IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 28, 2019, accepted September 16, 2019, date of publication September 18, 2019, date of current version October 3, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2942154

VAA: Visual Aligning Attention Model for Remote Sensing Image Captioning

ZHENGYUAN ZHANG[1,2,3], WENKAI ZHANG[1,2], WENHUI DIAO[1,2], MENGLONG YAN[1,2], XIN GAO[1,2], AND XIAN SUN[1,2]

[1]Institute of Electronics, Chinese Academy of Sciences, Beijing 100190, China
[2]Key Laboratory of Network Information System Technology (NIST), Institute of Electronics, Chinese Academy of Sciences, Beijing 100190, China
[3]School of Electronics, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100190, China

Corresponding author: Xian Sun (sunxian@mail.ie.ac.cn)

This work was supported by the National Natural Science Foundation of China under Grant 41701508.

ABSTRACT Owing to the effectiveness in selectively focusing on regions of interest of images, the attention mechanism has been widely used in image caption task, which can provide more accurate image information for training deep sequential models. Existing attention-based models typically rely on top-down attention mechanism. While somewhat effective, attention masks in these attention-based models are queried from image features by hidden states of LSTM, rather than optimized by the objective functions. This indirectly supervised training approach cannot ensure that attention layers accurately focus on regions of interest. To address the above issue, in this paper, a novel attention model, Visual Aligning Attention model (VAA), is proposed. In this model, the attention layer is optimized by a well-designed visual aligning loss during the training phase. The visual aligning loss is obtained by explicitly calculating the feature similarity of attended image features and corresponding word embedding vectors. Besides, in order to eliminate the influence of non-visual words in training the attention layer, a visual vocab used for filtering out non-visual words in sentences is proposed, which can neglect the non-visual words when calculating the visual aligning loss. Experiments on UCM-Captions and Sydney-Captions prove that the proposed method is more effective in remote sensing image caption task.

INDEX TERMS Image captioning, remote sensing image captioning, attention mechanism, visual aligning.

I. INTRODUCTION

Image captioning is a complicated task that bridges both the visual and linguistic domains. In this task, image captioning models are required to understand the content of input images to generate sentences with human languages. Unlike most of other existing models, designed for classification [1]–[4], object detection [5]–[12], and semantic segmentation [13]–[17], [17], [18] tasks, image caption models are able to satisfy the demand of refinement retrieval. Simultaneously, with the rapid development of remote sensing technology, remote sensing images with high resolution can be easily accessed. However, the increase of quantity of remote sensing images bring more difficulty for managing such big remote sensing images. Therefore, remote sensing image captioning (RSIC) is quite meaningful for this problem. What's more, many applications, such as remote sensing image retrieval [19], scene classification [20], military intelligence generation [21], require technologies related to image captioning to interpret remote sensing images.

Compared to the template-based models [22]–[25] and retrieval-based models [26]–[28], encoder-decoder based models are good at generating length-variable and syntax-variable sentences. Therefore, encoder-decoder based models [29]–[31] are widely used in image caption tasks. In this model, a convolutional neural network (CNN) is selected as the encoder for extracting image features, and the decoder is composed of a recurrent neural network (RNN) for generating sentences. However, in common sense, in order to predict different words, different image information related to the predicted words is required. If all the image information is fed into the decoder without filtering out useless information, the decoder cannot capture purer information, which is useful for generating more accurate sentences. So, this large amount of interference information in the images greatly limits the robustness of the existing model. In order to filter out the

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

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

useless image information, the attention mechanism has been applied to the image caption task and has attracted more attention. The attention layers can learn to focus on different regions of interest in images. The attended image features will be purer and useful for generating more accurate sentences to describe the content of input images.

Although the attention mechanism has certain effects in suppressing image useless information and improving image caption results, there are still some defects for processing remote sensing images. Compared with natural images, remote sensing images cover a wide area. The background in remote sensing images occupies a considerable proportion. Especially when a specific visual word, for example airplane, is wanted to be generated. The useless information in image features is more than that in natural images. Therefore, it is necessary to exclude the useless information by further improving the conventional attention mechanism. In detail, the training of attention layer is not directly constrained by the loss function. The propagation of function loss passes through the MLPs, RNN and embedded layer to reach the attention layer. When predicting visual words, this implicit constraint training process cannot ensure that the attention layer is accurately focused on interested region on the image. At the same time, they cannot guarantee that the useless information in the image features, which will be sent to the following RNN to predict words, can be filtered out at each time step.

In summary, the main contributions of this paper are as follows:

1) A novel attention model, Visual Aligning Attention model (VAA), is proposed to align the visual words and their corresponding image features for constraining the attention layers.
2) A novel visual mask is proposed to filter out non-visual words in captions. It is helpful for caption models to choose visual words at each time step in the training phase. In addition to this, a method of automatically constructing a visual mask is also proposed.
3) A novel attention loss function, visual aligning loss, is proposed for visual words to constrain the formation of attention masks. Optimized by this well-designed visual aligning loss, better sentences can be generated and better scores can be obtained, separately, by the visual aligning attention models.

Finally, by a series of experiments on UCM-Captions and Sydney-Captions, the well-trained Visual Aligning Attention model can obtain better results.

II. RELATED WORKS

A. NATURAL IMAGE CAPTIONING (NIC)

With the rapid development of computer vision and natural language processing. Many methods [32]–[36] have been proposed in the field of natural image captioning. There are mainly three ways to generate descriptions for natural images: template-based methods, retrieval-based methods, and encoder-decoder based methods.

