IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 16, 2019, accepted December 13, 2019, date of publication December 25, 2019,
date of current version January 6, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2962195

Exploring Multi-Level Attention and Semantic Relationship for Remote Sensing Image Captioning

ZHENGHANG YUAN, XUELONG LI, (Fellow, IEEE), AND QI WANG, (Senior Member, IEEE)
School of Computer Science, Center for Optical Imagery Analysis and Learning, Northwestern Polytechnical University, Xi'an 710072, China

Corresponding author: Xuelong Li (li@nwpu.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFB1002202, in part by the National Natural Science Foundation of China under Grant 61871470, Grant 61761130079 and Grant U1801262, and in part by the Project of Special Zone for National Defense Science and Technology Innovation.

ABSTRACT Remote sensing image captioning, which aims to understand high-level semantic information and interactions of different ground objects, is a new emerging research topic in recent years. Though image captioning has developed rapidly with convolutional neural networks (CNNs) and recurrent neural networks (RNNs), the image captioning task for remote sensing images still suffers from two main limitations. One limitation is that the scales of objects in remote sensing images vary dramatically, which makes it difficult to obtain an effective image representation. Another limitation is that the visual relationship in remote sensing images is still underused, which should have great potential to improve the final performance. In order to deal with these two limitations, an effective framework for captioning the remote sensing image is proposed in this paper. The framework is based on multi-level attention and multi-label attribute graph convolution. Specifically, the proposed multi-level attention module can adaptively focus not only on specific spatial features, but also on features of specific scales. Moreover, the designed attribute graph convolution module can employ the attribute-graph to learn more effective attribute features for image captioning. Extensive experiments are conducted and the proposed method achieves superior performance on UCM-captions, Sydney-captions and RSICD dataset.

INDEX TERMS Remote sensing image, image captioning, deep learning, graph convolutional networks (GCNs), semantic understanding.

I. INTRODUCTION

With the great progress of remote sensing technology, high-quality remote sensing images are captured more easily, which provides a large number of available data for researches [1], [2]. The most common remote sensing image processing tasks, such as object detection [3], semantic classification [4] and change detection [5], [6], are dealing with objects or class labels, which can only provide limited object- or image-level information. However, generating natural-language descriptions for remote sensing image can provide richer high-level semantic information, such as scene structures or object relationships. Remote sensing image captioning, which aims to understand the high-level semantic information and the interactions of different ground objects, is distinct from the above-mentioned tasks. It provides far richer descriptions of remote sensing scene in a higher-semantic level by generating a corresponding sentence to abstract the content. Specifically, accurate and flexible sentences are generated automatically to describe the content of remote sensing images. Remote sensing image captioning identifies the ground objects under different levels and analyzes their attributes and spatial relationships in the aerial view [7]. Spatial relationships, also called visual relationships, contain the major elements of an image, which include the interactions between objects and the geometric information of objects. The interactions between objects are visual relationships which are embedded in image captions. For example, the caption ''Some white planes are in an airport'' describes the visual relationship between planes and airport.

Image captioning is a challenging task, which is a combination of Computer Vision (CV) and Natural Language

The associate editor coordinating the review of this manuscript and approving it for publication was Po Yang.

Processing (NLP). For the CV part, image representation is explored by handcrafted feature designing [8], [9] or deep feature leaning [10]–[12]. For the NLP part, sentence generation is conducted by grammar template-based generation models or deep neural networks based methods. Generally, the methods for image captioning can be roughly divided into three categories: object detection-based method, retrieval-based method, and encoder-decoder network based method [13].

The first type of image captioning methods is based on object detection [14], which first extracts image information by identifying objects and the relationships between them. The extracted information of the image includes three parts: 1) The detected objects (things and stuff). The localized areas contain the object-level information. 2) The visual attributes (word embedding). The attributes of each detection object, such as color or shape. 3) The visual relationships, such as ''on, besides, under, near, . . . ''. Then sentences are generated by the sentence generating model with the extracted information. Reference [15] took advantage of the labeled images from object recognition datasets to describe object categories that were not present in existing image captioning datasets. [16] first generated a sentence template with slot locations tied to specific image regions. Then these slots were filled by identifying the regions with object detectors. Therefore, the performance of object detection plays an important role in the final sentence generation. This two-stage method suffers from the error accumulation problem, and has low computation efficiency.

The second type is retrieval-based [17], which utilizes retrieval to generate the corresponding sentence. In this case, retrieval-based methods search for similar images with the query image, and then generate sentence according to the given sentences of similar images. However, the performance will degrade when the query image is different from the given training images. Retrieval-based method may generate sentences which are not relevant to the specific image content. This problem occurs because of the data sparsity. The number of remote sensing images is not enough to guarantee similar image matches, so the query image may have no similar match in the training set.

The third type of method is based on encoder-decoder [18], which first encodes the input image into a vector and then decodes this representation to generate the sentence. Generally, the encoder process usually uses CNNs to extract representative features, while the decoder process employs RNNs or Long-Short Term Memory (LSTM) to output the desired sentence. Current state-of-the-art image captioning methods are based on deep encoder-decoder framework. Reference [18] aimed to maximize the likelihood of the target description sentence for training the model. Reference [19] proposed a novel dual-stream RNN framework to integrate the hidden states of visual and semantic streams for caption generation. Reference [20] proposed an effective online positive recall and missing concepts mining method to address the dataset imbalance and incomplete labeling problems.

FIGURE 1. (a) Multi-scale objects in remote sensing images. (b) Relationships of multiple attributes in remote sensing images.

Reference [21] designed an effective decoder named multimodal RNN, which includes a bidirectional RNN to represent sentences, and a structured objective for leaning the multimodal representation. A convolutional image captioning model was proposed in [22], which employed a novel CNN model instead of LSTM for sentence generation.

Though considerable methods have been proposed for natural image captioning, remote sensing image-based captioning has not been fully studied. The main differences between natural image captioning and remote sensing image captioning are from three aspects: 1) There are apparent directional distinctions in remote sensing images, as they are captured from satellites or airplanes with the aerial view; 2) There is usually a lack of certain salient objects as remote sensing scene may contain many types of land-cover objects; 3) Scales of objects may vary significantly in remote sensing images. For example, the scales of airplanes vary greatly in the same scene, as shown in Fig. 1. Since the image features are encoded with CNNs, images containing objects of different scales are input to the same CNN for feature extraction. At the last convolution layer, the features of the same type of object will be quite different owing to the fixed receptive field. This inconsistency makes the image representation less effective. Due to the above-mentioned differences, it is difficult to directly apply the natural image captioning methods to the remote sensing image.

In recent years, many methods have been proposed for remote sensing image captioning task. A multimodal neural network [23] was proposed to understand the high spatial resolution (HSR) remote sensing images in semantic level, where different types of CNNs with RNNs or LSTMs were combined to select the best combination. Also, this paper constructed two datasets for remote sensing image captioning and there was no such dataset before this work. Shi *et al.* [7] proposed a two-stage based framework to generate human-like descriptions for remote sensing images, which were multilevel image understanding and language generation.

