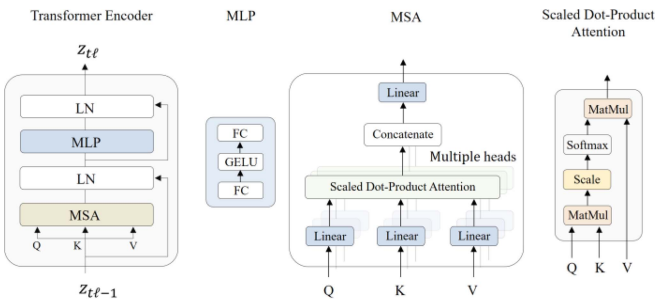


Fig. 2. Language transformer layer.

mark the start and end of the sequence. Thus, the sentence representation z_{t0} is expressed as

$$z_{t0} = [w_{\text{class}};\ w_1 E_t; w_2 E_t; \ldots;\ w_m E_t]\ + E_{\text{pos}} \qquad (1)$$

where w_{class} is a special classification token that provides a general representation of all the tokens, and $E_{\text{pos}} \in \mathbb{R}^{(m+1)\times d_t}$ is the positional embedding. The initial representation z_{t0} is fed as input through multiple identical layers of the encoder to generate the final representation z_{tL} at the last layer L. Each layer in the encoder contains a multihead self-attention (MSA) block followed by a multilayer perceptron (MLP) block. Which is a simple feed-forward network consisting of two fully connected layers with GELU activation function in between as shown in Fig. 2. The MSA and MLP blocks are connected by residual skip connections and each layer is followed by a normalization layer (LN):

$$z'_\ell = \text{MSA}\left(\text{LN}\left(z_{t\ell-1}\right)\right) + z_{t\ell-1},\ \ell = 1 \ldots L \qquad (2)$$

$$z_{t\ell} = \text{MLP}\left(\text{LN}\left(z'_{t\ell}\right)\right) + z'_{t\ell},\qquad \ell = 1 \ldots L. \qquad (3)$$

The main goal of the MSA is to manage the complex relationships within the sequential data by modeling the long-range dependencies between a specific token and all other tokens in the sequence. It comprises multiple independent self-attention heads operating in parallel, each head computes a different attention score using the scaled dot-product similarity between the queries (Q), keys (K) and values (V) expressed by

$$\text{Attention} = \text{softmax}\left(\frac{QK}{\sqrt{d_K}}\right)V \qquad (4)$$

where d_K is the dimension of the key. The outputs of all heads are concatenated and then projected with learnable weights matrix to the desired dimension.

B. Image Transformer Encoder

After the tremendous success of transformer in NLP [34], they have been extended recently to computer vision tasks leading to the so-called ViT [35]. This last showed competitive results compared to CNN for several image processing tasks, thanks to the self-attention mechanism. In vanilla ViT, the sequence of word tokens is replaced with a sequence of image patches. The input image X_{i} of size 224×224×3 pixels is first divided into N nonoverlapping patches $(x_p^1;\ x_p^2; \ldots;\ x_p^N)$. Each patch

in the sequence has the dimension of $(3p^2)$, where p represents the width/height of the patch and N is the total number of patches $N = (224 \times 224)/p^2$. This sequence of patches is flattened and projected via a linear projection layer E_v, to the encoder dimension d_v. Then, in a way similar to the language transformer, position embeddings are added to keep the position information. Also, x_{class} token is appended to the patch representations. The resulting image representation z_{v0} that is then fed into the encoder

$$z_{v0} = \left[x_{\text{class}};\ x_p^1 E_v;\ x_p^2 E_v; \ldots;\ x_p^N E_v\right] + E_{\text{pos}} \tag{5}$$

where $E_v \in \mathbb{R}^{(p^2 \cdot c) \times d_v \cdot}$ is the linear embedding layer and $E_{\text{pos}} \in \mathbb{R}^{(N+1) \times d_v \cdot}$ is the positional encoding. Finally, by applying the same operations as in (2) and (3), we obtain the final image representation z_{vL} at the last layer L. It is worth recalling, that the architecture of the vision encoder is similar to the language encoder. Except that the normalization layer comes before the MSA and the MLP blocks.

C. Network Optimization

In order to learn the weights of the model, we apply global average pooling to the representation matrices $z_{tL} \in \mathbb{R}^{(m+1) \times d_t}$ and $z_{vL} \mathbb{R}^{(N+1) \times d_v}$ obtained from the text and the image encoders, respectively, yielding a feature $f_t \in \mathbb{R}^{d_t}$ and $f_v \in \mathbb{R}^{d_v}$, with w_{class} and x_{class} tokens representing the whole sentence and image representations ignored. Then, the visual features are further mapped using a linear projection layer to the same dimension of the textual feature. We note that the dimension of the visual and textual features are $d_v = 768$, $d_t = 512$. Afterward, we apply L2-normalization to the resulting textual and visual features.

If we consider $B_k = \{X_{\text{i}},\ t_{\text{i}}\}_{i=1}^{b}$ as the kth mini-batch of size b sampled from the archive $D^{(l)}$. Feeding this mini-batch as input to the model yields the following normalized visual and textual feature representations $\{f_{vi}\}_{i=1}^{b}$ and $\{f_{ti}\}_{i=1}^{b}$. The main learning objective is to jointly train the image and text transformer encoders to maximize the similarity of truly corresponding image-text features pairs while simultaneously minimizing the similarity of mismatched image-text features pairs within the kth mini-batch. To achieve this objective, we rely on contrastive loss, which is a popular loss in self-supervised learning that has shown an excellent performance in pairwise similarity measurement tasks [36], [37].