1) TEMPLATE-BASED METHODS

Template-based methods need people to pre-define sentence templates for images of each category in the datasets [22]–[25]. Then, by detecting or classifying the entities in the images, the detected or classified results will be filled in the blank of pre-defined templates. In this way, sentences can be generated for the corresponding images. It is not hard to find that template-based methods need several stages for processing images and combining sentences. Hence, the models applied these methods need well-trained detection models, instead of RNNs, to detect objects in images and cannot be trained end-to-end. More seriously, if the objects are detected wrongly, the generated sentences will not be semantically correct. The whole natural image caption models' performance relies mainly on the detection models. The number of types of sentence templates are limited and the length of generated sentences are not variable.

2) RETRIEVAL-BASED METHODS

Retrieval-based methods first try to find similar images in the training dataset according to the query image [26]–[28]. Then, the sentence for the query image can be chosen from the similar images' corresponding captions. Although the styles and content of the retrieved sentences are different, the generated sentences cannot totally correctly represent the content of the query image. The retrieved sentences often cannot describe the details of the query image well. What's more, models applied these methods cannot be trained end-to-end, either.

3) ENCODER-DECODER BASED METHODS

Encoder-decoder based methods have been widely used in natural image captioning task and achieve many great performances. Generally, the encoders are used to extract image features from input images, of which the backbones are composed of CNN part of the classification networks, such as AlexNet [1], VGG [2], Inception [3], and ResNet [4]. The decoders are made of RNN, GRU or LSTM to generate sentences for describing the query images. Vinyals *et al.* [37] propose a neural image captioning (NIC) method by using LSTM to generate sentences instead of RNN to avoid long term gradient dissipation. The words input to LSTM at different time steps are different. But the image features are only input to LSTM at the first time step. Hence, the remained image features will decrease as time flows. Besides, the local image features are abandoned and not utilized in this method. The information contained in the local image features are richer than that in the global image features. Thus, abandoning the local image features will lead that the generated sentences cannot describe the details in images. Hence, Xu *et al.* [29] introduced attention mechanisms into encoder-decoder frameworks to solve the problem above. At each time step, attention models will focus on different regions in images and obtain different attended image features according to hidden state of LSTM at last

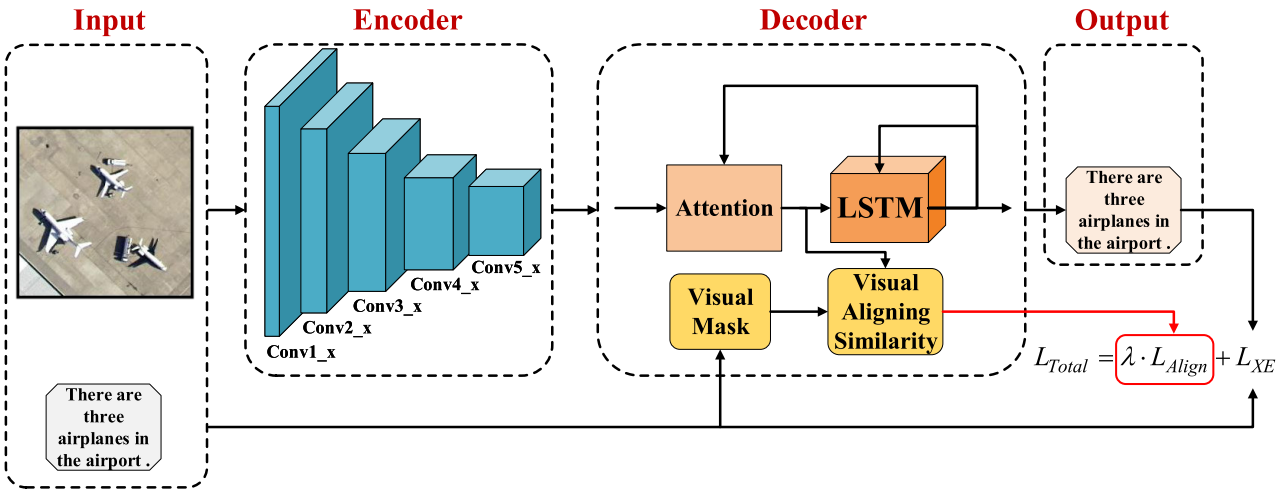


FIGURE 1. It shows the overview of proposed Visual Aligning Attention model, which is based on encoder-decoder framework. The Visual Mask module is used for filtering out non-visual words in sentences in the training phase. The Visual Aligning Similarity module are utilized to calculate the feature similarity between image features and word embedding vectors. The red line indicates how to obtain the visual aligning loss, which is incorporated into the total loss function.

time step. Then the attended image features will be changed at each time step and be sent to LSTM to predict the next word. However, the process of training attention layers is implicit and does not be directly constrained by the loss function. Lu *et al.* [30] propose a method to adaptively adjust to attend image features and word features. If the according word is a visual word, the attention weight will larger than the weights for attending word features. Although this method utilizes visual words to adaptively attend image features and word features, the attention layers are not constrained directly by the loss function, either. Chen and Zhao [31] utilized stimulus-based Attention method to obtain the saliency map rather than directly optimizing the attention mechanism.

B. REMOTE SENSING IMAGE CAPTIONING (RSIC)

Qu *et al.* [38] adopt a multimodal encoder-decoder based method, utilizing CNN to extract image features and fusing the image features with hidden state at last time step. Then, this fused feature vectors will be processed by MLP to predict the word at current time step. Finally, captions for describing remote sensing images can be generated. This is the earliest work in remote sensing image captioning task and they have disclosed two remote sensing image captioning datasets, UCM-Captions and Sydney-Captions, for further research. Lu *et al.* [39] propose an attention-based encoder-decoder framework for remote sensing image captioning task. This prove that attention mechanism is adaptive in remote sensing image captioning task although there is a big difference between natural images and remote sensing images. In addition, they also create a new dataset, in which the number of images is larger and the sentence pattern and content of captions are more various than UCM-Captions and Sydney-Captions. Zhang *et al.* [40] propose a novel attribute attention mechanism. Models applied the attribute attention mechanism can dynamically attend to the image regions according to the attributes at each time step. In detail, a align model is used to concatenate the high-level image features and the hidden states from decoders to calculate the attention masks. The calculation process is changed indeed, nevertheless, the process of training attention layers is implicit and not changed.