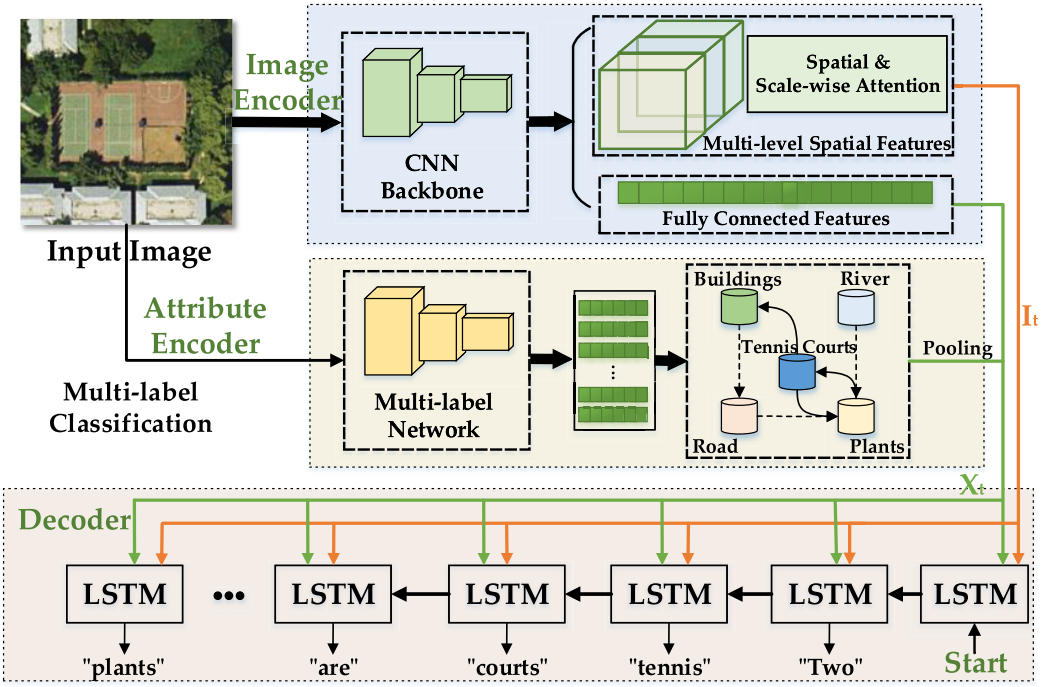

FIGURE 2. The overall network architecture of the proposed method. The image encoder extracts fully connected image features and multi-level attention-derived image features. The attribute encoder learns the attribute features by GCN and attribute-graph. All the extracted features are input to the LSTM decoder for sentence generation step by step. The Ground Truth label: "Two tennis courts are surrounded by some buildings and plants" is used for training the model in an end-to-end fashion.

Zhang *et al.* [24] leveraged CNNs to detect the main objects and utilized RNN language model to generate the descriptions of the detected objects. The work [13] presented some annotated instructions for the better description in terms of special characteristics such as scale ambiguity, rotation ambiguity and category ambiguity. Additionally, this work also constructed a large-scale benchmark dataset for remote sensing image captioning. In another work [25], a training mechanism of multi-scale cropping was proposed, which could collect more fine-grained information and improve the generalization performance. Wang *et al.* [26] proposed a collective semantic metric learning framework, which used semantic embedding to measure the presentation of image and sentence. The work [27] presented a framework based on attribute attention mechanism, which assigned different weights to different areas of remote sensing images.

Visual attention mechanism is widely used in natural image captioning. Natural images usually contain several salient objects, and the sentences mainly describe these objects and their relationships. Thus, attention mechanism can be used to adaptively focus on different objects to generate object-related words. SCA-CNN [28] exploited channel-wise and spatial-wise attention mechanism simultaneously for image captioning. Spatial-channel attention module was also employed and proved to be effective for image classification [29] and semantic segmentation tasks [30].

Although spatial attention is effective for natural image captioning, it still has limitation for remote sensing image-based captioning task. The main reason is that scales of the same type of object may vary significantly. As shown in Fig. 1(a), the scales of airplanes are quite different in different remote sensing images. Moreover, the relationships of objects in remote sensing scene shown in Fig. 1(b) are usually neglected in the task of captioning.

Considering the scale-variability problem depicted above, a multi-level attention-based method is proposed to adaptively focus on features of different scales for a more flexible remote sensing image representation. The motivations of this work are two-fold: 1) The proposed multi-level attention module includes a scale-wise attention and a spatial-wise attention subnetwork, which allows the model to learn features at specific positions and scales; 2) The visual relationships of local objects are critical for improving the performance of image captioning [31], [32]. However, exploiting object detection for local object-level extraction is time-consuming and needs extra annotations. Thus, in this work we aim to explore the visual relationships with image semantic attributes.

To sum up, the main contributions of this work can be summarized as follows.

(1) Different from previous attention-based methods, a novel multi-level attention module is proposed to focus on different spatial positions and different scales. The proposed module can not only extract specific spatial features adaptively, but also aim to learn features of specific scales.

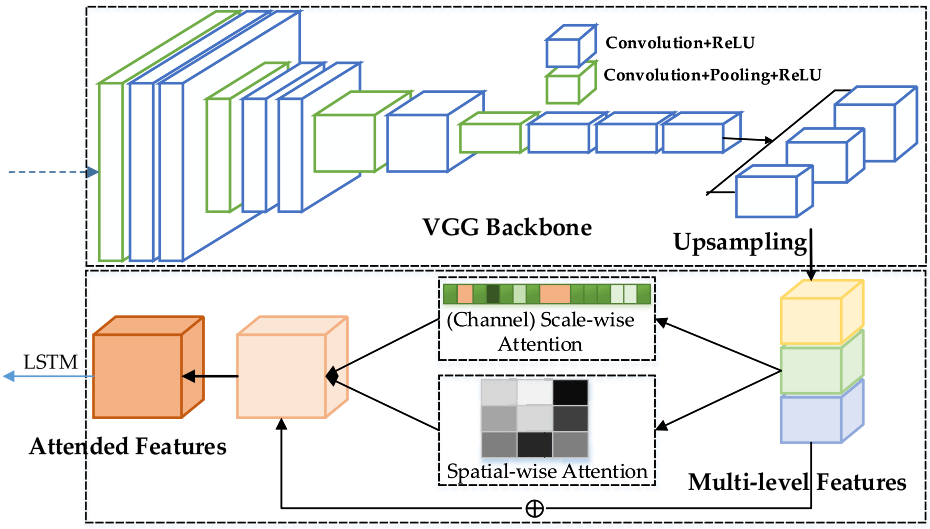

FIGURE 3. Illustration of the multi-level attention network, including multi-level features extraction, spatial- and scale-wise attention module.

(2) Multi-label classification task is introduced with no need for extra annotation in this work. By utilizing image attributes, more semantic information can be exploited for improving the performance of remote sensing image captioning.
(3) Attribute graph-based graph convolutional network (GCN) is designed to explore the semantic visual relationships of remote sensing images. By considering the attribute relationships, more robust and effective semantic features can be used for image captioning.

The rest of this paper is organized as follows. In Section II, the details of the proposed method are described. Then, extensive experimental results are shown and analyzed in Section III. Finally, we conclude this paper in Section IV.

II. METHODOLOGY

The overall network architecture of the proposed method is shown in Fig. 2. The whole framework consists of two main components. One is the deep learning-based encoding module, and the other is the LSTM-based decoding module. The encoding module includes two parts: 1) The image encoder extracts image representations by the proposed multi-level attention module; 2) The attribute encoder learns attribute features by GCN and pre-computed attribute-graph. The decoding module takes three image features as input: the multi-level attention-derived features, the fully connected (FC) features and the attribute features. It is worth mentioning that the multi-label classification network is trained in advance of the image captioning model, and its parameters are fixed when training the captioning model.

During the training stage, all the remote sensing images are first input into the image encoder for feature extraction. Then the obtained FC features and multi-level CNN feature maps are fed to LSTM cells for sentence generation. Cross entropy loss is used to train the network. Meanwhile, a multi-label classification network is pre-trained to generate image attributes. By utilizing image attributes, the attribute-graph can be constructed and the attribute relationships can be learned by GCN. After the mean pooling operation, the learned attribute features are also input to LSTM cells for caption generation. As shown in Fig. 2, ground truth labels are used for training by minimizing the cross entropy loss. During the inference stage, the input image is first encoded into image features and attribute features, then these features are input to LSTM to decode word step by step.

