# A Deep Semantic Alignment Network for the Cross-Modal Image-Text Retrieval in Remote Sensing

Qimin Cheng, Yuzhuo Zhou, Peng Fu, Yuan Xu, and Liang Zhang

***Abstract*—Because of the rapid growth of multimodal data from the internet and social media, a cross-modal retrieval has become an important and valuable task in recent years.The purpose of the cross-modal retrieval is to obtain the result data in one modality (e.g., image), which is semantically similar to the query data in another modality (e.g., text).In the field of remote sensing, despite a great number of existing works on image retrieval, there has only been a small amount of research on the cross-modal image-text retrieval, due to the scarcity of datasets and the complicated characteristics of remote sensing image data. In this article, we introduce a novel cross-modal image-text retrieval network to establish the direct relationship between remote sensing images and their paired text data. Specifically, in our framework, we designed a semantic alignment module to fully explore the latent correspondence between images and text, in which we used the attention and gate mechanisms to filter and optimize data features so that more discriminative feature representations can be obtained. Experimental results on four benchmark remote sensing datasets, including UCMerced-LandUse-Captions, Sydney-Captions, RSICD, and NWPU-RESISC45-Captions, well showed that our proposed method outperformed other baselines and achieved the state-of-the-art performance in remote sensing image-text retrieval tasks.**

***Index Terms*—Convolutional neural network (CNN), cross-modal remote sensing image-text retrieval, recurrent neural network (RNN), semantic alignment.**

## I. Introduction

With the rapid development of Earth observation technology, the quantity and quality of remote sensing data have increased rapidly. Aiming at the complicated characteristics such as diversity, complexity, and massiveness of remote sensing image data, the predecessors have conducted a great number of researches on the remote sensing image retrieval task [1]–[6]. Besides, the types of data (such as text, images, and videos) are manifold now, and these different types of data are referred to as multimodal data. Data of different modalities can be used to describe the same entity (see Fig. 1). Instead of retrieving in unimodal data, people are more inclined to search for the required information in multimodal data with richer semantics. For instance, given a query textual description, people may want to find out all the semantically similar hyperspectral images or videos that are captured by satellite sensors. Furthermore, cross-modal retrieval technology can mine effective information and has broad application prospects in many fields, such as early warning of disasters and resource management. In fact, satisfactory accuracy has been observed in the cross-modal retrieval of natural images [7]–[9]. However, it is difficult to implement an effective and efficient cross-modal retrieval of remote sensing images since these images have complicated characteristics such as multiscale, small targets, high resolution, and lack of annotated information.

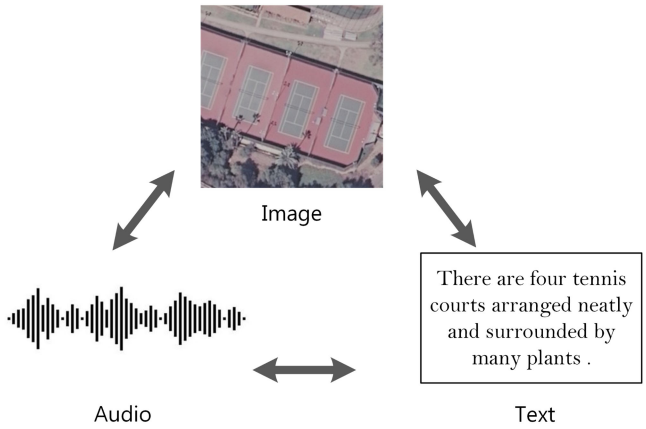


Fig. 1. Here, we give an example to show that a single entity usually requires data from multiple modalities to properly describe it.

In recent three years, there have been a small amount of cross-modal retrieval researches in the field of remote sensing. For example, Chaudhuri *et al.* [10] studied the remote sensing cross-modal retrieval between multispectral images and panchromatic images. Besides, based on label annotations, they also researched on the cross-modal retrieval between very-high-resolution (VHR) images and speech; Lu *et al.* [11] designed a

Manuscript received November 18, 2020; revised January 1, 2021 and March 16, 2021; accepted March 31, 2021. Date of publication April 5, 2021; date of current version May 3, 2021. This work was supported in part by the National Key Research, and Development Program of China under Grant 2018YFB0505401; in part by the National Natural Science Foundation of China under Grant 41771452; and in part by the Director Fund of Institute of Remote Sensing, and Digital Earth under Grant Y5SJ1500CX. *(Corresponding author: Qimin Cheng.)*

Qimin Cheng, Yuzhuo Zhou, and Yuan Xu are with the School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: chengqm@hust.edu.cn; 723965376@qq.com; yuanxu96@hust.edu.cn).

Peng Fu is with the Department of Plant Biology, University of Illinois at Urbana-Champaign, Champaign, IL 61801 USA (e-mail: pengfu@illinois.edu).

Liang Zhang is with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430072, China (e-mail: szhangliang2016@whu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2021.3070872

deep visual-audio network (DVAN), which was used to find the latent relationship between image and audio; Chen *et al.* [12] proposed a deep image-voice retrieval approach in the field of remote sensing, in order to explore the semantic information of remote sensing images in the multiscale level, thereby generating hash codes that occupy little memory space and can achieve rapid retrieval. However, in practical application scenarios, there are still some unfavorable factors such as improper pronunciation and blurring of words that are difficult to overcome and can directly reduce the accuracy of the image-voice retrieval. Therefore, in order to express semantic information and implement the cross-modal retrieval more accurately, text description is still necessary. Hence, we aim to research on the remote sensing cross-modal retrieval between images and text.

All of the aforementioned methods in the cross-modal retrieval in remote sensing have achieved an appealing performance. However, these methods mapped the features of different modalities into a latent embedding space, treating different kinds of semantics (e.g., words with different parts of speech, such as nouns, verbs, adjectives, etc.) equally, and then, implementing semantic alignment without exploring the subtle difference between them. For example, the preposition "of" will be assigned the same importance score as the noun "airport," ignoring the different importance of semantics they contain. As a consequence, it is difficult to model the fine-grained relationships between different modalities, and thus, the properties such as accuracy and efficiency of the cross-modal retrieval model are degraded.

Actually, in the field of natural images, many cross-modal retrieval frameworks with an attention mechanism have been proposed to explore the latent relationships of different modalities. For example, Huang *et al.* [13] used a multimodal attention mechanism based on context in order to generally explore the object-level saliency maps between images and sentences. Gu *et al.* [14] proposed a hashing method for the cross-modal retrieval task, which is called AGAH. They used an attention mechanism to generate discriminative features, guided by adversarial learning, thus ensuring the robustness of the architecture. These works have achieved a high retrieval accuracy on several benchmark natural image datasets, and show that the attention mechanism can quickly pick out and retain salient information by distributing different attention scores to each word and image region, and thus, can help implement the semantic alignment more accurately.

Motivated by this idea, in this article, we proposed to use the attention mechanism to explore fine-grained semantic correspondence between remote sensing images and text. Besides, inspired by [15], we also designed a gate function in our proposed semantic alignment module (SAM), in order to make the visual and textual features more discriminative. Moreover, since the size of the available remote sensing image dataset is usually much smaller than that of the natural image dataset, overfitting will be seen if fine tuning the pretrained CNN on remote sensing datasets. To solve this problem, a pretrained CNN to extract visual features, without fine tuning, is highly recommended. The experimental results on four benchmark remote sensing datasets highlight the advantages of our method and well demonstrate that our proposed method can achieve a high retrieval accuracy between remote sensing images and text.

The main contributions of this article can be summarized as follows.

1) We proposed a cross-modal image-text retrieval network for remote sensing to deal with the complexity of semantics and refine the correspondence between remote sensing images and text. To the best of our knowledge, this is the first attempt to research on the cross-modal image-text retrieval in the field of remote sensing.
2) The SAM introduced in this article is designed to discover and strengthen the underlying semantic relationships between remote sensing images and text by updating the visual and textual features. Specifically, in this module, an attention mechanism is employed to enhance the corresponding relationships between images and text, and then, we design a gate function to filter out as much unnecessary information as possible, and finally, obtain discriminative visual and textual features.

The remainder of this article is organized as follows: In Section II, we summarized the development and the related works of this field in recent years. In Section III, we elaborated the details of each process in our method. Experimental results are shown and analyzed in Section IV. Finally, Section V concludes this article.

## II. Related Work

In this section, we primarily summarize and analyze the current research status from the following three aspects: 1) deep learning and neural network; 2) cross-modal image-text retrieval; and 3) semantic alignment between different modalities.