III. VISUAL ALIGNING ATTENTION MODEL

In this paper, a novel attention model, Visual Aligning Attention model is proposed to solve the problem of indirectly constraining the training process of attention layers. In this way, more constraint can be introduced into training the attention layers. And then the well-trained attention layers can attend more useful information in images. The details about the Visual Aligning Attention model are presented as follows. Firstly, the overall framework of VAA model are presented in 1, which is composed of a visual model and a language model, used to extract image features and generate sentences, separately. Then, in the training phase, visual words in captions can be chosen automatically by a novel visual mask. Finally, these visual words will be used to constrain the attention layers by a well-designed visual aligning loss function, calculated by the feature similarity between attended image features and visual word embedding vectors.

A. OVERALL FRAMEWORK

Analogous to natural image captioning task, the definition of remote sensing image captioning task is generating sentence descriptions for input remote sensing images. The purpose of remote sensing image captioning models is to maximize the probability of correctly generated sentences given the images. The optimization of caption models can be defined by using the following formulation:

$$\boldsymbol{\theta}^{\star} = \arg\max_{\boldsymbol{\theta}} \sum_{(\boldsymbol{I},Y)} \log p(Y|Y;\boldsymbol{\theta}) \tag{1}$$

where $\boldsymbol{\theta}$ are the parameters of the proposed model, $\boldsymbol{I}$ is the input remote sensing images, and Y stand for the generated sentences according to the content of input images. Owing to the length of generated sentences are variable. Thus, applying the chain rule to the formulation, the following formulation can be obtained.

$$\log p(Y|\boldsymbol{I}) = \sum_{t=0}^{N} \log p\left(y_t|\boldsymbol{I}, y_0, \ldots, y_{t-1}\right) \tag{2}$$

where y_0,âĂę,y_{t-1} are the generated word sequence in previous $t-1$ time steps.

Most of existing methods utilize CNN (e.g., VGG16 [2]), pre-trained on ImageNet [41], to extract image features from the input remote sensing image $\boldsymbol{I}$. The extracted image features are extracted from the *conv*5_3 layer before the last fully connected layers and can be flattened and represented as a set of feature vectors $\boldsymbol{V} \in \mathbf{R}^{k\times d}$ as defined in (3), each of which is a d-dimension feature vector $\boldsymbol{v}_i \in \mathbf{R}^d$ and corresponds to one region of the input image $\boldsymbol{I}$. For VGG16, the dimension of feature maps from conv5_3 layer of VGG16 is $512\times 14\times 14$. 512 is the number of feature maps. 14×14 is the size of one feature map. Then, all these feature maps can be flattened into 512×196 features ($k=196$,$d=512$).

$$\mathbf{V} := \{\boldsymbol{v}_i\}_{i=1}^{k}, \quad \text{where} \quad \boldsymbol{v}_i \in \mathbb{R}^d \tag{3}$$

Compared with RNN, LSTM has the benefit that it can avoid long term gradient dissipation. Hence, as many other methods, LSTM is selected to be the decoder in this paper. Besides, the attention mask $\hat{\boldsymbol{\alpha}}_t$ at each time step is calculated by the function $f_{att}(\boldsymbol{V}, \boldsymbol{h}_{t-1})$. Concretely, spatial image features combined with hidden states of LSTM at the previous time step $\boldsymbol{h}_{t-1}$ are fed into a fully connected layer followed by a softmax layer to generated attention distributions over k regions of the input image. The following formulation is shown to illustrate the calculation process.

$$\begin{aligned}\hat{\boldsymbol{\alpha}}_t &= f_{att}\left(\mathbf{V}, \mathbf{h}_{t-1}\right)\\ &= \varphi\left[\boldsymbol{\omega}_c^T \delta\left(\mathbf{W}_v\mathbf{V} + \mathbf{W}_h\mathbf{h}_{t-1}\right)\right]\end{aligned} \tag{4}$$

where $\boldsymbol{W}_v \in \mathbf{R}^{k\times d}$, $\boldsymbol{W}_h \in \mathbf{R}^{k\times n}$, and $\boldsymbol{\omega}_c \in \mathbf{R}^k$ are the trainable parameters of MLP.

According to the attention masks $\hat{\boldsymbol{\alpha}}_t$, the attention distributions for the image regions will be re-adjusted. Thus, the importance of the different regions will be changed. To collect the information of all image regions, the context vector $\boldsymbol{c}_t$ can be computed by summing up the attended image features. The context vector contains the image information and the word embedding vector contains the word information. With the help of LSTM, the hidden state $\boldsymbol{h}_t$ at the time step t can be predicted. The details are shown in the following formulation.

$$\boldsymbol{c}_t = \sum_i \boldsymbol{\alpha}_t \boldsymbol{v}_i \tag{5}$$

$$\boldsymbol{h}_t = \mathrm{LSTM}\left(\boldsymbol{c}_{t-1}, \boldsymbol{W}_e^T \boldsymbol{x}_{t-1}\right) \tag{6}$$

$$y_t = \arg\max\left(\mathrm{softmax}[\boldsymbol{W}_o^T \boldsymbol{h}_t]\right) \tag{7}$$

where $\boldsymbol{x} \in \mathbf{R}^m$ is the one-hot vector representation for the vocab of size m. $\boldsymbol{W}_e \in \mathbf{R}^{m\times e}$ is a parameter for e dimension word embedding. $\boldsymbol{W}_o \in \mathbf{R}^{n\times m}$ is composed of a MLP for mapping hidden states into vectors with m dimension, convenient for the subsequent word prediction.

