High Resolution Remote Sensing Image Captioning based on Structured Attention

Rui Zhao, Zhenwei Shi*, *Member IEEE,* and Zhengxia Zou

***Abstract*—Automatically generating language descriptions of remote sensing images has become an emerging research hot spot in the remote sensing field. Attention-based captioning, as a representative group of recent deep learning based captioning methods, share the advantage of generating the words while highlighting corresponding object locations in the image. Standard attention-based methods generate captions based on coarse-grained and unstructured attention units, which fails to exploit structured spatial relations of semantic contents in remote sensing images. Although the structure characteristic makes remote sensing images widely divergent to natural images and poses a greater challenge for the remote sensing image captioning task, the key of most remote sensing captioning methods is usually borrowed from the computer vision community without considering the domain knowledge behind. To overcome this problem, a fine-grained, structured attention-based method is proposed to utilize the structural characteristics of semantic contents in high resolution remote sensing images. Our method learns better descriptions and can generate pixel-wise segmentation masks of semantic contents. The segmentation can be jointly trained with the captioning in a unified framework without requiring any pixel-wise annotations. Evaluations are conducted on three remote sensing image captioning benchmark datasets with detailed ablation studies and parameter analysis. Compared with the state-of-the-art methods, our method achieves higher captioning accuracy and can generate high-resolution and meaningful segmentation masks of semantic contents at the same time.**

***Index Terms*—Remote sensing image, structured attention, image captioning, image segmentation.**

I. Introduction

IMAGE captioning is an important computer vision task that emerged in recent years that aims to automatically generate language descriptions of an input image [1], [2]. In the remote sensing field, image captioning also has attracted increasing attention recently due to its broad application prospects both in civil and military usages, such as remote sensing image retrieval and military intelligence generation [3]. Different from other tasks in remote sensing field such as object detection [4]–[8] classification [9]–[11] and segmentation [12]–[15], remote sensing image captioning focuses more on generating comprehensive sentence descriptions rather than predicting individual category tags or words. To generate accurate and detailed descriptions, the captioning model needs to not only determine the semantic contents that exist in the image but also have a good understanding of the relationship between them and what activities they are engaged in [2].

In most recent deep learning-based image captioning methods, the models are built based on the "Encoder-Decoder" network architecture [1], [2], [16]–[18]. In the encoding stage, deep convolutional neural networks (CNNs) are used to extract high-level internal representations of the input image. In the decoding stage, a recurrent neural network (RNN) is typically trained to decode the representations to sentence descriptions. More recently, the visual attention mechanism, a technique derived from automatic machine translation [19], [20], has greatly promoted the research progress in image captioning [21]–[26]. The attention mechanism was originally introduced to improve the performance of an RNN model by taking into account the input from several time steps to make one prediction [19]. In image captioning, visual attention can help the model better exploit spatial correlations of semantic contents in the image and highlight those contents while generating corresponding words [21].

For the remote sensing image captioning task, Qu *et al.* [27] first proposed a deep multimodal neural network model for high resolution remote sensing image caption generation. Shi *et al.* [3] proposed a Fully Convolutional Networks (FCN) captioning model which mainly focuses on the multi-level semantics and semantic ambiguity problems. Lu *et al.* [28] explored several encoder-decoder based methods and their attention based variants, and published a remote sensing image caption dataset named RSICD. Wang *et al.* [29] introduced the multi-sentence captioning task and proposed a framework using semantic embedding to measure the image representation and the sentence representation to improve captioning results. Lu *et al.* [30] proposed a sound active attention framework for more specific caption generation according to the interest of the observer. Wang *et al.* [31] proposed the retrieval topic recurrent memory network that first retrieves the topic words of input remote sensing images from the topic repository, and then generate the captions by using a recurrent memory network [32] based on both the topic words and the image features. Ma *et al.* [33] proposed two multi-scale captioning methods to grab multi-scale information for generating better captions. Cui *et al.* in [34] proposed an attention based remote sensing image semantic segmentation and spatial relationship recognition method. However, the captioning module in their

The work was supported by the National Key R&D Program of China under the Grant 2019YFC1510905, the National Natural Science Foundation of China under the Grant 61671037 and the Beijing Natural Science Foundation under the Grant 4192034. *(Corresponding author: Zhenwei Shi.)*

Rui Zhao (e-mail: ruizhaoipc@buaa.edu.cn) and Zhenwei Shi (Corresponding author, e-mail: shizhenwei@buaa.edu.cn) are with Image Processing Center, School of Astronautics, Beihang University, Beijing 100191, China, and with the Beijing Key Laboratory of Digital Media, Beihang University, Beijing 100191, China, and also with the State Key Laboratory of Virtual Reality Technology and Systems, School of Astronautics, Beihang University, Beijing 100191, China.

Zhengxia Zou (e-mail: zzhengxi@umich.edu) is with the Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, U.S.A.

method just follows the classical model [21] without modification based on the characteristic of remote sensing images. The captioning module is independent of other modules, and the accuracy of caption generation is not improved by other modules. Sumbul *et al.* [35] proposed a summarization driven image captioning method, which integrated the summarized ground truth captions to generate more detailed captions for remote sensing images. Li *et al.* [36] proposed a truncation cross entropy to deal with the overfitting problem. Wang *et al.* [37] proposed a word-sentence framework to extract the valuable words firstly and then organize them into a well formed caption. Huang *et al.* [38] proposed a denoising based multiscale feature fusion mechanism to enhance the image feature extraction. Li *et al.* [39] proposed a multi-level attention model to enhance the effect of attention through a hierarchical structure. Wu *et al.* [40] proposed a scene attention mechanism which tried to catch the scene information to improve the captions.

These remote sensing image captioning methods are all based on encoder-decoder architecture, which can be roughly divided into two groups, 1) methods without visual attention mechanisms that constructed between caption and image space [3], [27]–[29], [31], [35]–[38] and 2) methods with visual attention mechanisms [28], [30], [33], [34], [39], [40]. The visual attention mechanisms in these methods are designed based on coarse-grained, unstructured attention units, which fails to exploit structured spatial relations of semantic contents in remote sensing images. For example, in the popular natural image captioning method "Show, attend and tell" [21], the authors uniformly divide the image feature map into 14x14 spatial units. However, in remote sensing images, the semantic contents are usually highly structured where narrow and irregularly shaped objects like roads, rivers, and structures usually occupy a large portion. The uniform division of the feature map inevitably leads to an under-exploit of the spatial structure of remote sensing semantic contents. Besides, due to the coarse division of attention units, these methods also fail to produce fine-grained attention maps of irregularly shaped semantic contents.

In this paper, we show that the structured and pixel-level regional information can be used to enhance the efficacy of attention based remote sensing image captioning. In computer vision, in-depth research has been made on the pixel-level description of irregularly shaped semantic contents, where a representative group of the method is semantic segmentation [41]–[45]. We thus introduce a structured attention module in our captioning model and propose a joint captioning and segmentation framework for high resolution remote sensing images by taking advantage of the structured attention mechanism. The structured attention module aims to focus on the semantic contents in the remote sensing images with structured geometry and appearance. Structured attention is performed on each structured unit obtained in the segmentation proposal generation, that is, the pixels within each structured unit receive the same attention weight, while different structured units get different attention weights. In this way, the proposed method can exploit the spatial structure of semantic contents and produce fine-grained attention maps to guide decoder

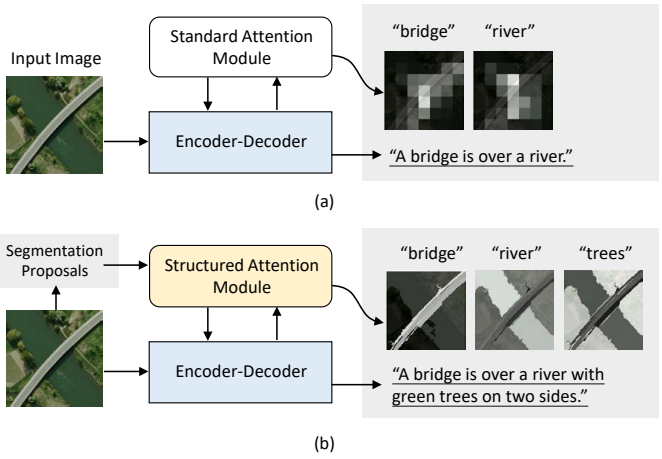

Fig. 1. A brief comparison between (a) the standard attention based captioning method and (b) the proposed structured attention based captioning method. In (a), the captions are learned from a set of coarse and unstructured image regions. As a comparison, in (b), our method exploits the fine-grained structure of the image and thus generates more accurate descriptions.

to more proper caption generation. We show our method generates better sentence descriptions and pixel-level object masks under a unified framework. It is worth mentioning that in our method, the segmentation is trained solely based on the image-level ground truth sentences and does not require any pixel-wise annotations.