In our context, we compute this loss in both textual and visual domains, respectively. In the visual domain, we aim at making the textual feature closer to its corresponding visual feature while being away from other visual features in the mini-batch. Similarly, in the text domain we aim at making the visual feature closer to its corresponding textual feature while pushing away other textual features in the mini-batch. This problem can be viewed as a multiclass classification problem with b classes, where b refers to the size of the mini-batch. Basically, we learn the weights of the model by minimizing the cross-entropy loss over the similarity matrix of size $b \times b$ in the horizontal and vertical directions. The text-to-image classification loss is given as follows:

$$\mathcal{L}_{t \to v} = -\frac{1}{b} \sum_{i=1}^{b} log \frac{\exp\left(f_{ti}^T f_{vi}/\tau\right)}{\sum_{j=1}^{b} \exp\left(f_{ti}^T f_{vj}/\tau\right)} \tag{6}$$

and likewise, the image-to-text classification loss is computed as

$$\mathcal{L}_{v \to t} = -\frac{1}{b} \sum_{i=1}^{b} log \frac{\exp\left(f_{vi}^T f_{ti}/\tau\right)}{\sum_{j=1}^{b} \exp\left(f_{vi}^T f_{tj}/\tau\right)} \tag{7}$$

where the learnable temperature parameter τ is set to 0.07 to control the sharpness of the distribution. Finally, the total loss function to optimize is

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{v \to t} + \lambda_2 \mathcal{L}_{t \to v} \tag{8}$$

where λ_1 and λ_2 are two hyper-parameters controlling the contributions of both losses. In all experiments, we set them to the value 0.5.

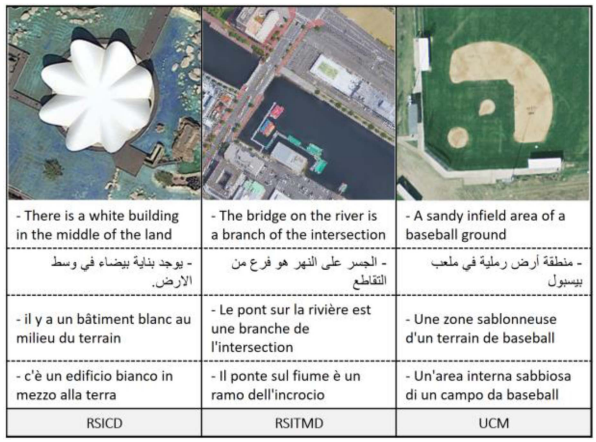


Fig. 3. Sample images from the three datasets used in the experiments with of their textual annotations.

IV. Experimental Results

In this section, Section A describes the three text-image datasets used in the experiments and the evaluation metrics. Section B introduces the details of the implementation and Section C presents the experimental results.

A. Datasets Description and Evaluation Metrics

1) Datasets: To validate the proposed multilanguage retrieval model, three RS cross-modal datasets were exploited to train and evaluate our model. All the datasets are annotated with five English textual descriptions. We used an online software to translate the descriptions from English to three different languages—Arabic, French, and Italian. Fig. 3 shows some images from the datasets with examples of their textual annotations in English, Arabic, French, and Italian, respectively, and Table I shows a comparison between the datasets. More details on the datasets are provided in the following.

a) RSICD [38]: This is the largest text-image dataset that contains 10921 images with various resolutions that belong to

TABLE I
TEXT-IMAGE RS DATASETS

Dataset	#Images	#Captions per image	Image size
RSICD	10921	5	224×224 pixels
RSITMD	4743	5	256×256 pixels
UCM	2100	5	256×256 pixels

30 different classes. Images were collected from various sources including Google Earth, MapABC, Baidu Map, and Tianditu. Each image in the dataset has 224×224 pixels and is described with five sentences, where each sentence is at least six words in length. In RSICD, the total number of captions is 24 333, which implies that not all images have five sentences. For consistency, captions have been duplicated for images with less than five sentences.

b) RSITMD [26]: This dataset consists of 4743 images belonging to 30 classes. The images were collected from the RSICD dataset and Google Earth. Each image has the size 256×256. Five different sentences were given to describe every image with a total of 23 715 captions. In addition to one to five keywords, the dataset was designed to have fewer images but more diverse captions compared to the RSICD datasets.

c) UCM [39]: This dataset comprises of 2100 images, each image with the size of 256×256 pixels with a spatial resolution of 30 cm. The dataset is based on the well-known Merced Land-use dataset [40] for scene classification that categorizes the 2100 images into 21 classes. Each image in this dataset is described with five different captions, resulting in 10 500 descriptions in total. Yet, there is a high similarity between the sentences of images that belong to the same class.

2) Evaluation Metrics: In this article, we present the results in terms of Recall@K (R@K) as it is the most adopted evaluation metric for cross-modal retrieval. The recall is a measure that represents the ratio of the correctly retrieved items to the total number of existing relevant items to the given query, and it is defined as follows:

$$R@k = \frac{TP@k}{TP@k + FN@k} \tag{9}$$

where TP is the true positive and FN is the false negative. We utilized the $R@k$ indicator with different values of k (1, 5, and 10) to measure the retrieval performance.

In addition, we provide the result in terms of mean recall (mR) to evaluate the overall performance of the model. The mR represents the average of R@1, R@5 and R@10 for both the text-to-image and image-to-text retrieval tasks. Besides the numerical metrics mentioned above, subjective metrics are also used to better understand the performance of models in retrieving the relevant images on different datasets.

B. Experimental Setup

Given the constraints that the current RS text-image datasets are of a small-scale type, we propose to transfer knowledge from backbones pretrained on a large-scale text-image dataset. To this end, we use the vision language model proposed in [36], which was trained on 400 million general text-image pairs. This model built upon two transformers for visual and textual feature representations. Specifically, ViT32 is adopted as the vision transformer, and a BERT-like model as the language transformer. ViT32 consists of $L = 12$ encoder layers. It divides the image of dimension 224×224×3 pixels into $n = 49$ patches each of dimension $(p, p) = (32 \times 32)$ pixels. These patches are flattened and mapped to the dimension $d_v = 768$. ViT32 has about 86M parameters. The language transformer encoder is a BERT-like model it has 63M parameters; and $L = 12$ layers. The vocabulary size is equal to 49,408. To facilitate batch processing, it provides a sequence with a fixed length equal to $m = 77$. Then, it uses a word embedding layer to embed the sequence into features of dimension $d_t = 512$. The resulting visual f_{vi} and textual f_{ti} feature representations after the GAP operation will be equal to 768 and 512, respectively. The visual features are further mapped using a linear projection layer to the same dimension as the textual feature, which is 512. Both visual and textual features are normalized using L2-Noramlization.

For data augmentation, we apply standard operations such as random crops, horizontal and vertical flips with 50% probability, and ColorJitter. For comparison purposes, we use the same split as in previous works. For RSICD and Merced, we consider 80% of the image-text pairs as training while 10% are left for validation and 10% for testing. For RSTIMD, we use 80% for training and 20% for testing.