In this section, the details of the proposed framework will be introduced. Firstly, the multi-level attention module is described in section II-A. Secondly, the multi-label classification network for attribute generation and GCN-based attribute relationship mining are presented in section II-B. Finally, the detailed architecture for generating image description using modified LSTM is introduced in section II-C.

A. MULTI-LEVEL ATTENTION MODULE

Image representation is crucial for computer vision tasks including image captioning. Owing to the high-level semantic features extracted by deep neural networks, the CNN intermediate features are used as representation for remote sensing images. As aforementioned, the scales of objects may vary significantly. To handle this problem, multi-resolution features are exploited in this work. For CNN, the feature maps of deeper layer contain features of higher semantic-level, while feature maps of shallower layer are with higher resolutions. Thus, features of deeper layer are more suitable for large scale objects and features of shallower layer are more suitable for representing small scale objects. In general, the motivation of multi-level attention module is to make use of features of different levels adaptively for more effective representation of remote sensing images.

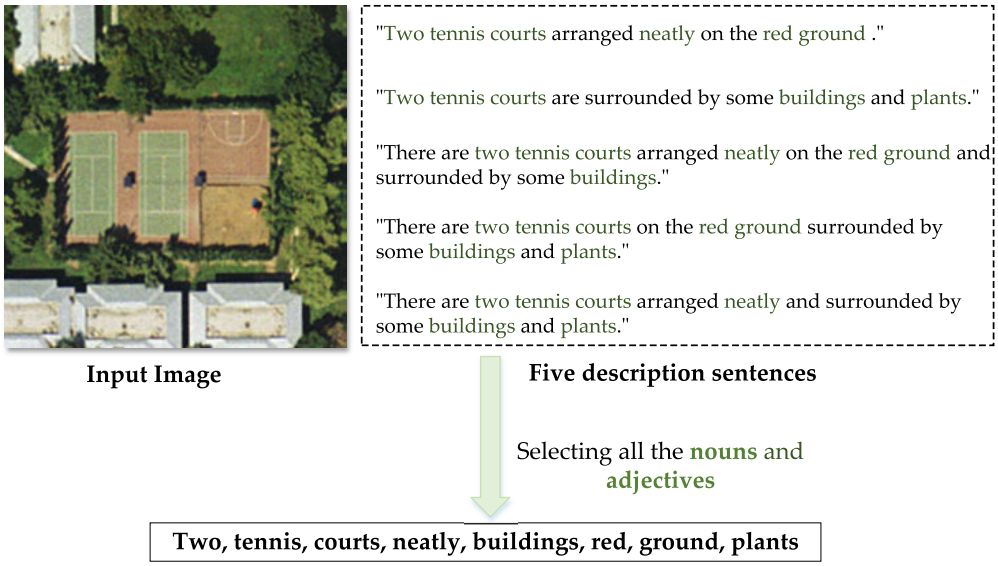

FIGURE 4. Illustration of the image attribute generation process. For each image, all the nouns and adjectives are selected as attributes.

Although there are several channel-attention and spatial attention-based methods [28]–[30], the multi-level attention method is proposed to learn dynamic scale-selective attention-derived features. By concatenating features from different layers (not only two), the multi-scale deep representation is embedded in different channels of the fused features. Then the channel-wise attention mechanism is employed to adaptively select features of different scales.

In this work, VGG network is chosen as our backbone for two reasons. Firstly, we mainly focus on exploring the effect of adaptive multi-scale features and more robust attribute features. Thus using different CNN backbones is not nessecery. Secondly, since VGG is widely used in prior works, we choose it as the CNN backbone to compare with other works fairly. Specifically, supposing that the input image is denoted as X, then the intermediate representation can be computed by:

$$\begin{aligned} F_{fc} &= VGG_{fc}(X), \\ F_{L1} &= VGG_{conv4}(X), \\ F_{L2} &= VGG_{conv5}(X), \end{aligned} \tag{1}$$

where F_{fc} is the feature of FC layer, F_{L1} and F_{L2} are the CNN feature maps of $conv4\_3$ and $conv5\_3$ of VGG16 respectively. Fig. 3 illustrates the details of multi-level attention network. Different from the illustration, we use two-level features to describe the proposed method for simplicity. As multi-level deep features contain information of different scales of images, these features are concatenated to focus on different spatial positions and scales adaptively during the LSTM decoding stage.

It is worth mentioning that the spatial resolution of multi-level feature maps for some network architectures may be different. In this case, the up-sampling layer can be used to enlarge the feature maps with lower spatial resolution, and make its spatial resolution to be the same with the larger one. Then the final multi-level features can be obtained by equation 2, i.e., concatenating the feature maps of different CNN layers:

$$F_{ml} = concat(F_{L1}, upsample(F_{L2})), \tag{2}$$

where F_{ml} is the final multi-level feature used for the description generation. After concatenating the feature maps of different CNN layers, the channels of F_{ml} contain CNN features of different levels. Thus channel-wise attention [33] on F_{ml} can learn selective features with different resolutions.

As different positions of CNN feature maps encode information of specific objects, using spatial-wise attention can enhance the network to focus on the appropriate objects during the LSTM decoding stage. In this work, spatial-wise and channel-wise attention are simultaneously exploited to focus not only on specific spatial position, but also on specific feature scales.

B. MULTI-LABEL ATTRIBUTE GRAPH CONVOLUTION

The image attribute is high-level concept, which contains global information of images, and has been exploited by previous works [27], [34]. However, the relationships between different attributes are not well exploited by prior remote sensing image captioning works, which are also critical for improving the image captioning performance.

To obtain the attributes of each remote sensing image, all the nouns and adjectives are selected from five sentences as illustrated in Fig. 4. The nouns and adjectives contain the global semantic information of the image, which are useful for description generation. Specifically, the sentences are first split into tokens, and then the stop words such as

"there, are, the, and, ..., by," are removed. After selecting the attributes for each image, a multi-label classification network is employed for the attribute generation. In this work, ResNet18 is used as the attribute classification network backbone. Supposing that I is the input image, the final output of ResNet18 is computed as:

$$V_{ml} = sigmoid(ResNet18(I)) \in \mathbb{R}^{N,K}, \tag{3}$$

where V_{ml} is the final output, and each dimension represents the probability of each attribute. N is the batch size, and K is the number of attributes. For the training of multi-label network, binary cross-entropy is used as the loss function, and Stochastic Gradient Descent (SGD) is used as the optimizer. It is worth mentioning that the multi-label classification network is trained independently, and it is fixed when training the image captioning network.

As relationships between different attributes are also important for image captioning task, a GCN-based feature learning module is designed to mine the attribute relationships in this work. GCN was introduced in [35] for semi-supervised classification task. The core idea of GCN is propagating information between nodes of the graph. Thus, in order to exploit GCN for learning features of relationships, the first step is to construct the adjacency matrix. In this work, the attribute embeddings are treated as the nodes in graph. Inspired by the work [36], we model the adjacency matrix with the conditional probability. Namely, $P(Attr_j|Attr_i)$ denotes the probability of occurrence of attribute $Attr_j$ when attribute $Attr_i$ appears. Note that $P(Attr_j|Attr_i)$ is not equal to $P(Attr_i|Attr_j)$. The probability $P(Attr_j|Attr_i)$ is computed as:

$$P(Attr_j|Attr_i) = \frac{N(Attr_j, Attr_i)}{N(Attr_i)}, \tag{4}$$

where $N(Attr_j, Attr_i)$ denotes the count of co-occurrence of $Attr_i$ and $Attr_i$, and $N(Attr_i)$ is the count of occurrence of attribute $Attr_i$. The computation for $P(Attr_i|Attr_j)$ is similar to $P(Attr_j|Attr_i)$, except that the denominator is changed to $N(Attr_j)$.