Fig. 1 shows the key differences between the proposed method and previous attention based methods. In our method, we first divide the input image into a set of class-agnostic segmentation proposals and then encode the structure of each of the segmentation proposal into our attention module. Structured attention can guide the model to accurately focus on highly structured semantic contents during the training, thereby improving the performance of the image captioning task. Although the class label of each proposal are not available during the training and are considered as latent variables, we show the correspondence between the predicted words and the attention weights for each proposal can be adaptively learned under a weakly supervised training process. Our method, therefore, produces much more accurate attention maps for the semantic contents than those unstructured attention methods.

Extensive evaluations of our method are made on three benchmark datasets. Our method achieves higher captioning accuracy than other state-of-the-art captioning methods and generates object masks with high quality. Detailed ablation studies and parameter analysis are also conducted which suggest the effectiveness of our method.

The contributions of this paper are summarized as follows:

- We propose a novel image captioning method for high resolution remote sensing images based on the structured attention mechanism. The proposed method deals with image captioning and pixel-level segmentation under a unified framework.
- We investigate the possibility of using structured attention for weakly supervised image segmentation. To our best

knowledge, such a topic has rarely been studied before.
- We achieve higher captioning accuracy over other state of the art methods on three remote sensing image captioning benchmark datasets.

The rest of this paper is organized as follows. In Section II, we will introduce the structured attention and the details of our method. Experimental results and analysis are given in section III. The conclusions are drawn in Section IV.

II. Methodology

In this section, we give a detailed introduction to the proposed structured attention method and how we build our image captioning model on top of it.

A. Overview of the Method

The captioning model proposed in this paper mainly consists of three parts: an encoder, a decoder, and a structured attention module. Fig. 2 shows the processing flow of the proposed model. From the natural image captioning literature, we borrow the encoder-decoder framework which has been shown to work well in the image captioning task. We use a deep Convolutional Neural Network as our encoder to extract high-level feature representations from the input image. It is worth mentioning that our method is indeed independent of the choice of the backbone model. Any deep convolutional neural network can be used as an encoder. We use a Long Short-Term Memory Network [46], [47] as our decoder to decode the image features to the sentence description. Before feeding features to the structured attention module, we use a pre-defined method "Selective Search" [48] to segment the input image to a set of class-agnostic segmentation proposals based on the color and texture features. The selective search module in our framework requires the remote sensing image be high resolution to extract available segmentation proposals. Fig. 3 shows some samples generated by the selective search. The proposals are then synchronously encoded to our attention module with the image features by using a newly proposed pooling method, named the "structured pooling" method. In this way, the original image features are re-calibrated and we thus can obtain a set of structured region descriptions for captioning and mask generation. The attention weights generated by the model for each region on predicting a certain word are considered as the probability that the region belongs to the word category (e.g., building, tree, bridge, etc.).

B. Encoder and Decoder

We use the 50-layers deep residual network (ResNet-50) [49] as our encoder. We remove the full connection layer (prediction layer) of the ResNet-50 and use the feature maps produced by the last convolution block "Conv_5" as our internal feature representations. Our decoder is a one-layer LSTM with 512 hidden units. The LSTMs are a special kind of RNN, capable of learning long-term dependencies. By selectively forgetting and updating information in the training process, the LSTM can achieve better performance in the complex sequential prediction problems than the vanilla recurrent neural networks. Our decoder is trained to generate the word score vector $\mathbf{y}_t \in \mathbb{R}^K$ at each time step t based on a context vector $\mathbf{z}_t$, a previous hidden state vector $\mathbf{h}_{t-1}$ and a previously generated word score vector $\mathbf{y}_{t-1}$, where K is the size of the vocabulary. The prediction of $\mathbf{y}_t$ can be written as follows:

$$\mathbf{y}_t = \mathbf{L}_o(\mathbf{L}_h\mathbf{h}_t + \mathbf{L}_y\mathbf{y}_{t-1} + \mathbf{L}_z\mathbf{z}_{t-1}), \tag{1}$$

where $\mathbf{L}_h$, $\mathbf{L}_y$ and $\mathbf{L}_z$ are a group of trainable parameters that transforms the input vectors to calibrate their dimensions. The $\mathbf{L}_o$ is a group of trainable parameters that transforms from the summarized vectors to the output score vectors. To compute $\mathbf{h}_t$, $\mathbf{z}_t$, and $\mathbf{y}_t$ at each time step, the LSTM gates and internal states are defined as follows:

$$\begin{aligned}
\mathbf{i}_t &= \sigma(\mathbf{W}_i\mathbf{x}_t + \mathbf{b}_i),\\
\mathbf{f}_t &= \sigma(\mathbf{W}_f\mathbf{x}_t + \mathbf{b}_f),\\
\mathbf{o}_t &= \sigma(\mathbf{W}_o\mathbf{x}_t + \mathbf{b}_o),\\
\tilde{\mathbf{c}}_t &= \tanh(\mathbf{W}_c\mathbf{x}_t + \mathbf{b}_c),\\
\mathbf{c}_t &= \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{c}}_t,\\
\mathbf{h}_t &= \mathbf{o}_t \odot \tanh(\mathbf{c}_t),
\end{aligned} \tag{2}$$

where $\mathbf{x}_t$ is the concatenation of previous hidden state $\mathbf{h}_{t-1}$, the previously generated word vector $\mathbf{y}_{t-1}$ and the context vector $\hat{\mathbf{z}}_t$: $\mathbf{x}_t = [\mathbf{h}_{t-1}; \mathbf{P}\mathbf{y}_{t-1}; \hat{\mathbf{z}}_t]$. $\mathbf{P} \in \mathbb{R}^{m\times K}$ is an embedding matrix, where m denotes the embedding dimension. The $\mathbf{i}_t$, $\mathbf{f}_t$, $\mathbf{o}_t$, $\mathbf{c}_t$, $\mathbf{h}_t$ are the outputs of the input gate, forget gate, output gate, memory, and hidden state of the LSTM, respectively. $\mathbf{W}_i$, $\mathbf{W}_f$, $\mathbf{W}_o$, $\mathbf{W}_c$ are trainable weight matrices and $\mathbf{b}_i$, $\mathbf{b}_f$, $\mathbf{b}_o$, $\mathbf{b}_c$ are their trainable biases. The $\sigma(\cdot)$, $\tanh(\cdot)$ and $\odot$ represent the logistic sigmoid activation, hyperbolic tangent function and element-wise multiplication operation, respectively.

Finally, to produce word probabilities $\mathbf{p}_t$, we use a "softmax" layer to normalize the generated score vectors to probabilities:

$$\begin{aligned}
\mathbf{p}_t &= \text{softmax}(\mathbf{y}_t)\\
&= \exp(\mathbf{y}_t) / \sum_{i=1}^{K} \exp(y_t^{(i)}).
\end{aligned} \tag{3}$$

C. Structured Attention

1) Structured Pooling: Most CNNs produce unstructured image feature representations. Here we propose a new pooling operation called "structured pooling" to generate structured feature representations given a set of region proposals of any shapes. The structured pooling can be considered as a modification of the standard "ROI (Region of Interest) pooling". The difference between the two operations is that the ROI pooling is only designed to pool the features from rectangular regions while the structured pooling applies to the regions of any shape. Suppose I is the input image and $\mathbf{F} \in \mathbb{R}^{h\times w\times c}$ represents the image features produced by the encoder, where h, w and c are the height, width, and the number of channels respectively. The region proposals R_i, $i = 1, ..., N$ produced by the selective search are considered as the base units when performing structured pooling.

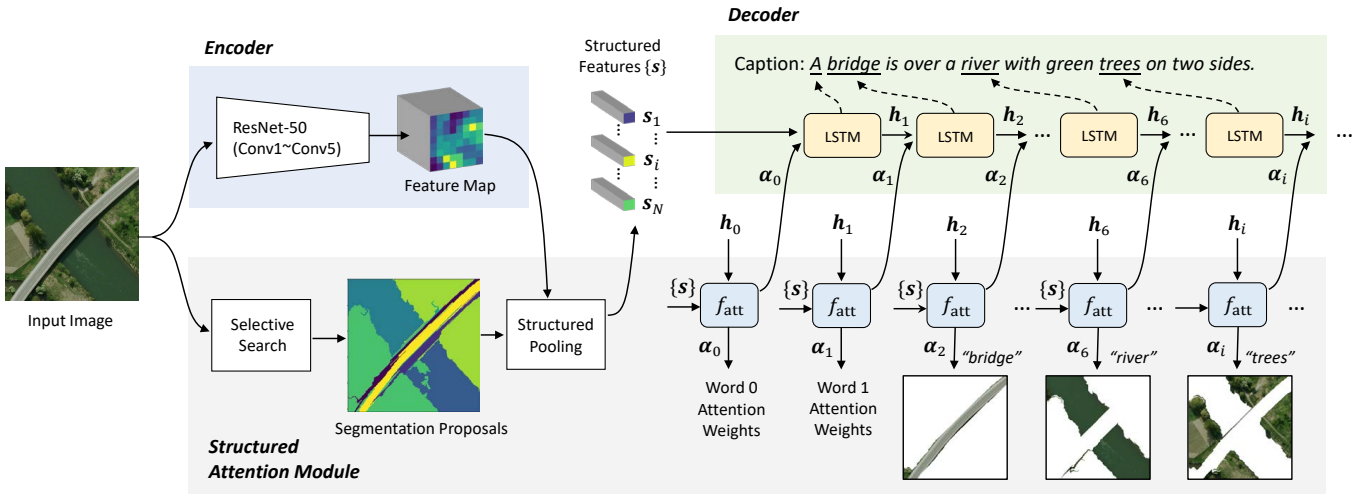

Fig. 2. An overview of the proposed captioning method. Our method consists of three parts: an encoder, which maps the input image to feature maps; a decoder, which generates the sentences based on the image feature; and a structured attention module, which interacts with the decoder during the captioning and at the same time generates pixel-level object masks.