Since each image is described by five sentences, in the single-language learning setting, we randomly select one of the five sentences for learning. For training the model on multiple languages, we pick each time a sentence from one of the four considered languages, which are English, Arabic, French, and Italian. It is worth-recalling that, we have used online translations tools from Google to generate the Arabic, French, and Italian captions from English sentences followed by manual correction.

As an optimizer, we use the SGD optimizer with Nesterov momentum. We set the initial learning rate to 0.1, and the momentum to the default value of 0.9. During training, we decrease the learning rate to 0.01 after 40 epochs, and to 0.001 for the last 20 epochs. For numerical stability, we found that it is useful to apply gradient clipping with a max norm of the gradients set to 0.1. The model is trained for 60 iterations with a mini-batch size set to $b = 120$.

The model was implemented in PyTorch and all the experiments were implemented on a station with a RAM of 32 GB and an NVIDIA GeForce GTX 1080 Ti Graphical Processing Unit (GPU) (with 11 GB GDDR5X memory).

C. Experimental Results

To evaluate the effectiveness of the proposed model, we report the results of two learning scenarios: the single-language and the multilanguage. In the first scenario, the model is trained on sentences in English language only to compare its performance with state-of-the-art methods. In the latter scenario, it is trained with sentences from multiple languages either independently

TABLE II
RETRIEVAL RESULTS USING ENGLISH LANGUAGE COMPARED TO STATE-OF-THE-ART METHODS

RSICD DATASET							
APPROACH	Text retrieval			Image retrieval			mR
	R@1	R@5	R@10	R@1	R@5	R@10	
VSE++ [42]	3.38	9.51	17.46	2.82	11.32	18.10	10.43
SCAN [43]	5.85	12.89	19.84	3.71	16.40	26.73	14.23
MTFN [44]	5.02	12.52	19.74	4.90	17.17	29.49	14.81
AMFMN [23]	5.39	15.08	23.40	4.90	18.28	31.44	16.42
LW-MRC-b [24]	4.57	13.71	20.11	4.02	16.47	28.23	14.52
LW-MRC-d [24]	3.29	12.52	19.93	4.66	17.51	30.02	14.66
LW-MRC-u [24]	4.39	13.35	20.29	4.30	18.85	32.34	15.59
Ours	**10.70**	**29.64**	**41.53**	**9.14**	**28.96**	**44.59**	**27.42**
RSITMD DATASET							
APPROACH	Text retrieval			Image retrieval			mR
	R@1	R@5	R@10	R@1	R@5	R@10	
VSE++ [42]	10.38	27.65	39.60	7.79	24.87	38.67	24.83
SCAN [43]	11.06	25.88	39.38	9.82	29.38	42.12	26.28
MTFN [44]	10.40	27.65	36.28	9.96	31.37	45.84	26.92
AMFMN [23]	10.63	24.78	41.81	11.51	34.69	54.87	29.72
LW-MRC-b [24]	9.07	22.79	38.05	6.11	27.74	49.56	25.55
LW-MRC-d [24]	10.18	28.98	39.82	7.79	30.18	49.78	27.79
LW-MRC-u [24]	9.73	26.77	37.61	9.25	34.07	54.03	28.58
Ours	**19.69**	**40.26**	**54.42**	**17.61**	**49.73**	**66.59**	**41.38**
UCM DATASET							
APPROACH	Text retrieval			Image retrieval			mR
	R@1	R@5	R@10	R@1	R@5	R@10	
VSE++ [42]	12.38	44.76	65.71	10.10	31.80	56.85	36.93
SCAN [43]	12.85	47.14	69.52	12.48	46.86	71.71	43.43
MTFN [44]	10.47	47.62	64.29	14.19	52.38	78.95	44.65
AMFMN [23]	16.67	45.71	68.57	12.86	53.24	79.43	46.08
LW-MRC-b [24]	12.38	43.81	59.52	12.00	46.38	72.48	41.10
LW-MRC-d [24]	15.24	51.90	62.86	11.90	50.95	75.24	44.68
LW-MRC-u [24]	18.10	47.14	63.81	13.14	50.38	79.52	45.35
Ours	**19.04**	**53.33**	**77.61**	**19.33**	**64.00**	**91.42**	**54.12**

TABLE III
RETRIEVAL RESULTS FOR SINGLE LANGUAGE AND MULTIPLE LANGUAGE TRAINING

RSICD DATASET							
APPROACH	Text Retrieval			Image Retrieval			mR
	R@1	R@5	R@10	R@1	R@5	R@10	
Single Language							
English	10.70	29.64	41.53	9.14	28.96	44.59	27.42
Arabic	9.05	25.06	36.50	6.40	22.03	35.38	22.40
French	8.69	24.33	37.14	7.02	24.33	39.03	23.42
Italian	10.06	26.44	39.70	8.05	26.34	41.55	25.35
Multiple Languages							
English	11.61	30.10	42.17	9.55	29.27	44.73	27.90
Arabic	10.06	26.98	37.05	7.77	23.93	39.17	24.16
French	8.78	26.89	40.07	8.19	26.34	43.11	25.56
Italian	10.15	26.98	37.69	8.43	26.56	42.08	25.31
RSITMD DATASET							
APPROACH	Text Retrieval			Image Retrieval			mR
	R@1	R@5	R@10	R@1	R@5	R@10	
Single Language							
English	19.69	40.26	54.42	17.61	49.73	66.59	41.38
Arabic	12.83	37.61	50.44	13.23	39.82	58.98	35.49
French	21.01	44.91	57.30	17.21	47.25	63.53	41.87
Italian	18.14	36.06	51.79	15.35	42.12	61.15	37.44
Multiple Languages							
English	22.56	43.80	57.30	19.29	51.37	68.18	43.75
Arabic	19.46	41.37	53.31	15.97	46.10	62.65	39.81
French	20.35	43.80	58.18	19.29	49.38	65.61	42.77
Italian	21.90	42.69	56.63	18.18	49.02	64.91	42.22
UCM DATASET							
APPROACH	Text Retrieval			Image Retrieval			mR
	R@1	R@5	R@10	R@1	R@5	R@10	
Single Language							
English	19.04	53.33	77.61	19.33	64.00	91.42	54.12
Arabic	12.38	45.23	73.33	13.42	59.23	92.19	49.29
French	20.00	56.66	81.90	18.05	60.95	92.38	54.99
Italian	17.61	52.38	78.09	15.90	57.80	85.04	51.13
Multiple Languages							
English	19.52	55.23	81.42	17.71	67.47	93.52	55.81
Arabic	18.57	49.04	78.52	15.42	60.19	91.23	52.16
French	12.85	53.80	80.95	16.28	59.61	91.61	52.51
Italian	13.33	47.14	75.23	14.66	60.57	92.38	50.55

or jointly to verify the retrieval performance on multilanguage settings.