B. VISUAL MASK

The embedding vectors of visual words are needed for calculating the feature similarity. But there are visual words and non-visual words in sentences. Therefore, a novel visual aligning mask is proposed to automatically filter out those non-visual words in the sentences. To achieve this purpose, a visual vocabulary $\boldsymbol{V}_{visual}$, only containing visual words, should be built according to the sentences in image caption datasets at first. In detail, the first step is to build a complete vocabulary for remote sensing image-caption datasets by using an open API pycocotools.[1] Next, separate visual words from the complete vocabulary by Stanford tools.[2] And then a new visual vocabulary is created. The reason why not directly creating visual vocabulary from remote sensing captions is that compared with building a visual vocabulary from remote sensing captions, separating the visual words part from the complete vocabulary is more time-saving.

For better measuring the feature similarity between attended image features and word embedding vectors, building a relationship between them is needed. Both of them, thus, need be mapped into a common vector space with the same dimension at first. Then their correlation can be measured by calculating the feature similarity. Owing to the vector space distributions of attended image features and word embedding vectors are different, two individual multi-layer perceptrons (MLPs), f_c and f_x, are adopted to project these two kinds of vectors into a common vector space, separately. The details are as follows:

$$\hat{\boldsymbol{c}} = f_c(\boldsymbol{c}) \tag{8}$$

$$\hat{\boldsymbol{x}} = f_x(\boldsymbol{x}) \tag{9}$$

where $\hat{\boldsymbol{c}} \in \mathbb{R}^c$ and $\hat{\boldsymbol{x}} \in \mathbb{R}^c$ are the mapped context vectors and the mapped word embedding vectors.

What's more, in order to better build the relationship between the mapped image features and mapped word embedding vectors, the dimension of the common multimodal vector space should be set to a little larger value [33]. Meanwhile, the dimension should not be too large, otherwise the weights of the MLPs will be too hard to be trained well and easy to overfit. Hence, the dimension of the common multimodal vector space is set to 1024 in this paper.

C. VISUAL ALIGNING LOSS

When the attended image features and word embedding vectors are mapped into a common vector space, they can be

[1] https://github.com/cocodataset/cocoapi
[2] https://nlp.stanford.edu/software/tagger.shtml

utilized for constraining the training phase of attention layers by maximizing the feature similarity of these two kinds of vectors. In this paper, the cosine function is chosen to calculate the feature similarity of the mapped image features and the mapped word embedding vectors. The equation is as follows:

$$\boldsymbol{Sim}_{align}(\hat{\boldsymbol{v}}_t, \hat{\boldsymbol{x}}_t) = \cos(\hat{\boldsymbol{v}}_t, \hat{\boldsymbol{x}}_t) \tag{10}$$

For each image, the cross-entropy loss is computed after the whole sentence is generated. Thus, in order to introduce the visual aligning loss into the total loss function, we need to calculate the visual aligning similarity for words in the generated sentence and sum them up. It is worth noting that only the visual words are chosen to calculate the visual aligning similarity, while those non-visual words are neglected. What's more, in order to prevent the final total loss value too large, mean operation is performed on all visual aligning similarity for visual words in each sentence. The equations are as follows:

$$\boldsymbol{L}_{align} = \begin{cases} 1 - \dfrac{1}{N}\sum\limits_{t} \boldsymbol{Sim}_{align}(\hat{\boldsymbol{v}}_t, \hat{\boldsymbol{x}}_t), & \text{if } \boldsymbol{x}_t \text{ in } \boldsymbol{V}_{visual} \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

where N is the number of visual words existing in captions at each epoch.

Before the visual aligning loss is added into the total loss function, a trade-off parameter is proposed to adjust the importance of visual aligning loss. In this way, the total loss can enhance the constraint of attention layers without bringing bad effects on the process of generating sentences. The formulation is as follows:

$$\boldsymbol{L}_{total} = \boldsymbol{L}_{XE} + \lambda \cdot \boldsymbol{L}_{align} \tag{12}$$

where $\boldsymbol{L}_{XE}$ is the cross-entropy loss for calculating loss of between generated captions and ground truth sentences. λ is a trade-off parameter.

IV. EXPERIMENTS

A. DATASETS

In these experiments, we choose two public remote sensing image captioning datasets, UCM-Captions and Sydney-Captions. Parts of images in UCM-Captions and Sydney-Captions are shown in Fig. 2, separately.

UCM-Captions is based on UC Merced Land Use Dataset [42], which is initially used for scene classification task and contains 21 classes, including agricultural, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks, and tennis court. There are 100 images in each class in UCM Dataset and every image is 256*256 pixels. In addition, Lu *et al.* [39] exploit natural sentences to describe the content of images in UCM dataset and supplement five different sentences for each image. The new created UCM Dataset with captions is called UCM-Captions.