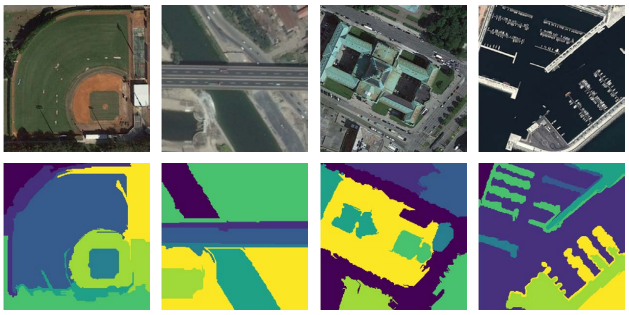

Fig. 3. Input images (first row) and the class-agnostic segmentation proposals generated by the selective search method (second row).

For the unit i, the structured feature representation $\mathbf{s}_i$ produced by the structured pooling can be represented as follows:

$$\mathbf{s}_i = \frac{1}{hw} \sum_{(x,y)\in R'_i} \mathbf{F}(x,y) \odot R'_i, \tag{4}$$

where R'_i is the projected region proposal which is resized from the size of the input image to the size of the feature map. The summation is performed among the pixels (x, y) within the region R'_i along the spatial dimensions. It should be noticed that when we average the feature values within a certain region R'_i, we divide the number of all spatial pixels in the feature map (hw) instead of the number of valid pixels in that region. The reason behind this is that we want to enhance the features according to their structured unit size, that is, the features of small structured units will be less weighted to reduce the noise effect from these regions.

Fig. 4 gives a simple illustration of the proposed structured pooling operation. To help understand, in this figure, we show an alternative but equivalent way of performing structured pooling, where we first pixel-wisely multiply the features on a set of resized region masks and then perform the global average pooling to produce the pooling output. To reduce the misalignment effect, we use the bilinear interpolation when we reduce the size of the binary region masks.

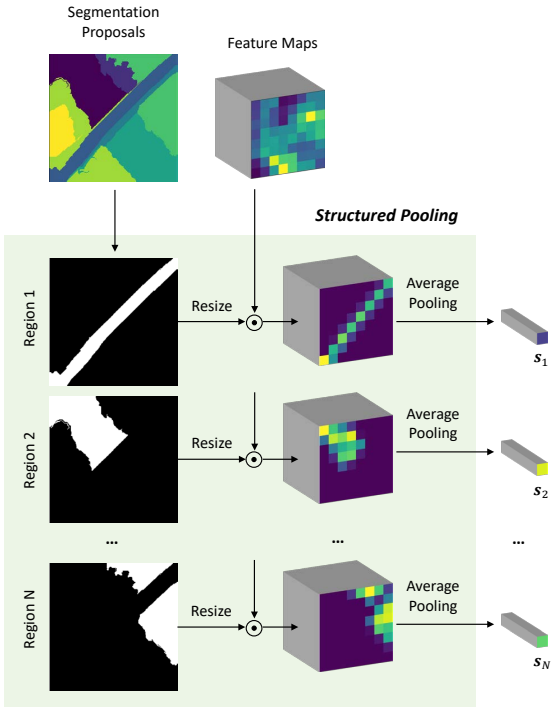

Fig. 4. An illustration of the proposed structured pooling method. The notation $\odot$ represents the element-wise product operation.

It is worth noting that although proposal-based approaches are widely used in computer vision tasks, the proposed structured attention method is designed specifically for remote sensing images. Since remote sensing images are captured from high above, many semantic contents in remote sensing images, such as rivers and bridges, show highly structured characteristics. For example, bridges are always long and straight, rivers are always winding and slender, while buildings

are mostly in the form of regular polygon aggregation. As a comparison, semantic contents in natural images often lack regular and structured shape outlines under different views and occlusions. Since the proposed structured attention mechanism relies on effective structure extraction of the semantic contents in images, it is more suitable for remote sensing images rather than natural images.

2) Context Vector Generation: The context vector $\hat{\mathbf{z}}_t$ in our method is a dynamic representation of the corresponding structured unit of the image at the time t. Given the feature representations $\mathbf{s}_i$ and the previous hidden state $\mathbf{h}_{t-1}$, we calculate an attention weight value $\alpha_t^{(i)}$, which represents the degree of correlation between the structured units and the generated word vector $\mathbf{y}_t$:

$$\tilde{\alpha}_t^{(i)} = f_{att}(\mathbf{s}_i, \mathbf{h}_{t-1}), \tag{5}$$

where $f_{att}(\cdot)$ represents a multi-layer perceptron (MLP) which is trained to generate the attention weights.

To build the the network $f_{att}(\cdot)$, we first adjust the dimensions of $\mathbf{s}_i$ and $\mathbf{h}_{t-1}$ to a same number by passing each of them through a fully connected layer. Then the transformed vectors are added together to fuse the information from both the structured unit and the context and the fusion vector is further fed to another fully connected layer to produce the attention weight $\alpha_t^{(i)}$:

$$f_{att}(\mathbf{s}_i, \mathbf{h}_{t-1}) = f_3(\text{ReLU}(f_1(\mathbf{s}_i) + f_2(\mathbf{h}_{t-1}))), \tag{6}$$

where f_1, f_2 and f_3 represent the three fully connected layers and the ReLU$(\cdot)$ represents the rectified linear unit activation function. Then, the attention weights of the N unique regions at the time step t are normalized with a softmax layer to produce the final attention vector $\boldsymbol{\alpha}_t$:

$$\boldsymbol{\alpha}_t = \text{softmax}([\tilde{\alpha}_t^{(1)}, ..., \tilde{\alpha}_t^{(N)}]). \tag{7}$$

Once we get the attention weighted vector, the context vector $\mathbf{z}_t$ can be finally computed as a linear combination of the structured feature represents $\mathbf{s}_i$ and their attention weights $\alpha_t^{(i)}$:

$$\mathbf{z}_t = \sum_{i=1}^{N} \alpha_t^{(i)} \mathbf{s}_i. \tag{8}$$

Note that at the time step t, the attention weight of each structured unit is computed based on the same context information $\mathbf{h}_{t-1}$, which ensures that the initial competitiveness of each structured unit is fair and reduces the possibility of introducing deviation. This is called the "context information broadcast" mechanism, which was introduced by Vinyals *et al*. [1] for the first time.

3) Object Masks Generation: We generate the object masks based on the attention weights of each structured region. In our attention module, the attention weights $\alpha_t^{(i)}$, $i = 1, ...N$ represent the semantic correlation between the t-th word of the sentence and each of the N regions. The larger the $\alpha_t^{(i)}$, the more relevant it is to the ith structured unit R_i. We use the attention weights as the category probability of the segmentation output. The nouns of the semantic contents of interest can be easily picked out from the generated captions by comparing each word to a pre-defined noun set. The pixel-wise object masks can be finally generated by binarizing the segmentation weights of each region.

D. Loss Functions

Since image captioning is a sequential prediction problem of each word in the sentence, we follow the previous works [1], [21] and formulate the prediction of each word as a regularized classification process. The loss can thus be written as a running sum of the regularized cross-entropy loss of each word in the sentence:

$$\mathcal{L}(\mathbf{x}) = -\sum_{t=1}^{C} \log(\sum_{j=1}^{K} \hat{y}_t^{(j)} p_t^{(j)}) + \beta r_d(\boldsymbol{\alpha}_t) + \gamma r_v(\boldsymbol{\alpha}_t), \tag{9}$$

where $\mathbf{p}_t = [p_t^{(1)}, ..., p_t^{(K)}]$ is the predicted word probability vector. $\hat{\boldsymbol{y}}_t = [\hat{y}_t^{(1)}, ..., \hat{y}_t^{(K)}]$ is the one-hot label of the tth word in the ground truth caption, and $\hat{y}_t^{(j)} \in \{0, 1\}$. C is the number of words in the generated sentence. $r_d(\boldsymbol{\alpha}_t)$ and $r_v(\boldsymbol{\alpha}_t)$ are the doubly stochastic regularization [21] and the proposed attention variance regularization, which we will introduce later. β and γ are the weight coefficients for balancing different loss terms.