1) English Language Retrieval: Table II presents a comparison between the proposed transformer-based model and the state-of-the-art retrieval methods published recently, namely, VSE++ [41], SCAN [42], MTFN [43], AMFMN [26], and three models of LW-MCR [27]. For reliable comparison, our model is trained and tested on descriptions in English language only. The results are shown for three RS text-image datasets where the best results are represented in bold. As shown in the Table II, for all metrics, the proposed model outperforms the current state-of-the-art results on all datasets in both text and image retrieval tasks by a considerable margin. Specifically, it achieves an improvement of 11%, 11.66%, and 8.04% on the mR indicator over the AMFMN, which is the second-best method on the RSICD, RSITMD, and UCM datasets, respectively.

By comparing the results of the three datasets, we observe that the results of the RSICD, which is the largest dataset, are lower compared to the other two datasets, and the results of UCM are the higher. This seems natural as it is easier for the retrieval model to find the relevant item in a smaller dataset. It is also interesting to notice, that the retrieval performance in both the text retrieval and image retrieval tasks are very close, which indicates that the matching is effective in both directions. Moreover, the recall shows a significant increase from R@1 to R@5 and from R@5 of R@10. This is because the retrieval task is very challenging that it is difficult to find the best match in the first retrieved results.

Generally, this experiment demonstrates the powerful expressive ability of transformer encoders for image and text and how they are effective in boosting the retrieval performance of RS data.

2) Multilanguage Retrieval: To assess the validity of the retrieval model on the multilanguage retrieval task, we trained it independently on sentences from each of the following languages: English, Arabic, French, and Italian. In addition, we train it jointly on sentences from the four considered languages. Table III shows the retrieval results in terms of R@k and mR on the three datasets. The first part of each table shows the results of the model trained on a single language, and the second part presents the results of learning on multiple languages.

By comparing the results of Tables II and III, it can be seen that the performance of the model trained using single or multiple languages yields better than all of the state-of-the-art methods. In most of the cases, the results show that the best retrieval performance is achieved with the model trained on

multiple languages. Specifically, in the RSICD dataset, the best results are achieved using an English query on the multilanguage model. The same observation can also be noticed on RSITMD, where the best results are achieved by using the English and French queries. The text retrieval task on the UCM dataset is an exception to this, where the results of the multilanguage model is lower than the single-language model trained on French language only. However, the mR of the image retrieval task for the UCM dataset shows slightly better performance over the single language model.

Next, we compare the performance of each language when the model is trained independently on this language and when the model is trained jointly with other languages. As shown in Table III, for English language, we can clearly see that the retrieval performance has increased using English queries when the model is trained on multiple languages compared to a model trained on English language only. For example, on RSICD and RSITMD datasets we can see an improvement in all metrics. However, on the UCM dataset, there is a slight decrease in multilingual performance compared to the single language model for R@1 image retrieval. One possible reason for this is the high similarity between the descriptions given to the images that belong to the same class in this dataset. In general, English has the highest scores on most indicators compared to other languages. This could be explained by the fact that the model is basically pretrained on English language and then finetuned on the other languages.

For the French language, the results on the RSICD dataset show an improvement in all metrics when the model is trained on multilanguage. However, the results of RSITMD dataset show a slight decrease in R@1 and R@5 of the text retrieval. For UCM dataset, the French single-language model has the highest scores in all metrics on the text retrieval. Yet, Table III shows a decrease in the performance in all metrics when the model is trained on multiple languages, especially in the R@1 text retrieval score, where the performance decrease is significant from 20.00% to 12.85%.

The Arabic language has relatively the lowest scores on most indicators in all datasets. The reason could be attributed to the uniqueness of its alphabets, and it is script direction as it is the only Semitic language in the group. The multiple languages model obtained an improved performance over the single language model, when given an Arabic query. In particular, the improvement of the multiple languages model in mR indicator are 1.76%, 4.32%, and 2.87% on the RSICD, RSITMD, and UCM datasets, respectively. On the UCM dataset, the Arabic multiple languages model has improved all the text retrieval, and the image retrieval indicators except the R@10 image retrieval score.

The Italian language has the second-best overall performance on the RSICD dataset for the single and the multiple languages models. The multiple languages model has improved the performance of all metrics compared to the single language model, except the mR and the R@10 text retrieval score. On RSITMD dataset, the Italian language has the highest improved performance on multiple languages learning with an increase of 4.78%, in the mR metric. On the UCM dataset, the single language model outperformed the multiple languages model except for R@5 and R@10 image retrieval.

Generally, the English and French languages have the highest results on the three datasets, and the RSITMD dataset shows the highest improvement percentage compared to other datasets when the multiple language model is used. UCM dataset which is the smallest dataset shows a decrease in multilingual performance compared to the single language model. The possible reason could be the high similarity between the dataset sentences.

To get an intuition of how the image and the text are aligned in the joint embedding space, Fig. 4 shows the features of both the image and text obtained from the two encoders projected into the 2-D space using the t-distributed stochastic neighbor embedding (t-SNE) method. We can observe that the model is able to aggregate data from each modality into well-separated clusters. Furthermore, it attempts to align the embedding of the images and their corresponding texts from the four considered languages to form larger clusters.

Aligning between two different modalities is challenging as we observe a slight shift between the image representations and their corresponding texts and some overlaps between the formed clusters. This is clearer in the RISTMD and the RSICD datasets, which are relatively larger datasets compared to the UCM. We recall that a good alignment is essential to retrieve an image that matches a given query text, and also to retrieve an image that matches a given query text, and also to retrieve the textual description that describes a given image.

D. Visual Explainability

In addition to the quantitative results, we performed another qualitative experiment to better understand the behavior of the attention mechanism employed by the transformer encoders. Fig. 5 shows some examples of textual queries and the retrieved images from RSITMD and RSICD datasets.

