# Retrieval Topic Recurrent Memory Network for Remote Sensing Image Captioning

Binqiang Wang, Xiangtao Zheng, *Member, IEEE*, Bo Qu, and Xiaoqiang Lu, *Senior Member, IEEE*

***Abstract*—Remote sensing image (RSI) captioning aims to generate sentences to describe the content of RSIs. Generally, five sentences are used to describe the RSI in caption datasets. Every sentence can just focus on part of images' contents due to the different attention parts of annotation persons. One annotated sentence may be ambiguous compared with other four sentences. However, previous methods, treating five sentences separately, may generate an ambiguous sentence. In order to consider five sentences together, a collection of words, which named topic words contained common information among five sentences, is jointly incorporated into a captioning model to generate a determinate sentence that covers common contents in RSIs. Instead of employing a naive recurrent neural network, a memory network in which topic words can be naturally included as memory cells is introduced to generate sentences. A novel retrieval topic recurrent memory network is proposed to utilize the topic words. First, a topic repository is built to record the topic words in training datasets. Then, the retrieval strategy is exploited to obtain the topic words for a test image from topic repository. Finally, the retrieved topic words are incorporated into a recurrent memory network to guide the sentence generation. In addition to getting topics through retrieval, the topic words of test images can also be edited manually. The proposed method sheds light on controllability of caption generation. Experiments are conducted on two caption datasets to evaluate the proposed method.**

***Index Terms*—Controllable caption, recurrent memory network (MN), remote sensing image (RSI) caption generation, retrieval topic.**

## I. Introduction

REMOTE sensing image (RSI) captioning aims to generate a concise sentence automatically given a high-resolution RSI [1]. RSI captioning has emerged as an important task for semantic understanding, which is a cross field between RSI processing and natural language processing [2]. Many traditional remote sensing tasks concentrate on image processing or low-level semantic information. For image processing, RSI denoising [3] deals with the noise of a RSI and output a denoised RSI. Based on RSIs with different spatial and spectral resolutions, RSI fusion [4] aims to generate a RSI with both high spatial and spectral resolutions. For low-level semantic information, RSI classification [5]–[9] endows a word level label to a RSI. Hyperspectral unmixing [10] provides endmember information and corresponding abundance information of each pixel in a hyperspectral RSI. Different from previous tasks, an RSI captioning task concentrates on generating high-level semantic information (a descriptive sentence) and has received a significant amount of attention [1], [11]–[13]. Many useful potential applications need to utilize the RSI captioning task, such as RSI retrieval, disaster assessment [14], [15], terrain scanning [16], [17], city planning, and military scout. The automatic caption generation can provide more semantic information about an RSI. This information can boost the RSI retrieval by providing more descriptive information except for a query RSI. For disaster assessment, images of the disaster area captured by an unmanned aerial vehicle or satellite can be translated into descriptive texts. These texts can be seen as a quick initial assessment report. The image captioning in a computer vision community aims to generate semantic descriptions for a natural image containing ordinary scene content [18], [19]. The generated descriptions present more semantic and logistic level information. Researchers in the remote sensing field have made an earnest endeavor to introduce RSI captioning [1], [11]–[13]. But for RSI captioning, the objects in RSIs are mountains, lakes, rivers, and artificial architectures. These are completely different perspectives from the daily life. The RSI captioning methods can be divided into three main categories according to the different ways of obtaining the descriptive sentence: template-based methods [12], retrieval-based methods [20], and *recurrent neural network* (RNN)-based methods. Template-based methods generate sentences based on object detection [12]. Specifically, the first step is to detect objects in the RSI. Then, sentences are generated based on the detection results by filling fixed sentence templates that lack a subject or an object. However, sentences generated by template-based methods are relatively simple and mode fixed. This can lead to generate some awkward and wrong sentences that should be avoided. Retrieval-based methods treat the captioning task as an image-sentence retrieval task [20]–[22]. The five sentences

Manuscript received July 27, 2019; revised October 18, 2019 and November 20, 2019; accepted December 3, 2019. Date of current version February 12, 2020. This work was supported in part by the National Key R&D Program of China under Grant 2017YFB0502900, in part by the National Natural Science Foundation of China under Grant 61925112, Grant 61806193, Grant 61702498, and Grant 61772510, in part by the Young Top-Notch Talent Program of Chinese Academy of Sciences under Grant QYZDB-SSW-JSC015, in part by the CAS "Light of West China" Program under Grants XAB2017B26 and XAB2017B15, and in part by the Xi'an Postdoctoral Innovation Base Scientific Research Project. *(Corresponding author: Xiangtao Zheng.)*

B. Wang is with the Key Laboratory of Spectral Imaging Technology CAS, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: binqiang2wang@gmail.com).

X. Zheng, B. Qu, and X. Lu are with the Key Laboratory of Spectral Imaging Technology CAS, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China (e-mail: xiangtaoz@gmail.com; bo.qu.opt@gmail.com; luxq666666@gmail.com).

Digital Object Identifier 10.1109/JSTARS.2019.2959208

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/

- there are some planes and cars in the airport
- four different kinds of airplanes are in the airport
- here are some airplanes and cars in the airport
- four different sizes of airplanes are in the airport
- four airplanes are parked at the airport

| Semantic Topics |
|---|
| planes |
| cars |
| airport |
| some |
| different |
| are |
| parked |
| |
| |

Fig. 1. Semantic Topics: all the nouns, verbs, adjectives, and their variants are extracted from the sentence of the caption dataset using natural language toolkit. The nouns and their variants are in red color, the adjective is in green, and the verbs and their variants are in blue color.

in caption dataset are represented as a collective sentence representation [20]. The collective sentence representation and image representation are embedded into a common semantic space. Then, the sentence is generated by a retrieval way in the common semantic space. RNN-based methods have become mainstream methods in RSI captioning problem [1], [11], [13] because of the strong sentence coding ability of RNN. These methods first extract the features from RSIs by *convolutional neural networks* (CNNs), and then use RNNs to generate sentences based on the extracted features. However, the complex structure of RNN makes it difficult to extend reasonably to add additional inputs. Generally, the additional inputs are combined with the origin inputs in the outsides of the structure of RNN. Furthermore, the long-term information is diluted, little by little, at every time step in RNN.

Although the RNN-based methods can generate reasonable sentences, there are still some challenges that need to be conquered. The backpropagation through time in the RNN may cause vanishing gradient during training. Furthermore, an image is usually annotated with five diverse sentences due to different subjective attentions of different annotation persons during the sentence labeling stage [13]. As shown in Fig. 1, some sentences are annotated to describe *some airplanes and cars*, while other sentences focus on *four airplanes*. Previous methods consider these five sentences separately and may generate ambiguous sentences. However, there are some common determinate information among the diverse sentences. Taking Fig. 1 as an example, all these five sentences contain the common determinate information: *airport*. It is beneficial to utilize the common determinate information for sentence generation. In general, an RSI contains many types of objects on the ground, which makes deterministic information important in the RSI processing field. The common determinate information is called topic words herein.

To incorporate the common determinate information into RSI caption generation, several subproblems need to be alleviated. The first problem is how to decide what is the common determinate information covered by the sentences in caption datasets. The common determinate information is called topic words in this article. The second problem is how to design a model that can naturally utilize the topic words as guide information to generate determinate sentences flexibly.

In this article, topic words are used to guide the captioning instead of utilizing sentence as guide information in [23]. Inspired by the development of *memory networks* (MNs) [24], in which memory cells can be read or written by memory neural networks to capture the long-term information, we proposed a *retrieval topic recurrent memory network* (RTRMN). In particular, a topic repository is constructed based on five sentences of an image in training datasets to capture the common determinate information. Thanks to the recent advances in MNs [25], topic words can be seen as memory cells, which can be seamlessly integrated into MNs. The key insight is that topic words are expected to serve as guiding and control information for RSI captioning. Details of the RTRMN are illustrated as follows.

First, a collection of words (named as topic words) are extracted from the five annotated sentences, which are utilized as guide information to assist the task of RSI captioning. To extract topic words of sentences in caption datasets, two alternative methods are designed: "semantic topics" and "statistical topics." Details of topic extraction methods are introduced in Section III-A. The topic words extracted from the training set make up the topic repository.

Then, the retrieval strategy is utilized to obtain the topic words of test images from topic repository. During the testing process, the annotated sentences are unknown. The topic words cannot be directly extracted from the origin captioning datasets. Therefore, the topic words extracted from the training datasets are utilized to obtain the topic words of the test images. Applying retrieval strategy to RSI captioning is nontrivial. It comes with significant challenge due to the huge quantity of the caption dataset. Random sampling strategy is used in this article to alleviate this challenge.

Finally, inspired by the work of [26] and [27], the *Convolutional-MaxPooling* (CMP) based on 1-D convolution is designed to capture the relationships of different memory cells in a recurrent MN. Recurrent MN is motivated by the developments of MNs and RNN. To utilize the topic words in a reasonable way, the topic words are inserted as the memory cells in the RTRMN. Instead of utilizing RNN, which suffers long-term information dilution during time steps, CNN is applied to the top of memory cells to generate sentence. The CNN can capture the related information in every parallel steps, which overcomes information dilution in RNN. It should be noted that the recurrent in RTRMN is totally different from the recurrent in RNN. Specifically, the recurrent in RNN means recurrent information passed in training while the recurrent in RTRMN means that the model should be recurrently run to generate a whole sentence.

Overall, the main contributions of this article can be briefly summarized as follows.

1) The five sentences of an image annotated in caption dataset are considered to contain ambiguous information and determinate information. Topic information extracted from

five sentences is utilized as guide information to generate a determinate sentence. To the best of our knowledge, no exploration work has been done in this aspect.

2) A novel retrieval topic recurrent MN is proposed to utilize the topic words as a part of extensible memory cells, which can overcome the shortcoming of long-term information dilution in RNN. In addition, CNN is applied on the top of MN to capture the relationship between the topics and the images. The topic words are imported into architecture naturally and flexibly because of the memory cells structure.
3) The topic information can be seen as a control signal for captioning. Given an RSI, the caption that focus on specific ground objects can be generated by the proposed method if the topic is available as a word format. The proposed method has great potential for controlling caption generation.