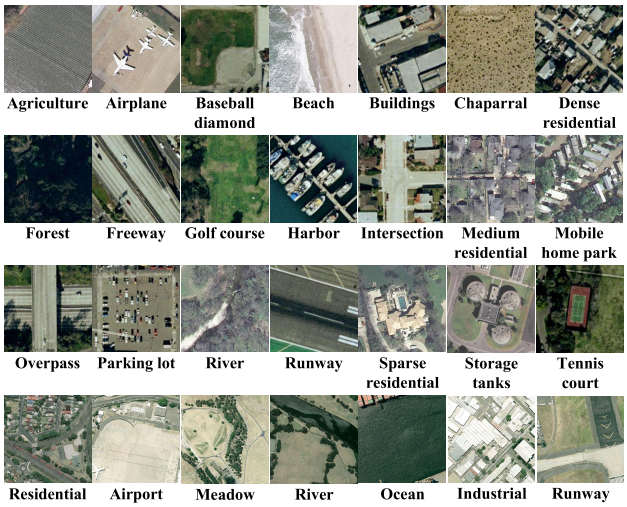


FIGURE 2. It shows one image of each category in the UCM-Captions and Sydney-Captions. The first row and the second row are images in UCM-Captions, while the third row are images in Sydney-Captions.

Sydney-Captions is based on Sydney Dataset [43], which is initially used for scene classification task and contains 7 classes, including residential, airport, meadow, river, ocean, industrial, and runway. There are totally 613 images in Sydney Dataset and every image is 500*500 pixels. In addition, Lu *et al.* [39] also add five different natural language descriptions for each image in Sydney dataset. This new remote sensing image caption dataset is called Sydney-Captions.

B. EVALUATION METRICS

In this paper, we utilize the pycocoevalcap tool[3] to compute the scores of the evaluation metrics. The pycocoevalcap module is able to improve the work efficiency and save much time for the image caption task. With the help of this tool, BLEU-n (range from 0 to 1), ROUGE_L (range from 0 to 1), METEOR (range from 0 to 1), CIDEr (range from 0 to 10) scores can be automatically computed. BLEU-n measures the co-occurrences of n-gram between the generated sentences and the ground truth captions. In this paper, n is set to from 1 to 4 [44]. ROUGE_L is designed for measuring the common subsequence with maximum length between the generated sentence and the reference sentence [44]. CIDEr measures the human consensus by adding a term frequency inverse document frequency weighting for every n-gram in the generated image sentence [39]. METEOR is computed by generating an alignment between the reference sentence and generated sentence [39].

The evaluation metrics for image captioning are so many that the performance of different models cannot be easily judged. Hence, referring to the equation (13) in AI

[3]https://github.com/tylin/coco-caption

TABLE 1. The scores of all metrics on UCM-Captions with different methods.

Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGH_L	CIDEr	S_m
CSMLF	0.4361	0.2728	0.1855	0.1210	0.1320	0.3927	0.2227	0.2171
Multimodal	0.7087	0.5966	0.5521	0.4599	0.3426	0.6612	2.9254	1.0973
Attention	0.8130	0.7430	0.6869	0.6369	0.4348	0.7634	3.2271	1.2656
VAA($\lambda = 1$)	**0.8192**	**0.7511**	**0.6927**	**0.6387**	**0.4380**	**0.7824**	**3.3946**	**1.3134**

TABLE 2. The scores of all metrics on Sydney-Captions with different methods.

Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGH_L	CIDEr	S_m
CSMLF	0.5998	0.4583	0.3869	0.3433	0.2475	0.5018	0.7555	0.4620
Multimodal	0.6966	0.6123	0.5431	0.5040	0.3588	0.6347	2.2022	0.9249
Attention	0.7258	0.6340	0.5665	0.5111	0.3714	0.6748	2.3659	0.9808
VAA($\lambda = 1$)	**0.7431**	**0.6646**	**0.6029**	**0.5495**	**0.3930**	**0.6999**	**2.4073**	**1.0124**

challenge 2017,[4] Mean score is beed defined as follows.

$$S_m = \frac{1}{4}\begin{pmatrix} BLEU_4 + METEOR \\ + ROUGE_L + CIDEr \end{pmatrix} \tag{13}$$

C. QUALITATIVE COMPARISON

Tables 1-2 show the results of remote sensing image captioning models on UCM-Captions and Sydney-Captions, respectively. There are big differences between these two datasets. UCM-Captions contains a greater number of images than Sydney-Captions, while the length of most of ground truth captions in the Sydney-Captions are longer than that in the UCM-Captions. As described above, the number of images in Sydney-Captions is 613 in total, while the UCM-Captions contains 2100 images. More importantly, the quantity between different categories is balanced in UCM-Captions and there are 100 images in each category, while the number of images in each category in Sydney-Captions is different. Airport category only contains 22 images, but residential category contains 242 images. The number of each category in Sydney-Captions is shown in Table 3. This unbalanced between categories can affect the training effect of models and the analysis of this phenomenon will be discussed below. Hence, it is normal to see that different scores of all metrics are obtained according to different datasets.

1) RESULTS ON UCM-CAPTIONS

In the experiment on UCM-Captions, we adopt the original way of dataset partition from the author [39]. 80% images are selected to be as the training dataset (1680 images), while 10% are used as the evaluation dataset (210 images) and the remaining 10% are used as the test dataset (210 images). Table 1 shows the results of seven scores on metrics by different methods, i.e., CSMLF method [45], multimodal method [38], attention-based method [39], Visual Aligning Attention model on the Sydney-Captions. In the table, the best scores are marked in bold. The scores of all metrics obtained by CSMLF method and Multimodal method are directly from [40].

From Table 1 we can see that the scores obtained by CSMLF method are the lowest. The multimodal method can get the average scores on all metrics, while attention-based method is able to improve the performance of models applied the multimodal method a little. It is noteworthy that the proposed Visual Aligning Attention model can obtain the best scores on all the metrics.

[4]https://challenger.ai/competition/caption

TABLE 3. The number of images in each category in Sydney-Captions.