On the left, Fig. 5 shows the input query with the textual attention map, and on the right the ground truth image and the retrieved images with the associated visual attention maps. The textual attention highlights the important words that the model pays attention to retrieve the image, and the visual attention maps show the spatial areas of the image that the model focuses on to make the retrieval. It is worth noting that both the textual and visual attention maps are generated by using attention rollout technique [44] on the attention scores for the top layers (form layer 8 up to the last layers of the encoder).

We can initially notice from Fig. 5 that the model can highlight the important keywords in the sentence no matter which language we use for the query. In addition, the visual attention maps of the retrieved images show high responses at the area related to the semantic meaning of the query.

To further analyze the retrieved results, Table IV shows the ground-truth images that match the queries in Fig. 5 and the images retrieved by the model. For the first case of the RSITMD dataset, the model failed to retrieve the “airport_3” image in the first ten results. However, all the retrieved images are from the “airport” category. The textual attention shows a high response

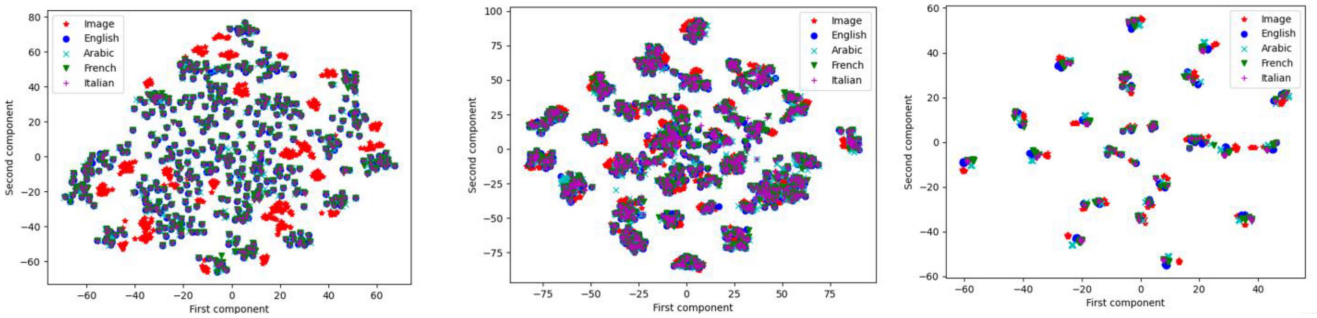
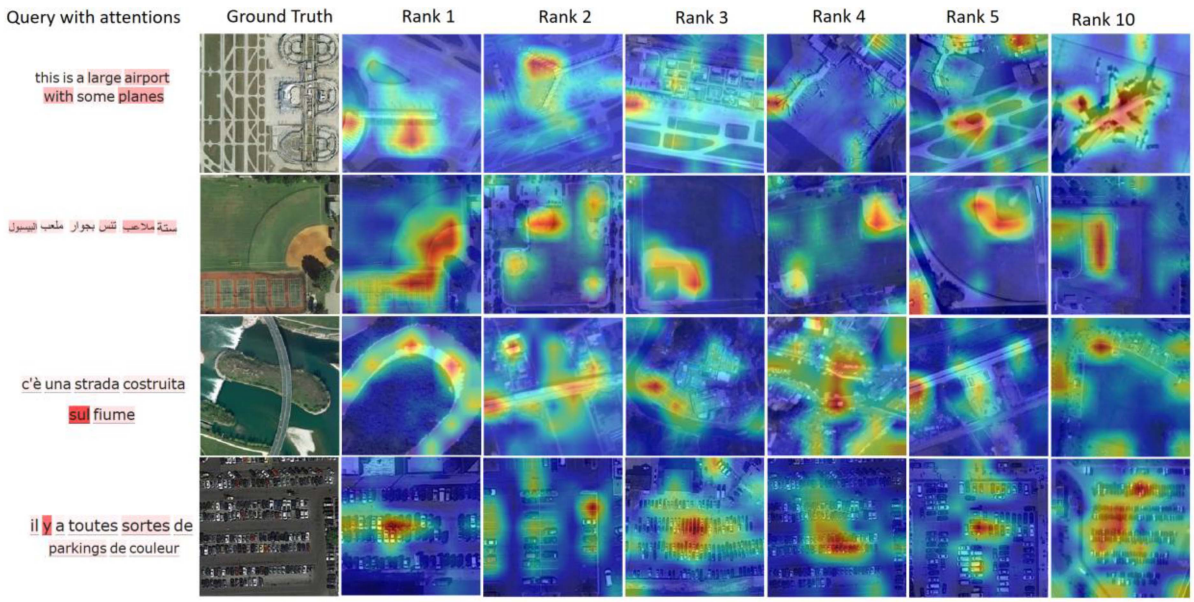
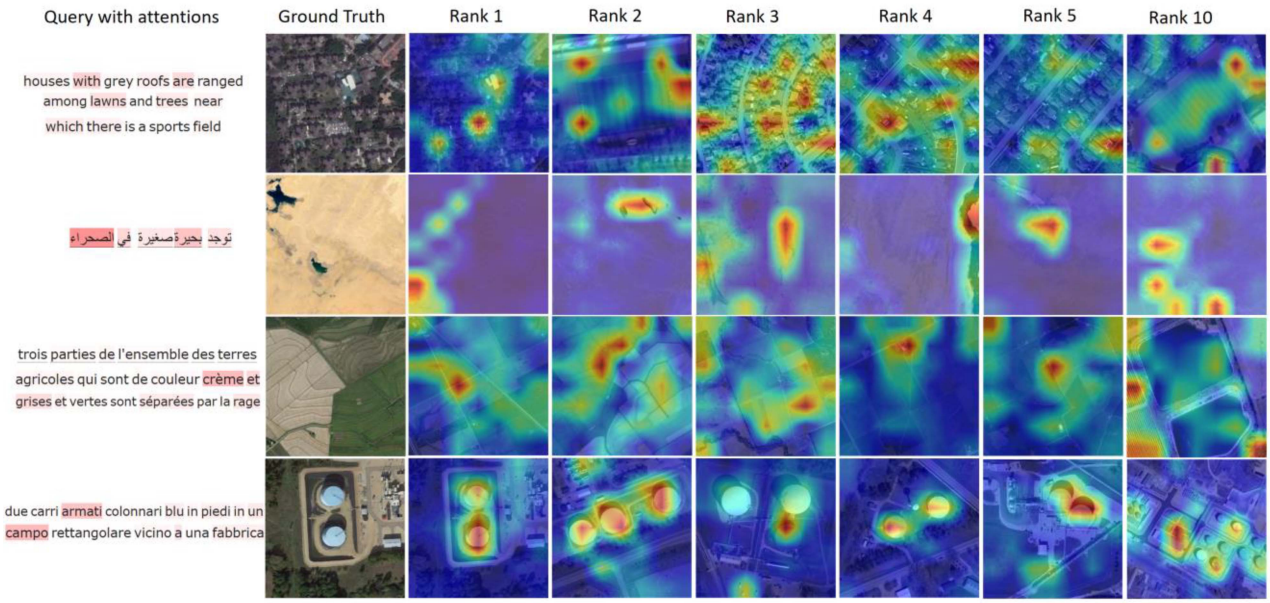