### A. Deep Learning and Neural Network

Before the advent of deep neural networks, the traditional hand-crafted features have been used for the cross-modal retrieval. For instance, Michael *et al.* [16] proposed a method to implement the distance learning. Specifically, they proposed to map the input data of arbitrary modalities into a latent common Hamming space, thereby converting the problem of the cross-modal retrieval into a problem of binary classification, and the output of the binary classifier determined whether these two query data were a positive pair or negative pair. Moreover, Michael *et al.* [16] suggested that this model should be trained by a boosting algorithm. In 2011, Kumar *et al.* [17] proposed a principled method for the multimodal retrieval, which was designed to encode the input data of different modalities into similar codes, thus the similarity between the heterogeneous data can be conveniently obtained. Zhang *et al.* [18] introduced a method to explore the correspondences between heterogeneous data by clustering data from the same modality with the help of the affinity propagation clustering algorithm, and strengthen the pair-wise corresponding relationships based on the canonical correlation analysis. All of these aforementioned studies have played an essential role in the field of computer vision. However, compared with automatic features learned by deep convolutional neural network (CNN), such hand-crafted features are incapable of describing the full content of visual data, thus failing to obtain satisfactory performance in computer vision tasks. Moreover,

there are many defects of hand-crafted features, such as large storage space and time cost.

Since 2006, deep CNNs have been widely used to solve manifold problems of computer vision [19], mainly because the CNN can imitate the human visual perception mechanism to effectively extract the visual features from the original images. For example, a deep CNN that has been pretrained on a large dataset (such as ImageNet [20]) can be used as a general feature extractor and applied on other datasets, and then, the convolutional layer features or fully connected layer features of the CNN model can be extracted to implement image feature extraction. Or, in order to achieve the best performance on the target dataset, the pretrained CNN model can also be migrated to the target dataset for fine tuning, and this method can effectively update the network weights in small increments, and thus, making the model more suitable for the target dataset. As an example, Radoi *et al.* [21] proposed a novel framework for multilabel classification of multispectral remote sensing images, which used a pretrained CNN to extract image feature. Yang *et al.* [22] proposed a hashing method to implement the cross-modal retrieval, which is called PRDH. They constructed an end-to-end architecture to extract features from heterogeneous data concurrently by applying CNNs, and then, learned their corresponding hashing codes for distance measurement and cross-modal retrieval. Particularly, they used the VGG-F architecture to fine tune their CNN module.

Similarly, the CNN can also extract text features, and it has gained great success in the natural language processing (NLP) field. Zhang *et al.* [23] proposed a creative network for text classification, in which they applied a character-level CNN to learn the textual representation. Kim *et al.* [24] used a pretrained CNN for sentence-level classification tasks. However, the ability of the CNN to capture the key information of text is still limited.

Another commonly used neural network is recurrent neural network (RNN). The RNN has a strong capacity to extract sequence features, so it has been widely used for text classification. However, since the RNN is prone to gradient disappearance and gradient explosion, in 1997, Hochreiter *et al.* [25] proposed a method based on gradient, called long short-term memory (LSTM), which was proved to be extremely convenient and efficient. In 2014, Cho *et al.* [26] proposed a simpler gate recurrent unit (GRU) network, which consists of two RNNs, and they showed that the proposed GRU can learn semantically smooth representations of text.

Benefiting from the prompt development of CNNs and RNNs, there has been extremely significant progress in the cross-modal retrieval. For example, Guo *et al.* [27] designed a DVAN for the image-audio retrieval in the field of remote sensing. Cao *et al.* [28] proposed to use the CNN and LSTM to obtain unified hashing codes in a separate way from images and text, and then, apply them into a cross-modal retrieval task. Niu *et al.* [29] introduced a hierarchical structured RNN, which was called hierarchical multimodal LSTM (HM-LSTM), to project the complete sentences and images into the latent common space by applying the dense visual-semantic embedding and project the complete textual phrases and salient image regions into the common space. Extensive experiments show that both CNN and RNN have a strong ability to extract feature representations from images and text, and prove capable of most computer vision tasks.

### B. Cross-Modal Image-Text Retrieval

Most existing image-text retrieval algorithms are instance-based methods, that is, to retrieve the predefined instances. For example, Wang *et al.* [30] proposed a novel network for the image-text retrieval, named MTFN. Instead of learning a latent common space for every image-sentence pair, they designed a similarity function to measure the distance between the input image and sentence accurately, and they trained this network with a ranking-loss function. However, in actual application scenarios, there are various factors that can affect the visual effect of instance objects, such as angle, illumination, and location. As a result, images containing the same one object may look very different, and consequently, the accuracy of the retrieval may decrease.

Beyond that, there are some image-text retrieval algorithms, which are based on class labels. For example, given a sentence as query, people may expect to retrieve all the images that are semantically similar to the query sentence in the dataset. Although these images are not exactly the same, they have a common label, which means they are similar to some extent. In this way, the labels-based retrieval is more likely to give users the desired results. Mason *et al.* [7] proposed a graphic retrieval model, which regarded the task as a summary extraction task and used scene attributes as the visual representation of the image, and reranked candidate text descriptions based on information of class labels to obtain the final textual retrieval result of the test image.

Due to the application of deep vision features, the retrieval accuracy was also greatly improved. Devlin *et al.* [31] proposed a deep-learning-based image-text retrieval model, which used a deep CNN to extract visual features of the query images and retrieve similar images in the visual space. Then, the candidate text descriptions were encoded by the RNN and reranked according to the semantic features of the text, thereby obtaining the final retrieval result. Socher *et al.* [8] used a Dependency Tree RNN (DT-RNN) to construct the semantic representation of text and a nine-layer neural network to extract the visual feature representation from image. As a result, the semantic features of text and the visual features of images were mapped to the same cross-modal embedding space for semantic alignment. In 2014, Karpathy *et al.* [9] extended and improved the model proposed by Socher *et al.* [8]. Different from Socher *et al.* [8], their study did not directly map the entire image and sentence to the cross-modal embedding space, but rather mapped the more fine-grained image features and text features to a latent cross-modal embedding space. In other words, the target object of image and the word of sentence were mapped to the same cross-modal space so that the model proposed by Socher *et al.* [8] was greatly improved. The all aforementioned methods involve integrating deep learning in the cross-modal image-text retrieval and suggest that deep learning can learn more abstract and discriminative features.

All of the aforementioned methods have achieved an appealing retrieval performance. However, due to the complicated characteristics of remote sensing images such as diversity, complexity, and massiveness, the methods that are proposed for natural images cannot well establish relationships between remote sensing images and text, and thus, fail to get a satisfactory performance on remote sensing datasets. Therefore, we mainly focus on how to explore the intrinsic correlations between remote sensing images and text descriptions, i.e., implement efficient semantic alignment between these two different modalities.

### C. Semantic Alignment

In recent years, most image-text retrieval algorithms have been implemented based on joint cross-modal embedding space, that is, the visual features of image and the semantic features of text are aligned into a common cross-modal hamming space to implement the cross-modal image-text retrieval in an accurate way. Therefore, how to achieve the alignment between image features and text semantic features in the cross-modal embedding space is the key issue of the entire image-text retrieval task. Recently, more and more image–text alignment models introduced an RNN architecture to implement text encoding and the construction of cross-modal embedding space. For example, Lee *et al.* [32] proposed a model, named SCAN, to implement an image-text retrieval. Specifically, in this method, the input region of image and word of sentence were used as the context of each other, and then, the semantic alignment relationship as well as the similarity score between them were explored. Chen *et al.* [33] proposed a model, called IMRAM, in which they creatively proposed an alignment mechanism iterated by multiple steps, in order to explore the correspondences between visual regions and words. Kiros *et al.* [34] proposed a visual-semantic embedding (VSE) model, which used an LSTM network to encode sentences to obtain text semantic feature representation, and used the CNN to extract visual feature representation of images. Under the supervision of a two-way hierarchical loss function, two mapping matrices are learned so that the two cross-modal features can be mapped into the same one embedding space for alignment. On the basis of the VSE proposed by Kiros *et al.* [34], Faghri *et al.* [35] proposed the enhanced VSE (VSE++), which improved a two-way loss function in the VSE model and introduced the concept of the most difficult negative sample so that the final image retrieval accuracy was greatly improved, and the current optimal performance of the current image retrieval was realized.

Besides, people have been attempted to introduce an attention mechanism to learn an aligned cross-modal embedding space. With the help of the attention mechanism, each region of the image and word of the sentence are assigned different weight scores, which depends on the importance of their semantic, thereby allocating different attention to different image regions or words. In 2014, the GoogleMind team first proposed to use the content attention mechanism for image classification, which effectively improved the accuracy of image recognition [36]. Subsequently, Bahdanau *et al.* [15] applied an attention mechanism into the NLP task, and successfully improved the accuracy of translation. Therefore, in addition to using the RNN to obtain textual features, people also began to consider adding attention mechanism to the semantic alignment models. Ba *et al.* [37] proposed an attention mechanism model based on the RNN, which focused on different regions related to image classification at different moments, so as to achieve the detection of multiple targets in the image. Huang *et al.* [13] also proposed an attention mechanism model, which used context information for encoding and decoding, and applied it to the image–text alignment model to compute the similarity between image regions and words. Nam *et al.* [38] applied the attention mechanism to both image and text to capture the fine-grained correspondence between image and text, that is, to achieve the semantic alignment between regions of image and words of sentence. Thus, the visual semantic alignment between different modalities could be realized. Wang *et al.* [39] proposed a model named PFAN to explore the relationship between the image region and blocks, then they used an attention mechanism to generate a particular feature, called position feature, which contains valuable spatial information of the image region. Furthermore, the position feature is capable to strengthen the correspondence between the image-sentence pair. Generally speaking, all of the aforementioned methods have achieved significant performance improvements in experiments, which demonstrated the effectiveness of the attention mechanism in semantic alignment models. Inspired by this idea, we proposed to use the attention and gate mechanism to implement semantic alignment between remote sensing images and text.