1) Doubly Stochastic Regularization: In Section II-C2, we show that $\sum_i \alpha_{t,i} = 1$ since the attention weights are finally normalized by a softmax function. Here we further regularize the attention weights from the time dimension and introduce the doubly stochastic regularization is as follows:

$$r_d(\boldsymbol{\alpha}_t) = \sum_{i=1}^{N} (1 - \sum_{t=1}^{C} \alpha_{t,i})^2. \tag{10}$$

This regularization term encourages the model to pay equal attention to each part of the image during the generation of captions. In other words, it can prevent some regions from always receiving strong attention while other regions from being ignore all the time.

2) Attention Variance Regularization: When we fixed the time step t and look at the attention weights of each structure regions, we usually hope to see these regions receive a highly diverse attentions. This means we don't want each region receive equal attentions. We thus design the attention variance regularization term to enforce the regions have a high variance in their attention weights:

$$\begin{aligned} r_v(\boldsymbol{\alpha}_t) &= -\sum_{t=1}^{C} \|\boldsymbol{\alpha}_t - \mathbb{E}\{\boldsymbol{\alpha}_t\}\|_2^2 \\ &= -\sum_{t=1}^{C} \|\boldsymbol{\alpha}_t - \frac{1}{N}\|_2^2, \end{aligned} \tag{11}$$

where we have $\mathbb{E}\{\boldsymbol{\alpha}_t\} = 1/N$ since the $\boldsymbol{\alpha}_t$ has been normalized by the softmax function. We take the negative value of the l_2 norm since we want to maximize the attention variance. It is easy to proof that when the $\boldsymbol{\alpha}_t$ is an one-hot vector, i.e., only one region contributes to the prediction of the current word, the value of $r_v(\boldsymbol{\alpha}_t)$ will reach its minimum. On the contrary, when every region receives equal attention, i.e., $\alpha_t^{(i)} = 1/N$, $i = 1, ..., N$, in which we do not hope to see, the $r_v(\boldsymbol{\alpha}_t)$ will be maximized.

E. Implementation details

1) Training Details: In the training phase, we use Adam optimizer [50] to train our model. We set regularization coefficients $\beta = \gamma = 1$. We set the learning rate of our encoder to $1e^{-4}$ and set the learning rate of our decoder learning rate to $4e^{-4}$. The batch size is set to 64 and the model is trained for 100 epochs. In our encoder, the ResNet-50 is pre-trained on the ImageNet dataset [51]. To speed up training, we only fine-tune the convolutional blocks 2-4 of the ResNet-50 during training. In our decoder, the memory and hidden state gate of the LSTM at the time step 0 are initialized separately based on the averaged image features. We use a fully connected layer to transform the features to produce their 0-time inputs.

2) Segmentation Proposal Generation: When we use the selective search to generate segmentation proposals, three key parameters need to be specifically tuned, including a smooth parameter σ of the Gaussian filter, a min_size parameter which controls the minimum bounding box size of the proposals, and a scale parameter s which controls the initial segmentation scales. We set $\sigma = 0.8$, min_size $= 100$, and $s = 100$. Besides, to prevent over-segmentation, we applied the guided image filter [52] to pre-process the image before the selective search. The guided image filter can effectively smooth the input image while keeping its edge and structures. The smoothed images are only used for generating segmentation proposals. When the encoder computes the image features, we still use the original images.

3) Beam Search: At the inference stage, instead of using a greedy search that chooses the word with the highest score and uses it to predict the next word, we apply the beam search [53] to generate more stabilized captions. The beam search selects the top k candidates in each time step and then predicts top k new words accordingly for each of these sequences in the next step. Then, the new top k sequences of the next time step are selected out of all $k \times k$ candidates. It is worth to mention that in consideration of computational efficiency, top k candidate sequences are selected for each time step, and the sequence with the highest score is selected as the final caption output at the last time step. Therefore, up to time t, k sequences are generated instead of k^t. The k is called the "beam size", which is set to 5 for our experiment.

III. Experiments

In this section, we will introduce in detail on our experimental datasets, metrics, and comparison results. We also provide ablation experiments, parameter analysis, and speed analysis to verify the effectiveness of the proposed structured attention module.

A. Experimental Setup

1) Datasets: We conduct experiments on three widely-used remote sensing image captioning datasets - UCM-Captions [27], Sydney-Captions [27], and RSICD [28]. For each dataset, we followed the standard protocols on splitting the dataset into training, validation, and test sets. In any of the three datasets, each image is labeled with five sentences as ground truth captions. The following are the details of the three datasets.

a) UCM-Captions: The UCM-Captions dataset [27] is built based on the UC Merced land use dataset [54]. It contains 2,100 remote sensing images from 21 types of scenes. Each image has a size of 256×256 pixels and a spatial resolution of 0.3m/pixel.

b) Sydney-Captions: The Sydney-Captions dataset [27] is built based on the Sydney land dataset [55]. It totally contains 613 remote sensing images collected from the Google Earth imagery in Sydney, Australia. Each image has a size of 500 × 500 pixels and a spatial resolution of 0.5m/pixel.

c) RSICD: The RSICD [28] is the most widely used dataset for remote sensing image caption generation task. It contains 10,921 remote sensing images collected from the AID dataset [56] and other platforms such as Baidu Map, MapABC, and Tianditu. The images are in various spatial resolutions. The size of each image is 224×224 pixels.

2) Evaluation Metrics: We use four different metrics to evaluate the accuracy of the generated captions, including the BLEU [57], ROUGE-L [58], METEOR [59], and CIDEr-D [60], which are all widely used in recent image captioning literature.

a) BLEU: The BLEU (BiLingual Evaluation Understudy) [57] measures the co-occurrences between the generated caption and the ground truth by using n-grams (a set of n ordered words). The key of the BLEU-n ($n = \{1, 2, 3, 4\}$) is the n-gram precision - the proportion of the matched n-grams out of the total number of n-grams in the evaluated caption.

b) METEOR: Since the BLEU does not take the recall into account directly, to address this weakness, the METEOR [59] is introduced to compute the accuracy based on explicit word-to-word matches between the captioning and the ground truth.

c) ROUGE-L: ROUGE-L [58] is a modified version of ROUGE, which computes an F-measure with a recall bias using the Longest Common Subsequence (LCS) between the generated and the ground truth captions.

d) CIDEr-D: CIDEr-D [60] is an improved version of CIDEr, which first converts the caption into the form of the Term Frequency Inverse Document Frequency (TF-IDF) vector [61], and then calculates the cosine similarity of the reference caption and the caption generated by the model. CIDEr-D penalizes the repetition of specific n-grams beyond the number of times they occur in the reference sentence.

For any of the above four metrics, a higher score indicates a higher accuracy. The scores of BLEU, ROUGE-L, and METEOR are between 0 and 1.0. The score of CIDEr-D is between 0-10.0.

B. Ablation Studies

The ablation studies are conducted to analyze the importance of three different technical components of the proposed method, including the structured attention module, doubly stochastic regularization, and attention variance regularization. The ablation studies and parameter analysis experiments are performed on the UCM-Captions dataset, Sydney-Captions dataset, and RSICD dataset. We found that the proposed method behaves similarly on these datasets. For brevity, we only report results on the UCM-Captions.

TABLE I
ABLATION STUDIES ON THE PROPOSED STRUCTURED ATTENTION MECHANISM. THE EVALUATION SCORES (%) ARE REPORTED ON THE UCM-CAPTIONS DATASET [27].

Soft Attention [21]	Structured Attention (ours)	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr-D
✓	×	83.21	76.78	71.09	66.02	42.93	77.63	314.78
×	✓	**85.38**	**80.35**	**75.72**	**71.49**	**46.32**	**81.41**	**334.89**

TABLE II
ABLATION STUDIES ON THE TWO REGULARIZATION TERMS: DOUBLY STOCHASTIC REGULARIZATION (DSR) AND ATTENTION VARIANCE REGULARIZATION (AVR). THE EVALUATION SCORES (%) ARE REPORTED ON THE UCM-CAPTIONS DATASET [27].

DSR	AVR	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr-D
×	×	80.45	74.48	69.64	64.99	42.68	76.89	302.06
✓	×	84.71	79.00	74.24	69.71	45.27	79.97	329.62
×	✓	83.73	77.64	72.74	68.21	44.02	78.79	312.28
✓	✓	**85.38**	**80.35**	**75.72**	**71.49**	**46.32**	**81.41**	**334.89**

TABLE III
THE EVALUATION SCORES (%) OF OUR METHODS WITH A DIFFERENT NUMBER OF PROPOSALS PER IMAGE. ALL MODELS ARE TRAINED AND EVALUATED ON UCM-CAPTIONS DATASET [27].