The remaining parts of this article are organized as follows. Section II discusses the related works of RSI captioning and MNs. Section III describes the proposed RTRMN. The experimental results on two RSI captioning datasets are shown in Section IV. Finally, Section V concludes this article.

## II. Related Work

### A. RSI Captioning

RSI captioning has been an interesting topic in RSI processing combined with the natural language processing [1], [11]–[13], [20], [28], [29]. According to the way of generating sentences, the main methods can be divided into three categories: *template-based methods*, *retrieval-based methods,* and *RNN-based methods*.

Template-based methods are based on object detection [12]. Blanks of predefined sentence templates are filled with the detected objects to generate complete sentence. To describe the content of RSI from different scales, the objects in RSI are decomposed to three levels: Key-Instance (such as airplane), Envi-Element (such as airport), and Landscape (such as city) in [12]. A fully convolutional network [30] is used to capture the objects in RSI. Then the descriptive sentence is generated by filling detected objects in predefined templates.

The second kind of methods are retrieval-based methods [20]. In this category, the captioning task is treated as an image-sentence retrieval task [20]–[22]. The image representation and the sentence representation are mapped to a common semantic space. In this semantic space, the distance of matched image-sentence is closer than that of mismatched image-sentence. In order to make better use of annotated sentences, a collective sentence representation is proposed in [20] to obtain better matching results in semantic space than individual sentence representation.

The third kind of methods are *RNN-based methods* [1], [11], [13], which get more and more attention. Compared with template-based methods, RNN-based methods take the sentence generation as a sequence generation process in a continuous way. The RSI captioning task is proposed first by Qu *et al.* [1] from the perspective of semantic level understanding of RSI. In addition, a model is proposed in which the feature of RSI is extracted by CNN and sentence is generated by the *long short-term memory* (LSTM), a popular variant of RNN. Based on the classification results of CNN, Zhang *et al.* [11] generate the sentence utilizing RNN whose input is the classification label of RSI. Considering the scale ambiguity, category ambiguity, and rotation ambiguity of RSI, Lu *et al.* [13] explore the RSI captioning task based on both the hand-crafted features and deep features. A captioning method is proposed by Wang *et al.* [28], which can provide more accurate object location via adding additional location and category tags to images. The information from the fully connected layer of CNN is treated as attribute information to improve the performance of captioning method based on convolutional features [29].

A very similar research field is the natural image captioning. Recent studies of natural image captioning pay more attention on the CNN [31]–[33]. Two of LSTM's major limitations, the complex structure and long-term dependence, are alleviated by the utilization of CNN in [32]. Besides, Chen *et al.* propose a group-based image captioning considering the structured relevance and diversity in [34]. Anderson *et al.* propose a model utilizing bottom-up and top-down attention for image captioning and visual question answering [35], in which the bottom-up attention is implemented by the Faster R-CNN [36]. Mathews *et al.* in [37] propose an unaligned text model to generate stylized image captions. For news image, a more specific captioning is generated by a proposed context-driven entity-aware captioning model [38]. Factual, humorous, romantic, positive, and negative captions are generated separately for a same input image in [39]. Instead of generating sentence directly based on image features, a dense relational captioning task is proposed in [40] to generate more than one sentence based on relationship information.

### B. Memory Networks

Many MNs have been designed to store the information needed for different tasks: visual question answering, question answering, and document understanding [27], [41]–[46]. Graves *et al.* [41] extend the capabilities of MNs by coupling them to external memory resources, which they can interact with by attentional processes. Weston *et al.* [42] propose MNs, which reasoned with inference components combined with a long-term memory component. Sukhbaatar *et al.* [43] modify the work [42] and make the model to be trained end-to-end. Graves *et al.* [44] introduce a machine model called a differentiable neural computer, which consists of a neural networks that can read from and write to an external memory matrix, analogous to the random-access memory in a conventional computer. A dynamic MN is introduced by Kumar *et al.* [45] to solve the question answering problems over a language input. Miller *et al.* [46] propose a key-value MNs that make reading documents more viable by utilizing different encodings in the addressing and output stages of the memory read operation.

Similar to [19] and [23], the retrieval strategy is explored during the captioning task. Instead of retrieval sentences in [23], the topic words, which contain the information from five sentences, is retrieved. The method [19] tries to generate discriminative

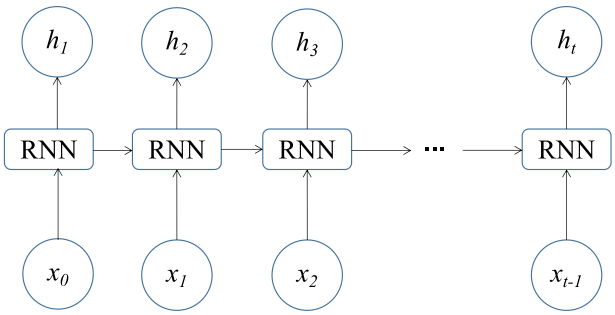



Fig. 2. Structure of RNN. There are coupling between adjacent steps.

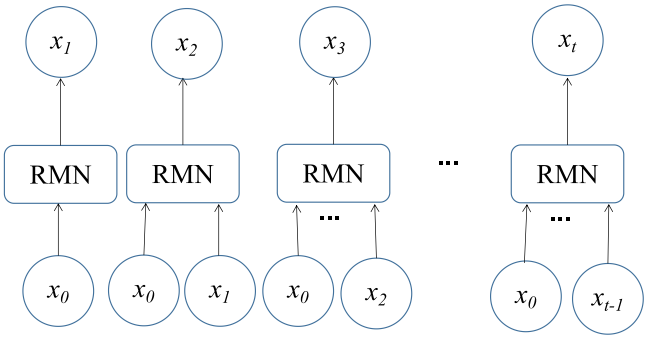



Fig. 3. Structure of proposed RMN. The coupling between adjacent steps is avoided.

captions by retrieval strategy with the help of the partially labeled data. Another similar work proposed a novel captioning model named context sequence MN [27] for personalized image captioning. Recurrent MN is motivated by the developments of MNs and RNN. RNN uses the gate mechanism to control the information flow between the steps of the RNN. However, the long-term information is diluted during the steps in RNN. This is due to the coupling between RNN steps. Specifically, as shown in Fig. 2, the input $x_0$, $x_1$,... represents the words in the sentence, the output $h_1$, $h_2$,...represents the hidden states of the RNN. The final generated word is generated based on $h_1$. All the previous steps' information is contained in $h$. When predicting the current step's output, all the previous hidden states are considered. The farther away from the current step, the less information of that step is provided for the generation of the current word. For example, when generating the $h_3$, the input $x_2$ provides more information than the input $x_0$ in this structure. The mixture of information makes the gradient backpropagation difficult during the steps. Gradient may disappear in the process of gradient backpropagation along the different time steps.

The coupling between time steps is avoided in the proposed method and long-term information dilution is averted by treating all the previous words equally. The rough diagram of the proposed method is shown in Fig. 3. Specifically, the input $x_0$, $x_1$,... represents the words in the sentence, the output $x_1$, $x_2$,... represents the words should be generated by the proposed model. Although all the previous word information is used to generate the current word during the training process, there is no coupling during the steps. The information of every previous words is treated equally. For example, when generating the $x_3$, the input $x_2$ and the input $x_0$ are treated equally in this structure. The details of the network are shown in Section III-B.

TABLE I
PHYSICAL MEANING OF SYMBOLS IN THIS ARTICLE

| symbol | physical meaning |
|---|---|
| $\mathbf{I}$ | a remote sensing image |
| $\mathbf{y}_i$ | $i$-th word in a sentence, $i = \{1, 2, ..., L\}$ |
| $L$ | the length of a sentence |
| $\mathbf{y}_0$ | start token of a sentence |
| $\mathbf{y}_{L+1}$ | end token of a sentence |
| $c$ | size of the dictionary |
| $\mathbf{I}^{p5}$ | ResNet-101 *pool5* feature of a RSI |
| $\mathbf{I}^{all}$ | the *pool5* and convolutional features of a RSI |
| $\mathbf{I}_{test}$ | the representation of a test image |
| $\mathbf{t}$ | the representation of a topic word |
| $\mathbf{m}$ | a memory cell |
| $\mathbf{p}$ | the output probability vector after softmax function |
| $\hat{\mathbf{y}}$ | the one-hot format of predicted word |

## III. RETRIEVAL TOPIC RECURRENT MEMORY NETWORK

In this section, the proposed RTRMN architecture is introduced. The physical meaning of symbols in this article is listed in Table I. Given a RSI $\mathbf{I}$, the RSI captioning task is going to generate an ordered sequence of words to form a sentence: $\mathbf{y}_0, \mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_L, \mathbf{y}_{L+1}$, where $L$ is the length of the sentence and $\mathbf{y}$ is a one-hot vector representing the word. $\mathbf{y}_0$ represents the start symbol START, and $\mathbf{y}_{L+1}$ represents the end symbol END. With the exception of two special symbols, a collection of words make up a dictionary of size $c$. The ordinary caption generation process is to generate a series of ordered word conditioned on the previous words and image features. This kind of model is influenced by the different annotated sentences containing ambiguous information. To generate more determinate sentences, RTRMN is proposed, which uses one-hot format topic $\mathbf{t}$ to assist the generation process. First, the topic in this article is introduced in Section III-A. Then, the RTRMN is introduced in detail in Section III-B.

### A. Topic Repository

Topic repository is a collection of words that extracted from the caption dataset. The definition and extraction methods of semantic topics and statistical topics are explained as below.

*1) Semantic Topics:* In the task of visual relation detection [47], the aim is to localize the objects and capture their interactions with a subject–predicated–object triplet. The "subject" and "object" are always nouns and their variants. The "predicated" is verb, adjective, and their variants. All the above-mentioned words are extracted as semantic topic that can be used as auxiliary information. These words are extracted and considered to be "semantic topics." As shown in Fig. 1, the nouns and their variants are in red color, the adjectives are in green, and the verbs and their variants are in blue color.