## III. Methodology

In this article, we intend to solve the issue of the cross-modal retrieval for the remote sensing image and text. We show the whole structure of our proposed deep image–text semantic alignment network in Fig. 2, which mainly includes the following three parts: 1) extraction of remote sensing image features; 2) extraction of text features; and 3) an SAM. For each input remote sensing image, we extract the high-level image representation by an image subnetwork, and as for the text part, we extract the word-level semantic feature of each sentence by an RNN. Throughout the whole architecture, our proposed SAM plays a role similar to the fusion layer in filtering out information that may reduce the discrimination of visual or textual features, and thereby, optimizing the representations, which is beneficial to improve the performance of the cross-modal retrieval.

And the whole retrieval process can be roughly divided into the following four steps:

1) enter the query image (or query text);
2) input the query image (or query text) into the trained model to obtain the visual feature (or text feature) of the query image (or query text);
3) use the obtained feature to calculate the similarity score between the query image (or query text) and all the samples in the test set; and
4) the samples in the test set are ranked according to the similarity score and returned as the search result.

Next, we will provide details of our proposed model for the cross-modal image-text retrieval in remote sensing. First,

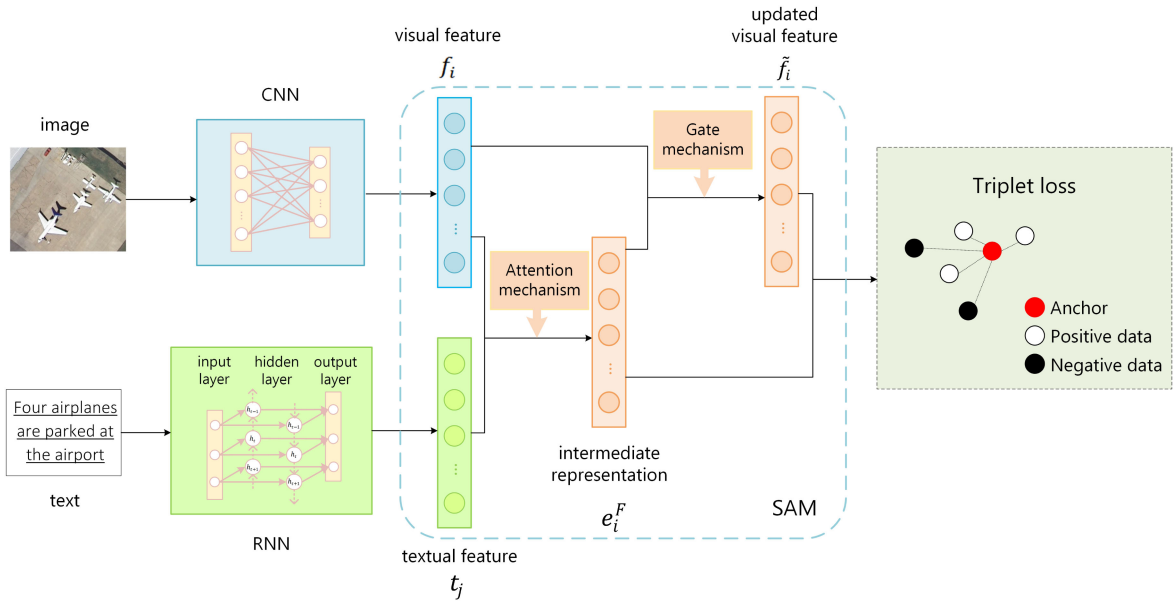


Fig. 2. Intuition of our proposed deep image-text semantic alignment network. The visual feature $f_i$ is extracted from a CNN, and the textual feature $t_j$ is extracted from an RNN. In our proposed SAM, we use the attention mechanism to distribute different attention score to each word, with which the intermediate representation $e_i^F$ can be generated. Besides, we design a gate function to filter out as much unnecessary information as possible, and finally, we can obtain the updated feature $\tilde{f}_i$, which has become more discriminative than $f_i$.

we introduce the image feature encoding in Section III-A, and we introduce text feature encoding in Section III-B. Then, we describe in detail about our proposed SAM in Section III-C. Finally, we discuss the objective function in Section III-D.

### A. Image Feature Representation

To implement the cross-modal image-text retrieval in the field of remote sensing, first we need to extract the image features. We choose to use a multilayer network architecture to extract features layer by layer from raw images. It has been proven that the CNN is capable to extract discriminative image features, because it expresses the high-level semantic information of images better than traditional hand-crafted features. Suppose that a set of training images are available: $I = \{x_1, x_2, \ldots, x_P\}$. Each image $x$ in $I$ is fed to the deep CNN that is pretrained on ImageNet, and $F = \{f_1, f_2, \ldots, f_K\}, f_i \in \mathbb{R}^D$ is the set of resulting visual features from the input remote sensing image.

### B. Text Feature Representation

After a sentence with $N$ words is input, each word is encoded into a one-hot vector that indicates the index in the vocabulary, denoted as $w_i$. Then, the one-hot vector $w_i$ is embedded into a 300-dimensional vector $y_i$ by a linear mapping function $y_i = W_y w_i, i \in \{1, \ldots, N\}$, where the $W_y$ is the embedding matrix. Then, the word vector $y_i$ is fed into a bidirectional GRU [40] to summarize the context information of sentence from the forward and the backward directions, respectively. Therefore, we can map the word vector into the word-level feature. Particularly, one bidirectional GRU contains a forward GRU, which reads the sentence word by word forwards from $w_1$ to $w_N$ as

$$\overrightarrow{h}_i = \overrightarrow{\text{GRU}}(\overrightarrow{h}_{i-1}, y_i) \tag{1}$$

as well as a backward GRU, which reads the sentence backwards from $w_N$ to $w_1$ as

$$\overleftarrow{h}_i = \overleftarrow{\text{GRU}}(\overleftarrow{h}_{i+1}, y_i) \tag{2}$$

where $\overrightarrow{h}_i$ represents the hidden state and is generated by the forward GRU, and $\overleftarrow{h}_i$ is the hidden state generated by the backward GRU. Then, we average these two hidden states of GRU, in order to obtain the textual feature $t_i$ as

$$t_i = \frac{\overrightarrow{h}_i + \overleftarrow{h}_i}{2}, i \in \{1, \ldots N\} \tag{3}$$

Finally, we obtain the word-level feature set for the sentence

$$T = \{t_i | i = 1, \ldots, N, t_i \in \mathbb{R}^D\} \tag{4}$$

where each $t_i$ contains the context information of the corresponding word $w_i$, and the dimension $D$ of the text feature is 2048, which is the same as the image feature.

### C. Semantic Alignment Module (SAM)

Our SAM expects two inputs: one is the image feature set $F = \{f_1, \ldots, f_K\}, f_i \in \mathbb{R}^D$, in which each region of the input remote sensing image is encoded into the image feature $f_i$; and the other one is the textual feature set $T = \{t_1, \ldots, t_N\}, t_i \in \mathbb{R}^D$, in which each word of the sentence is encoded into the textual feature $t_i$. The output of our SAM is a similarity score, which is derived from the similarity of the input image-text pair. While computing the similarity score, our SAM summarizes context information in sentence for each feature $f_i$ in image, or vice

**Algorithm 1:** Optimization Algorithm for Our Network (I2T).

**Input :** Image set $F$, Text set $T$.
**Output:** Weight matrix $W_1$ and $W_2$, bias $b_1$ and $b_2$.
**repeat**
1. Compute the cosine similarity score $sim(i,j)$ between $f_i$ and $t_j$;
2. Generate intermediate representation $e_i^w$;
3. $g_i^r \leftarrow \sigma[l_i^r(W_1, b_1)]$
   $c_j^r \leftarrow \sigma[l_i^r(W_2, b_2)]$;
4. $\tilde{f}_i \leftarrow (1\text{-g}_i^r) \cdot f_i + g_i^r \cdot c_i^r$;
5. Compute the matching score $S(F,T)$ between $\tilde{f}_i$ and $e_i^w$;
6. Minimize the triplet loss function $L$ by SGD;

**until** *convergence or max training iter T is reached*;

versa. We introduce two different forms of semantic alignment below: the image–text semantic alignment and the text–image semantic alignment. We summarize the learning process of the image–text semantic alignment network in Algorithm 1.