Categories	Number
Residential	242
Airport	22
Meadow	50
River	45
Ocean	92
Industrial	96
Runway	66
Total	613

2) RESULTS ON SYDNEY-CAPTIONS

In the experiments on Sydney-Captions, we also adopt the original way of dataset partition from the author [39], while 80% images are chosen to be regarded as training dataset (497 images), 10% for evaluation dataset (58 images), and the remaining 10% for test dataset (58 images). Table 2 presents the results of all metrics by methods mentioned above on the Sydney-Captions. And the best scores are marked in bold as well.

From Table 2, it is easy to find that methods applied attention mechanisms can obtain better scores that those not applied attention mechanisms. More importantly,

Visual Aligning Attention model can achieve the best results than all of the other methods in the table.

3) RESULTS ANALYSIS

From Tables 1-2, it is easy to find that CSMLF method gets the worst performance on UCM-Captions and Sydney-Captions. This indicates that the encoder-decoder framework is effective for remote sensing image captioning and LSTM is good at generating semantically correct sentences. Additionally, scores of all metrics obtained by models applied attention mechanisms are much higher than CSMLF method. This is owing to the attention-based methods are capable of adaptively focusing on the regions in the remote sensing images and provide more purer visual information for LSTM to generate sentences. Specifically, the Visual Aligning Attention model can get higher scores than the conventional attention models. Hence, the conclusions can be drawn that the proposed method in this paper is able to describe the content of the images better.

In addition, it is easy to find that, generally, the scores in Table 1 are higher than those in Table 2. This is because that the unbalance between categories in Sydney-Captions will make models learn more information of image-caption pairs with higher frequency of occurrence. This will lead that the models cannot generate sentence description for unfrequent images. What's more, longer captions in Sydney-Captions also add a burden to training the decoder. All of these makes the scores obtained by models on Sydney-Captions are lower than those on UCM-Captions. Simultaneously, this also indicates that more image-caption pairs with balance between categories are important for training remote sensing image caption models.

D. QUANTITATIVE COMPARISON

Fig. 3 shows part of experimental results on UCM-Captions and Sydney-Captions generated by the conventional attention model and the proposed Visual Aligning Attention model in this paper. From the point of view of generated sentences, the sentences generated by the model proposed in this paper are more accurate on describing image categories and object numbers. What's more, the generated sentences contain more detail descriptions than those generated by the conventional attention models.

From Fig. 3, it is easy to see that, for the first image in the first row, the main object described in the sentence generated by the conventional attention model is wrongly distinguished. The real category of this image is farmland, while the attention model regard it as freeways. For the second image in the first row, there are many airplanes with the same size stopped at the airport. Well, the conventional attention model only detects one airplane. In addition, there are some buildings in the third image, however, the sentence generated by the conventional attention model only contains a storage tank in the right corner of the image. For the images in the second row, the proposed method is also able to detect more scene information in the images, such as road and water's color, while the conventional attention-based methods are not able to generate description for these details in remote sensing images. For the images in the third row, the sentences generated by the conventional attention models on Sydney-Captions are also easy to make mistakes, such as wrongly distinguishing categories and omitting some scene descriptions.

Above all, the proposed method in this paper can generate sentence with accurate descriptions on image categories and number of objects. What's more, the propose method can catch more detail information and describe more content of different scenes in images. Moreover, owing to the proposed method is able to acquiring more information from images, the generated sentences are generally longer than those generated by the conventional attention models.

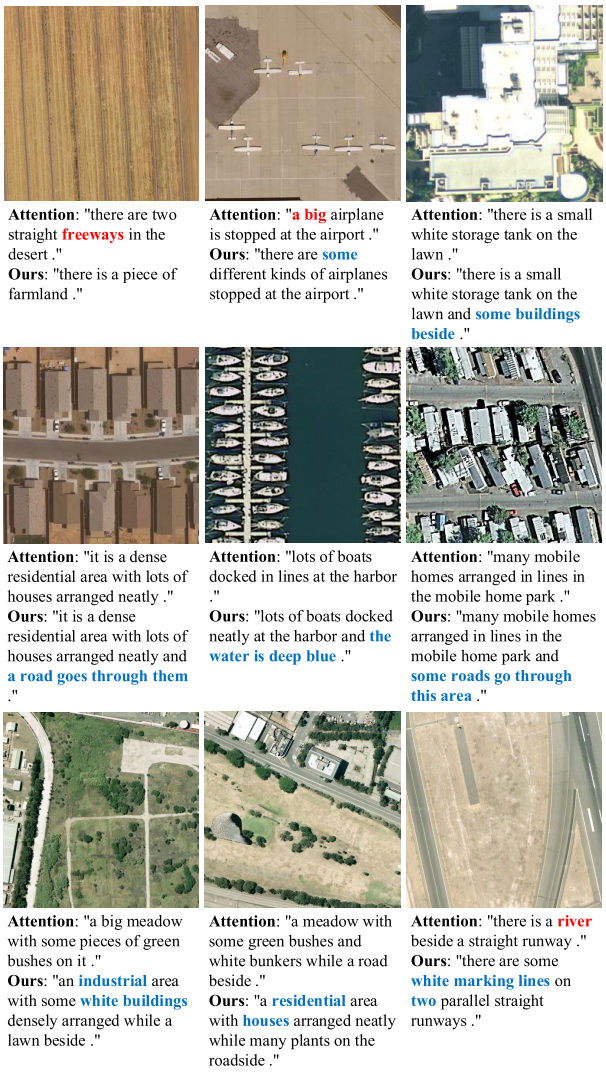


FIGURE 3. Images with generated sentences by attention models and VAA-models on UCM-Captions and Sydney-Captions. The first row and the second row are images in UCM-Captions. The third row are images in Sydney-Captions.