*2) Statistical Topics: Term frequency-inverse document frequency* (TF-IDF) [48] can represent the importance of words in the literature from a statistical point of view. These words are called "statistical topics." These topic words are used as

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} (k = 1, 2, ...5)$$

$$IDF_i = \lg \frac{|D|}{|j : t_i \in d_j|}$$

| Statistical Topics | IF-IDF |
|---|---|
| airport | 0.41812 |
| are | 0.41812 |
| the | 0.41812 |
| airplanes | 0.33449 |
| in | 0.33449 |
| four | 0.25087 |
| cars | 0.16725 |
| some | 0.16725 |
| ... | |

- there are some planes and cars in the airport
- four different kinds of airplanes are in the airport
- here are some airplanes and cars in the airport
- four different sizes of airplanes are in the airport
- four airplanes are parked at the airport

Fig. 4. Statistical topics: one sentence is treated as a document, and five sentences are seen as a corpus to show the method to extract statistical topics.

guidance information to generate sentences of images. For image captioning, the five sentences annotated to an image can be considered as a document. All the sentences of a dataset can be seen as a corpus, which contains $D$ documents as

$$tf_{i,j} = \frac{n_{i,j}}{n_j} \tag{1}$$

$$idf_i = \log \frac{|D|}{|j : t_i \in d_j|} \tag{2}$$

$$\text{TF} - \text{IDF}_{i,j} = tf_{i,j} \cdot idf_i \tag{3}$$

where $t_i$ is a word in document $d_j$, $n_{i,j}$ is the number of $i$th word $t_i$ appeared in document $d_j$, $n_j$ is the number of all the words appeared in document $d_j$, $|D|$ represents the number of documents in a corpus, and $|j : t_i \in d_j|$ represents the number of documents that contain the word $t_i$. The top $C$ words of TF-IDF score are used as statistical topics in the topic repository. As shown in Fig. 4, an example of method to calculate the TF-IDF score: the one sentence is treated as a document, and five sentences are seen as a corpus to show the method to extract statistical topics.

### B. Retrieval Topic Recurrent Memory Network

RTRMN is an architecture that generate a sentence conditioned on image representation and topic representation. The processes of training stage and testing stage are different. For the training stage, the image is represented by ResNet-101. The topic is extracted by the method mentioned in Section III-A. The ground-truth annotated sentence is combined with the image representation and topic representation to train the generation model. A block diagram of the testing stage is shown in Fig. 5. A testing image is represented by ResNet-101. Because the topic representation cannot be extracted from the ground-truth sentences of a testing image, retrieval strategy is utilized to get the potential topic from the topic repository. The retrieved topic representation and testing image representation are fed into the trained generation model to generate a sentence.

The detailed structure of ResNet-101 can be found in [49]. The topic extraction has been introduced in Section III-A. The structure of the generation model is described as below. The memory cell is used to store the representation of image and topic. The temporary memory cell is used to store the previous generated word representation. After embedding, these representations are embedded into the same length. Then, the CMP is utilized to capture the relations between memory cells, which will be introduced later. The representations of different memory cells passed through CMP are concatenated together. The concatenated representation is fed into a fully connected layer with softmax to generate a word. The generated word representation is inserted into temporary memory cells to generate next word until the end of a sentence.

*1) Memory Cells:* Memory cells, just like human brain cells storing information, are introduced in this section. Three types of memory cells are designed in our architecture: image memory cells, topic memory cells, and temporary memory cells. The image memory cells are used to store the image features. The topic memory cells contain the information of topics. The temporary memory cells are used to store the temporary words that generated during the generation of a sentence.

*Image memory cells:* Features extracted by CNN have been successfully applied in many applications. To explore the semantic information and the spatial information at the same time, the *pool5* feature of ResNet-101 [49] (denoted as $\mathbf{I}^{p5}$) and the convolutional feature before *pool5* are extracted to represent the content of RSI. The dimension of *pool5* feature is $2048 \times 1$ and the size of the convolutional feature size is $2048 \times 7 \times 7$. The convolutional feature, preserved more spatial information, is reshaped to $2048 \times 49$. Combining the two features, the image representation of the RSI is denoted as $\mathbf{I}^{\text{all}}$ whose size is $2048 \times 50$. There are 50 memory cells for image representation. Processes of 50 image features are just the same, one image feature $\mathbf{I}_i^{\text{all}}$ is shown as an example. The image features are embedded to dimension of memory cell (300) by

$$\mathbf{m}_i = \mathbf{W}_i \mathbf{I}_i^{all} + \mathbf{b}_i \quad i = 1, 2, \ldots, 50 \tag{4}$$

where $\mathbf{m}_i$ represents the $i$th image memory cell, and $\mathbf{W}_i \in \mathbb{R}^{300 \times 2048}$ and $\mathbf{b}_i \in \mathbb{R}^{300}$ are learnable parameters.

*Topic memory cells:* For the training stage, topic words are extracted from the ground-truth sentences. But for the testing stage, topic words are obtained by retrieval. There is a large gap between image feature and topic feature. Instead of retrieval topic words directly, the most similar image in training set is found as a pivot. To retrieve in training images, the random sampling is needed since the number of training images is large. $K$ images are picked up randomly from training images, then the matching score is defined by calculating distance between the image representation vectors. Given a test image $\mathbf{I}_{\text{test}}$, the most similar image from sampled $K$ images is obtained by

$$\mathbf{I}_s^{p5} = \arg\min_{\mathbf{I}_k^{p5}} \text{dis}(\mathbf{I}_{\text{test}}^{p5}, \mathbf{I}_k^{p5}) \quad k = 1, 2, \ldots, K \tag{5}$$

where $\mathbf{I}_{\text{test}}^{p5}$ and $\mathbf{I}_k^{p5}$ are the features of the testing image and sampled images, respectively. The most similar image is named $\mathbf{I}_s^{p5}$. $\text{dis}(\mathbf{I}_{\text{test}}^{p5}, \mathbf{I}_k^{p5})$ represents the distance between $\mathbf{I}_{\text{test}}^{p5}$ and $\mathbf{I}_k^{p5}$. Finally, topic words for testing images are extracted from corresponding sentences of retrieved image $\mathbf{I}_s^{p5}$.

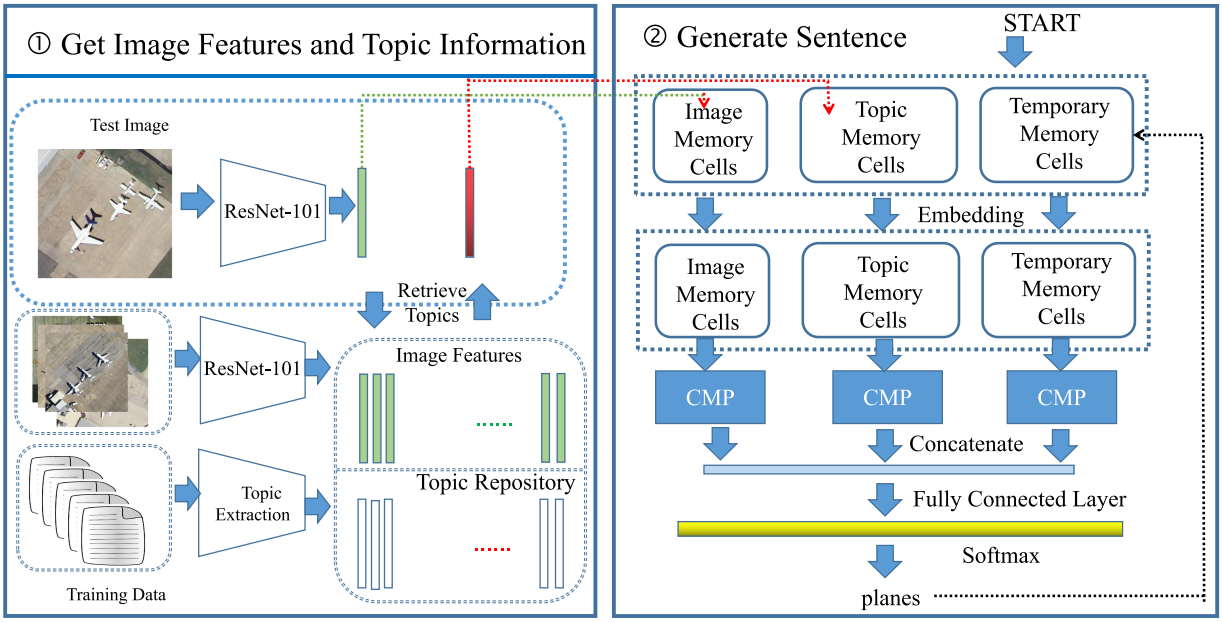



Fig. 5. RTRMN architecture. First the image feature is extracted by ResNet-101 and topic words are obtained from topic repository by random sample retrieval. The information is copied to memory cells. Then start from a symbol "START," the CNN is utilized to capture the information of different memory cells. Finally, the output of softmax is the corresponding word. The word is copied to temporary memory cell. The next word can be generated by replacing the "START" with the "planes."

The number of topics is restricted to 10. Zero is padded if the number of topics is less than 10. Processes of all the topics are just the same, one topic memory cell is shown as an example. $\mathbf{t}_i$ is the one-hot vector of a topic. The one-hot vector is a vector to represent a word and the dimension of $\mathbf{t}_i$ is $c+2$. There is only one nonzero number to represent the existence of the word. The $\mathbf{t}_i$ is embedded into cell length dimension (300)

$$\mathbf{m}_t = \mathrm{ReLU}(\mathbf{W}_e \mathbf{t}_i + \mathbf{b}_h) \tag{6}$$

where $\mathbf{m}_t$ represents a topic memory cell, and $\mathbf{W}_e \in \mathbb{R}^{300\times(c+2)}$ and $\mathbf{b}_h \in \mathbb{R}^{300}$ are parameters.

*Temporary memory cells:* The image captioning is a process that predicts next word given the previous words and the image features. The temporary memory cells are utilized to store the words that have been predicted. When there is no previous word, a token represents start is stored in the first cell. The number of temporary memory cells is 25, meaning that the length of generated sentence is within 25.