*1) Image-Text Semantic Alignment:* The overview of the image-text semantic alignment is shown in Fig. 2. First of all, for each region of the input remote sensing image, our SAM pays attention to words of the sentences, which are semantically associated with the image region. Then, the SAM calculates the attention score of each word. The higher the attention score is, the more semantically related the word is to the image region. Specifically, given a sentence with $N$ words and a remote sensing image with $K$ regions, first, we calculate the cosine similarity score for all possible region-word pairs as

$$\text{sim}(i,j) = \frac{f_i^T t_j}{\|f_i\| \cdot \|t_j\|} \; \forall i \in [1,K] \, \forall j \in [1,N] \tag{5}$$

where the $\text{sim}(i,j)$ means the similarity between the $i$th image region and the $j$th word. In order to speed up the process of training and enable our algorithm to rapidly get the best solution, we further normalize the similarity score $\text{sim}(i,j)$ as

$$\overline{\text{sim}}(i,j) = \frac{[\text{sim}(i,j)]_+}{\sqrt{\sum_i = 1^K [\text{sim}(i,j)]_+^2}} \tag{6}$$

$$[\text{sim}(i,j)]_+ = \max\{[\text{sim}(i,j)], 0\}. \tag{7}$$

In order to get a feature representation with more valuable information and explore the fine-grained correspondence between image regions and words, we construct different attention weights according to the similarity between them, and then, distribute them to each word-level textual feature $t_j$, thereby obtaining the sentence-level attended textual feature $e_i^w$, where the subscript $i$ indicates that the sentence feature $e_i^w$ is related to the $i$th image region

$$e_i^w = \sum_{j=1}^{N} t_j \cdot \left( \frac{\exp(\alpha \cdot \overline{\text{sim}}(i,j))}{\sum_{j=1}^{N} \exp(\alpha \cdot \overline{\text{sim}}(i,j))} \right) \tag{8}$$

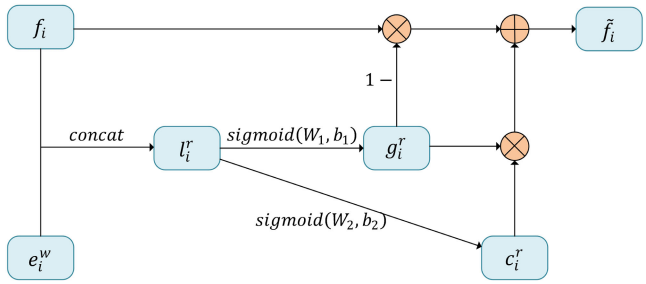


Fig. 3. Specific architecture of our designed SAM.

where $i \in [1,K]$, $e_i^w \in \mathbb{R}^D$. The hyperparameter $\alpha$ is the inverse temperature parameter of the softmax function to control the distribution of feature attention.

In order to filter out as much information as possible that may reduce the discrimination of features, after constructing the attended representation $e_i^w$, we propose to make further improvements in the image features.

First of all, we define a function as an intermediate representation, which concatenates two inputs $f_i$ and $e_i^w$ as following:

$$l_i^r(W,b) = \text{concat}(f_i, e_i^w) \cdot W + b. \tag{9}$$

The intermediate representation $l_i^r(W,b)$ will be used later to generate the update gate and the new memory cell, and the superscript $r$ indicates the region of an image.

Then, inspired by the gate mechanism in GRU, we design a feature updating function to further optimize the visual feature, which contains an update gate $g_i^r$ and a new memory cell $c_i^r$ as

$$\tilde{f}_i = (1 - g_i^r) \cdot f_i + g_i^r \cdot c_i^r \tag{10}$$

the $g_i^r$ and $c_i^r$ are defined as following according to (9):

$$\begin{cases} g_i^r = \text{sigmoid}[l_i^r(W_1, b_1)] \\ c_i^r = \text{sigmoid}[l_i^r(W_2, b_2)] \end{cases} \tag{11}$$

where $W_1, b_1, W_2$, and $b_2$ are hyperparameters to be learned during the process of training, and $\tilde{f}_i$ is the updated image region feature. $g_i^r$ performs as the update gate to discard the trivial information, and $c_i^r$ performs as the new memory cell to retain the discriminative information. Specifically, the expression $(1 - g_i^r) \cdot f_i$ represents the "selectively forgetting" of the original feature $f_i$. The closer $g_i^r$ is to 1, the more original information in $f_i$ is forgotten. And correspondingly, the expression $g_i^r \cdot c_i^r$ represents the "selectively remembering" of $c_i^r$. Thereby, the updated visual feature $\tilde{f}_i$ is obtained. The detailed structure of the image-text SAM is shown in Fig. 3.

Finally, we define a similarity metric function $S(F,T)$ to calculate the matching score between the image and sentence as following:

$$S(F,T) = \frac{1}{K} \sum_{i=1}^{K} \frac{\tilde{f}_i^T e_i^w}{||\tilde{f}_i|| \cdot ||e_i^w||}. \tag{12}$$

Specifically, if the $i$th image region is not relevant to the text, the updated region feature $\tilde{f}_i$ will contribute little to the similarity score $S(F,T)$, and vice versa. Therefore, the similarity score

determines how important the remote sensing image is to the sentence.

*2) Text–Image Semantic Alignment:* Similar to the image–text semantic alignment, we first compute the attended image representation $e_j^r$ with respect to the related word-level textual feature $t_j$ as following:

$$e_j^r = \sum_{i=1}^{K} f_i \cdot \left( \frac{\exp(\alpha \cdot \overline{\text{sim}}(i,j))}{\sum_{i=1}^{K} \exp(\alpha \cdot \overline{\text{sim}}(i,j))} \right) \tag{13}$$

where the subscript $j$ indicates that the attended image feature $e_j^r$ is related to the $j$th word in sentence. We design the textual intermediate function $l_j^w(W,b)$ and further get the updated textual feature $\tilde{t}_j$ as following:

$$l_j^w(W,b) = \text{concat}(t_j, e_j^r) \cdot W + b \tag{14}$$

$$\tilde{t}_j = (1 - g_j^w) \cdot t_j + g_j^w \cdot c_j^w \tag{15}$$

and the definitions of $g_j^w$ and $c_j^w$ are

$$\begin{cases} g_j^w = \text{sigmoid}[l_j^w(W_3, b_3)] \\ c_j^w = \text{sigmoid}[l_j^w(W_4, b_4)]. \end{cases} \tag{16}$$

And likewise, the similarity score $S(F,T)$ should be computed by the following expression:

$$S(F,T) = \frac{1}{N} \sum_{j=1}^{N} \frac{\tilde{t}_j^T e_j^r}{||\tilde{t}_j|| \cdot ||e_j^r||}. \tag{17}$$

Similarly, if the $j$th word is not relevant to the image, the updated textual feature $\tilde{t}_j$ will contribute little to the similarity score $S(F,T)$.

### D. Objective Function

The loss function we utilized to train our model is the triplet loss [8],[34], which is based on the idea that instances with the same label (i.e., sharing the same semantics) should lie closer to each other than those having different labels in the learned common space. Triplet loss is a common and popular objective function in a cross-modal retrieval. For a query instance, there will be several mismatched samples in a minibatch, which are called negative samples. We select the one that is closest to the query instance among these negative samples, that is, the hardest negative. It is very common to train the model with the hardest negatives to improve the performance of the model, and thus, we suggest employing the triplet loss function with emphasis on the hardest negatives. By minimizing the triplet ranking loss function, the model is trained to ensure that the ground-truth positive image-sentence pairs always keep higher similarity scores than negative pairs. The loss function is defined as

$$\begin{aligned} L = \sum_{i=1}^{M} \{ &\text{relu}[\beta - S(F_i, T_i^p) + S(F_i, T_i^n)] \\ &+ \text{relu}[\beta - S(F_i^p, T_i) + S(F_i^n, T_i)] \} \end{aligned} \tag{18}$$

where $\text{relu}[x] \equiv \max(x, 0)$. The $S(F_i, T_i^p)$ and $S(F_i^p, T_i)$ are positive pairs, and the $S(F_i, T_i^n)$ and $S(F_i^n, T_i)$ are the hardest negative pairs. $\beta$ is the margin threshold for triplet loss. For a certain anchor point $F_i$ (and $T_i$), the positive sample $T_i^p$ (and $F_i^p$) is the sample that has the same class label as the anchor, and the label of the negative sample $T_i^n$ (and $F_i^n$) is different from the anchor. For computational efficiency, we divide the training data into $M$ minibatches, and find the hardest negatives in each minibatch instead of selecting in the entire dataset. For training the network, we suggest choosing the minibatch gradient descent mechanism [41]. Compared with batch gradient descent [42] and stochastic gradient descent [43], the minibatch gradient descent mechanism can update parameters faster and make the model converge more robustly.

TABLE I
STATISTICS OF THE DATASETS USED IN OUR EXPERIMENTS

| Dataset | Images | Sentences | Classes | Size |
|---|---|---|---|---|
| UCM-Captions | 2100 | 5/image | 21 | 256×256 |
| Sydney-Captions | 613 | 5/image | 7 | 500×500 |
| RSICD | 10921 | 5/image | 31 | 224×224 |
| NWPU-Captions | 31500 | 5/image | 45 | 256×256 |

## IV. EXPERIMENT

In order to prove the effectiveness of our proposed method, we evaluate our SAM on four public datasets: UCMerced-LandUse-Captions, Sydney-Captions, RSICD, and NWPU-RESISC45-Captions. We compare our method with several state-of-the-art methods impartially and objectively, and we fully observe the performance of our SAM. The deep CNN we employ to extract visual feature is Inception V3 [44], and the RNN we use to extract text semantic feature is Bi-GRU [40].