Region num	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr-D	Training Time
4	84.63	78.84	73.85	69.40	45.62	80.59	328.23	155min
8	85.38	**80.35**	**75.72**	**71.49**	46.32	81.41	334.89	175min
12	85.70	80.03	75.27	70.75	**46.68**	**81.99**	332.93	193min
16	**85.78**	79.95	74.96	70.35	46.37	81.38	**338.80**	210min

TABLE IV
THE EVALUATION SCORES (%) OF OUR METHODS WITH DIFFERENT BEAM SIZE. ALL MODELS ARE EVALUATED ON THE UCM-CAPTIONS DATASET [27].

Beam Size	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr-D	Inference speed
1	83.48	76.60	70.78	65.51	43.15	78.93	321.25	3.23
2	85.27	79.73	74.70	70.11	45.72	81.01	332.15	2.08
3	85.21	79.95	75.14	70.71	45.89	81.15	333.49	1.59
4	85.33	80.26	75.64	71.42	46.25	81.36	334.30	1.20
5	**85.38**	**80.35**	75.72	71.49	46.32	**81.41**	**334.89**	1.09
6	85.28	80.34	**75.73**	**71.54**	**46.33**	81.23	333.93	0.94

We first remove the proposed structured attention module of our method and replace it with a standard soft-attention module [21] while keeping other configurations unchanged. Table I shows the comparison results. The best scores are marked as bold. The results show that compared with the baseline method, the structured attention improves the accuracy with a large margin in terms of all evaluation metrics (+5.47% on BLEU-4, +3.39% on METEOR, +3.78% on ROUGE-L, and +20.11% on CIDEr-D).

We then gradually remove the regularization terms from our loss function and train the corresponding captioning model separately. Table II shows their evaluation accuracy. We show that the doubly stochastic regularization and the attention variance regularization can both yield noticeable improvements in the captioning accuracy. Particularly, the proposed method trained with both of these two regularization terms achieves the best accuracy on all metrics.

C. Parameter Analysis

We also analyze two important parameters in our method: 1) the number of segmentation proposals N and 2) the beam size k.

We set the number of segmentation proposals N in the selective search to 4, 8, 12, 16, train each model and then evaluate the captioning accuracy accordingly. Table III shows the accuracy of our model with different segmentation proposals. The result shows that when the number of regions increases from 4 to 8, the evaluation scores are greatly improved. However, when the number of regions further increases from 8 to 16, the improvement of evaluation scores becomes less significant and some scores even decrease. This is because when the number of segmentation proposals set by N is much larger than the actual number of regions in the remote sensing image, the over-segmentation of the image will destroy the structures of the semantic contents. We also report the models' training time in the last column of Table III. We show that

TABLE V
EVALUATION SCORES (%) OF DIFFERENT METHODS ON THE UCM-CAPTIONS DATASET [27].

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr-D
VLAD+RNN [28]	63.11	51.93	46.06	42.09	29.71	58.78	200.66
VLAD+LSTM [28]	70.16	60.85	54.96	50.30	34.64	65.20	231.31
mRNN [27]	60.10	50.70	32.80	20.80	19.30	-	214.00
mLSTM [27]	63.50	53.20	37.50	21.30	20.30	-	222.50
mGRU [62]	42.56	29.99	22.91	17.98	19.41	37.97	124.82
mGRU-embedword [30]	75.74	69.83	64.51	59.98	36.85	66.74	279.24
ConvCap [63]	70.34	56.47	46.24	38.57	28.31	59.62	190.15
Soft-attention [28]	74.54	65.45	58.55	52.50	38.86	72.37	261.24
Hard-attention [28]	81.57	73.12	67.02	61.82	42.63	76.98	299.47
CSMLF [29]	36.71	14.85	7.63	5.05	9.44	29.86	13.51
RTRMN (semantic) [31]	55.26	45.15	39.62	35.87	25.98	55.38	180.25
RTRMN (statistical) [31]	80.28	73.22	68.21	63.93	42.58	77.26	312.70
SAA [30]	79.62	74.01	69.09	64.77	38.59	69.42	294.51
baseline	83.21	76.78	71.09	66.02	42.93	77.63	314.78
structured attention (ours)	**85.38**	**80.35**	**75.72**	**71.49**	**46.32**	**81.41**	**334.89**

* The "-" means that the scores are not reported in the reference papers.

TABLE VI
EVALUATION SCORES (%) OF DIFFERENT METHODS ON THE SYDNEY-CAPTIONS DATASET [27].

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr-D
VLAD+RNN [28]	56.58	45.14	38.07	32.79	26.72	52.71	93.72
VLAD+LSTM [28]	49.13	34.12	27.60	23.14	19.30	42.01	91.64
mRNN [27]	51.30	37.50	20.40	19.30	18.50	-	161.00
mLSTM [27]	54.60	39.50	22.30	21.20	20.50	-	186.00
mGRU [62]	69.64	60.92	52.39	44.21	31.12	59.17	171.55
mGRU-embedword [30]	68.85	60.03	51.81	44.29	30.36	57.47	168.94
ConvCap [63]	74.72	65.12	57.25	50.12	34.76	66.74	214.84
Soft-attention [28]	73.22	66.74	62.23	58.20	39.42	71.27	**249.93**
Hard-attention [28]	75.91	66.10	58.89	52.58	38.98	71.89	218.19
CSMLF [29]	59.98	45.83	38.69	34.33	24.75	50.18	75.55
SAA [30]	68.82	60.73	52.94	45.39	30.49	58.20	170.52
baseline	73.05	64.37	56.67	52.80	36.50	69.79	215.21
structured attention (ours)	**77.95**	**70.19**	**63.92**	**58.61**	**39.54**	**72.99**	237.91

* The "-" means that the scores are not reported in the reference papers.

increasing the number of regions leads to an increase in training time. To balance the accuracy of different metrics, we finally set the number of regions to $N = 8$.

The beam size k will affect the captioning accuracy as well as the inference time. We set different beam sizes in our method and analyze their accuracy and speed. All models are trained and evaluated on UCM-Captions dataset [27]. Table IV shows the evaluation results of our method with different beam sizes. We can see that when the beam size increases from 1 to 5, the evaluation scores are improved but are saturated at 6. We also show that increasing the beam size leads to a slower inference speed. To balance the captioning accuracy and the inference speed, we set the beam size to $k = 5$.

Fig. 5 and Fig. 6 show the percentage of accuracy improvement of the proposed method with different number of segmentation proposals and different beam size. The percentage of improvement is defined as: $\frac{\text{acc}-\text{acc}_0}{\text{acc}_0} \times 100\%$. We define the accuracy on $N = 4$ and $k = 1$ as the baseline accuracy acc_0.

D. Comparison with Other Methods

In this section, we evaluate our method on three datasets and compared our method with a variety of recent image captioning methods. The comparison methods include the VLAD+RNN [28], VLAD+LSTM [28], mRNN [27], mLSTM [27], mGRU [62], mGRU-embedword [30], ConvCap [63], Soft-attention [28], Hard-attention [28], CSMLF [29], RTRMN [31], and Sound Active Attention (SAA) [30]. Among these methods, most of them (except mGRU and ConvCap) are initially designed for the remote sensing image captioning task. However, their basic ideas are mainly borrowed from the natural image captioning [1], [21]. The details of these models are described as follows.

a) VLAD+RNN: VLAD+RNN [28] uses the handcrafted feature descriptor "VLAD" [64] as its encoder to compute image representations and use a naive RNN as its decoder to generate captions.

b) VLAD+LSTM: VLAD+LSTM [28] also uses VLAD to compute the image features, but the difference is that it uses

TABLE VII
EVALUATION SCORES (%) OF DIFFERENT METHODS ON THE RSICD DATASET [28].

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr-D
VLAD+RNN [28]	49.38	30.91	22.09	16.77	19.96	42.42	103.92
VLAD+LSTM [28]	50.04	31.95	23.19	17.78	20.46	43.34	118.01
mRNN [27]	45.58	28.25	18.09	12.13	15.69	31.26	19.15
mLSTM [27]	50.57	32.42	23.19	17.46	17.84	35.02	31.61
mGRU [62]	42.56	29.99	22.91	17.98	19.41	37.97	124.82
mGRU-embedword [30]	60.94	46.24	36.80	29.81	26.14	48.20	159.54
ConvCap [63]	63.36	51.03	41.74	34.52	33.25	57.70	166.48
Soft-attention [28]	67.53	53.08	43.33	36.17	32.55	**61.09**	**196.43**
Hard-attention [28]	66.69	51.82	41.64	34.07	32.01	60.84	179.25
CSMLF [29]	51.06	29.11	19.03	13.52	16.93	37.89	33.88
RTRMN (semantic) [31]	62.01	46.23	36.44	29.71	28.29	55.39	151.46
RTRMN (statistical) [31]	61.02	45.14	35.35	28.59	27.51	54.52	148.20
SAA [30]	67.60	54.30	44.33	36.45	31.09	55.36	193.96
baseline	68.39	53.64	43.69	36.37	30.35	55.76	154.80
structured attention (ours)	**70.16**	**56.14**	**46.48**	**39.34**	**32.91**	57.06	170.31

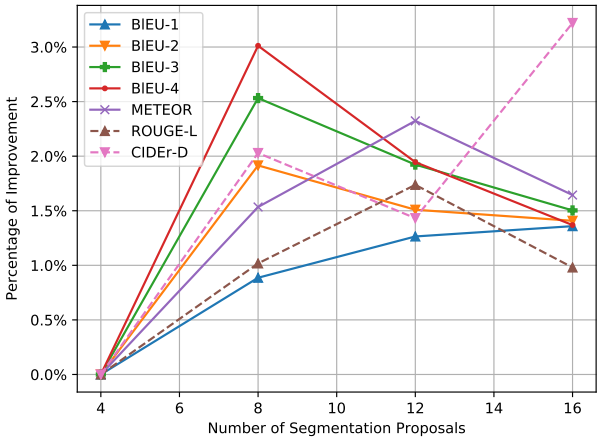

Fig. 5. (Better viewed in color) The percentage of accuracy improvement of the proposed method with a different number of segmentation proposals $N = \{4, 6, 8, 10, 12, 16\}$. All models are trained and evaluated on UCM-Captions dataset [27].