Overall, there are 50 image memory cells, 10 topic memory cells, and 25 temporary memory cells. The number of all the memory cells is 85.

*2) Sentence Generation:* The generation process of a sentence is started from image features, topic words, and a start symbol. After the first word is generated, the generation of the second word is based on three kinds of information: image feature, topic words, and previous generated words. The process of sentence generation is introduced in detail as below.

A start symbol of a sentence is denoted as START. START is represented as a one-hot vector $\mathbf{y}_0$ whose dimension is $c+2$. $\mathbf{y}_0$ is embedded into 300

$$\mathbf{m} = \mathrm{ReLU}(\mathbf{E}_z \mathbf{y}_0 + \mathbf{b}_z) \tag{7}$$

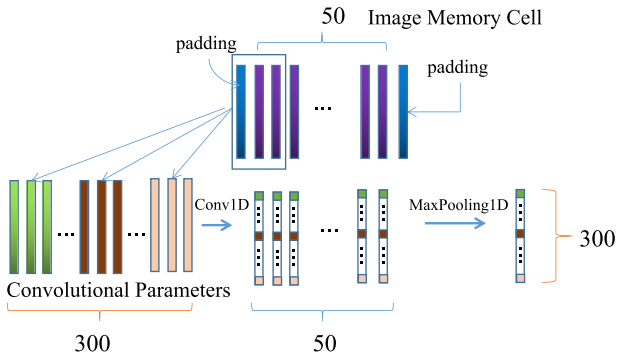



Fig. 6. Details of the CMP on image memory cells. The blue in the two ends of the image memory cells are padding vector. The different colors convolutional parameters decide the output dimension of the 1-dimensional (1-D) convolution. The final output is 300 dimension.

where $\mathbf{m}$ represents a temporary memory cell and, $\mathbf{E}_z \in \mathbb{R}^{300\times(c+2)}$ and $\mathbf{b}_z \in \mathbb{R}^{300}$ are parameters.

CMP is designed to capture the information between different memory cells, the operations are as follows. 1-D convolution is operated along with the dimension of memory cells. The window size is 3. The channel number is 300 and the stride is 1. The padding strategy is "same." To get the vectorization feature without adding parameters, max pooling is utilized to obtain a vector (the dimension of the vector is 300) after 1-D convolution according to the channel dimension. For a more specific presentation, the image memory cells are taken as an example to show the operations. The structure of CMP is shown in Fig. 6, the operation is taken on image memory cells. In order to keep the output to be the same with the input, the padding is added to two ends of image memory cells. After the

1-D convolution and maxpooling, the output feature of image memory cell is 300. As the 1-D convolution is conducted on three types of memory cells independently. Hence, the output dimension of concatenate operation is $300 \times 3$.

The output of operation mentioned above is denoted as $\mathbf{h}$ whose dimension is 900. The output word is obtained by

$$\mathbf{p}_1 = \text{softmax}(\mathbf{W}_o \mathbf{h}) \tag{8}$$

where $\mathbf{W}_o \in \mathbb{R}^{(c+2)\times 900}$ is learned parameters, and $\mathbf{p}_1 \in \mathbb{R}^{(c+2)\times 1}$ is the predicted vector. $\hat{\mathbf{y}_1}$ is the one-hot word vector where only the index of the maximum value in $\mathbf{p}_1$ is one, the others are zero.

The generated $\hat{\mathbf{y}_1}$ replaces zeros in the second temporary memory cell. In addition, the $\hat{\mathbf{y}_2}$ is generated based on the temporary memory cells containing $\mathbf{y}_0$ and $\hat{\mathbf{y}_1}$. The generated $\hat{\mathbf{y}_2}$ can be utilized to generate $\hat{\mathbf{y}_3}$ in the same way. The operation is looped until the end of the sentence END is generated.

The training set is divided into batches to train the model. The size of batch is set to 100. The cross-entropy loss function [50] is calculated in a batch in this article as

$$L = -\frac{1}{100} \sum_{i=0}^{100} \sum_{k=0}^{L+1} \sum_{m=0}^{c+1} y_{i,k,m} \log p_{i,k,m} \tag{9}$$

where $y_{i,k,m}$ is the $k$th word in $i$th sentence of a batch. $m$ in $y_{i,k,m}$ represents the index of $k$th word in the dictionary. There is only one ground-truth $m$ that makes $y_{i,k,m} = 1$. $p_{i,k,m}$ is the predicted probability of $y_{i,k,m}$.

## IV. Experiments

To validate the effectiveness of RTRMN, we conduct experiments on two datasets. First, the datasets used in our experiments are introduced in Section IV-A. Second, the metrics of RSI captioning are introduced in Section IV-B. Third, the compared approaches in this article are introduced in Section IV-C. Fourth, a thorough study on the main parameters of the proposed method is conducted in Section IV-D. Fifth, to analyze the performance of the proposed method with different types of topic words, ablation experiments are conducted on two RSI captioning datasets in Section IV-E. Sixth, to show the tremendous potential of the proposed method in controlling caption generation, some preliminary testing experiments are conducted in Section IV-F.

### A. Datasets

The experimental datasets include two RSI captioning datasets: UCM-captions and *remote sensing image captioning dataset* (RSICD).

*1) UCM-Captions:* The first dataset, UCM-captions [1], is a RSI captioning dataset whose image size is 256×256. This dataset contains 21-class land use image, including airplane, baseball diamond, beach, building, storage tank, tennis court, etc. For each class, 100 images are collected to construct the dataset and there are five sentences for every image annotated by the author.

*2) RSICD:* The second dataset, RSICD [13], is acquired from Google Earth, Baidu Map, etc., where the image size is $224 \times 224$. The number of images in RSICD is totally 10 921. Unlike UCM-captions in which sentences are annotated by one person, the five sentences in this dataset are annotated by different persons.

### B. Metrics

Simulating the evaluation metrics of machine translation, several metrics including BLEU [51], ROUGE_L [52], METEOR [53], CIDEr [54], and SPICE [55] are used as evaluations of the captioning task, which are also adopted in this article. From an intuitive view, the co-occurrences of $n$-gram between two sentences can be used to measure the similarity of two sentences. $n$-gram is a set of ordered words and $n$ is 1, 2, 3, or 4. BLEU is an evaluation applying co-occurrence strategy. According to the length of *longest common subsequence*, F-measure is computed as a similarity metric called ROUGE_L. METEOR obtains an alignment between two sentences on the basis of harmonic mean of unigram recall and precision. CIDEr can reflect the consensus of sentences by adding a TF-IDF weight coefficient for every $n$-gram. By emphasizing the semantic propositional content, SPICE is designed to evaluate the captioning task utilizing scene graphs [55].

And the higher the metrics are, the more likely the generated sentences are to the reference sentences in datasets.

### C. Compared Approaches

To verify the performance of RTRMN, comparative experiments are conducted on several captioning methods.

1) *BOW+cos* [21] is a retrieval-based method in which the distance between image representation and sentence representation is measured by cosine distance. The image is represented by hand-crafted features. The sentence is represented by *Bags Of Words* (BOW). For fair comparison, the hand-crafted features here are replaced with deep features.
2) *Deep Visual-SEmantic* (DeViSE) [22] is a retrieval-based method. What is different from [21] is that the sentences in [22] are represented by the mean of their word embedding vectors. The word embedding vectors used here are *Global Vectors* (GloVe) [56].
3) *Collective Semantic Metric Learning Framework* (CSMLF) [20] is a retrieval-based method. Five sentences in dataset are represented by a collective sentence proposed in [20]. The collective sentence representation and image representation are embedded into a common semantic space by the metric learning method.
4) *mRNN* [1] first uses RNN to generate caption. The image feature in mRNN is imported into RNN at every step. The implementation is based on naive Python.
5) *mLSTM* [1] utilizes LSTM, instead of RNN, to generate sentences. The image feature in mLSTM is imported into LSTM only in the first step. The implementation is based on naive Python.
6) *mGRU* encodes the image with CNN, specifically, VGG-16. The decoder is GRU, a variant of the naive RNN. The

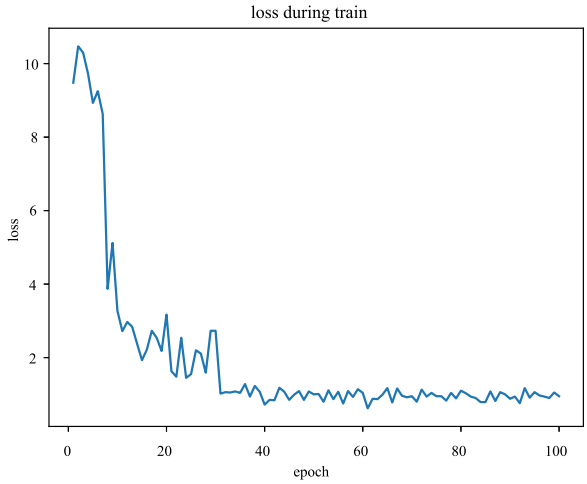



Fig. 7. Validation loss after different epochs during training.

implementation of this method is based on Keras, one of the most popular deep learning frameworks.

7) *Convcap* [32] encodes the image with CNN, specifically, VGG-16. The attention weights are computed by using convolutional layer activations. The decoder is CNN designed in [32]. The implementation of this method is based on PyTorch, one of the most popular deep learning frameworks. The default parameters of the proposed code in GitHub are utilized to generate the results in this article.
8) *Soft-attention* encodes the image with CNN, specifically, VGG-16. The decoder is LSTM, a variant of the naive RNN. The feature maps of *conv5_3* sized $14 \times 14 \times 512$ are used. The attention parameters are computed by using hidden states of LSTM. Different parts of feature maps are given different weights to decide the attention part. The implementation of this method is based on Theano, one of the most popular deep learning frameworks.
9) *Hard-attention* encodes the image with CNN, specifically, VGG-16. The decoder is LSTM, a variant of the naive RNN. The feature maps of *conv5_3* sized $14 \times 14 \times 512$ are used. The attention parameters are computed by using hidden states of LSTM. The sampling strategy is used to decide the attention part. The implementation of this method is based on Theano, one of the most popular deep learning frameworks.