GCN [35] is an approximation of spectral graph convolution, and spectral graph convolution is operated in the Fourier domain. Supposing that the constructed graph is $G = (V, A)$, where V is the set of vertex (attribute word embedding) and A is the adjacency matrix. Then the Laplacian matrix L can be computed as:

$$L = D - A, \tag{5}$$

where D is the degree matrix of the graph, which can be defined as:

$$D_{ii} = \sum_j A_{ij}, \tag{6}$$

where i and j are the indexes of adjacency matrix A. Since L is a positive semidefinite matrix, it can be decomposed as:

$$L = U \Lambda U^T, \tag{7}$$

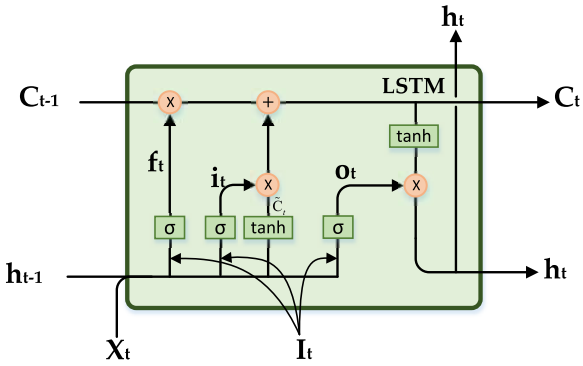

FIGURE 5. Illustration of the multi-level attention based LSTM decoding network. The input x_t is the combination of FC features and attribute features. I_t is the attention-derived multi-level feature map at time step t. At each step t during decoding, the feature map I_t will focus on different positions and scales. $h_{t-1}, h_t, c_{t-1}, c_t$ are the hidden state and cell state, which carry the context information from step $t-1$ to t. i_t, f_t, o_t are the output of input gate, forget gate and output gate respectively.

where U is the combination of eigenvectors, and $\Lambda = diag([\lambda_1, \ldots, \lambda_k])$ is the combination of eigenvalues. With these notations, the graph convolution in Fourier domain can be defined as:

$$Y = \sigma(U g_\theta(\Lambda) U^T X), \tag{8}$$

where g_θ is the convolution filter and σ is the non-linear function. X is the input and Y is the output. However, this form of graph convolution is time-consuming. The reason is that spectral decomposition needs high computational complexity when the graph is large. Considering this, [35] approximates the spectral GCN as equation 9:

$$\begin{aligned} Y &= (D+I)^{-\frac{1}{2}}(A+I)(D+I)^{-\frac{1}{2}} X W, \\ Y &= \sigma(Y), \end{aligned} \tag{9}$$

where W is the trainable weight matrix.

The proposed network consists of two layers of GCN. The adjacency matrix $A \in \mathbb{R}^{K,K}$ is pre-computed with all the training data. During the training stage, the pre-trained multi-label network is used to predict the attributes $Attr \in \mathbb{R}^{N,K}$. Then the $Attr$ can be represented with the word embedding $Attr\_vector \in \mathbb{R}^{N,K,voc}$. With these notations, one layer of GCN can be defined as:

$$Y = \sigma((D+I)^{-\frac{1}{2}}(A+I)(D+I)^{-\frac{1}{2}} Attr\_vector \cdot W), \tag{10}$$

where voc is the vocabulary size and $W \in \mathbb{R}^{voc,voc}$ is the learnable weight matrix. Finally, mean pooling is used to obtain the final attribute features $V_{attr} \in \mathbb{R}^{N,voc}$ in consideration of the relationship.

C. MULTI-LEVEL ATTENTION LSTM FRAMEWORK

In this work, LSTM [37] is employed for sequence learning and description generation. The detailed computation process of attention-based LSTM is illustrated in Fig. 5. The vanilla LSTM contains three gates (supposing t as the time step): input gate i_t, forget gate f_t, and output gate o_t. Hidden state h_t

and cell state c_t are used to propagate information from time step $t-1$ to t. For each gate, three learnable weight matrices W_x, W_h, and b are used for transforming the input x_t and h_t.

In this work, the input x_t is the combination of FC feature F_{fc} and attribute feature V_{attr} computed by VGG and GCN. The computation for the conventional LSTM decoder is defined as follows.

$$\begin{aligned} i_t &= \sigma\left(W_{ix}x_t + W_{ih}h_{t-1} + b_i\right), \\ f_t &= \sigma\left(W_{fx}x_t + W_{fh}h_{t-1} + b_f\right), \\ o_t &= \sigma\left(W_{ox}x_t + W_{oh}h_{t-1} + b_o\right), \\ c_t &= i_t \odot \phi\left(W^{\otimes}_{zx}x_t + W^{\otimes}_{zh}h_{t-1} + b^{\otimes}_{z}\right) + f_t \odot c_{t-1}, \\ h_t &= o_t \odot \tanh\left(c_t\right), \\ s_t &= softmax(W_s h_t), \end{aligned} \tag{11}$$

where σ is the sigmoid activation function. ϕ is the maxout non-linearity function and $\otimes$ denotes the units (number of units k is 2 in this work).

Specifically, given the multi-level CNN feature F_{ml} and the spatial attention size P, the attention-derived image feature Fa_{ml} at time step t is defined as

$$Fa_{ml} = \sum_{i=1}^{P} \alpha_t^i F_{ml_i}, \tag{12}$$

where $\alpha_t \in \mathbb{R}^P$ is the attention vector at time t. For each time step, the attention vector is computed by equation 13:

$$\alpha_t = softmax(W_a \cdot tanh(W_{af}F_{ml_i} + W_{ah}h_{t-1} + b_a) + b_\alpha) \tag{13}$$

where W_{af}, W_{ah}, and W_a are learnable weight matrices for attention vector generation. Besides the spatial attention mechanism, we further enhance $Fa_{ml} \in \mathbb{R}^{H,W,C}$ with the channel-wise self-attention mechanism. The final attention-derived feature Fac_{ml} can be computed as

$$\begin{aligned} M_{ml} &= GP(Fa_{ml}) \in \mathbb{R}^{1,1,C}, \\ \alpha_{ml} &= sigmoid(FC_2(ReLU(FC_1(M_{ml})))), \\ Fac_{ml} &= \alpha_{ml} * Fa_{ml} + Fa_{ml}, \end{aligned} \tag{14}$$

where GP is the global pooling operation. FC_1 and FC_2 are FC layers for transforming the globally pooled features. α_{ml} is the learned channel-wise attention vector.