### A. Dataset and Metric

*1) Datasets:* We perform experiments on four benchmark remote sensing datasets for the cross-modal image-text retrieval: UCMerced-LandUse-Captions, Sydney-Captions, RSICD, and NWPU-RESISC45-Captions. We evaluate the performance of the SAM and validate the effectiveness of our network by comparing with other state-of-the-art methods. Here, we give the statistics of these four benchmark remote sensing datasets in Table I.

*a) UCMerced-LandUse-Captions:* This dataset is constructed by Qu *et al.* [45], and it is based on the UCMerced-LandUse dataset [46]. It contains land use images in 21 classes, including agricultural, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks, and tennis court, with 100 images for each class. Each image has 256 × 256 pixels. The pixel resolution of these images is 0.3048 m. The images in UCMerced-LandUse dataset were manually extracted from a large amount of remote sensing images from the United States Geological Survey National Map Urban Area Imagery. Based on [45], five different sentences were exploited to describe every image.

1. A big basketball court and a small basketball court next to a badminton court and a road beside.
2. There are two basketball courts, One of which is twice the size of the other.
3. There are two basketball courts beside the road, with many trees nearby.
4. There are two basketball courts next to the road.
5. Some cars are on the road next to two basketball courts.

(a)

1. The cirrus clouds are located above the land covered by dense vegetation and a blue river.
2. There is a little thin cloud above the farmland.
3. Large thick white clouds are above the sea-land junction.
4. The white cloud is above green land and blue sea.
5. There are a few slices of white clouds above a green land.

(b)

Fig. 4. Here, we give two examples in the NWPU-RESISC45-Captions dataset. These two image-text pairs belong to different classes that are not collected in other three datasets: (a) basketball court and (b) cloud, respectively.

*b) Sydney-Captions:* Sydney-Captions dataset is also provided by [45], which is based on the Sydney dataset [47]. The image of Sydney, Australia, at 18 000 × 14 000 pixels, was got from Google Earth. The pixel resolution of each image is 0.5 m. Similar to the UCMerced-LandUse dataset, five different sentences were given to describe each image [45].

*c) RSICD:* RSICD dataset is used for remote sensing image captioning task [48]. More than ten thousand of remote sensing images are collected from Google Earth, Baidu Map, MapABC, and Tianditu. The images are fixed to 224 × 224 pixels with various resolutions. The total number of remote sensing images are 10 921, with five sentences of descriptions per image.

*d) NWPU-RESISC45-Captions:* This dataset is provided by Prof. Z. Shao [6],[49], [50] and his Urban Remote Sensing team from Wuhan University based on the NWPU-RESISC45 dataset [51], which is a publicly available benchmark for Remote Sensing Image Scene Classification (RESISC), created by Northwestern Polytechnical University (NWPU). This dataset contains 31 500 images, covering 45 scene classes with 700 images in each class. Each image in the NWPU-RESISC45 dataset is annotated with five sentences, and each sentence is not shorter than six words. To the best of our knowledge, this is the largest dataset for remote sensing captioning. Thus, this dataset provides the scientific community a data resource to advance the task of remote sensing captioning. We show two example image-text pairs in Fig. 4.

*2) Evaluation Metric:* We conduct two kinds of image–text matching tasks: 1) sentence retrieval, i.e., retrieving ground-truth sentences related to the query image (I2T); and 2) image retrieval, i.e., retrieving ground-truth images related to the query text (T2I). The commonly used evaluation metric for retrieval tasks is Recall at $K$ ($R@K$), which is defined as the percentage of queries in which the ground-truth matchings are contained in the first $K$ retrieved results. The higher value of $R@K$ means better performance. Based on [52], we use two rank metrics to evaluate our proposed method, i.e., $MedR$ and $MeanR$. $MedR$ is the median rank of the first retrieved ground-truth sentence or image. The lower its value, the better. And correspondingly, the Mean Rank ($MeanR$) is used as a metric in our experiment. Both of these two rank metrics are statics over the position of the ground-truth term in the retrieval order. We also compute another score, denoted as "$R@$sum," to evaluate the overall performance for the cross-modal retrieval, which is the summation of all $R@1$, $R@5$, and $R@10$ scores defined as follows:

$$R@\text{sum} = \underbrace{R@1 + R@5 + R@10}_{\text{Image-to-Text}} + \underbrace{R@1 + R@5 + R@10}_{\text{Text-to-Image}}. \tag{19}$$

### B. Experimental Settings

For the UCMerced-LandUse-Captions dataset, we randomly select 1 680 images as the training set and the rest 420 images as the testing set. For the Sydney-Captions dataset, we collect 490 images for training, and the rest 123 images for testing. In the RSICD dataset, we randomly select 8 737 images for training and 2 184 images for testing. For the NWPU-RESISC45-Captions dataset, we randomly choose 25 200 images as the training set and the rest 6 300 images as the testing set. Before extracting visual features, we resize the raw images into 224 × 224 as a fixed size.

In the experiment, we set the learning rate to 0.0005, the batch size to 16, the temperature parameter $\alpha$ to 9.0, the margin threshold $\beta$ to 0.2, and the number of epochs to 120. We use the Inception V3 network, which has been pretrained on ImageNet to extract 64 visual features for each remote sensing image, and we set the dimension of visual feature to 2 048. For each sentence, we initialize the word vector by random weights (the dimension of word vector is 300), and then, fed them into the Bi-GRU whose hidden dimension is set to 2048.

### C. Results on Benchmark Datasets

To assess the effectiveness of the proposed method, we conduct experiments and compare our proposed model with several published state-of-the-art models on four benchmark datasets.

*Results:* Comparison results are reported in Tables II–V for UCMerced-LandUse-Captions, Sydney-Captions, RSICD, and NWPU-RESISC45-Captions, respectively. Compared with the state-of-the-art methods, the developed method not only fused the visual and textual features by the concatenate function to establish a direct correspondence between two different modalities, but also filtered and optimized the features by the gate function, which further strengthened the interaction between different modalities and improved the retrieval accuracy. Experimental results on these four benchmark datasets well demonstrate that our proposed method is capable to achieve a superior performance over other state-of-the-art methods. Particularly, our t-i model performs better than i-t model in general.

TABLE II
COMPARISON OF THE CROSS-MODAL IMAGE-TEXT RETRIEVAL PERFORMANCE IN TERMS OF $R@K$ AND RANKS ON UCMERCED-LANDUSE-CAPTIONS DATASET

| Method | Text Retrieval | | | | | Image Retrieval | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $R@1$ | $R@5$ | $R@10$ | $MedR$ | $MeanR$ | $R@1$ | $R@5$ | $R@10$ | $MedR$ | $MeanR$ | $R@sum$ |
| IMRAM [33] | 3.8 | 36.2 | 60.5 | 9.0 | 17.7 | 7.6 | 36.2 | 65.2 | 8.0 | 20.0 | 209.5 |
| PFAN [39] | 8.1 | 28.6 | 53.8 | 9.0 | 23.9 | 8.1 | 38.1 | 70.0 | 7.0 | 13.6 | 206.7 |
| MTFN [30] | 9.0 | 35.2 | 58.1 | 9.0 | 22.8 | 9.5 | 44.3 | 78.6 | 6.0 | 9.8 | 234.7 |
| SCAN [32] | **14.3** | 45.2 | 76.2 | **6.0** | 9.2 | 11.4 | 49.5 | 91.9 | 6.0 | 5.8 | 288.5 |
| (ours)SAM t-i | 9.5 | **47.6** | **78.6** | **6.0** | **7.4** | **13.8** | **55.2** | **94.8** | **5.0** | **5.5** | **299.5** |
| (ours)SAM i-t | 11.9 | 47.1 | 76.2 | **6.0** | 9.9 | 10.5 | 47.6 | 93.8 | 6.0 | 5.8 | 287.1 |

The best results are marked in bold font.

TABLE III
COMPARISON OF THE CROSS-MODAL IMAGE-TEXT RETRIEVAL PERFORMANCE IN TERMS OF $R@K$ AND RANKS ON SYDNEY-CAPTIONS DATASET

| Method | Text Retrieval | | | | | Image Retrieval | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $R@1$ | $R@5$ | $R@10$ | $MedR$ | $MeanR$ | $R@1$ | $R@5$ | $R@10$ | $MedR$ | $MeanR$ | $R@sum$ |
| IMRAM [33] | 3.8 | 26.9 | 46.2 | 14.0 | 36.0 | 1.9 | 21.2 | 42.3 | 13.0 | 16.8 | 142.3 |
| PFAN [39] | 3.8 | 15.4 | 30.8 | 30.0 | 42.2 | 5.8 | 23.1 | 46.2 | 11.0 | 14.5 | 125.1 |
| MTFN [30] | 7.7 | 25.0 | 38.5 | 13.0 | 28.7 | 5.8 | 28.8 | 53.8 | 10.0 | 13.5 | 159.6 |
| SCAN [32] | **9.6** | 23.1 | 40.4 | 22.0 | 45.1 | 7.7 | 21.2 | 48.1 | 15.0 | 17.0 | 150.1 |
| (ours)SAM t-i | 5.8 | 32.7 | 48.1 | 11.0 | 24.8 | **9.6** | **34.6** | 55.8 | **9.0** | 12.5 | 186.6 |
| (ours)SAM i-t | **9.6** | **34.6** | **53.8** | **10.0** | **20.8** | 7.7 | 28.8 | **59.6** | **9.0** | **12.1** | **194.1** |

The best results are marked in bold font.