### D. Experimental Setup

The method is implemented using Keras. The $C$ in the process of extracting statistical topics is set to 10. The number of temporary memory cells is set to 25 in this article. Adam optimizer is utilized to train the model. The initial learning rate is set to 0.001. When validation loss does not descend for five epochs, the learning rate is divided by 10. To make the model concentrate on captioning generation part, the learning rate of image embedding part is set to 0.0001 and the word is represented by GloVe. The batch size is set to be 200. The validation loss is shown in Fig. 7. The validation loss reaches a relative stable value after about 40 epochs. In order to ensure that the model converges to a stable value, the number of epoch is set 100.

TABLE II
RESULTS OF DIFFERENT DISTANCE MEASURE METHODS ON UCM-CAPTIONS

| | manhan | cos | cheby | eucli |
|---|---|---|---|---|
| BLEU-1 | 63.06 | **63.52** | 59.92 | 63.28 |
| BLEU-2 | 56.71 | 57.06 | 53.14 | **57.07** |
| BLEU-3 | 51.62 | 51.83 | 48.12 | **52.04** |
| BLEU-4 | 46.97 | 47.15 | 43.70 | **47.37** |
| METEOR | 35.47 | **35.99** | 32.55 | 35.82 |
| ROUGE_L | 67.91 | **68.36** | 63.73 | 68.21 |
| CIDEr | 273.96 | 271.62 | 244.32 | **274.67** |
| SPICE | 41.65 | 41.58 | 38.26 | **41.59** |

"encli" represents Euclidean Distance, "cos" represents Cosine Distance, "cheby" represents Chebyshev Distance and "manhan" represents Manhattan Distance.
The bold font shows the best score for every evaluation metric.

The retrieval strategy is used to get the topic words for testing images and different retrieval strategies will affect the results. Here, the following different distance measure methods are tested. The $\mathrm{dis}(\mathbf{I}_{\mathrm{test}}^{p5}, \mathbf{I}_{k}^{p5})$ in (5) could be the Euclidean Distance, Manhattan Distance, Chebyshev Distance, or Cosine Distance. In addition, for the Cosine Distance, (5) should be maximal. In order to determine the proper type of the distance measure method, experiments of different distances are conducted on UCM-captions. The model is trained for 100 epochs with the previous setting, the trained model is fixed. For the testing set, the noun words are picked up as topic information. Different distance measure methods are adopted for getting the noun words topics for testing images. With the model fixed, the testing results of different distance measure methods are listed in Table II. It can be seen from Table II that the captioning results of the Euclidean Distance method are best among four types of distance measure methods according to BLEU-2, BLEU-3, BLEU-4, CIDEr, and SPICE. The captioning results of the Cosine Distance are best according to BLEU-1, METEOR, and ROUGE. In conclusion, the Euclidean Distance is selected as the distance measure method to get the topic information for testing images based on the experimental results.

The sampling strategy in retrieval has an important influence on efficiency. With the increase of a sampling number from the dataset, comparison time of the testing phase will be longer and longer. How to compromise between time and effect needs to be verified by experiments. For UCM-captions, 1680 images are used for training the model, which means 1680 times of comparison is needed for every testing image. This order of magnitude comparison is acceptable to hardware. So sampling strategy is not used on UCM-captions. For RSICD, 8734 images are used for training the model, which means the number of comparisons is 5.19 times that of the UCM-captions. In order to provide reference results in the case of insufficient computing resources, comparison experiments of different samples are conducted on RSICD. In order to ensure that the sampling images of different topics are fixed, the random seed is set before sampling. As shown in Figs. 8–12, the overall performance of the algorithm shows a trend of improvement with the increase of the sampling number. The number of topic words is an

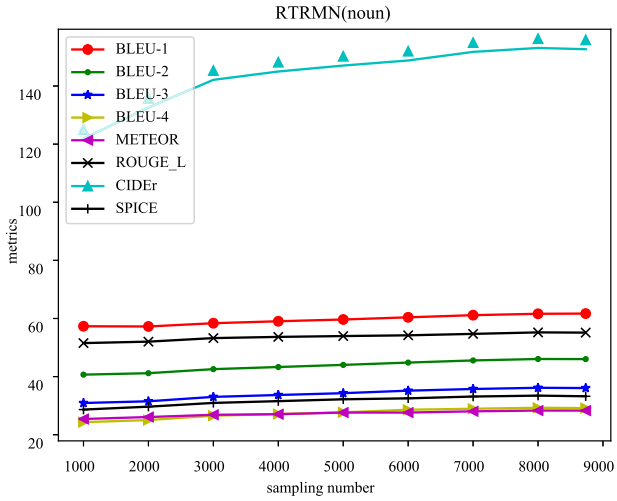



Fig. 8. RTRMN (noun) results of different sampling number images.

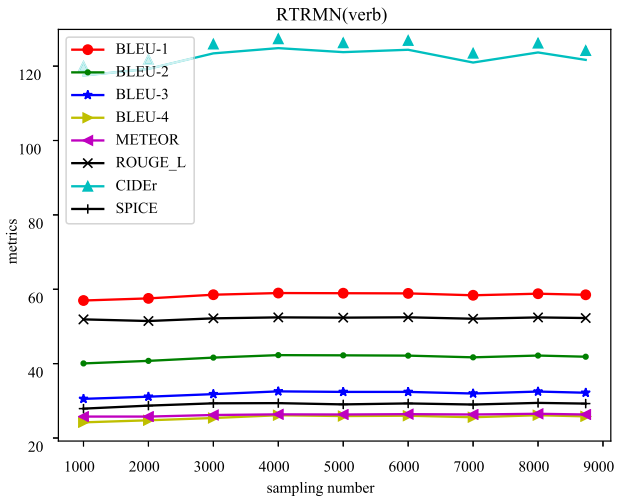



Fig. 9. RTRMN (verb) results of different sampling number images.

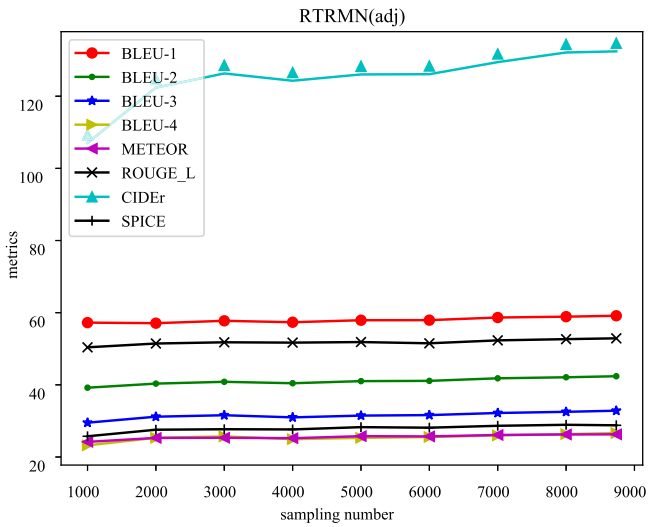



Fig. 10. RTRMN (adj) results of different sampling number images.

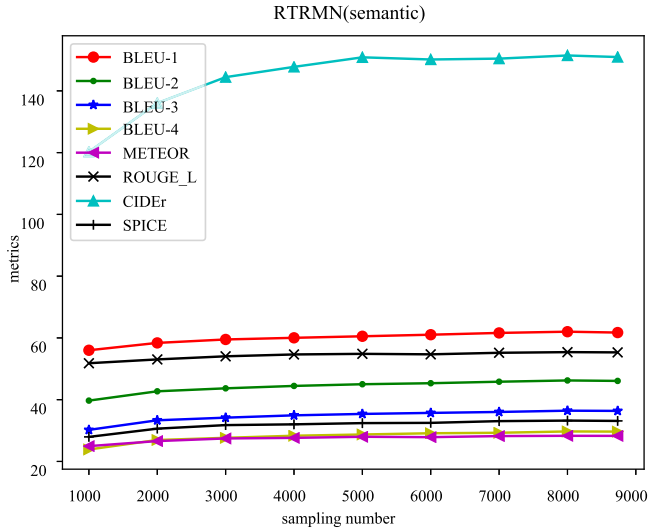



Fig. 11. RTRMN(semantic) results of different sampling number images.

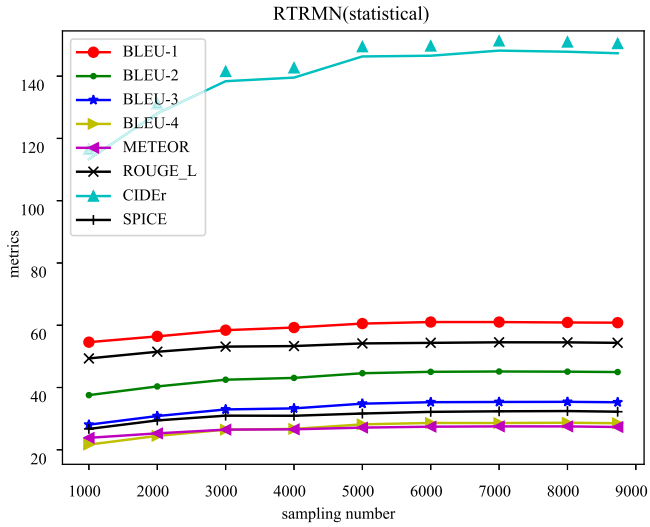



Fig. 12. RTRMN(statistical) results of different sampling number images.

important hyperparameter in the proposed method. To verify the best choice of topic number, experiments are conducted on UCM-captions. Note that the image representation in this experiment is VGG-16. The captioning evaluation results of different topic numbers are shown in Fig. 13. It can be seen that the captioning evaluation metrics and the number of topic words remain positively correlated. When the topic number increase from 0 to 2, captioning metric CIDEr increases from 170.31 to 320.78. This shows the importance of topics in our method. As the topic number increases, the evaluation still remains an upward trend. To get the best metrics, the number of topic words is set to 10 in this article.