With the spatial and scale-wise attention-derived feature Fac_{ml}, the LSTM decoding process can adaptively re-weight the features to dynamically focus on specific regions and scales. To encode the learned attention feature Fac_{ml} into LSTM, we input the attended feature into all the gates in LSTM. This process can be defined as

$$\begin{aligned} i_t &= \sigma\left(W_{ix}x_t + W_{ih}h_{t-1} + \boldsymbol{W_{ia}Fac_{ml_t}} + b_i\right), \\ f_t &= \sigma\left(W_{fx}x_t + W_{fh}h_{t-1} + \boldsymbol{W_{fa}Fac_{ml_t}} + b_f\right), \\ o_t &= \sigma\left(W_{ox}x_t + W_{oh}h_{t-1} + \boldsymbol{W_{oa}Fac_{ml_t}} + b_o\right), \\ c_t &= i_t \odot \phi\left(W^{\otimes}_{zx}x_t + W^{\otimes}_{zh}h_{t-1} + \boldsymbol{W^{\otimes}_{za}Fac_{ml_t}} + b^{\otimes}_{z}\right) \\ &\quad + f_t \odot c_{t-1}, \\ h_t &= o_t \odot \tanh\left(c_t\right), \end{aligned}$$

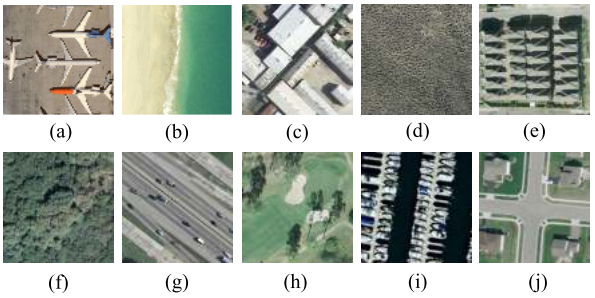

FIGURE 6. Parts of scene categories of UCM-captions dataset. (a) airplane, (b) beach, (c) building, (d) chaparral, (e) dense residential, (f) forest, (g) freeway, (h) golf course, (i) harbor, (j) intersection. The five captions of (a) are: 1) There are many airplanes at the airport; 2) Many different kinds of airplanes are stopped at the airport; 3) Four airplanes are stopped dispersedly at the airport; 4) Four airplanes scattered at the airport; 5) Four different airplanes are stopped dispersedly at the airport.

$$s_t = softmax(W_s h_t), \tag{15}$$

where W_{ia}, W_{fa}, W_{oa} and W_{za} are learnable parameters for encoding the attention-derived features. Finally, the cross entropy loss is exploited for training the LSTM as well as the GCN, which is defined in equation 16:

$$L(\theta) = -\sum_{t=1}^{T} \log\left(p_\theta\left(s_t | s_1^*, \ldots s_{t-1}^*\right)\right), \tag{16}$$

where θ denotes the parameters of the proposed model and $s_1^*, s_2^*, \ldots, s_{t-1}^*, s_t^*$ is the label sequence. It is worth noting that the parameters of CNN backbone and multi-label network are pre-trained and fixed, which are not included in θ.

III. EXPERIMENTS

In this section, extensive experiments on three datasets are conducted to demonstrate the effectiveness of the proposed method. First of all, three datasets for remote sensing image captioning task are introduced. Then, evaluation indexes and parameter settings are also given. In the end, the proposed method is compared with the prior state-of-the-art methods, and the experimental results are analyzed in detail.

A. DATASETS

In order to evaluate the proposed method and compare it with other methods, three popular remote sensing image captioning datasets are used to perform the experiments. All of the three datasets are online available [38], and the descriptions of datasets are as follows.

(1) *UCM-Captions Dataset:* This dataset is provided by Qu *et al.* [23], which extends the UC Merced Land Use dataset [39] by annotating detailed descriptions of each image manually. Parts of scene categories of this dataset are shown in Fig. 6. There are 21 different scene categories and 100 images for each category. Specifically, each image has a size of 256 × 256 pixels with the pixel resolution of 0.3048 m. It is worth mentioning that five sentences are given for the descriptions of each

FIGURE 7. Parts of scene categories of Sydney-captions dataset. (a) residential, (b) airport, (c) meadow, (d) industrial, (e) runway. The five captions of (a) are: 1) Lots of houses with red and white roofs arranged neatly; 2) A residential area with houses arranged neatly and some roads go through this area; 3) A town with many houses arranged neatly while some cars on the roads; 4) A residential area with houses arranged neatly while many plants on the roadside; 5) A residential area with houses arranged neatly and some roads go through this area.

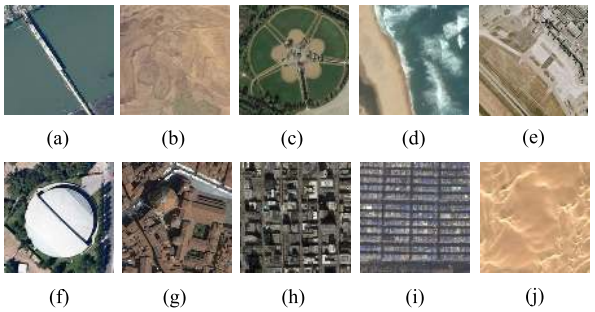

FIGURE 8. Parts of scene categories of RSICD dataset. (a) bridge, (b) bare land, (c) baseball field, (d) beach, (e) airport, (f) center, (g) church, (h) commercial, (i) denser residential, (j) desert. The five captions of (a) are: 1) The lathy bridge is across the broad river; 2) Here is a bridge under construction; 3) A white bridge starts from a forest and ends in the city; 4) A white narrow bridge spans the river with an island in it; 5) Some buildings and green plants are in two sides of a bridge respectively.

image in this dataset. Totally, there are 2,100 images with 10,500 descriptions.

(2) *Sydney-Captions Dataset:* This dataset is also proposed in [23], which is based on a remote sensing scene classification dataset named Sydney [40]. Parts of scene categories of this dataset are shown in Fig. 7. 613 images are given and they are classified into 7 scene categories. Moreover, each image is composed of 500 × 500 pixels with the pixel resolution of 0.5 m. Five sentences are provided for the descriptions of each image from multiple aspects. Totally, there are 613 images with 3065 descriptions.

(3) *RSICD Dataset:* The RSICD dataset is constructed by [13] and there are 10,921 images with low interclass difference and high intraclass diversity. Parts of scene categories of this dataset are shown in Fig. 8. These remote sensing images have a size of 224 × 224 pixels with different resolutions. Originally, 24,333 sentences are given to describe the images. To provide five sentences for each image, the captions are extended by randomly making a copy of existing sentences when there are less than five captions for an image.

As displayed in Fig. 9, the scale-variation of remote sensing images can be dramatically large. For example, the scale of the airplane: the larger ariplane can be 100 times larger (eg: 160 × 160 *vs.* 16 × 16) than the small one. Generally, the scale-variation of UCM-Captions and RSICD dataset is more obvious than Sydney-Captions dataset.

B. EXPERIMENTAL DETAILS

1) EVALUATION MEASURES

Proper metrics need to be selected to evaluate the proposed method and compare the results from different image captioning methods. Following the existing works [13], [27], [41], we adopt four evaluation indexes: BLEU (BiLingual Evaluation Understudy) [42], ROUGE_L (Recall-Oriented Understudy for Gisting Evaluation) [43], METEOR (Metric for Evaluation of Translation with Explicit ORdering) [44], and CIDEr (Consensus-based Image Description Evaluation) [45]. For the detailed formulas of these metrics, readers can refer to [46]. The reason why they are selected as metrics is that they evaluate the image captioning methods from different aspects. BLEU compares the overlap ratio of N-gram (a set of N ordered words) between the generated sentence and the corresponding reference sentence. N is set to 1, 2, 3 and 4 in this paper. ROUGE_L utilizes the longest common subsequence (LCS) based F-measure to evaluate the similarity between the candidate sentence and the reference sentence. In this case, LCS automatically contains the longest in-sequence N-gram. METEOR evaluates the generated sentence by a score based on word-to-word matches and computes a word alignment between two sentences. CIDEr takes into both precision and recall to compute how well a generated sentence matches the consensus of a collection of image captions. The BLEU, ROUGE_L and METEOR metric all range from 0 to 1, while the CIDEr ranges from 0 to 5. The higher the score is, the better the performance of the image captioning algorithm is.

2) PARAMETERS SETUP

The proposed method is implemented with Pytorch [47] based on [48]. ResNet18 pre-trained on ImageNet is used for the multi-label classification network. We fine-tune it with the generated multi-label attribute annotations. For the training of multi-label network, the input images are first resized to 224×224. The batch size is set to 64, and 25 epochs are used for fine-tuning the network. During the training of image captioning network, the parameters of multi-label network are fixed. Adam is used as the optimizer with initial learning rate 1e-4 for optimization.