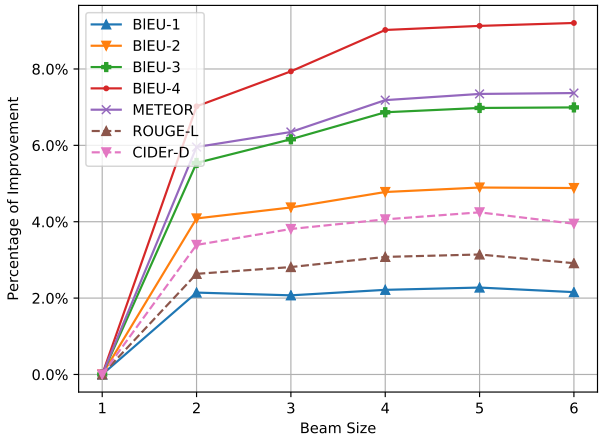

Fig. 6. (Better viewed in color) The percentage of accuracy improvement of the proposed method with different beam sizes $k = \{1, 2, 3, 4, 5, 6\}$. All models are evaluated on UCM-Captions dataset [27].

an LSTM as its decoder.

c) mRNN, mLSTM and mGRU: These three methods [27], [27], [62] all use the VGG-16 [65] as their encoders but use different RNNs (naive RNN, LSTM, and GRU) as their decoders.

d) mGRU-embedword: Similar to the mGRU [62], the mGRU-embedword [30] also uses the VGG-16 as its encoder and the GRU as its decoder. The difference is that mGRU-embedword uses a pre-trained global vector, namely GloVe [66], to embed words.

e) ConvCap: The ConvCap [63] uses the VGG-16 as its encoder and computes the attention weights based on the activations of the last convolutional layer. Instead of using the RNN based decoder, this method generates captions by using a CNN based decoder [63].

f) Soft-attention and Hard-attention: Soft-attention [28] and Hard-attention [28] are two methods using VGG-16 as the encoder and LSTM as the decoder. The decoders are build based on soft attention and hard attention mechanism [21], respectively.

g) CSMLF: CSMLF [29] is a retrieval-based method that uses latent semantic embedding to measure the similarity between the image representation and the sentence representation in a common semantic space.

h) RTRMN: RTRMN [31] uses Resnet-101 as its encoder and then uses the topic extractor to extract topic information. A retrieval topic recurrent memory network is used to generate captions based on the topic words. "RTRMN (semantic)" and "RTRMN (statistical)" are two variants of the RTRMN, which are based on semantic topics repository and statistical topics repository respectively.

i) SAA: SAA [30] introduces a sound active attention framework to combine the sound information during the generation of captions. SAA uses the VGG-16 and a sound

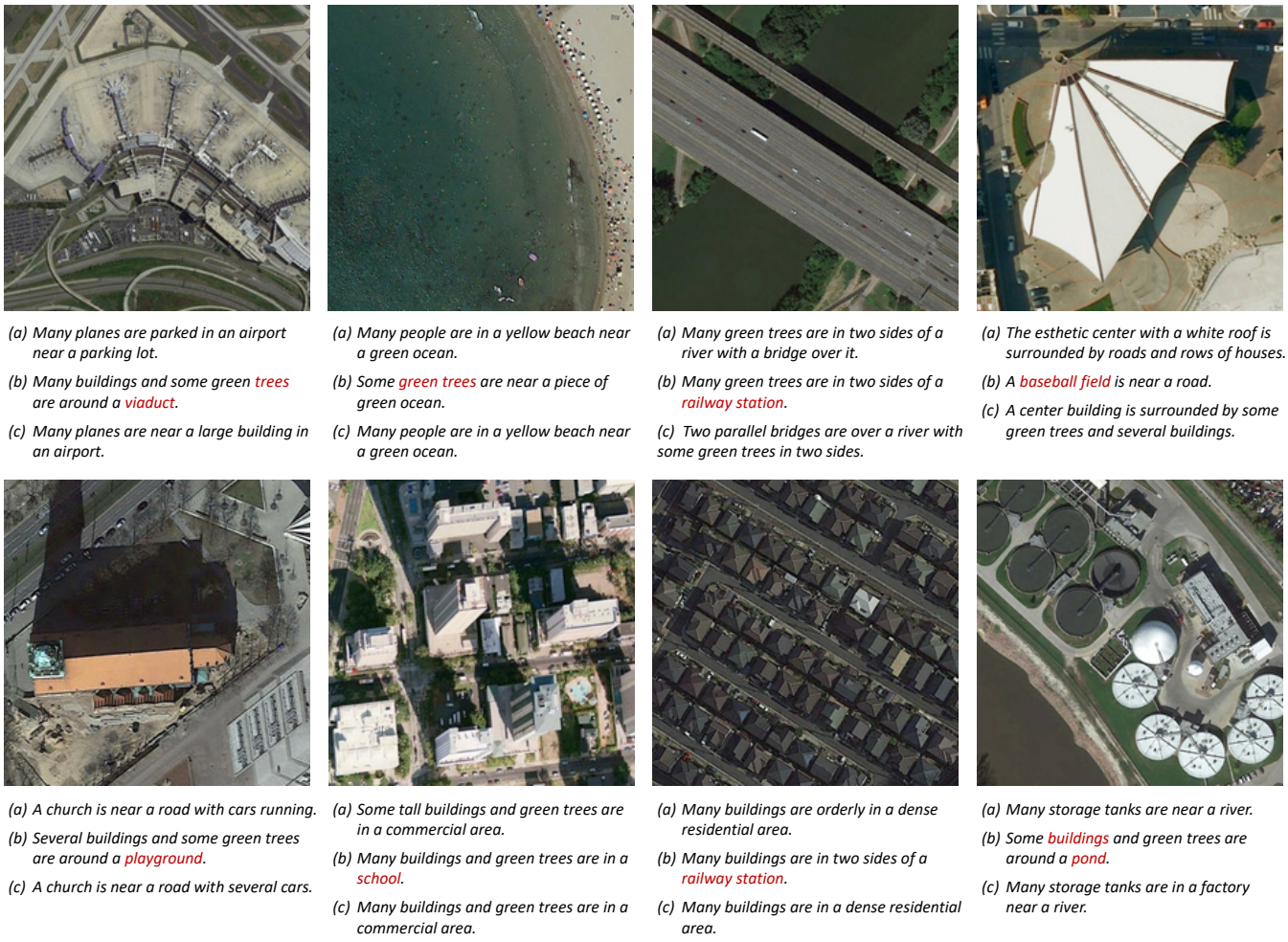

Fig. 7. Captioning results of the baseline method and the proposed method on the RSICD dataset [28]. In each group of the image, (a) shows one of the five ground truth captions, (b) shows the captions generated by the baseline method, and (c) shows the captions generated by the proposed method. The words that do not match the images are marked in red.

GRUs as its encoder and uses another GRUs as its decoder.

j) baseline: We first remove the proposed structured attention module of our method and replace it with a standard soft-attention module [21] while keeping other configurations unchanged as the baseline method.

Table V, Table VI and Table VII show the accuracy of our method and the above comparison ones on the three different datasets. All comparison methods follow the same fixed partitions of data (80% for training, 10% for validation, and 10% for test) which makes the comparison fair. In these tables, the best scores are marked as bold. For the comparison methods, the metric scores are taken from the papers that proposed them. Since Qu et al. [27] did not report the ROUGE-L scores on UCM-Captions and Sydney-Captions datasets, these numbers are missing in Table V and Table VI. We can see our method achieves the best accuracy in most of the entries. For example, on the RSICD dataset, our baseline method (Resnet50 + LSTM + soft attention), which also applies beam search with the same beam size during the inference stage, is already better than most of the other methods, as shown in Table VII. When we integrate the structured attention, we further improve our baseline by 2.97% on BLEU-4, 2.56% on METEOR, 1.30% on ROUGE-L, and 15.51% on CIDEr-D. While the baseline and proposed structured attention method both use ResNet-50 as the encoder and LSTM as the decoder, the proposed approach always achieves a score higher than the baseline, which means the achieved improvement is due to the proposed structured attention method.