### *E. Experimental Results*

To validate the proposed RTRMN more specifically, the experiments without topic words are conducted, which is denoted

TABLE III
RESULTS OF DIFFERENT METHODS ON UCM-CAPTIONS

| | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | METEOR | ROUGE_L | CIDEr | SPICE |
|---|---|---|---|---|---|---|---|---|
| BOW+cos [21] | 40.59 | 25.47 | 18.42 | 14.35 | 14.39 | 36.59 | 41.59 | 8.35 |
| DeViSE [22] | 36.97 | 17.40 | 9.77 | 5.98 | 9.77 | 29.73 | 9.70 | 3.22 |
| CSMLF [20] | 36.71 | 14.85 | 7.63 | 5.05 | 9.44 | 29.86 | 13.51 | 2.85 |
| mRNN [1] | 60.10 | 50.70 | 32.80 | 20.80 | 19.30 | - | 214.00 | - |
| mLSTM [1] | 63.50 | 53.20 | 37.50 | 21.30 | 20.30 | - | 222.50 | - |
| mGRU [57] | 42.56 | 29.99 | 22.91 | 17.98 | 19.41 | 37.97 | 124.82 | 30.58 |
| convcap [32] | 70.34 | 56.47 | 46.24 | 38.57 | 28.31 | 59.62 | 190.15 | 29.48 |
| soft-attention [13] | 74.54 | 65.45 | 58.55 | 52.50 | 38.86 | 72.37 | 261.24 | - |
| hard-attention [13] | **81.57** | 73.12 | 67.02 | 61.82 | **42.63** | 76.98 | 299.47 | - |
| RTRMN (none) | 67.03 | 57.22 | 51.74 | 48.26 | 31.05 | 61.86 | 209.19 | 32.35 |
| RTRMN (noun) | 76.41 | 69.80 | 64.78 | 60.46 | 40.84 | 74.33 | 293.64 | 44.54 |
| RTRMN (adj) | 60.56 | 55.55 | 51.74 | 48.46 | 35.67 | 66.96 | 283.24 | 40.43 |
| RTRMN (verb) | 73.22 | 64.49 | 58.75 | 54.20 | 36.68 | 70.79 | 260.05 | 38.63 |
| RTRMN (semantic) | 55.26 | 45.15 | 39.62 | 35.87 | 25.98 | 55.38 | 180.25 | 26.57 |
| RTRMN (statistical) | 80.28 | **73.22** | **68.21** | **63.93** | 42.58 | **77.26** | **312.70** | **45.35** |

The results of BOW+cos, DeViSE, and CSMLF are from paper [20]. The results of mRNN and mLSTM are from paper [1]. The results of mGRU are gained by our own reimplementation of methods in paper [57] with Keras. The results of soft-attention and hard-attention are from paper [13]. The results of convcap are gained by the released code in [32]. "-" means the metric is not available in [1] and [13]. The bold font shows the best score for every evaluation metric.

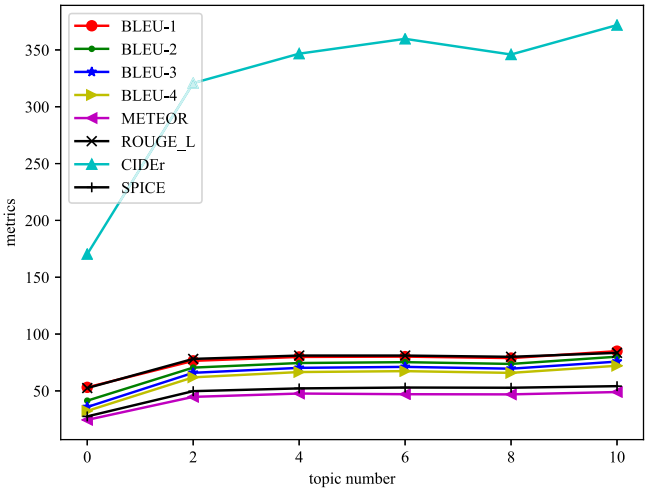



Fig. 13. Results of different topic numbers.

as RTRMN (none). Following the same strategies, the RTRMN (noun) means the topic words of nouns, the RTRMN (adj) means the topic words of adjectives and the RTRMN (verb) means the topic words of verbs. The RTRMN (semantic) means all the semantic topics information. The RTRMN (statistical) means the statistical topics, which represents the importance of information in the brain.

*1) Results on UCM-Captions:* For UCM-captions, the number of training images is not tremendous; thus, the sampling strategy is not involved. Table III lists the experimental results of different captioning methods on UCM-captions. It can be seen that the performance of the proposed method is much better than that of BOW+cos, DeViSE, and CSMLF. This is because BOW+cos, DeViSE, and CSMLF are retrieval-based methods. Generally speaking, the performance of the generation-based method is better than that of the retrieval-based method in the RSI captioning field. The performance of RTRMN (noun) is better than that of convcap. Both RTRMN (noun) and convcap are based on CNN. RTRMN (noun) is not only simple in structure, but also has better performance. The performance of RTRMN (noun) can be compared to that of mRNN, mLSTM, and mGRU. Even without adding the topic information, the proposed method in this article can achieve good performance in caption generation.

*2) Ablation Analysis on UCM-Captions:* In order to verify the role of topic words, RTRMN (none) is used as a comparison. When nouns are added as topic words, CIDEr of the RTRMN is raised from 209.19 to 293.64. This performance gain indicates that the noun is very helpful to improve the performance of the algorithm. When adjectives are added as topic words, CIDEr of the RTRMN is raised from 209.19 to 283.24. This performance gain indicates that the adjective is very helpful to improve the performance of the algorithm. When verbs are added as topic words, CIDEr of the RTRMN is raised from 209.19 to 260.05. This performance gain indicates that the verb is very helpful to improve the performance of the algorithm. When nouns, adjectives, and verbs are used as topic words, the performance gains are 84.45, 74.05, and 50.86, respectively. The performance gain of nouns is greater than that of verbs and adjectives, which is consistent with common sense. The RTRMN (semantic) in Table III gives a poor performance even poorer than that of RTRMN (none). To explore this abnormal phenomena, some semantic topics are shown in Fig. 14. The semantic topics in the middle of the Fig. 14 are messy and redundancy. Specifically, the theme of the first image in Fig. 14 is medium residential. However, the semantic topics of this image, which related to medium residential, are not included. This may be the retrieval process cannot capture the right corresponding topics. Even though, the generated caption captures the words "medium residential." But for the second image, RTRMN (semantic) mistakenly identifies the forest as a golf course. For the third image, RTRMN (semantic) mistakenly identifies the airplane as a house. Thus, semantic topics, considering the nouns, adjectives, and verbs, may cause

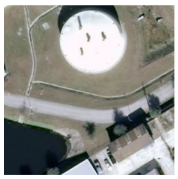
| Image | Topics | Generated caption |
|---|---|---|
|  | **semantic topics:** desolate golf parallel emerald scattered cropland sands course colors spaces<br>**edited semantic topics:** green medium residential<br>**noun topics:** medium houses plants area<br>**edited noun topics:** medium residential | a medium residential area with houses and plants .<br>a house with verdant lawn surrounded and a road beside .<br>medium residential area with plants surrounded .<br>this is a medium residential area . |
|  | **semantic topics**: desolate golf parallel emerald scattered cropland course spaces parked go<br>**edited semantic topics**: green plants<br>**noun topics**: forest dense plants lots plants<br>**edited noun topics**: forest plants | a part of a golf course with green turfs and some bunkers and trees .<br>this is a dense forest with dark green plants .<br>lots of emerald green plants in the dense forest.<br>there are some grey plants scattered on the ground . |
|  | **semantic topics**: desolate scattered cropland parked go ground compose green-blue residential black<br>**edited semantic topics**: airplane runway parked<br>**noun topics**: airplanes airport luggage cars airplane grey<br>**edited noun topics**: airplane airport runway | a house with blue roofs is surrounded by verdant lawn in the sparse residential area .<br>there are two airplanes stopped at the airport .<br>airplane beside beside beside beside beside beside.<br>an airplane with some plants beside beside beside beside. |
|  | **semantic topics**: desolate scattered cropland parked go ground compose green-blue residential black<br>**edited semantic topics**: storage tank cropland<br>**noun topics**: storage tank lawn houses<br>**edited noun topics**: storage tank | roadside the roadside roadside roadside roadside roadside roadside.<br>composed of boats docked neatly at the harbor .<br>beside beside beside beside beside beside beside.<br>beside beside beside beside beside beside beside. |

Fig. 14. Results of editable captioning examples. The red captions in right are the captions generated by the edited topics.

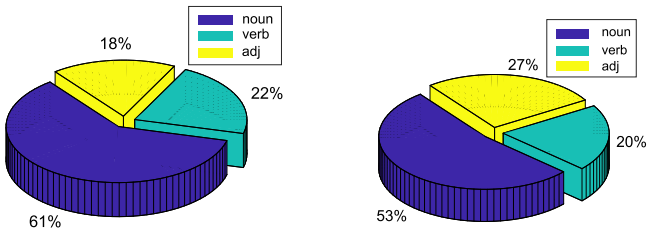



Fig. 15. Statistics illustration of different attribute words on two datasets. Left: UCM-captions. Right: RSICD.

confusion and redundancy of information, which results in poor performance. One potential solution is to edit topic words, as detailed in Section IV-F. Specially, the generated caption for the fourth image in Fig. 14 is just a repeated series of a same word. The reason for this phenomenon may be that the prediction is carried out separately and the correlation between the steps is not captured enough.

The distribution of words in different datasets is different. For a more fair comparison, the proportions of words in different attributes in the datasets are shown in Fig. 15. As shown in Fig. 15, 61% of the topic words are nouns in UCM-captions. This shows the importance of nouns to the generation of sentences. The results of the RTRMN (noun) are better than the results of RTRMN (adj) and RTRMN (verb), as listed in Table III. The experimental results are consistent with the statistical results of nouns. The performance of RTRMN (statistical) is better than that of soft-attention according to all metrics. The performance of RTRMN (statistical) is better than that of hard-attention according to BLEU-2, BLEU-3, BLEU-4, ROUGE_L, CIDEr, and SPICE. In summary, the proposed RTRMN in this article can enhance the performance of caption generation.