TABLE IV
COMPARISON OF THE CROSS-MODAL IMAGE-TEXT RETRIEVAL PERFORMANCE IN TERMS OF $R@K$ AND RANKS ON RSICD DATASET

| Method | Text Retrieval | | | | | Image Retrieval | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $R@1$ | $R@5$ | $R@10$ | $MedR$ | $MeanR$ | $R@1$ | $R@5$ | $R@10$ | $MedR$ | $MeanR$ | $R@sum$ |
| IMRAM [33] | 7.1 | 9.5 | 26.7 | 27.0 | 44.7 | 4.8 | 17.1 | 35.2 | 20.0 | 20.2 | 100.4 |
| PFAN [39] | 4.8 | 14.3 | 29.0 | 28.0 | 43.9 | 4.3 | 18.1 | 37.1 | 18.0 | 19.9 | 107.6 |
| MTFN [30] | 2.4 | 7.1 | 33.8 | 25.0 | 46.8 | 3.8 | 21.4 | 39.0 | 19.0 | 19.6 | 107.5 |
| SCAN [32] | 7.6 | 25.0 | 41.4 | 11.0 | 31.6 | 6.2 | 24.8 | 46.8 | 13.0 | 19.6 | 151.8 |
| (ours)SAM t-i | **12.8** | **31.6** | 47.3 | 9.0 | **21.5** | **11.5** | **35.7** | **53.4** | **8.0** | **11.8** | **192.3** |
| (ours)SAM i-t | 10.3 | 29.8 | **47.6** | **8.0** | 23.6 | 10.5 | 34.2 | 52.1 | **8.0** | 14.6 | 184.5 |

The best results are marked in bold font.

TABLE V
COMPARISON OF THE CROSS-MODAL IMAGE-TEXT RETRIEVAL PERFORMANCE IN TERMS OF $R@K$ AND RANKS ON NWPU-RESISC45-CAPTIONS DATASET

| Method | Text Retrieval | | | | | Image Retrieval | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $R@1$ | $R@5$ | $R@10$ | $MedR$ | $MeanR$ | $R@1$ | $R@5$ | $R@10$ | $MedR$ | $MeanR$ | $R@sum$ |
| IMRAM [33] | 7.1 | 25.3 | 51.1 | 13.0 | 14.7 | 4.3 | 31.0 | 65.3 | 11.0 | 13.4 | 184.1 |
| PFAN [39] | 7.6 | 22.9 | 47.5 | 17.0 | 17.5 | 11.2 | 32.6 | 71.0 | 9.0 | 12.5 | 192.8 |
| MTFN [30] | 5.2 | 21.8 | 56.1 | 12.0 | 22.4 | 11.9 | 37.0 | 74.2 | 9.0 | 17.2 | 206.2 |
| SCAN [32] | 9.5 | 35.7 | 63.2 | 9.0 | 10.2 | 12.6 | 40.1 | 81.6 | 7.0 | 16.7 | 242.7 |
| (ours)SAM t-i | **13.3** | **38.8** | **65.5** | **7.0** | **9.4** | 14.8 | **45.2** | **82.9** | **6.0** | **10.2** | **260.5** |
| (ours)SAM i-t | 8.5 | 37.6 | 64.3 | 8.0 | 11.3 | **15.3** | 44.7 | 81.8 | 8.0 | 14.7 | 252.2 |

The best results are marked in bold font.

As shown in Table II, on the UCMerced-LandUse-Captions dataset, our model can surpass other methods on metrics including $R@5$, $R@10$, $MedR$, and $MeanR$. It can be seen that SCAN [32] performs the best among all these state-of-the-art methods, but our method can even outperform SCAN [32]. Specifically, compared with SCAN [32], our model can exhibit an increase of 2.4% ($R@10$) for the text retrieval and 5.7% ($R@5$) for the image retrieval, respectively. Although for the text retrieval, the proposed method SAM t-i is only second to SCAN [32] in terms of R@1 for the text retrieval, it displays a significantly better performance than the other techniques.

As shown in Table III, on the Sydney-Captions dataset, it can be seen that our model is able to achieve outstanding result on all metrics. In particular, compared with the best baseline SCAN [32], our model can achieve the remarkable improvements of 13.4% ($R@10$) and 11.5% ($R@10$) for the text retrieval and the image retrieval, respectively. Besides, on this dataset, we find that the MTFN [30] method performs slightly better than SCAN [32] in terms of $R@sum$, and our model can even surpass the MTFN [30] method in the text retrieval by 15.3% ($R@10$) and in the image retrieval by 5.8% ($R@10$), respectively.

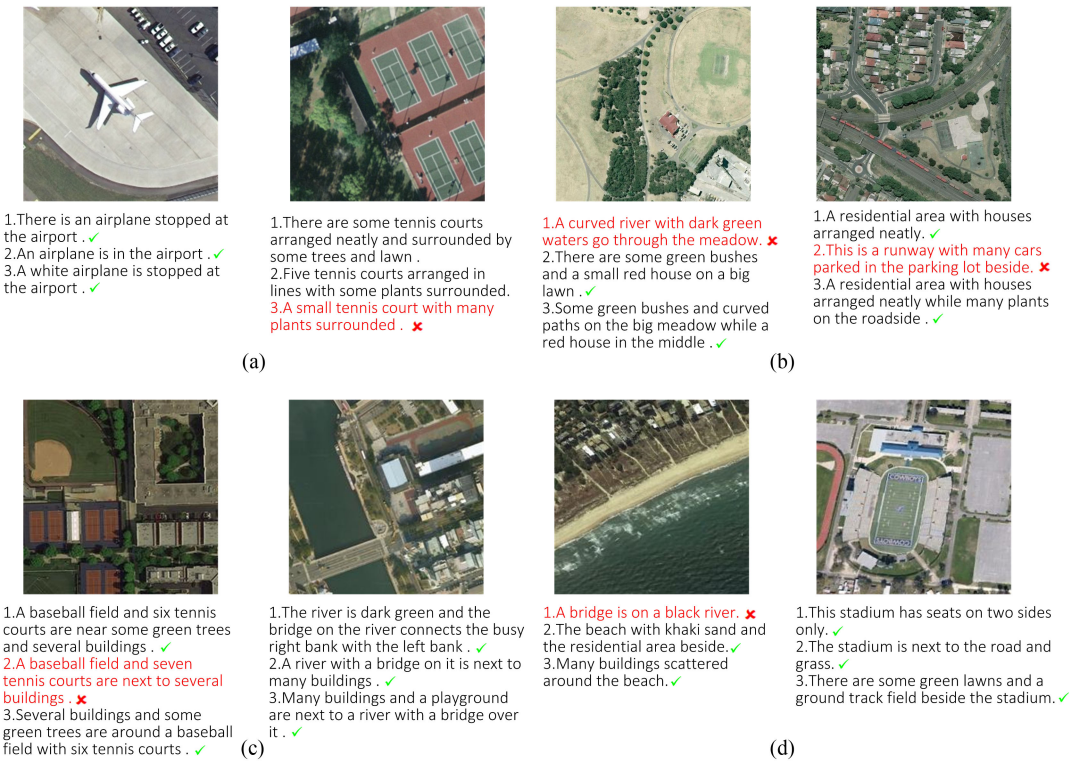


Fig. 5. Visual results of the image-to-text retrieval on four datasets. (a) UCMerced-LandUse-Captions. (b) Sydney-Captions. (c) RSICD. (d) NWPU-RESISC45-Captions. For each query image, we show the top-three ranked sentences.

As shown in Table IV, on the RSICD dataset, our model still provides the best numbers on all metrics. Compared with SCAN [32], our model can obtain an increase of 6.6% in $R@5$ for the text retrieval and an increase of 10.9% in $R@5$ for the image retrieval.

As is shown in Table V, on the NWPU-RESISC45-Captions dataset, our model can also surpass other methods on all metrics. Our model can surpass the best baseline SCAN [32] in the text retrieval by 3.3% ($R@10$) and in the image retrieval by 1.3% ($R@10$), respectively.

For both the image-query-text and the text-query-image tasks, our proposed method has achieved the best performance in almost all metrics on four remote sensing datasets except for the metric $R@1$. It is worth noting that all methods including our method have not achieved satisfactory performance on this metric. This can be probably explained by the fact that remote sensing images generally present much more complex spectral and structure information compared with natural images, and thus, restrict the robustness of cross-modal retrieval models.