TABLE 4. The scores of all metrics on UCM-Captions with different value of the trade-off parameter.

Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGH_L	CIDEr	S_m
$L1$	0.8116	0.7454	0.6915	0.6428	0.4446	0.7741	3.3892	1.3127
$L2$	0.8150	0.7435	0.6852	0.6354	0.4446	0.7732	3.3757	1.3072
$cosine$	**0.8192**	**0.7511**	**0.6927**	**0.6387**	0.4380	**0.7824**	**3.3946**	**1.3134**

TABLE 5. The scores of all metrics on Sydney-Captions with different value of the trade-off parameter.

Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGH_L	CIDEr	S_m
$L1$	0.7389	0.6549	0.5924	0.5427	0.3832	0.6894	**2.4299**	1.0113
$L2$	0.7259	0.6373	0.5718	0.5186	0.3685	0.6739	2.3898	0.9877
$cosine$	**0.7431**	**0.6646**	**0.6029**	**0.5495**	**0.3930**	**0.6999**	2.4073	**1.0124**

TABLE 6. The scores of all metrics on UCM-Captions with different value of the trade-off parameter λ.

Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGH_L	CIDEr	S_m
$\lambda = 100$	0.7973	0.7190	0.6554	0.5987	0.4191	0.7634	3.0522	1.2084
$\lambda = 10$	0.7931	0.7168	0.6566	0.6031	0.4273	0.7507	3.0788	1.2150
$\lambda = 1$	**0.8192**	**0.7511**	**0.6927**	**0.6387**	**0.4380**	**0.7824**	**3.3946**	**1.3134**
$\lambda = 0.1$	0.8165	0.7454	0.6857	0.6308	0.4397	0.7854	3.2697	1.2814

TABLE 7. The scores of all metrics on Sydney-Captions with different value of the trade-off parameter λ.

Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGH_L	CIDEr	S_m
$\lambda = 100$	0.7184	0.6286	0.5585	0.5001	0.3664	0.6722	2.4473	0.9965
$\lambda = 10$	0.7271	0.6437	0.5873	0.5303	0.3740	0.6740	**2.4644**	1.0107
$\lambda = 1$	**0.7431**	**0.6646**	**0.6029**	**0.5495**	**0.3930**	**0.6999**	2.4073	**1.0124**
$\lambda = 0.1$	0.7118	0.6214	0.5576	0.5049	0.3599	0.6705	2.3903	0.9814

E. SIMILARITY FUNCTION CHOICES

Three different functions for calculating the feature similarity have been used in the Visual Aligning Attention model. The results obtained by these methods are shown in Table 4-5.

It is not hard to find that models utilizing the cosine function to measure the feature similarity between feature vectors can obtain the best scores on most of metrics in these three methods on UCM-Captions and Sydney-Captions. This proves that, compared with $L1$ function and $L2$ function, the cosine function is more able to measure the feature similarity between feature vectors. More importantly, the Visual Aligning Attention model adopting all these similarity functions can obtain higher scores than the conventional attention models.

F. TRADE-OFF PARAMETER ANALYSES

Table 6-7 show the results obtained by the Visual Aligning Attention model with different trade-off parameter settings. In these experiments, four values have been used to be as the value of λ. It is easy to find that, on both UCM-Captions and Sydney-Captions, best scores can be obtained by models with $\lambda = 1$.

If λ is set to 100, the total loss will be heavily leaning on the visual aligning loss. The original cross-entropy loss will be insignificant compared with the original cross-entropy loss and that's not feasible. The ability of generating grammatically correct sentences is more important, although the visual words is wanted to be present in the generated sentences. Increasing the value of λ for the visual aligning loss without limitation is not correct. And the experimental results on both datasets prove that larger value of λ will not bring better results. And if λ is set to 0.1, the visual aligning loss value is too small that it can almost be negligible. Hence, the scores obtained by VAA-models with $\lambda = 0.1$ will a little higher than those obtained by the conventional attention models. And results from Table 6-7 can prove that. In addition, if λ is set to 10, compared with the original cross-entropy loss's weight, the visual aligning loss's weight is a little too larger. The purpose of introducing the visual aligning loss is to improve the effect of attention process without affecting the original loss too much. Hence, the value of λ should be

close to the original cross-entropy loss's weight, otherwise, the process of generating sentences will be affected in a bad way.

V. DISCUSSION

The purpose of VAA is to utilize the visual aligning method to improve the attention masks' ability of focusing on regions of interest in input images. Thus, the proposed VAA is an extension work of the conventional attention model. However, there are some limitations in the proposed VAA model. The existing remote sensing datasets, UCM-Captions and Sydney-Captions, are initially used for scene classification. The difference between different categories of images are obvious. For example, the airport category of images mainly contains airplanes. But, if the given image contains many objects, such as airplanes, cars, buildings, and so on. The proposed VAA model maybe cannot generate sentences to describe all the objects in the image. Therefore, our future work may try to improve VAA model and make it more adaptive to complicated images.

VI. CONCLUSION

In this paper, a novel Visual Aligning Attention model is proposed for remote sensing image captioning task. This model is created to solve the problem of not explicitly training the attention layers. CNN is as the encoder to extract image features and LSTM is regarded as the decoder to generate sentences for describing the content of input images. It is noteworthy that, in this method, an approach of automatically filtering out non-visual words is proposed. What's more, a novel visual aligning loss is designed to explicitly constrain attention layers in the phase of training remote sensing image caption models. More importantly, the trained attention layers are able to focus on the regions more accurately and provide purer and more useful image information for the decoder to generate sentences for describing the content of input images. In addition, by doing a series of experiments on remote sensing image caption datasets, the conclusion can be draw that the proposed method in this paper can obtain higher scores on most of metrics than other methods in the experiments and describe the content of the input images well.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, vol. 141, no. 5, pp. 1097–1105.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 346–361.