The input data of the multi-label classification network is the image for generating captions, and the output of the multi-label network are the pre-defined attributes. Then the attribute features are extracted by GCN and the pre-computed attribute-graph.

For the image captioning model, VGG16 pre-trained on ImageNet is set as the backbone for feature extraction, and it is fixed during training. The number of hidden nodes of LSTM is 512 in all of the experiments. Beam search is used in this work with the beam size of 2. Adam with initial learning

FIGURE 9. Example images of the small and large objects in UCM-Captions, Sydney-Captions and RSICD dataset. As illustrated in this figure, the scale-variation in Sydney-Captions is not as large as the other two datasets.

rate 4e-4 is employed as the optimizer. The drop probability is set to 0.5 for dropout layers. Moreover, gradients are clipped if they are larger than 0.1. The image attribute relationship learning module consists of two GCN layers for all the experiments.

The input data of the captioning model are the image and the learned attribute features. The multi-level attention CNN module takes the image as input, and outputs FC features and multi-level feature maps. Then FC features, multi-level feature maps and the attribute features are input to the LSTM decoder for sentence generation.

C. COMPARISON RESULTS

In order to verify the performance of the proposed method, some state-of-the-art methods are selected for performance comparing. They are: 1) Collective Semantic Metric Learning Framework (CSMLF) [41]; 2) FC-ATT+LSTM [27]; 3) SM-ATT+LSTM [27]; 4) Attention-based (soft) [13]; 5) Attention-based (hard) [13]. Among these methods, CSMLF [41] employed metric learning for latent semantic embedding, and the collective sentence representation was designed to improve the captioning performance. SM-ATT+LSTM and FC-ATT+LSTM [27] enhanced the performance of image captioning by extending the attention mechanism with high-level image attributes. In [13], different attention mechanisms including the ''soft'' attention and ''hard'' attention were explored for remote sensing image captioning. Attention-based (soft) is a deterministic method, which assigns a weight to different positions of an image to focus on important parts. Attention-based (hard) is a stochastic method, which uses a sampling strategy to focus on different positions of the image. Then reinforcement learning is employed to optimize the model for better result. Moreover, three state-of-the-art captioning methods [49]–[51] for natural image are also selected for performance comparing.

All the comparison results are presented in Table 1, 2, 3, 4 and 5, including performance on UCM-captions dataset, Sydney-captions dataset and RSICD dataset. The Sydney-captions dataset contains only 613 images and UCM-captions dataset has 2,100 images, while there are more than 10,000 images in RSICD dataset. Since the number of images in three datasets is significantly different, the training hyper-parameters such as batch size and max training epochs are also tuned for each dataset.

1) UCM-CAPTIONS DATASET

On UCM-captions dataset, five current state-of-the-art methods are included for comparison with the proposed method. Following previous works, about 80% (1680) images are used for training, and 10% (210) images for validation. The remaining 10% (210) images are used for test. To fairly compare with prior works, VGG16 backbone pre-trained on ImageNet is employed in all of our experiments. There are 293 image attributes selected for constructing the attribute graph, which is visualized in Fig. 10. Since there are too many attributes, the attribute names are not presented in this figure. The nodes in Fig. 10 are the attributes. The relationships are represented by the edges. If two attributes (words) co-occur in one image caption for many times, the relationships between them are strong. Correspondingly, the edges between them are wider. The densely connected nodes are important nodes with higher degree, and the sparse nodes have weak relationship with other nodes.

Seven widely used image captioning metrics are employed for evaluating all the methods and the results are shown in Table 1. The best performances are marked in bold font for more clear presentation. Generally, the proposed method outperforms all existing state-of-the-art methods. Specifically, in terms of four BLEU metrics, the proposed framework achieves a significant improvement compared with other methods. In terms of METEOR and ROUGH_L, improvement is also achieved with the proposed method, which indicates the effectiveness of the designed framework. For CIDEr metric, although not the best, the proposed method is still competitive to the highest performance.

To further evaluate the proposed method comprehensively, three state-of-the-art captioning methods for natural image are also compared. The results are shown in Table 2. Although

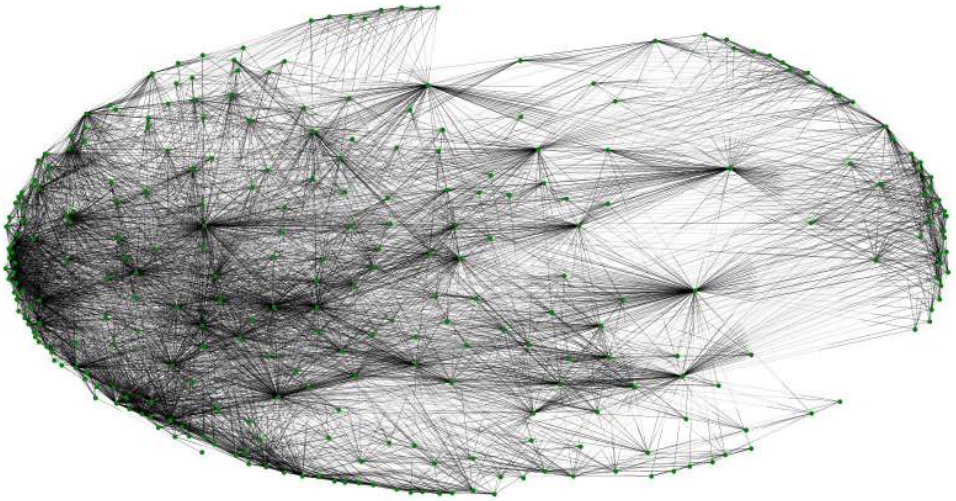

FIGURE 10. Illustration of the constructed attribute graph on UCM-captions dataset. There are 293 nodes (green spots) in the graph, and each node represents one attribute. The relationships between different attributes are displayed by the edges. As there are too many nodes in the graph, node names can not be displayed.

TABLE 1. Experimental results of different methods on UCM-captions dataset.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGH_L	CIDEr
CSMLF [41]	0.4361	0.2728	0.1855	0.1210	0.1320	0.3927	0.2227
FC-ATT+LSTM [27]	0.8135	0.7502	0.6849	0.6352	0.4173	0.7504	2.9958
SM-ATT+LSTM [27]	0.8154	0.7575	0.6936	0.6458	0.4240	0.7632	**3.1864**
Attention-based (soft) [13]	0.7454	0.6545	0.5855	0.5250	0.3886	0.7237	2.6124
Attention-based (hard) [13]	0.8157	0.7312	0.6702	0.6182	0.4263	0.7698	2.9947
Our Proposed	**0.8330**	**0.7712**	**0.7154**	**0.6623**	**0.4371**	**0.7763**	3.1684

[49]–[51] can achieve satisfactory results for natural image captioning task, directly applying them to remote sensing image still has limitation. This comparison demonstrates the effectiveness of the proposed multi-level attention and attribute GCN modules. Moreover, attention masks of some samples selected from UCM-captions dataset are visualized in Fig. 11. For each image, the first five decoding iterations of LSTM are shown.