Fig. 4. t-SNE representation of image and text embedding features. (a) RISTMD, (b) RSICD, and (c) UCM datasets.



(a)



(b)

Fig. 5. Visualization of attention on image and query text. (a) RSITMD and (b) RSICD datasets.

TABLE IV
RANKS OF THE RETRIEVED IMAGES OBTAINED FOR RSITMD AND RSICD DATASETS

RSITMD DATASET							
GROUND-TRUTH	RANK OF THE RETRIEVED IMAGES						
	Predicted	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 10
airport_3	Rank 12	airport_10	airport_5	airport_11	airport_9	airport_4	airport_448
baseballfield_26	Rank 1	baseballfield_26	baseballfield_24	baseballfield_21	baseballfield_19	baseballfield_22	playground_1
bridge_55	Rank 17	river_322	bridge_47	river_332	port_288	railwaystation_292	river_329
parking_235	Rank 4	parking_231	parking_232	parking_225	parking_235	parking_224	parking_226
RSICD DATASET							
GROUND-TRUTH	RANK OF THE RETRIEVED IMAGES						
	Predicted	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 10
denseresidential_60	Rank 1	denseresidential_60	playground_41	denseresidential_407	denseresidential_398	denseresidential_397	playground_39
desert_52	Rank 1	desert_52	desert_63	desert_70	desert_53	desert_51	desert_61
farmland_38	Rank 6	farmland_51	farmland_37	farmland_47	farmland_48	farmland_43	pond_410
storagetanks_36	Rank 1	storagetanks_36	storagetanks_44	Storagetanks_45	storagetanks_60	storagetanks_5	storagetanks_353

on the words "planes" and "airport." The visual attention maps focus on an area containing airplanes, which indicates that there is some confusion duo to the high similarity within the samples of the "airport" category. The second case shows a successful retrieval where the image "baseballfield_26" is predicted in the first rank using an Arabic query with the meaning "six tennis courts besides a baseball field". The corresponding attention maps are in line with the semantic meaning of the query as we see the model paying high attention to the baseball diamond and the tennis courts area, and to the words "six," "tennis courts," and "baseball" in the query. The third example shows another failure case in retrieving the image "bridge_55" with a query given in Italian language. The query which means "A bridge built on a river," shows a high attention to the words "sul" and "fiume" which means "on the river." The model wrongly retrieved an image of a river in the first rank and retrieved other images from categories that have similar visual features to the "bridge" category, such as "river," "port," and "railway station." This is also confirmed by the attention maps of the retrieved images that show a high response in the river and railway areas which look similar to the bridge. The last case shows another successful retrieval of the "parking_235" image with a query given in French. The query has the meaning of "There are a number of colored cars in the parking." The target image is correctly retrieved in the fourth rank and all the matched images are from the "parking" category.

For the RSICD dataset, Table IV shows three examples of accurate retrieval of the "denseresidential_60," "desert_52," and "storagetanks_36," where the target image is correctly retrieved in the first rank. In the first example, which shows a query in English, the model retrieved the "denseresidential_60" in the first rank, but the results show some confusion with images from the "playground" category in the second and tenth ranks. The possible reason for this is that many dense residential images contain playground areas. The second example shows a query in Arabic with the meaning "There is a small lake in the desert." This example shows that the model was successful in retrieving the target image and that all the matched images are from the category of the "desert." The textual attention map shows high responses to the words "desert" and "lake" in Arabic, which is consistent with the visual attention maps of the retrieved images that show high focus on the lake areas.

The third example shows the result of a query given in French in which the model retrieved the "farmland_38" image in the sixth rank, but all the retrieved images are from the same "farmland" category. This is because in the RSICD dataset, there is a high intraclass similarity which can be the reason for this confusion. Finally, the last example shows a successful retrieval for a query given in Italian with the meaning "two blue storage tanks in a rectangular field near a factory." Both the textual and the visual attention maps show that the model highlights the information that is relevant to the prediction.

According to the above observations, it can be seen that the model is generally effective in retrieving RS images using queries from different languages. Even though the model fails to retrieve the targeted images in some cases, in many of the fail cases it retrieves images that belong to the same category or a very relevant category. In addition, the model can successfully capture the keywords in the query text and the fine-grained details areas on the retrieved images as well.

V. CONCLUSION

In this article, we have proposed an approach for multilanguage RS text-image retrieval based on language and vision transformers. We used vision and language transformer encoders for generating visual and textual representations, respectively. We have aligned these representations by optimizing a bidirectional contrastive loss related to text-to-image and image-to-text classification. In contrast to previous retrieval methods, which restrict the language of the query to English, the model allows queries to be formulated in English, Arabic, French, and Italian. The qualitative and quantitative results on three RS datasets show that the model is capable of dealing with multilanguage queries while still achieving better performances than the current state-of-the-art methods.

ACKNOWLEDGMENT

The authors would like to thank the support from the Distinguished Scientist Fellowship Program at King Saud University.

REFERENCES

[1] M. Sudmanns et al., "Big earth data: Disruptive changes in earth observation data management and analysis?," *Int. J. Digit. Earth*, vol. 13, no. 7, pp. 832–850, Jul. 2020, doi: 10.1080/17538947.2019.1585976.

[2] Y. Li, J. Ma, and Y. Zhang, "Image retrieval from remote sensing big data: A survey," *Inf. Fusion*, vol. 67, pp. 94–115, Mar. 2021, doi: 10.1016/j.inffus.2020.10.008.