*3) Results on RSICD:* For RSICD, the number of training images is relative tremendous; thus, the sampling number is considered. The best performance of different sampling numbers for different topics is different, the best performances are reported in this article. Table IV describes the experimental results of different captioning methods on RSICD. It can be seen that the performance of the proposed method is much better than that of BOW+cos, DeViSE, and CSMLF. This is because BOW+cos, DeViSE, and CSMLF are retrieval-based methods. Generally speaking, the performance of the generation-based method is better than that of the retrieval-based method in the RSI captioning field. The performance of RTRMN (noun) and convcap is comparable. But the structure of RTRMN is much simpler than that of convcap. The performance of RTRMN (noun) can be compared to that of mGRU. Even without adding the topic information, the proposed method in this article can achieve good performance in caption generation.

The performance of the proposed method is better on UCM-captions compared with that of RSICD. We argue that the reason is due to mismatches between retrieved topic words and testing images caused by the retrieval process in RTRMN. This is related to properties of dataset. For UCM-captions, the scene is relatively simple and the resolution difference of the image is relatively small compared with RSICD. Some

TABLE IV
RESULTS OF DIFFERENT METHODS ON RSICD

| | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | METEOR | ROUGE_L | CIDEr | SPICE |
|---|---|---|---|---|---|---|---|---|
| BOW+cos [21] | 29.68 | 11.28 | 5.81 | 3.39 | 9.61 | 25.09 | 12.89 | 6.47 |
| DeViSE [22] | 30.68 | 11.38 | 5.58 | 3.07 | 9.73 | 25.63 | 12.44 | 5.64 |
| CSMLF [20] | 51.06 | 29.11 | 19.03 | 13.52 | 16.93 | 37.89 | 33.88 | 13.49 |
| mRNN [1] | 45.58 | 28.25 | 18.09 | 12.13 | 15.69 | 31.26 | 19.15 | 10.94 |
| mLSTM [1] | 50.57 | 32.42 | 23.19 | 17.46 | 17.84 | 35.02 | 31.61 | 16.80 |
| mGRU [57] | 42.56 | 29.99 | 22.91 | 17.98 | 19.41 | 37.97 | 124.82 | 30.58 |
| convcap [32] | 63.36 | 51.03 | 41.74 | 34.52 | 33.25 | 57.70 | 166.48 | **39.33** |
| soft-attention [13] | **67.53** | **53.08** | **43.33** | **36.17** | **32.55** | **61.09** | **196.43** | - |
| hard-attention [13] | 66.69 | 51.82 | 41.64 | 34.07 | 32.01 | 60.84 | 179.25 | - |
| RTRMN (none) | 57.98 | 39.37 | 29.04 | 22.51 | 24.76 | 51.24 | 88.45 | 24.99 |
| RTRMN (noun) | 61.62 | 46.06 | 36.15 | 29.28 | 28.32 | 55.22 | 153.10 | 33.45 |
| RTRMN (adj) | 59.15 | 42.39 | 32.83 | 26.62 | 26.28 | 52.89 | 132.43 | 28.79 |
| RTRMN (verb) | 58.80 | 42.17 | 32.49 | 26.13 | 26.52 | 52.43 | 123.68 | 29.44 |
| RTRMN (semantic) | 62.01 | 46.23 | 36.44 | 29.71 | 28.29 | 55.39 | 151.46 | 33.22 |
| RTRMN (statistical) | 61.02 | 45.14 | 35.35 | 28.59 | 27.51 | 54.52 | 148.20 | 32.36 |

The results of BOW+cos, DeViSE, and CSMLF are from paper [20]. The results of mRNN and mLSTM are from paper [1]. The results of mGRU are gained by our own reimplementation of methods in paper [57] with Keras. The results of soft-attention and hard-attention are from paper [13]. The results of convcap are gained by the released code in [32]. "-" means the metric is not available in [13]. The bold font shows the best score for every evaluation metric.

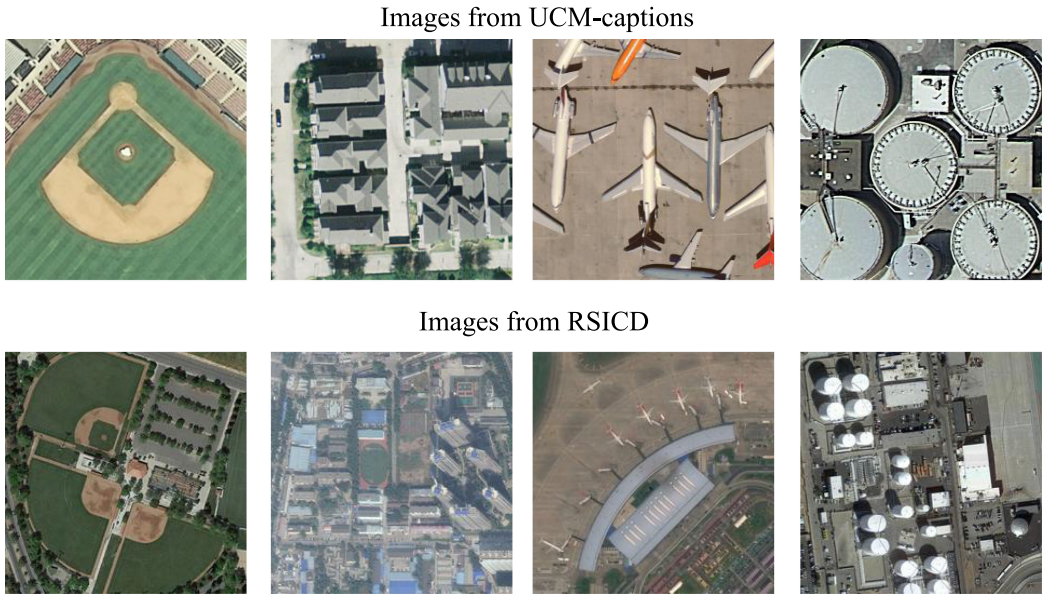



Fig. 16. Images from two datasets. The images from RSICD are relatively more complicated than that from UCM-captions.

examples are shown in Fig. 16. It can be seen that the scenarios are more complex in RSICD, even if the categories of images are the same. Furthermore, the number of scenarios in RSICD is 30 while the number of scenarios in UCM-captions is 21. The topics of the test images are obtained by retrieval. The two reasons mentioned above cause the retrieval accuracy to be lower for RSICD. To validate that the performance drop is caused by the retrieval process, the ground truth topic words are utilized as a comparison. Comparing the results of Tables IV and V, the ground truth topic words can boost the CIDEr from 148.20 to 377.52 compared to retried topic words for RTRMN (statistical). Note that the retrieval strategy in this article is relatively simple and practicable, which can be replaced by more powerful retrieval methods.

*4) Ablation Analysis on RSICD:* In order to verify the role of topic words, RTRMN (none) is used as a comparison. When nouns are added as topic words, CIDEr of the RTRMN is raised from 88.45 to 153.10. This performance gain indicates that the noun is very helpful to improve the performance of the algorithm. When adjectives are added as topic words, CIDEr of the RTRMN is raised from 88.45 to 132.43. This performance gain indicates that the adjective is very helpful to improve the performance of the algorithm. When verbs are added as topic words, CIDEr of the RTRMN is raised from 88.45 to 123.68. This performance gain indicates that the verb is very helpful to improve the performance of the algorithm. When nouns, adjectives, and verbs are used as topic words, the performance gains are 64.65, 43.98, and 35.23, respectively. The performance

TABLE V
RESULTS OF GROUND TRUTH TOPIC WORDS ON RSICD

| | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | METEOR | ROUGE_L | CIDEr | SPICE |
|---|---|---|---|---|---|---|---|---|
| RTRMN (noun) | 76.44 | 63.37 | 53.67 | 46.30 | 38.33 | 69.19 | 302.78 | 51.87 |
| RTRMN (adj) | 69.23 | 53.06 | 43.55 | 37.05 | 32.06 | 61.09 | 210.33 | 36.46 |
| RTRMN (verb) | 64.61 | 47.79 | 37.50 | 30.47 | 28.97 | 58.38 | 156.35 | 32.14 |
| RTRMN (semantic) | 81.54 | 70.47 | 62.43 | 56.22 | 42.40 | 74.75 | 361.47 | 53.94 |
| RTRMN (statistical) | 84.55 | 72.31 | 63.08 | 56.02 | 40.56 | 73.46 | 377.52 | 52.72 |


Fig. 17. Results of captioning examples. Red: reference sentence. Black: sentence generated by mLSTM. Blue: sentence generated by RTRMN.

gain of nouns is greater than that of verbs and adjectives, which is consistent with common sense.

The distribution of words in different datasets is different. For a more fair comparison, the proportions of words in different attributes in the datasets are shown in Fig. 15. As shown in Fig. 15, 53% of the topic words are nouns in RSICD. This shows the importance of nouns to the generation of sentences. The results of the RTRMN (noun) are better than the results of RTRMN (adj) and RTRMN (verb), as listed in Table IV. The experimental results are consistent with the statistical results of nouns. The performance of RTRMN (noun) is not comparable to that of soft-attention and hard-attention. This may be due

to the addition of attention mechanisms in soft-attention and hard-attention. In summary, the proposed RTRMN in this article can enhance the performance of caption generation.

To give a more intuitive illustration, some examples are presented, as shown in Fig. 17. The proposed RTRMN can capture the main topic of the image more accurately. For example, the second image of the first row in Fig. 17, the determinate topic of the image is school. But the caption generated by mLSTM shows the topic is residential area, not school. While the proposed RTRMN can generate caption that capture the school.

### *F. Controllability of Caption Generation*

The controllability of caption generation is a valuable research direction, because many times, sentences generated by the model may not be what we want. Changing the generated caption by changing the input topic is a potential advantage of the method proposed in this article. To show this controllability, some examples are shown in Fig. 14. Two types of topic words are edited in Fig. 14: semantic topics, and noun topics. Take the first image as an example, the original noun topics are "medium," "houses," "plants," and "area." The corresponding generated caption of the original noun topic is "medium residential area with plants surrounded." After editing, the edited noun topics are "medium" and "residential." The corresponding generated caption of edited noun topics is "this is a medium residential area." It can be seen that the "plants" disappears in the generated caption utilizing edited noun topics. This is because the "plants" is wiped off after editing. In summary, the generation of the caption can be controlled by editing the topics.