E. Qualitative Analysis

1) Caption Generation Results: In Fig. 7, we show some captioning results of our method on the RSICD dataset. In each group of the result, we show one of the five ground truth sentences, the sentence generated by our method (Resnet50 + LSTM + structured attention), and that generated by our baseline (Resnet50 + LSTM + soft attention) accordingly. The generated words that do not match the images are marked in red. As we can see, the baseline method tends to generate false descriptions of small objects or irregularly shaped objects. This is because the regions captured by the soft attention is coarse-grained and unstructured, which leads to insufficient use of structured information and low-level visual information. As a

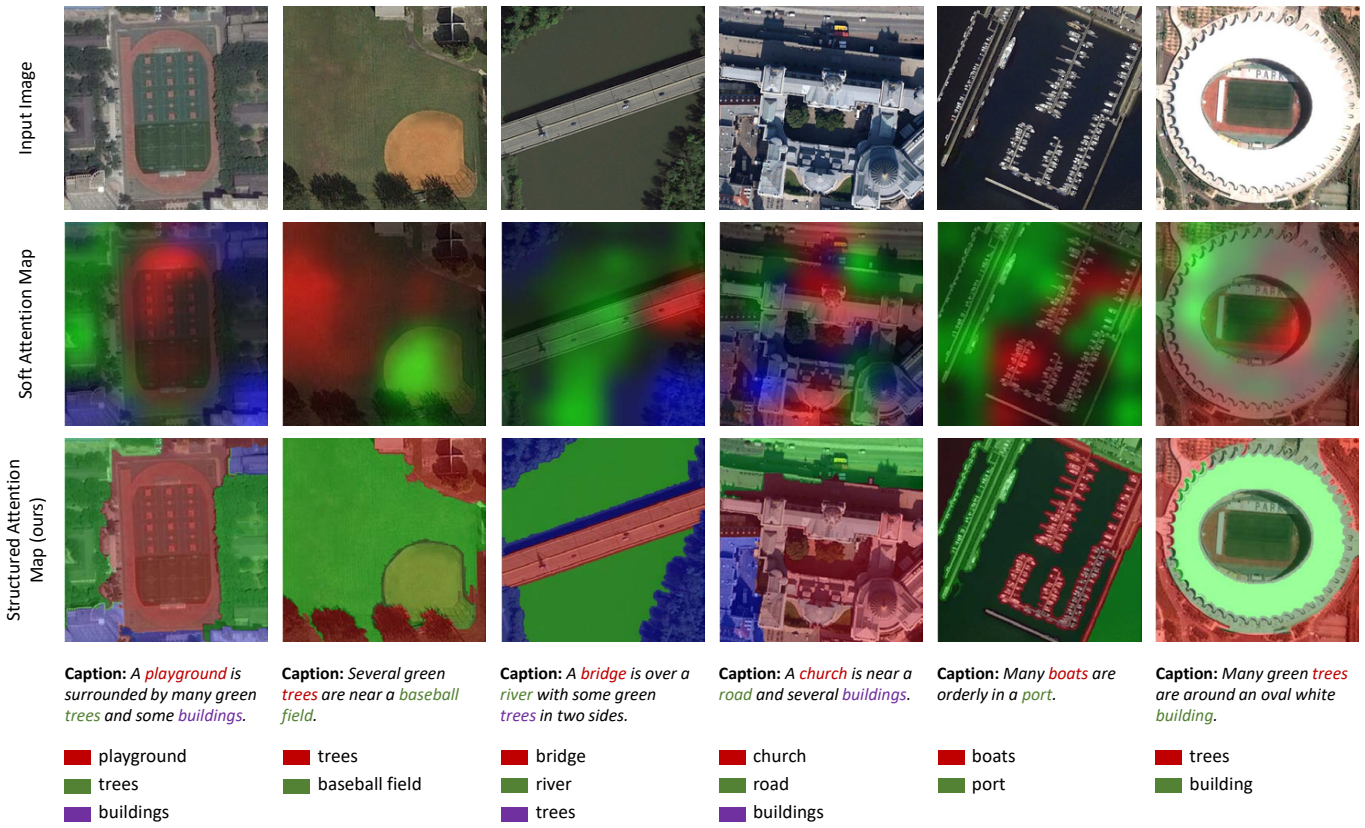

Fig. 8. Visualization of the attention weights on the RSICD dataset [28]. The attention weights are displayed in different colors and are overlaid on the original image for a better view. The proposed structured attention based method produces much more detailed and structural rational attention weights compared to the standard soft attention based method [21]. To reduce the mosaic effect, the soft attention maps are smoothed by bilinear interpolation for multiple times as suggested by previous literature [21].

comparison, our method generates more accurate descriptions, which is owning to the structured attention can fully exploit the structured information of remote sensing semantic contents.

2) Visualization of the Attention Weights: In Fig. 8, we visualize the weights produced by the standard soft attention and by the proposed structured attention. For each image, we visualize the attention weights of the attention module when the decoder generating corresponding words, such as trees, playgrounds, buildings, etc. For each word, the attention weights are displayed in different colors and are overlaid on the original image for a better view. As we can see that the structured attention can produce much more detailed and structural rational attention weights compared to the standard soft attention. Although we only train our method with image-level annotations and do not use any pixel-wise annotations during the training, our method still produces accurate and meaningful segmentation results. The attention maps generated by our method thus can be considered as a new way for weakly supervised image segmentation.

3) Visualization of Object Masks: Fig. 9 visualizes the regions in some images where the attention is heavily weighted as the decoder generates specific classes of words, such as the river, bridge, building, tree, etc. We can also see that some generated object masks usually have specific structures. For example, the river is winding and the bridge is long and straight. By leveraging the low level vision from the segmentation proposals with typical structures, the proposed structured attention can help the decoder to generate more accurate captions.

F. Speed performance

In Table VIII, we report the number of model parameters, the number of floating-point operations (FLOPs), training time, and the inference speed (images per second) of our method. The results are computed on the RSICD dataset. The training time and inference speed are tested on an NVIDIA TITAN X (Pascal) graphics card. Comparing with the baseline method, the proposed method does not increase the number of model parameters. This is because the basic network structure does not need to be changed when we modify the standard attention module to the proposed structured attention module. The proposed method even decreases the number of floating-point operations, because the baseline method generates 14x14=196 groups attention weights, each corresponding to a uniform grid in the image while the proposed method only need to generates 8 groups attention weights since we set the number of structured regions to 8 in our method. We extract and save region proposals of images locally before training the model, and directly load them from the local disk during training. Therefore, the training time is not affected by the region proposals extraction. Since the structured attention

Fig. 9. Visualization of the object masks generated by our method from the RSICD dataset. There are 10 classes of semantic contents where their masks can be generated by our method, including the baseball field, beach, bridge, center, church, playground, pond, port, river, and storage tanks.

TABLE VIII
A COMPARISON BETWEEN OUR METHOD AND THE BASELINE ON THE NUMBER OF PARAMETERS, FLPOS, TRAINING TIME, AND INFERENCE SPEED (IMAGES PER SECOND). ALL RESULTS ARE REPORTED BASED ON THE UCM-CAPTIONS DATASET.

Method	#parameters	FLPOs	training time	inference speed
baseline	34.6M	7.02G	171min	1.16
ours	34.6M	5.58G	175min	1.09

module produces only 8 structured attention weights instead of 196, which is used in a standard soft attention module, if we do not consider the segmentation time, the inference speed even becomes faster.

IV. CONCLUSION

We proposed a new image captioning method for remote sensing images based on the structured attention mechanism. The proposed method achieves captioning and weakly supervised segmentation under a unified framework. Different from the previous methods that are based on coarse-grained and soft attentions, we show the proposed structured attention based method can exploit the structured information of semantic contents and generate more accurate sentence descriptions. Experiments on three public remote sensing image captioning datasets suggest the effectiveness of our method. Compared with other state of the art captioning methods, our method achieves the best results on most evaluation metrics. The visualization experiments also show that our method can generate much more detailed and meaningful object masks than the soft attention based method.

Our method also has limitations. Our method is only applicable to high resolution remote sensing images. If the image is not high resolution, the selective search method will fail to extract available segmentation proposals to support the proposed structured attention mechanism. The number of segmentation proposals used in our structured attention module is fixed. When the input image contains more semantic contents than the predefined number of proposals, the structured attention module may fail to focus on the most salient regions. In future work, we will make the number of segmentation proposals adaptive. Another future direction is to combine remote sensing image captioning with object

detection. Particularly, we will focus on weakly-supervised detection, i.e., to train the detector only based on the sentence annotation.