2) SYDNEY-CAPTIONS DATASET

Sydney-captions is a relatively small dataset with 613 images in total. In this dataset, 497 images are used for training, 58 images are used for validation and another 58 images are used for test. Similar to UCM-captions dataset, five state-of-the-art methods are compared with the proposed method. The results are presented in Table 3 in detail. To prevent the model from overfitting, the batch size is set to 8 for training. Experimental results in Table 3 indicate that all the attention-based methods can perform better than CSMLF. In terms of BLEU, although SM-ATT+LSTM achieves quite good results compared with others, the proposed method can still outperform it. This shows the superiority of the proposed multi-level attention-derived features and the effectiveness of the GCN-based attribute feature learning module. As for CIDEr metric, although the result of the proposed method is not the best, it is still competitive to the best performance. In general, the comparison results on Sydney-captions dataset reveal that better performance can be obtained with the proposed method.

FIGURE 11. Visualization of some attention masks from UCM-captions dataset.

TABLE 2. Comparison results of natural image captioning methods and the proposed method on UCM-captions dataset.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGH_L	CIDEr
Show,Attend,and Tell [49]	0.787	0.725	0.670	0.621	0.371	0.702	2.76
SCST-Att2all [50]	0.810	0.747	0.694	0.647	0.417	0.750	3.05
Up-Down Attention [51]	0.823	0.759	0.707	0.661	0.424	0.763	3.02
Our Proposed	**0.833**	**0.771**	**0.715**	**0.662**	**0.437**	**0.776**	**3.17**

TABLE 3. Experimental results of different methods on Sydney-captions dataset.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGH_L	CIDEr
CSMLF [41]	0.5998	0.4583	0.3869	0.3433	0.2475	0.5018	0.7555
FC-ATT+LSTM [27]	0.8076	0.7160	0.6276	0.5544	0.4099	0.7114	2.2033
SM-ATT+LSTM [27]	0.8143	0.7351	0.6586	0.5806	0.4111	0.7195	2.3021
Attention-based (soft) [13]	0.7322	0.6674	0.6223	0.5820	0.3942	0.7127	**2.4993**
Attention-based (hard) [13]	0.7591	0.6610	0.5889	0.5258	0.3898	0.7189	2.1819
Our Proposed	**0.8233**	**0.7548**	**0.6587**	**0.6003**	**0.4202**	**0.7237**	2.3110

3) RSICD DATASET

Following the default settings, 8,734 images (about 80%) are used for training the proposed framework on RSICD-captions dataset. There are 1,094 images used for validation, which are selected randomly from RSICD dataset. The remaining 1,093 images (about 10%) are employed for test. Experimental results on RSICD dataset are displayed in Table 4. For each metric, the best results are marked in bold. Compared to UCM-captions and Sydney-captions dataset, RSICD is the largest one and contains five times images more than UCM-captions dataset. Though the maximum length of sentences in RSICD is 34, the output length on RSICD dataset is limited to 30 in this work. This is maily due to the fact that only a minority of sentences have length longer than 30. Some image captioning examples on RSICD dataset are displayed in Fig. 12. The image attributes are highlighted with green color.

From the results in Table 4, it is clear that the proposed method achieves better performance than other methods in terms of BLEU, METEOR and ROUGH_L. When compared with FC-ATT-LSTM, the proposed method yields slightly lower (0.005) CIDEr. However, in general, the proposed method achieves the best performance in terms of the other six metrics, which indicates the effectiveness of the proposed framework.

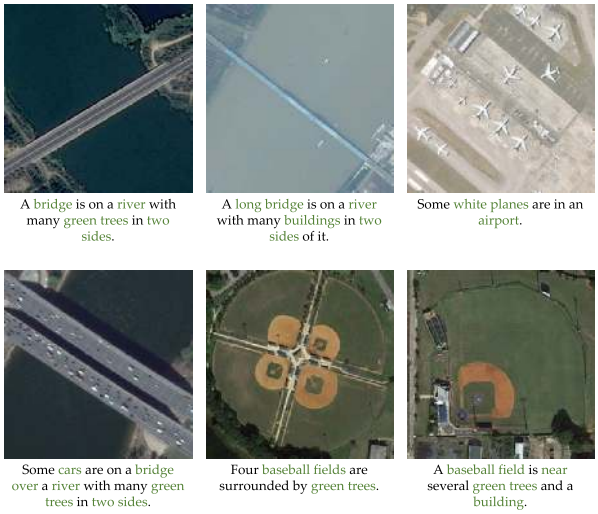

FIGURE 12. Visualization of remote sensing image captioning examples on the RSICD dataset. The image attributes are marked into green color. (Best viewed in color).

D. ABLATION STUDY

To comprehensively evaluate the effect of the designed submodules, ablation experiments are conducted on UCM-captions dataset. There are two reasons for choosing UCM dataset for ablation study. One is that it contains 2,100 images, so it is sufficient for evaluating submodules. The other one is that it does not need too much computation resource to conduct a series of experiments. Specifically, five methods with different submodules are evaluated to study the effect of the proposed submodules. *VGG16 Baseline* is the model with FC feature and LSTM for decoding. *Attribute GCN* is the model that employs GCN-transformed attribute features. *Multi-level Attention* denotes the model with the proposed multi-level attention module. *Multi-level Attention+Attribute* means that the model uses *Multi-level Attention* module and attribute feature for image captioning. Finally, *Multi-level Attention+Attribute GCN* denotes the model with *Multi-level Attention* and *Attribute GCN*.

In Fig. 9, the images of each column contain the same type of objects, which shows that the object-scale varies dramatically in UCM-Captions and RSICD dataset. The results in Table 5 indicate that the proposed method with multi-level attention module can improve the image captioning performance (BLEU-1) from 0.795 to 0.815 on UCM-Captions dataset. This clear performance improvement demonstrates that more powerful multi-scale features are learned by alleviating the scale-variation problem with the proposed multi-level attention module. Moreover, the *Multi-level Attention+Attribute* model is also evaluated to explore whether the feature enhancement and attribute features are complementary for improving the performance. The results reveal that using these two types of features together indeed improves the performance. Finally, the full model with both proposed sub-modules is also evaluated, which achieves state-of-the-art results on UCM-captions dataset.

TABLE 4. Experimental results of different methods on RSICD dataset.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGH_L	CIDEr
CSMLF [41]	0.5759	0.3859	0.2832	0.2217	0.2128	0.4455	0.5297
FC-ATT+LSTM [27]	0.7459	0.6250	0.5338	0.4574	0.3395	0.6333	**2.3664**
SM-ATT+LSTM [27]	0.7571	0.6336	0.5385	0.4612	0.3513	0.6458	2.3563
Attention-based (soft) [13]	0.6753	0.5308	0.4333	0.3617	0.3255	0.6109	1.9643
Attention-based (hard) [13]	0.6669	0.5182	0.4164	0.3407	0.3201	0.6084	1.7925
Our Proposed	**0.7597**	**0.6421**	**0.5517**	**0.4623**	**0.3543**	**0.6563**	2.3614

TABLE 5. Ablation study on UCM-captions dataset.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGH_L	CIDEr
VGG16 Baseline	0.795	0.740	0.691	0.643	0.402	0.736	3.05
Attribute GCN	0.810	0.748	0.697	0.650	0.419	0.751	3.06
Multi-level Attention	0.815	0.756	0.710	0.661	0.408	0.738	3.09
Multi-level Attention + Attribute	0.819	0.761	0.712	0.668	0.418	0.757	**3.17**
Multi-level Attention + Attribute GCN	**0.833**	**0.772**	**0.715**	**0.676**	**0.437**	**0.776**	**3.17**

In addition, since the attributes are obtained from the multi-label classification network, the predicted attributes may contain noise. This will affect the final captioning performance. While with the proposed attribute GCN, the predicted attributes are more robust and confident by exploiting the pre-computed graph information. As shown in Table 5, the proposed Attribute GCN can improve the performance of Attribute baseline obviously. From the comparison results, it is clear to see that Attribute GCN sub-module can boost the image captioning performance by introducing more confident attribute features.