## V. Conclusion

In this article, a novel RTRMN is proposed to generate a determinate sentence, which can overcome the shortcoming of long-term information dilution in RNN. A topic word strategy is a direct way to utilize the existing five sentences in caption datasets. This is a flexible way, which can be expanded up to more sentences and even documents. Furthermore, the method proposed in this article can control the generation result of caption by editing topic words. Thus, the proposed method sheds light on controllability of caption generation. Although the proposed approach has achieved good captioning results, in some cases sentences with no practical meaning will be generated, such as the fourth image captioning results in Fig. 14. How to reduce the occurrence of these conditions will be left for future work.

## References


[1] 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. 124–128.

[2] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *IEEE Comput. Intell. Mag.*, vol. 13, no. 3, pp. 55–75, Aug. 2018.

[3] X. Zheng, Y. Yuan, and X. Lu, "Hyperspectral image denoising by fusing the selected related bands," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 5, pp. 2596–2609, May 2019.

[4] Z. Shao and J. Cai, "Remote sensing image fusion with deep convolutional neural network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 5, pp. 1656–1669, May 2018.

[5] F. Luo, B. Du, L. Zhang, L. Zhang, and D. Tao, "Feature learning using spatial-spectral hypergraph discriminant analysis for hyperspectral image," *IEEE Trans. Cybern.*, vol. 49, no. 7, pp. 2406–2419, Jul. 2019.

[6] F. Luo, L. Zhang, X. Zhou, T. Guo, Y. Cheng, and T. Yin, "Sparse-adaptive hypergraph discriminant analysis for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, to be published, doi: 10.1109/LGRS.2019.2936652.

[7] X. Zheng, Y. Yuan, and X. Lu, "A deep scene representation for aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4799–4809, Jul. 2019.

[8] X. Lu, H. Sun, and X. Zheng, "A feature aggregation convolutional neural network for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7894–7906, Oct. 2019.

[9] W. Li, G. Wu, F. Zhang, and Q. Du, "Hyperspectral image classification using deep pixel-pair features," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 844–853, Feb. 2017.

[10] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1923–1938, Apr. 2019.

[11] X. Zhang, X. Li, J. An, L. Gao, B. Hou, and C. Li, "Natural language description of remote sensing images based on deep learning," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2017, pp. 4798–4801.

[12] Z. Shi and Z. Zou, "Can a machine generate humanlike language descriptions for a remote sensing image?" *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 6, pp. 3623–3634, Jun. 2017.

[13] 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.

[14] C. T. Recchiuto and A. Sgorbissa, "Post-disaster assessment with unmanned aerial vehicles: A survey on practical implementations and research approaches," *J. Field Robot.*, vol. 35, no. 4, pp. 459–490, 2018.

[15] Q. Liu, C. Ruan, S. Zhong, J. Li, Z. Yin, and X. Lian, "Risk assessment of storm surge disaster based on numerical models and remote sensing," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 68, no. 6, pp. 20–30, 2018.

[16] Y. Choi, Y. Choi, S. Briceno, and D. N. Mavris, "Three-dimensional UAS trajectory optimization for remote sensing in an irregular terrain environment," in *Proc. Int. Conf. Unmanned Aircr. Syst.*, 2018, pp. 1101–1108.

[17] T. N. Bhagwat, V. Hegde, and A. Shetty, "Application of remote sensing and GIS for identification of potential ground water recharge sites in semi-arid regions of hard-rock terrain, in north Karnataka, south India," *Sustain. Water Resour. Manage.*, vol. 4, pp. 1063–1076, 2018.

[18] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[19] X. Liu, H. Li, J. Shao, D. Chen, and X. Wang, "Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 338–354.

[20] B. Wang, X. Lu, X. Zheng, and X. Li, "Semantic descriptions of high-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1274–1278, Aug. 2019.

[21] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *J. Artif. Intell. Res.*, vol. 47, no. 8, pp. 853–899, 2013.

[22] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, and T. Mikolov, "Devise: A deep visual-semantic embedding model," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2121–2129.

[23] J. Mun, M. Cho, and B. Han, "Text-guided attention model for image captioning," in *Proc. Assoc. Adv. Artif. Intell.*, 2017, pp. 4233–4239.

[24] K. Tran, A. Bisazza, and C. Monz, "Recurrent memory networks for language modeling," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol.*, 2016, pp. 321–331.

[25] K. Xu, Y. Lai, Y. Feng, and Z. Wang, "Enhancing key-value memory neural networks for knowledge based question answering," in *Proc. Conf. North Amer. Ch. Assoc. Comput. Linguistics: Human Lang. Technol.*, 2019, pp. 2937–2947.

[26] L. Ma, Z. Lu, L. Shang, and H. Li, "Multimodal convolutional neural networks for matching image and sentence," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2623–2631.

[27] C. C. Park, B. Kim, and G. Kim, "Attend to you: Personalized image captioning with context sequence memory networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6432–6440.

[28] S. Wang, J. Chen, and G. Wang, "Intensive positioning network for remote sensing image captioning," in *Proc. Int. Conf. Intell. Sci. Big Data Eng.*, 2018, pp. 567–576.
[29] X. Zhang, X. Wang, X. Tang, H. Zhou, and C. Li, "Description generation for remote sensing images using attribute attention mechanism," *Remote Sens.*, vol. 11, no. 6, p. 612, 2019.
[30] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.
[31] J. Gu, G. Wang, J. Cai, and T. Chen, "An empirical study of language CNN for image captioning," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1222–1231.
[32] J. Aneja, A. Deshpande, and A. G. Schwing, "Convolutional image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5561–5570.
[33] Q. Wang and A. B. Chan, "CNN+CNN: Convolutional decoders for image captioning," 2018. [Online]. Available: https://arxiv.org/abs/1805.09019v1
[34] F. Chen, R. Ji, X. Sun, Y. Wu, and J. Su, "Groupcap: Group-based image captioning with structured relevance and diversity constraints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1345–1353.
[35] P. Anderson *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6077–6086.
[36] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
[37] A. Mathews, L. Xie, and X. He, "Semstyle: Learning to generate stylised image captions using unaligned text," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 8591–8600.
[38] A. F. Biten, L. Gomez, M. Rusinol, and D. Karatzas, "Good news, everyone! Context driven entity-aware captioning for news images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12 466–12 475.
[39] L. Guo, J. Liu, P. Yao, J. Li, and H. Lu, "MSCap: Multi-style image captioning with unpaired stylized text," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4204–4213.
[40] D.-J. Kim, J. Choi, T.-H. Oh, and I. S. Kweon, "Dense relational captioning: Triple-stream networks for relationship-based captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6271–6280.
[41] A. Graves, G. Wayne, and I. Danihelka, "Neural turing machines," 2014. [Online]. Available: https://arxiv.org/abs/1410.5401
[42] J. Weston, S. Chopra, and A. Bordes, "Memory networks," 2014. [Online]. Available: https://arxiv.org/abs/1410.3916
[43] S. Sukhbaatar, J. Weston, and R. Fergus, "End-to-end memory networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2440–2448.
[44] A. Graves *et al.*, "Hybrid computing using a neural network with dynamic external memory," *Nature*, vol. 538, no. 7626, pp. 471–476, 2016.
[45] A. Kumar *et al.*, "Ask me anything: Dynamic memory networks for natural language processing," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1378–1387.
[46] A. H. Miller, A. Fisch, J. Dodge, A.-H. Karimi, A. Bordes, and J. Weston, "Key-value memory networks for directly reading documents," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 1400–1409.
[47] Y. Li, W. Ouyang, and X. Wang, "VIP-CNN: A visual phrase reasoning convolutional neural network for visual relationship detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1347–1356.
[48] H. Christian, M. P. Agus, and D. Suhartono, "Single document automatic text summarization using term frequency-inverse document frequency (TF-IDF)," *ComTech: Comput., Math. Eng. Appl.*, vol. 7, no. 4, pp. 285–294, 2016.
[49] 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.
[50] X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-encoding scene graphs for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10 685–10 694.
[51] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. Assoc. Comput. Linguistics*, 2002, pp. 311–318.
[52] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. Assoc. for Comput. Linguistics Workshop*, 2004, pp. 74–81.
[53] M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in *Proc. Assoc. Comput. Linguistics*, 2014, pp. 376–380.
[54] R. Vedantam, C. L. Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4566–4575.
[55] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 382–398.
[56] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1532–1543.
[57] X. Li, A. Yuan, and X. Lu, "Multi-modal gated recurrent units for image description," *Multimedia Tools Appl.*, vol. 77, no. 22, pp. 29 847–29 869, 2018.

**Binqiang Wang** received the B.S. degree from the School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an, China. He is currently working toward the Ph.D. degree in signal and information processing with the Key Laboratory of Spectral Imaging Technology CAS, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China.

His current research interests include pattern recognition, computer vision, and machine learning.

**Xiangtao Zheng** (Member, IEEE) received the M.Sc. and Ph.D. degrees in signal and information processing from the Chinese Academy of Sciences, Xi'an, China, in 2014 and 2017, respectively.

He is currently an Assistant Professor with the Key Laboratory of Spectral Imaging Technology CAS, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China. His main research interests include computer vision and pattern recognition.

**Bo Qu** received the B.S. degree in automation from Xi'an Jiaotong University, Xi'an, China, in 2014, and the M.S. degree in control engineering from the University of Chinese Academy of Sciences, Beijing, China, in 2017.

He is currently an Assistant Researcher with the Key Laboratory of Spectral Imaging Technology CAS, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China. His current research interests include image processing and pattern recognition.

**Xiaoqiang Lu** (Senior Member, IEEE) is a Full Professor with the Key Laboratory of Spectral Imaging Technology CAS, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, P. R. China. His current research interests include pattern recognition, machine learning, hyperspectral image analysis, cellular automata, and medical imaging.