### D. Ablation Study

In order to evaluate the performance of each module in our proposed SAM method, we compare our current model (built using Inception V3 features) with several variants, which are as follows:

1) SAM-1 is built using the SIFT [53] features;
2) SAM-2 is constructed using the BoVW [54] features;
3) SAM-3 is configured using the ResNet-152 [55] network to extract visual features;
4) SAM-4 is established using the fine-tuned Inception V3 to extract visual features;
5) SAM-5 is assembled as the current model but without the attention mechanism; and
6) SAM-6 is composed as the current model but without the gate mechanism.

SIFT and BoVW are both handcrafted image features.

TABLE VI
RESULTS OF ABLATION STUDY ON NWPU-RESISC45-CAPTIONS DATASET

| Method | Text Retrieval | | | Image Retrieval | | |
|---|---|---|---|---|---|---|
| | $R@1$ | $R@5$ | $R@10$ | $R@1$ | $R@5$ | $R@10$ |
| SAM-1 | 1.8 | 14.2 | 37.6 | 2.2 | 15.7 | 46.8 |
| SAM-2 | 1.9 | 14.6 | 38.1 | 2.6 | 16.2 | 47.3 |
| SAM-3 | 13.3 | **39.2** | 65.3 | **15.3** | 45.1 | 81.8 |
| SAM-4 | 6.2 | 27.3 | 56.9 | 8.1 | 31.5 | 68.4 |
| SAM-5 | 4.8 | 18.9 | 48.7 | 5.2 | 26.7 | 59.3 |
| SAM-6 | 9.1 | 32.6 | 63.1 | 11.5 | 39.6 | 75.6 |
| SAM | **13.6** | 38.8 | **65.5** | 14.8 | **45.2** | **82.9** |

The best results are marked in bold font.

Table VI shows the performance of each model variant on the NWPU-RESISC45-Captions dataset. $R@5$ and $R@10$ provided by the SAM method are almost twice as much as those provided by SAM-1 and SAM-2, which demonstrates that CNN features can represent high-level semantic much better than handcraft


Fig. 6. Visual results of text-to-image retrieval on four datasets. (a) UCMerced-LandUse-Captions. (b) Sydney-Captions. (c) RSICD. (d) NWPU-RESISC45-Captions. For each query sentence, we show the top-three ranked images, ranking from left to right. We outline the true results in green boxes and false matches in red boxes.

features. Our model also achieves better performance than SAM-4, which proves that overfitting can be avoided by not fine tuning the CNN on remote sensing datasets. The performance of SAM-6 is better than that of SAM-5 but still less well than our proposed SAM method. The results of SAM-3 are very close to ours, indicating that using different CNNs to extract visual features have little effect on the results, which means that our SAM model has great robustness.

### E. Result Visualization

Fig. 5 shows the qualitative results of the text retrieval with given query images on four datasets. For each query image, we show the top-three retrieved sentences ranked by the similarity scores predicted by our model.

Fig. 6 shows the qualitative image retrieval results with given query sentences on four datasets. Each sentence corresponds to a ground-truth image, and for each query sentence, we show the top-three retrieved images, ranking from left to right. We outline the true results in green and false results in red.

From these results, we find that our method can return the correct results in the top-ranked sentences even for cases of clutter scenes. The model outputs some reasonable mismatches, for example, in Fig. 5, there are incorrect results such as (b.1) because the bushes in this picture look very similar to the green river. From the overall visualized results, we can see that our model is capable to discover the comprehensive and fine-grained correspondence between images and sentences by enhancing cross-modal interactions.

### F. Further Analysis

We show the classification accuracy of Inception V3 on four remote sensing datasets in Figs. 7, 8, 10, and 11, respectively.

We can see from Fig. 7 that the classification accuracy of Inception V3 can reach more than 90% in most categories. However, the classification accuracy on buildings, dense residential, and intersection are slightly lower. The images in these categories are similar in appearance (see Fig. 9), and thus, they are often misclassified. This phenomenon also appears in Figs. 10 and 11. In Fig. 8, the classification accuracy of the

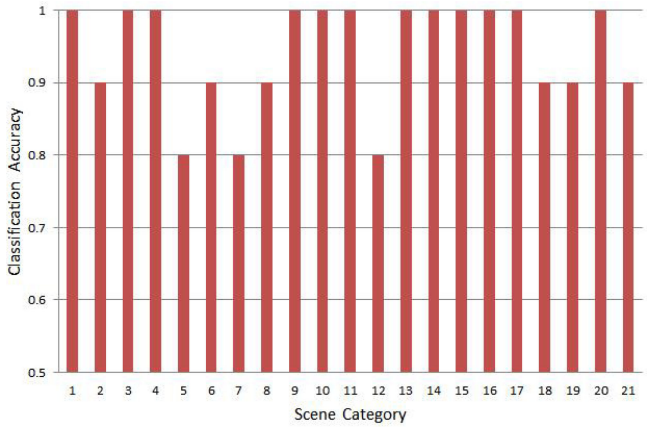


Fig. 7. Classification result on UCMerced-LandUse-Captions dataset. The categories from left to right are as follows: 1) agricultural; 2) airplane; 3) baseball diamond; 4) beach; 5) buildings; 6) chaparral; 7) dense residential; 8) forest; 9) freeway; 10) golf course; 11) harbor; 12) intersection; 13) medium residential; 14) mobile home park; 15) overpass; 16) parking lot; 17) river; 18) runway; 19) sparse residential; 20) storage tanks; and 21)tennis court.

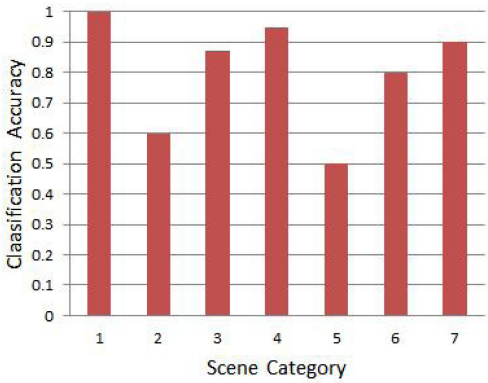


Fig. 8. Classification result on Sydney-Captions dataset. The categories from left to right are as follows: 1) airport; 2) bushes; 3) industrial; 4) residential; 5) river; 6) runway; and 7) sea.


Fig. 9. Here, we take three images, for example, from the categories of buildings, dense residential, and intersection in UCMerced-LandUse-Captions dataset, respectively, which show that images from these three different categories are similar in appearance.

category 5 (river) is 0.5, which is significantly lower than that of other categories. This low accuracy should be attributed to the relatively small number of images (45 images in total) in this category that may easily lead to overfitting. In the future, we may focus on solving this issue by data augmentation [56]. In general, Inception V3 is capable of achieving satisfactory

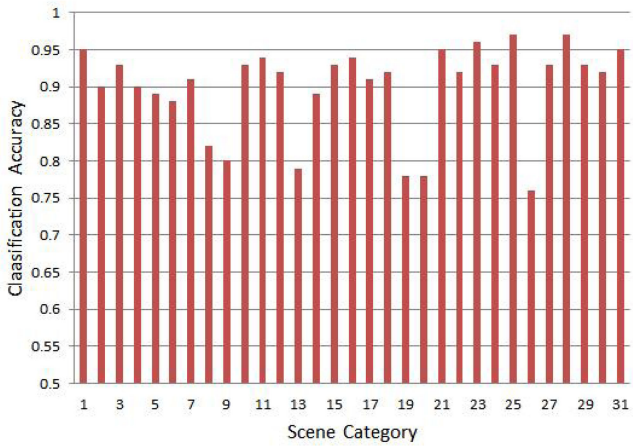


Fig. 10. Classification result on RSICD dataset. The categories from left to right are as follows: 1) airport; 2) bare land; 3) baseball field; 4) beach; 5) bridge; 6) center; 7) church; 8) commercial; 9) dense residential; 10) desert; 11) farmland; 12) forest; 13) industrial; 14) meadow; 15) medium residential; 16) mountain; 17) park; 18) parking; 19) play fields; 20) playground; 21) pond; 22) port; 23) railway station; 24) resort; 25) river; 26) school; 27) sparse residential; 28) square; 29) stadium; 30) storage tanks; and 31) viaduct.