IV. CONCLUSION

In this work, a remote sensing image captioning framework based on multi-level attention and multi-label attribute graph convolution is proposed to improve the performance from two aspects. Different from previous method, the multi-level attention module is designed to learn scale-adaptive and position-adaptive image representations simultaneously. By focusing on specific spatial positions and features of specific scales, the proposed framework can learn more discriminative features for image captioning. Besides, the attribute feature learning module is proposed based on multi-label classification and GCN. By employing the semantic attribute graph, the relationships between attributes are leveraged to learn more robust attribute features for image captioning. Experiments on three widely used public datasets are conducted for performance evaluation. The superior results of the proposed method indicate three main conclusions: 1) learning more representative image features is critical for improving image captioning performance; 2) How to learn effective multi-scale image features is important for remote sensing image captioning task; 3) Making better use of the relationships between image attributes is also helpful for image captioning.

REFERENCES

[1] Q. Wang, X. He, and X. Li, "Locality and structure regularized low rank representation for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 911–923, Feb. 2019.

[2] Q. Wang, F. Zhang, and X. Li, "Optimal clustering framework for hyperspectral band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 10, pp. 5910–5922, Oct. 2018.

[3] Y. Hu, X. Li, N. Zhou, L. Yang, L. Peng, and S. Xiao, "A sample update-based convolutional neural network framework for object detection in large-area remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 6, pp. 947–951, Jun. 2019.

[4] J. Zhang, T. Li, X. Lu, and Z. Cheng, "Semantic classification of high–resolution remote–sensing images based on mid-level features," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 6, pp. 2343–2353, Jun. 2016.

[5] Q. Wang, Z. Yuan, Q. Du, and X. Li, "GETNET: A general end-to-end 2-D CNN framework for hyperspectral image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 3–13, Jan. 2019.

[6] X. Li, Z. Yuan, and Q. Wang, "Unsupervised deep noise modeling for hyperspectral image change detection," *Remote Sens.*, vol. 11, no. 3, p. 258, Jan. 2019.

[7] 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.

[8] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, San Diego, CA, USA, Jun. 2005, pp. 886–893.

[9] D. G. Lowe, "Distinctive image features from scale–invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst., 26th Annu. Conf. Neural Inf. Process. Syst., Proc. Meeting Held*, Lake Tahoe, NV, USA, Dec. 2012, pp. 1106–1114.

[11] Z. Xiong, Y. Yuan, and Q. Wang, "RGB–D scene recognition via spatial–related multi–modal feature learning," *IEEE Access*, vol. 7, pp. 106739–106747, 2019.

[12] Y. Yuan, Z. Xiong, and Q. Wang, "VSSA–NET: Vertical spatial sequence attention network for traffic sign detection," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3423–3434, Jul. 2019.

[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] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi, "Composing simple image descriptions using Web-scale n-grams," in *Proc. 15th Conf. Comput. Natural Lang. Learn. (CoNLL)*, Portland, OR, USA, Jun. 2011, pp. 220–228.

[15] S. Venugopalan, L. A. Hendricks, M. Rohrbach, R. Mooney, T. Darrell, and K. Saenko, "Captioning images with diverse objects," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 1170–1178.

[16] J. Lu, J. Yang, D. Batra, and D. Parikh, "Neural baby talk," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 7219–7228.

[17] V. Ordonez, G. Kulkarni, and T. L. Berg, "Im2text: Describing images using 1 million captioned photographs," in *Proc. Adv. Neural Inf. Process. Syst., 25th Annu. Conf. Neural Inf. Process. Syst., Meeting Held*, Granada, Spain, Dec. 2011, pp. 1143–1151.

[18] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 3156–3164.
[19] N. Xu, A.-A. Liu, Y. Wong, Y. Zhang, W. Nie, Y. Su, and M. Kankanhalli, "Dual–stream recurrent neural network for video captioning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 8, pp. 2482–2493, Aug. 2019.
[20] M. Zhang, Y. Yang, H. Zhang, Y. Ji, H. T. Shen, and T.-S. Chua, "More is better: Precise and detailed image captioning using online positive recall and missing concepts mining," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 32–44, Jan. 2019.
[21] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3128–3137.
[22] J. Aneja, A. Deshpande, and A. G. Schwing, "Convolutional image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 5561–5570.
[23] 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. (CITS)*, Kunming, China, Jul. 2016, pp. 1–5.
[24] 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. (IGARSS)*, Fort Worth, TX, USA, Jul. 2017, pp. 4798–4801.
[25] X. Zhang, Q. Wang, S. Chen, and X. Li, "Multi–scale cropping mechanism for remote sensing image captioning," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul./Aug. 2019, pp. 10039–10042.
[26] 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.
[27] 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.
[28] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "SCA–CNN: Spatial and channel–wise attention in convolutional networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 6298–6306.
[29] S. Woo, J. Park, J. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. 15th Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 2018, pp. 3–19.
[30] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 3146–3154.
[31] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 684–699.
[32] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang, "Scene graph generation from objects, phrases and region captions," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1261–1270.
[33] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[34] Q. Wu, C. Shen, P. Wang, A. Dick, and A. Van Den Hengel, "Image captioning and visual question answering based on attributes and external knowledge," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1367–1381, Jun. 2018.
[35] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*. [Online]. Available: https://arxiv.org/abs/1609.02907
[36] Z.-M. Chen, X.-S. Wei, P. Wang, and Y. Guo, "Multi-label image recognition with graph convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 5177–5186.
[37] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," *Neural Comput.*, vol. 12, no. 10, pp. 2451–2471, Oct. 2000.
[38] *Datasets for Remote Sensing Images*. [Online]. Available: https://github.com/201528014227051/RSICD_optimal
[39] Y. Yang and S. D. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th ACM SIGSPATIAL Int. Symp. Adv. Geographic Inf. Syst. (ACM-GIS)*, San Jose, CA, USA, Nov. 2010, pp. 270–279.
[40] 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.
[41] 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.
[42] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, Philadelphia, PA, USA, Jul. 2002, pp. 311–318.
[43] C. Flick, "Rouge: A package for automatic evaluation of summaries," in *Proc. Workshop Text Summarization Branches Out*, 2004.
[44] A. Lavie and A. Agarwal, "METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments," in *Proc. 2nd Workshop Stat. Mach. Transl.*, Prague, Czech Republic, Jun. 2007, pp. 228–231.
[45] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 4566–4575.
[46] X. Chen, H. Fang, T. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft COCO captions: Data collection and evaluation server," 2015, *arXiv:1504.00325*. [Online]. Available: https://arxiv.org/abs/1504.00325
[47] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *Proc. NIPS Autodiff Workshop*, 2017.
[48] R. Luo. (2017). *An Image Captioning Codebase in Pytorch*. [Online]. Available: https://github.com/ruotianluo/ImageCaptioning.pytorch
[49] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. 32nd Int. Conf. Mach. Learn. (ICML)*, Lille, France, Jul. 2015, pp. 2048–2057.
[50] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self–critical sequence training for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 1179–1195.
[51] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 6077–6086.

ZHENGHANG YUAN received the B.E. degree in information security from Northwestern Polytechnical University, Xi'an, China, in 2017, where she is currently pursuing the M.S. degree in computer science with the Center for Optical Imagery Analysis and Learning, School of Computer Science. Her research interests include computer vision and machine learning.

XUELONG LI (Fellow, IEEE) is currently a Full Professor with the Center for Optical Imagery Analysis and Learning (OPTIMAL), School of Computer Science, Northwestern Polytechnical University, Xi'an, China.

QI WANG (Senior Member, IEEE) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Professor with the Center for Optical Imagery Analysis and Learning, School of Computer Science, Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and pattern recognition.