REFERENCES

[1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.

[2] X. Chen and C. L. Zitnick, "Learning a recurrent visual representation for image caption generation," *arXiv preprint arXiv:1411.5654*, 2014.

[3] Z. Shi and Z. Zou, "Can a machine generate humanlike language descriptions for a remote sensing image?" *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 6, pp. 3623–3634, 2017.

[4] Z. Zou, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *arXiv preprint arXiv:1905.05055*, 2019.

[5] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren, "Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 6, pp. 3325–3337, 2014.

[6] Z. Zou and Z. Shi, "Ship detection in spaceborne optical image with svd networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 10, pp. 5832–5845, 2016.

[7] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 12, pp. 7405–7415, 2016.

[8] Z. Zou and Z. Shi, "Random access memories: A new paradigm for target detection in high resolution aerial remote sensing images," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1100–1111, 2017.

[9] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proceedings of the IEEE*, vol. 105, no. 10, pp. 1865–1883, 2017.

[10] X. Lu, X. Zheng, and Y. Yuan, "Remote sensing scene classification by unsupervised representation learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 9, pp. 5148–5157, 2017.

[11] B. Pan, X. Xu, Z. Shi, N. Zhang, H. Luo, and X. Lan, "Dssnet: A simple dilated semantic segmentation network for hyperspectral imagery classification," *IEEE Geoscience and Remote Sensing Letters*, 2020.

[12] M. Pesaresi and J. A. Benediktsson, "A new approach for the morphological segmentation of high-resolution satellite imagery," *IEEE transactions on Geoscience and Remote Sensing*, vol. 39, no. 2, pp. 309–320, 2001.

[13] A. A. Farag, R. M. Mohamed, and A. El-Baz, "A unified framework for map estimation in remote sensing image segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 7, pp. 1617–1634, 2005.

[14] P. Ghamisi, M. S. Couceiro, F. M. Martins, and J. A. Benediktsson, "Multilevel image segmentation based on fractional-order darwinian particle swarm optimization," *IEEE Transactions on Geoscience and Remote sensing*, vol. 52, no. 5, pp. 2382–2394, 2013.

[15] Z. Zou, T. Shi, W. Li, Z. Zhang, and Z. Shi, "Do game data generalize well for remote sensing image segmentation?" *Remote Sensing*, vol. 12, no. 2, p. 275, 2020.

[16] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, "Explain images with multimodal recurrent neural networks," *arXiv preprint arXiv:1410.1090*, 2014.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[18] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra, "Draw: A recurrent neural network for image generation," *arXiv preprint arXiv:1502.04623*, 2015.

[19] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[21] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048–2057.

[22] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4651–4659.

[23] L. Zhou, C. Xu, P. Koch, and J. J. Corso, "Image caption generation with text-conditional semantic attention," *arXiv preprint arXiv:1606.04621*, vol. 2, 2016.

[24] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 375–383.

[25] L. Li, S. Tang, L. Deng, Y. Zhang, and Q. Tian, "Image caption with global-local attention," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[26] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6077–6086.

[27] B. Qu, X. Li, D. Tao, and X. Lu, "Deep semantic understanding of high resolution remote sensing image," in *2016 International Conference on Computer, Information and Telecommunication Systems (CITS)*. IEEE, 2016, pp. 1–5.

[28] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 4, pp. 2183–2195, 2017.

[29] B. Wang, X. Lu, X. Zheng, and X. Li, "Semantic descriptions of high-resolution remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 8, pp. 1274–1278, 2019.

[30] X. Lu, B. Wang, and X. Zheng, "Sound active attention framework for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, 2019.

[31] B. Wang, X. Zheng, B. Qu, and X. Lu, "Retrieval topic recurrent memory network for remote sensing image captioning," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 256–270, 2020.

[32] K. Tran, A. Bisazza, and C. Monz, "Recurrent memory networks for language modeling," *arXiv preprint arXiv:1601.01272*, 2016.

[33] X. Ma, R. Zhao, and Z. Shi, "Multiscale methods for optical remote-sensing image captioning," *IEEE Geoscience and Remote Sensing Letters*, 2020.

[34] W. Cui, F. Wang, X. He, D. Zhang, X. Xu, M. Yao, Z. Wang, and J. Huang, "Multi-scale semantic segmentation and spatial relationship recognition of remote sensing images based on an attention model," *Remote Sensing*, vol. 11, no. 9, p. 1044, 2019.

[35] G. Sumbul, S. Nayak, and B. Demir, "Sd-rsic: Summarization-driven deep remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[36] X. Li, X. Zhang, W. Huang, and Q. Wang, "Truncation cross entropy loss for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[37] Q. Wang, W. Huang, X. Zhang, and X. Li, "Word-sentence framework for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[38] W. Huang, Q. Wang, and X. Li, "Denoising-based multiscale feature fusion for remote sensing image captioning," *IEEE Geoscience and Remote Sensing Letters*, 2020.

[39] Y. Li, S. Fang, L. Jiao, R. Liu, and R. Shang, "A multi-level attention model for remote sensing image captions," *Remote Sensing*, vol. 12, no. 6, p. 939, 2020.

[40] S. Wu, X. Zhang, X. Wang, C. Li, and L. Jiao, "Scene attention mechanism for remote sensing image caption generation," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–7.

[41] J. Xiao and L. Quan, "Multiple view semantic segmentation for street view images," in *2009 IEEE 12th international conference on computer vision*. IEEE, 2009, pp. 686–693.

[42] Y. Liu, J. Liu, Z. Li, J. Tang, and H. Lu, "Weakly-supervised dual clustering for image semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2075–2082.

[43] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint arXiv:1412.7062*, 2014.

[44] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3234–3243.

[45] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

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

[47] W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent neural network regularization," *arXiv preprint arXiv:1409.2329*, 2014.

[48] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[50] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[51] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[52] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 6, pp. 1397–1409, 2012.

[53] A. Graves, "Sequence transduction with recurrent neural networks," *arXiv preprint arXiv:1211.3711*, 2012.

[54] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*. ACM, 2010, pp. 270–279.

[55] F. Zhang, B. Du, and L. Zhang, "Saliency-guided unsupervised feature learning for scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 4, pp. 2175–2184, 2014.

[56] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, "Aid: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 3965–3981, 2017.

[57] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.

[58] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[59] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.

[60] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.

[61] S. Robertson, "Understanding inverse document frequency: on theoretical arguments for idf," *Journal of documentation*, vol. 60, no. 5, pp. 503–520, 2004.

[62] X. Li, A. Yuan, and X. Lu, "Multi-modal gated recurrent units for image description," *Multimedia Tools and Applications*, vol. 77, no. 22, pp. 29 847–29 869, 2018.

[63] J. Aneja, A. Deshpande, and A. G. Schwing, "Convolutional image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5561–5570.

[64] H. Jegou, F. Perronnin, M. Douze, J. Sánchez, P. Perez, and C. Schmid, "Aggregating local image descriptors into compact codes," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 9, pp. 1704–1716, 2011.

[65] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[66] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

Rui Zhao received his B.S. degree from the Image Processing Center, School of Astronautics, Beihang University. He is currently working toward his master's degree in the Image Processing Center, School of Astronautics, Beihang University. His research interests include computer vision, deep learning, and image captioning.

Zhenwei Shi (Member, IEEE) received his Ph.D. degree in Mathematics from Dalian University of Technology, Dalian, China, in 2005. He was a Postdoctoral Researcher in the Department of Automation, Tsinghua University, Beijing, China, from 2005 to 2007. He was a visiting scholar in the Department of Electrical Engineering and Computer Science, Northwestern University, U.S.A., from 2013 to 2014. He is currently a Professor and the Dean of the Image Processing Center, School of Astronautics, Beihang University, Beijing. He has authored or coauthored over 100 scientific articles in refereed journals and proceedings, including the IEEE Transactions on Pattern Analysis and Machine Intelligence, the IEEE Transactions on Image Processing, the IEEE Transactions on Geoscience and Remote Sensing, the IEEE Conference on Computer Vision and Pattern Recognition, and the IEEE International Conference on Computer Vision. His research interests include remote sensing image processing and analysis, computer vision, pattern recognition, and machine learning. Dr. Shi serves as an Associate Editor for Infrared Physics and Technology and Editorial Advisory Board member for ISPRS Journal of Photogrammetry and Remote Sensing.

Zhengxia Zou received his B.S. degree and his Ph.D. degree from the Image Processing Center, School of Astronautics, Beihang University in 2013 and 2018. He is now working at the University of Michigan, Ann Arbor, as a postdoc research fellow. His research interests include computer vision and the related applications in remote sensing, self-driving vehicles, and video games. He serves as a Senior Program Committee member/Reviewer for a number of top conferences and top journals, including the NeurIPS, CVPR, AAAI, IEEE TIP, IEEE SPM, and IEEE TGRS. His personal website is http://www-personal.umich.edu/~zzhengxi/.