[3] Y. Yang and S. Newsam, "Geographic image retrieval using local invariant features," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 2, pp. 818–832, Feb. 2013, doi: 10.1109/TGRS.2012.2205158.

[4] E. Aptoula, "Remote sensing image retrieval with global morphological texture descriptors," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 3023–3034, May 2014, doi: 10.1109/TGRS.2013.2268736.

[5] X.-Y. Tong, G.-S. Xia, F. Hu, Y. Zhong, M. Datcu, and L. Zhang, "Exploiting deep features for remote sensing image retrieval: A systematic investigation," *IEEE Trans. Big Data*, vol. 6, no. 3, pp. 507–521, Sep. 2020, doi: 10.1109/TBDATA.2019.2948924.

[6] G. Sumbul, S. Nayak, and B. Demir, "SD-RSIC: Summarization-driven deep remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 6922–6934, Aug. 2021, doi: 10.1109/TGRS.2020.3031111.

[7] G. Hoxha and F. Melgani, "A novel SVM-Based decoder for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5404514, doi: 10.1109/TGRS.2021.3105004.

[8] Q. Wang, W. Huang, X. Zhang, and X. Li, "Word–Sentence framework for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10532–10543, Dec. 2021, doi: 10.1109/TGRS.2020.3044054.

[9] X. Zheng, B. Wang, X. Du, and X. Lu, "Mutual attention inception network for remote sensing visual question answering," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5606514, doi: 10.1109/TGRS.2021.3079918.

[10] S. Lobry, D. Marcos, J. Murray, and D. Tuia, "RSVQA: Visual question answering for remote sensing data," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8555–8566, Dec. 2020, doi: 10.1109/TGRS.2020.2988782.

[11] Y. Li, Y. Zhang, X. Huang, and J. Ma, "Learning source-invariant deep hashing convolutional neural networks for cross-source remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6521–6536, Nov. 2018, doi: 10.1109/TGRS.2018.2839705.

[12] U. Chaudhuri, B. Banerjee, A. Bhattacharya, and M. Datcu, "CMIR-NET : A deep learning based model for cross-modal retrieval in remote sensing," *Pattern Recognit. Lett.*, vol. 131, pp. 456–462, Mar. 2020, doi: 10.1016/j.patrec.2020.02.006.

[13] W. Xiong, Z. Xiong, Y. Cui, and Y. Lv, "A discriminative distillation network for cross-source remote sensing image retrieval," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1234–1247, 2020, doi: 10.1109/JSTARS.2020.2980870.

[14] W. Xiong, Y. Lv, X. Zhang, and Y. Cui, "Learning to translate for cross-source remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4860–4874, Jul. 2020, doi: 10.1109/TGRS.2020.2968096.

[15] T.-B. Jiang, G.-S. Xia, Q.-K. Lu, and W.-M. Shen, "Retrieving aerial scene images with learned deep image-sketch features," *J. Comput. Sci. Technol.*, vol. 32, no. 4, pp. 726–737, Jul. 2017, doi: 10.1007/s11390-017-1754-7.

[16] F. Xu, W. Yang, T. Jiang, S. Lin, H. Luo, and G.-S. Xia, "Mental retrieval of remote sensing images via adversarial sketch-image feature learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 7801–7814, Nov. 2020, doi: 10.1109/TGRS.2020.2984316.

[17] G. Mao, Y. Yuan, and L. Xiaoqiang, "Deep cross-modal retrieval for remote sensing image and audio," in *Proc. 10th IAPR Workshop Pattern Recognit. Remote Sens.*, Aug. 2018, pp. 1–7, doi: 10.1109/PRRS.2018.8486338.

[18] Y. Chen and X. Lu, "A deep hashing technique for remote sensing image-sound retrieval," *Remote Sens.*, vol. 12, no. 1, Jan. 2020, Art. no. 1, doi: 10.3390/rs12010084.

[19] Y. Chen, X. Lu, and S. Wang, "Deep cross-modal image–voice retrieval in remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7049–7061, Oct. 2020, doi: 10.1109/TGRS.2020.2979273.

[20] M. Guo, C. Zhou, and J. Liu, "Jointly learning of visual and auditory: A new approach for RS image and audio cross-modal retrieval," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 11, pp. 4644–4654, Nov. 2019, doi: 10.1109/JSTARS.2019.2949220.

[21] H. Ning, B. Zhao, and Y. Yuan, "Semantics-Consistent representation learning for remote sensing image-voice retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4700614, doi: 10.1109/TGRS.2021.3060705.

[22] T. Abdullah, Y. Bazi, M. M. Al Rahhal, M. L. Mekhalfi, L. Rangarajan, and M. Zuair, "TextRS: Deep bidirectional triplet network for matching text to remote sensing images," *Remote Sens.*, vol. 12, no. 3, Jan. 2020, Art. no. 3, doi: 10.3390/rs12030405.

[23] G. Hoxha, F. Melgani, and B. Demir, "Toward remote sensing image retrieval under a deep image captioning perspective," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4462–4475, 2020, doi: 10.1109/JSTARS.2020.3013818.

[24] M. M. A. Rahhal, Y. Bazi, T. Abdullah, M. L. Mekhalfi, and M. Zuair, "Deep unsupervised embedding for remote sensing image retrieval using textual cues," *Appl. Sci.*, vol. 10, no. 24, Jan. 2020, Art. no. 24, doi: 10.3390/app10248931.

[25] Q. Cheng, Y. Zhou, P. Fu, Y. Xu, and L. Zhang, "A deep semantic alignment network for cross-modal image-text retrieval in remote sensing," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4284–4297, 2021, doi: 10.1109/JSTARS.2021.3070872.

[26] Z. Yuan et al., "Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4404119, doi: 10.1109/TGRS.2021.3078451.

[27] Z. Yuan et al., "A lightweight Multi-scale crossmodal text-image retrieval method in remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5612819, doi: 10.1109/TGRS.2021.3124252.

[28] F. Tan, J. Yuan, and V. Ordonez, "Instance-level image retrieval using reranking transformers," pp. 12105–12115, (2021). Accessed: Aug. 7, 2022. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2021/html/Tan_Instance-Level_Image_Retrieval_Using_Reranking_Transformers_ICCV_2021_paper.html

[29] A. El-Nouby, N. Neverova, I. Laptev, and H. Jégou, "Training vision transformers for image retrieval," Feb. 2021, *arXiv:2102.05644*.