[7] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 936–944.

[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 779–788.

[9] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[11] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2015, pp. 1440–1448.

[12] K. Fu, T. Zhang, Y. Zhang, M. Yan, Z. Chang, Z. Zhang, and X. Sun, "Meta-SSD: Towards fast adaptation for few-shot object detection with meta-learning," *IEEE Access*, vol. 7, pp. 77597–77606, 2019.

[13] P. O. Pinheiro, R. Collobert, and P. Dollar, "Learning to segment object candidates," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1990–1998.

[14] J. Dai, K. He, and J. Sun, "Instance-aware semantic segmentation via multi-task network cascades," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3150–3158.

[15] J. Dai, K. He, Y. Li, S. Ren, and J. Sun, "Instance-sensitive fully convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 534–549.

[16] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.

[17] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[18] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*. [Online]. Available: https://arxiv.org/abs/1706.05587

[19] X. Lu, X. Zheng, and X. Li, "Latent semantic minimal hashing for image retrieval," *IEEE Trans. Image Process.*, vol. 26, no. 1, pp. 355–368, Jan. 2017.

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

[21] 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.

[22] V. Ordonez, G. Kulkarni, and T. L. Berg, "Im2Text: Describing images using 1 million captioned photographs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 1143–1151.

[23] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *J. Artif. Intell. Res.*, vol. 47, no. 1, pp. 853–899, 2013.

[24] C. Sun, C. Gan, and R. Nevatia, "Automatic concept discovery from parallel text and visual corpora," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2015, pp. 2596–2604.

[25] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik, "Improving image-sentence embeddings using large weakly annotated photo collections," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 529–545.

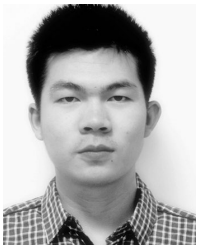
[26] V. Ordonez, X. Han, P. Kuznetsova, G. Kulkarni, M. Mitchell, K. Yamaguchi, K. Stratos, A. Goyal, J. Dodge, A. Mensch, H. Daumé, III, A. C. Berg, Y. Choi, and T. L. Berg, "Large scale retrieval and generation of image descriptions," *Int. J. Comput. Vis.*, vol. 119, no. 1, pp. 46–59, 2016.

[27] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, "Every picture tells a story: Generating sentences from images," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 15–29.

[28] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, "Baby talk: Understanding and generating simple image descriptions," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 1601–1608.

[29] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[30] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3242–3250.
[31] S. Chen and Q. Zhao, "Boosted attention: Leveraging human attention for image captioning," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 72–88.
[32] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-RNN)," 2015, *arXiv:1412.6632*. [Online]. Available: https://arxiv.org/abs/1412.6632/#
[33] 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.
[34] X. Chen and C. L. Zitnick, "Learning a recurrent visual representation for image caption generation," 2014, *arXiv:1411.5654*. [Online]. Available: https://arxiv.org/abs/1411.5654
[35] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollar, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick, and G. Zweig, "From captions to visual concepts and back," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, May 2015, pp. 1473–1482.
[36] A. Karpathy, A. Joulin, and L. Fei-Fei, "Deep fragment embeddings for bidirectional image sentence mapping," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1889–1897.
[37] 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.*, Jun. 2015, pp. 3156–3164.
[38] 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.*, Jul. 2016, pp. 1–5.
[39] 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.
[40] 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.
[41] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[42] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst.*, 2010, pp. 270–279.
[43] 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.
[44] L. Li, S. Tang, L. Deng, Y. Zhang, and Q. Tian, "Image caption with global-local attention," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 4133–4139.
[45] 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.



ZHENGYUAN ZHANG received the B.Sc. degree from Harbin Engineering University, Harbin, China, in 2016. He is currently pursuing the Ph.D. degree with the Institute of Electronics, Chinese Academy of Sciences, Beijing, China. His research interests include computer vision, pattern recognition, and remote sensing image processing, especially on image captioning.

WENKAI ZHANG received the B.Sc. degree from the China University of Petroleum, Shandong, China, in 2013, and the Ph.D. degree from the Institute of Electronics, Chinese Academy of Sciences, Beijing, in 2018, where he is currently a Research Assistant. His research interests include remote sensing image semantic segmentation and multi-media information processing.

WENHUI DIAO received the B.Sc. degree from Xidian University, Xi'an, China, in 2011, and the M.Sc. and Ph.D. degrees from the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, in 2016, where he is currently an Assistant Professor. His research interests include computer vision and remote sensing image analysis.

MENGLONG YAN received the B.Sc. degree from Wuhan University, Wuhan, China, in 2007, and the M.Sc. and Ph.D. degrees from Peking University, Beijing, China, in 2012. He is currently an Associate Professor with the Institute of Electronics, Chinese Academy of Sciences, Beijing. His research interests include LiDAR data processing and high-resolution remote sensing image processing.

XIN GAO received the Ph.D. degree from Beijing Normal University. He is currently a Professor with the Institute of Electronics, Chinese Academy of Sciences, Beijing, China. He has published more than 30 professional articles. His research interests include scene classification in SAR imagery, object detection and recognition, and interpretation and annotating of remote sensing imagery.

XIAN SUN received the B.Sc. degree from Beihang University, Beijing, China, in 2004, and the M.Sc. and Ph.D. degrees from the Institute of Electronics, Chinese Academy of Sciences, Beijing, in 2006 and 2009, respectively, where he is currently a Professor. His research interests include computer vision and remote sensing image understanding.