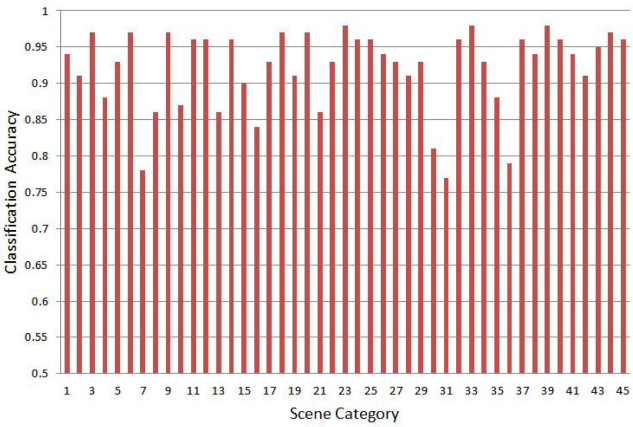


Fig. 11. Classification result on NWPU-RESISC45 dataset. The categories from left to right are: 1) airplane; 2) airport; 3) baseball diamond; 4) basketball court; 5) beach; 6) bridge; 7) chaparral; 8) church; 9) circular farmland; 10) cloud; 11) commercial area; 12) dense residential; 13) desert; 14) forest; 15) freeway; 16) golf course; 17) ground track field; 18) harbor; 19) industrial area; 20) intersection; 21) island; 22) lake; 23) meadow; 24) medium residential; 25) mobile home park; 26) mountain; 27) overpass; 28) palace; 29) parking lot; 30) railway: 31) railway station; 32) rectangular farmland; 33) river; 34) roundabout; 35) runway; 36) sea ice; 37) ship; 38) snow berg: 39) sparse residential; 40) stadium; 41) storage tank; 42) tennis court; 43) terrace; 44) thermal power station; and 45) wetland.

classification accuracy on these four remote sensing datasets, which demonstrates that we can use the image features extracted by Inception V3 for the cross-modal image-text retrieval task.

## V. Conclusion

In this article, we proposed a deep semantic alignment network for the cross-modal image-text retrieval in the field of remote sensing. Specifically, in order to sufficiently discover the

latent correspondences and preserve the semantic similarities between two different modalities, we designed an SAM to optimize the visual and textual features and strengthen the interaction between remote sensing images and text. We conducted extensive experiments on four benchmark remote sensing datasets, i.e., UCMerced-LandUse-Captions, Sydney-Captions, RSICD, and NWPU-RESISC45-Captions, to validate the effectiveness of our proposed method. It can be well demonstrated by experimental results that our proposed SAM is capable to improve the performance substantially on the cross-modal image-text retrieval task and outperform the state-of-the-art methods in the field of remote sensing. In the future work, we will extend our semantic alignment network to more cross-modal retrieval tasks, such as integrating VHR images and LiDAR data to implement the cross-modal image-text retrieval task and the 2D–3D image retrieval task.

## ACKNOWLEDGMENT

The authors would like to thank the editors and anonymous reviewers for their valuable comments, which helped them improve this work.

## REFERENCES

[1] O. E. Dai, B. Demir, B. Sankur, and L. Bruzzone, "A novel system for content-based retrieval of single and multi-label high-dimensional remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 7, pp. 2473–2490, Jul. 2018.

[2] 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, Aug. 2020.

[3] 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, Sep. 2020.

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

[5] C. Liu, J. Ma, X. Tang, X. Zhang, and L. Jiao, "Adversarial hash-code learning for remote sensing image retrieval," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 4324–4327.

[6] Z. Shao, W. Zhou, X. Deng, M. Zhang, and Q. Cheng, "Multilabel remote sensing image retrieval based on fully convolutional network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 318–328, Jan. 2020.

[7] R. Mason and E. Charniak, "Nonparametric method for data-driven image captioning," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, 2014, pp. 592–598.

[8] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng, "Grounded compositional semantics for finding and describing images with sentences," 2013. [Online]. Available: https://nlp.*stanford.edu*.

[9] A. Karpathy, A. Joulin, and L. Fei-Fei, "Deep fragment embeddings for bidirectional image sentence mapping," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, vol. 3, pp. 1889–1897.

[10] 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, 2020.

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

[12] 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.

[13] Y. Huang, W. Wang, and L. Wang, "Instance-aware image and sentence matching with selective multimodal LSTM," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2310–2318.

[14] W. Gu, X. Gu, J. Gu, B. Li, and W. Wang, "Adversary guided asymmetric hashing for cross-modal retrieval," in *Proc. Int. Conf. Multimedia Retrieval*, 2019, pp. 159–167.

[15] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *Comput. Sci.*, 2014, pp. 1–15.

[16] M. M. Bronstein, A. M. Bronstein, F. Michel, and N. Paragios, "Data fusion through cross-modality metric learning using similarity-sensitive hashing," in *Proc. Comput. Vis. Pattern Recognit.*, 2010, pp. 3594–3601.

[17] S. Kumar and R. Udupa, "Learning hash functions for cross-view similarity search," in *Proc. Int. Joint Conf. Artif. Intell.*, 2011, pp. 1360–1365.

[18] H. Zhang, Y. Zhuang, and F. Wu, "Cross-modal correlation learning for clustering on image-audio dataset," in *Proc. 15th ACM Int. Conf. Multimedia*, 2007, pp. 273–276.

[19] Y. Wei *et al.*, "Cross-modal retrieval with CNN visual features: A new baseline," *IEEE Trans. Cybern.*, vol. 47, no. 2, pp. 449–460, Feb. 2017.

[20] J. Deng, W. Dong, R. Socher, L. J. Li, and F. F. Li, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[21] A. Radoi and M. Datcu, "Multilabel annotation of multispectral remote sensing images using error-correcting output codes and most ambiguous examples," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 7, pp. 2121–2134, Jul. 2019.

[22] E. Yang, C. Deng, W. Liu, X. Liu, D. Tao, and X. Gao, "Pairwise relationship guided deep hashing for cross-modal retrieval," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 1618–1625.

[23] X. Zhang, J. Zhao, and Y. Lecun, "Character-level convolutional networks for text classification," in *Proc. Neural Inf. Process. Syst.*, 2015, pp. 1–9.

[24] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Nat. Lang. Process.*, 2014, doi: 10.3115/v1/D14-1181.

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

[26] K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1724–1734.

[27] 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.

[28] Y. Cao, M. Long, J. Wang, Q. Yang, and P. S. Yu, "Deep visual-semantic hashing for cross-modal retrieval," in *Proc. Assoc. Comput. Machinery*, 2016, pp. 1445–1454.

[29] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua, "Hierarchical multimodal LSTM for dense visual-semantic embedding," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1881–1889.

[30] 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*, 2019, pp. 12–20.

[31] J. Devlin, H. Cheng, H. Fang, S. Gupta, and M. Mitchell, "Language models for image captioning: The quirks and what works," *53rd Annu. Meeting Assoc. Comput. Linguistics/7th Int. Joint Conf. Natural Lang. Process. Asian Fed. Natural Lang. Process.*, 2015, pp. 100–105.

[32] K. H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 201–216.

[33] H. Chen, G. Ding, X. Liu, Z. Lin, and J. Han, "IMRAM: Iterative matching with recurrent attention memory for cross-modal image-text retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12655–12663.

[34] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," in *Proc. 31st Int. Conf. Mach. Learn. (ICML)*, 2014.

[35] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "VSE: Improving visual-semantic embeddings with hard negatives," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2018.

[36] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 3, 2014, pp. 1–12.

[37] J. Ba, V. Mnih, and K. Kavukcuoglu, "Multiple object recognition with visual attention," in *Proc. Int. Conf. Learn. Representations (ICLR)* 2014.

[38] H. Nam, J. W. Ha, and J. Kim, "Dual attention networks for multimodal reasoning and matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2156–2164.

[39] Y. Wang, H. Yang, X. Qian, L. Ma, and X. Fan, "Position focused attention network for image-text matching," *Comput. Res. Repository*, 2019, *arXiv*.

[40] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

[41] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. Mach. Learn. Res.*, vol. 11, no. 1, pp. 19–60, 2009.
[42] B. Pearlmutter, "Gradient descent: Second order momentum and saturating error," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 4, 1992, pp. 887–894.
[43] M. Zinkevich, M. Weimer, A. J. Smola, and L. Li, "Parallelized stochastic gradient descent," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2011, pp. 1–9.
[44] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 2818–2826, doi: 10.1109/CVPR.2016.308.
[45] B. Qu, X. Li, D. Tao, and X. Lu, "Deep semantic understanding of high resolution remote sensing image," in *Proc. Int. Conf. Comput., Inf. Telecommun. Syst.*, 2016, pp. 1–5.
[46] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th ACM SIGSPATIAL Int. Symp. Adv. Geographic Inf. Syst.*, San Jose, CA, USA, Nov. 3–5, 2010, , 2010, pp. 270–279.
[47] F. Zhang, B. Du, and L. Zhang, "Saliency-guided unsupervised feature learning for scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 2175–2184, Apr. 2015.
[48] 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.
[49] W. Zhou, S. Newsam, C. Li, and Z. Shao, "PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval," *ISPRS J. Photogrammetry Remote Sens.*, vol. 145, pp. 197–209, 2018.
[50] Z. Shao, K. Yang, and W. Zhou, "Performance evaluation of single-label and multi-label remote sensing image retrieval using a dense labeling dataset," *Remote Sens.*, vol. 10, no. 6, pp. 964–977, 2018.
[51] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state-of-the-art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.
[52] I. Vendrov, R. Kiros, S. Fidler, and R. Urtasun, "Order-embeddings of images and language," in *Proc. Int. Conf. Learn. Representations*, 2016, pp. 1–12.
[53] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, pp. 91–110, 2004.
[54] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2003, pp. 1470–1477.
[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[56] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, 2002.