[30] W. Xiong, Z. Xiong, Y. Zhang, Y. Cui, and X. Gu, "A deep cross-modality hashing network for SAR and optical remote sensing images retrieval," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 5284–5296, 2020, doi: 10.1109/JSTARS.2020.3021390.

[31] H. Zhou, C. Tian, Z. Zhang, Q. Huo, Y. Xie, and Z. Li, "Multispectral fusion transformer network for RGB-thermal urban scene semantic segmentation," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 7507105, doi: 10.1109/LGRS.2022.3179721.

[32] Z. Zhang, J. Li, C. Tian, Z. Zhong, Z. Jiao, and X. Gao, "Quality-driven deep active learning method for 3D brain MRI segmentation," *Neurocomputing*, vol. 446, pp. 106–117, Jul. 2021, doi: 10.1016/j.neucom.2021.03.050.

[33] A. Vaswani et al., "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, pp. 6000–6010, 2017, *arXiv:1706.03762*.

[34] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *North Amer. Chapter Assoc. Comput. Linguistics*, 2019. doi: 10.48550/arXiv.1810.04805.

[35] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

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

[37] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," 2020, *arXiv:2002.05709*.

[38] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2183–2195, Apr. 2018, doi: 10.1109/TGRS.2017.2776321.

[39] B. Qu, X. Li, D. Tao, and X. Lu, "Deep semantic understanding of high resolution remote sensing image," in *Proc. Int. Conf. Comput., Inf. Telecommunication Syst.*, Jul. 2016, pp. 1–5, doi: 10.1109/CITS.2016.7546397.

[40] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for Land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, 2010, pp. 270–279, doi: 10.1145/1869790.1869829.

[41] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "VSE++: Improving visual-semantic embeddings with hard negatives," Jul. 2017, Accessed: Nov. 23, 2021. [Online]. Available: https://arxiv.org/abs/1707.05612v4

[42] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," Jul. 2018, *arXiv:1803.08024* [cs], Accessed: Nov. 23, 2021. [Online]. Available: http://arxiv.org/abs/1803.08024

[43] T. Wang, X. Xu, Y. Yang, A. Hanjalic, H. T. Shen, and J. Song, "Matching images and text with multi-modal tensor fusion and re-ranking," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 12–20, doi: 10.1145/3343031.3350875.

[44] S. Abnar and W. Zuidema, "Quantifying attention flow in transformers," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 4190–4197, doi: 10.18653/v1/2020.acl-main.385.

Mohamad M. Al Rahhal (Senior Member, IEEE) received the B.Sc. degree in computer engineering from Aleppo University, Aleppo, Syria, in 2002, the M.Sc. degree in information technology from Hamdard University, New Delhi, India, in 2005, and the Ph.D. degree in computer engineering from King Saud University, Riyadh, Saudi Arabia, in 2015.

From 2006 to 2012, he was a Lecturer with Al-Jouf University, Sakakah, Saudi Arabia. He is currently an Associate Professor with the College of Applied Computer Engineering, King Saud University. His research interests include signal/image medical analysis, remote sensing, and computer vision.

Yakoub Bazi (Senior Member, IEEE) received the State Engineer and M.Sc. degrees in electronics from the University of Batna, Batna, Algeria, in 1994 and 2000, respectively, and the Ph.D. degree in information and communication technology from the University of Trento, Trento, Italy, in 2005.

From 2000 to 2002, he was a Lecturer with the University of M'Sila, M'Sila, Algeria. In 2006, he joined the University of Trento, as a Postdoctoral Researcher. From 2006 to 2009, he was an Assistant Professor with the College of Engineering, Al-Jouf University, Sakakah, Saudi Arabia. He is currently a Full Professor of Computer Engineering with the College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia. He is a referee for several international journals. His research interests include remote sensing, signal/image medical analysis, and computer vision.

Dr. Bazi is an Associate Editor of *IEEE Geoscience and Remote Sensing Letters*, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, and *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.

Norah A. Alsharif received the B.Sc. degree from Taif University, Taif, Saudi Arabia, in 2018, and the M.Sc. degree from King Saud University, Riyadh, Saudi Arabia, in 2022, both in computer engineering.

Her research interests include artificial intelligence and remote-sensing image analysis.

Laila Bashmal (Graduate Student Member, IEEE) received the B.S. degree in computer science from the University of Dammam, Dammam, Saudi Arabia, in 2011, and the M.Sc. degree in computer engineering in 2018 from King Saud University, Riyadh, Saudi Arabia, where she is currently working toward the Ph.D. degree in computer engineering.

Her research interests include machine learning and image processing with applications to remote sensing image analysis.

Naif Alajlan (Senior Member, IEEE) received the bachelor's and master's degrees in electrical engineering from the Electrical Engineering Department, King Saud University, Riyadh, Saudi Arabia, in 1998 and 2002, respectively, and the Ph.D. degree in computer engineering from the Electrical and Computer Engineering Department, University of Waterloo, ON, Canada, in 2006.

He is currently a Full Professor of AI in Computer Engineering Department, King Saud University, Saudi Arabia. He has authored and coauthored more than 130 referred journal papers in machine learning, pattern recognition, biomedical engineering, remote sensing, and other fields. In 2009, he founded ALISR, a research lab in intelligent systems where several research and consultation projects were conducted with public and private organizations.

Farid Melgani (Fellow, IEEE) received the State Engineer degree in electronics from the University of Batna, Batna, Algeria, in 1994, the M.Sc. degree in electrical engineering from the University of Baghdad, Baghdad, Iraq, in 1999, and the Ph.D. degree in electronic and computer engineering from the University of Genoa, Genoa, Italy, in 2003.

He is a Full Professor of Telecommunications with the Department of Information Engineering and Computer Science, University of Trento, Trento, Italy, where he teaches pattern recognition, machine learning, and digital transmission. He is the Head of the Signal Processing and Recognition Laboratory, the Coordinator of the Doctoral School in Industrial Innovation, and the Dean of Undergrad and Grad Studies with the same department. He has coauthored more than 250 scientific publications. His research interests include remote sensing, signal/image processing, pattern recognition, machine learning, and computer vision.

Dr. Melgani is currently an Associate Editor of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, *International Journal of Remote Sensing*, and *IEEE Journal on Miniaturization for Air and Space Systems*.