GLCM: Global-Local Captioning Model for Remote Sensing Image Captioning

Qi Wang, *Senior Member, IEEE*, Wei Huang, Xueting Zhang, and Xuelong Li, *Fellow, IEEE*

***Abstract*—Remote sensing image captioning (RSIC), which describes a remote sensing image with a semantically related sentence, has been a cross-modal challenge between computer vision and natural language processing. For visual features extracted from remote sensing images, global features provide the complete and comprehensive visual relevance of all the words of a sentence simultaneously, while local features can emphasize the discrimination of these words individually. Therefore not only global features are important for caption generation, but also local features are meaningful for making the words more discriminative. In order to make full use of the advantages of both global and local features, in this paper, we propose an attention-based Global-Local Captioning Model (GLCM) to obtain global-local visual feature representation for RSIC. Based on the proposed GLCM, the correlation of all the generated words and the relation of each separate word and the most related local visual features can be visualized in a similarity-based manner, which provides more interpretability for RSIC. In the extensive experiments, our method achieves comparable results in UCM-captions and superior results in Sydney-captions and RSICD which is the largest remote sensing image captioning dataset.**

***Index Terms*—remote sensing, image captioning, deep learning, Global-Local Captioning Model**

I. Introduction

NOWADYS, there are many applications based on optical remote sensing images, including scene classification [1]–[3], change detection [4]–[6], object detection [6]–[8], semantic segmentation [9], and geographical image retrieval [10], [11]. However, these tasks mainly concentrate on scene labels and object locations, without the exploration of their semantic relationship. To further study the structural relation of features, objects, and scenes in remote sensing images, researchers have put more attention on remote sensing image captioning (RSIC) [12]–[17], which attempts to describe the content of a remote sensing image with a semantic-relevant sentence.

In RSIC, the mainstream methods obey a common encoder-decoder framework, where the encoder extracts visual features from a given remote sensing image and the decoder generates a corresponding sentence based on the extracted features. For visual features, it can be divided into global and local features according to the receptive field it covers. As shown in Fig. 1, global features can express the complete and comprehensive relevance of the features and objects in the remote sensing images, while local features can make each word of the generated caption more discriminative and flexible. Both of them are beneficial for improving the quality of captions. For the existing RSIC methods, however, usually only global features are considered with the neglect of separated local features.

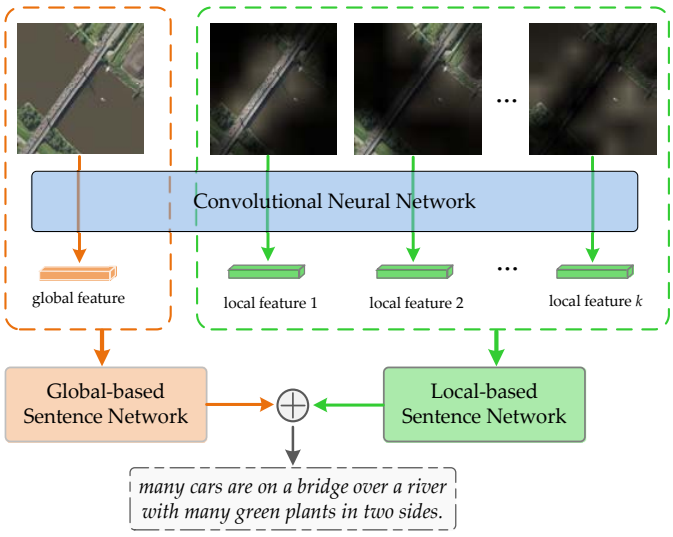


Fig. 1. Illustration of global and local features extraction and fusion for RSIC. In this case, the global features contain the entire structural relationship of the features and objects, and the local features are more independent and pure for their corresponding words. For local features, taking 'local feature 1' as an example, it pays more attention to the words 'cars' and 'bridge' and it can help these words to get more discriminative representation.

In order to make full use of the advantages of both global and local visual features simultaneously, in this paper, we propose an attention-based Global-Local Captioning Model (GLCM) to learn joint global and local visual feature representation for RSIC in an end-to-end manner. The proposed GLCM is in accordance with an encoder-decoder architecture, but several innovations are presented in GLCM for the first time for RSIC. Firstly, we design a simple but effective local feature aggregation module, which can be easily embedded into the encoder, *i.e.*, *convolutional neural network* (CNN) extractor, and be used to obtain multiple local visual features from a remote sensing image. Secondly, Self-Attention is first designed as the decoder to generate some RSIC captions based on global visual features, which plays the same role as *recurrent neural network* (RNN) and *long short-term memory*

This work was supported by the National Natural Science Foundation of China under Grant U21B2041, U1864204.

Q. Wang, W. Huang, X. Zhang and X. Li are with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, Shaanxi, China (e-mail: crabwq@gmail.com, hw2hwei@gmail.com, zxt@mail.nwpu.edu.cn, li@nwpu.edu.cn).

X. Li is the corresponding author.

(LSTM) [18]. Compared with RNN/LSTM which predicts a word w_t at time step s_t only utilizing hidden state h_{t-1} and a single previous word x_{t-1}, Self-Attention is beyond the limitation of time steps and can consider all the generated words $[x_1, ..., x_{t-1}]$ at the same time [19]. Thirdly, Co-Attention is developed to separately enhance the discrimination of all the generated words with the introduction of both global and local features. Hence when predicting the current word w_t in GLCM, the global visual information, multiple local visual features, and all the generated words will be together taken into consideration in a reasonable and effective manner.

In order to verify the effectiveness of the proposed GLCM, some extensive ablation experiments are performed on three widely used RSIC data sets including Sydney-captions [14], UCM-captions [14] and RSICD [12]. In comparison with some state-of-the-art methods, GLCM achieves comparable results in UCM-captions and superior results in Sydney-captions and RSICD. Besides, based on Self-Attention and Co-Attention, the similarity degree among words beyond the limitation of time steps is quantitatively visualized and the most matched pairs of words and local visual features are provided.

In general, the main contributions of this paper could be summarized as the following three aspects:

(1) In this paper, we propose an attention-based Global-Local Captioning Model (GLCM), an end-to-end trainable architecture, to learn joint global and local visual feature representation for RSIC. The learned global-local features contain not only the complete and comprehensive relation of features and objects but also the more independent and discriminative visual information for each word when generating captions.
(2) Based on GLCM, the relation between words and words, words and local visual features are visualized in an attention manner, which provides more interpretability for RSIC.
(3) The proposed GLCM achieves excellent experimental results in RSIC data sets, including superior results in Sydney-captions and RSICD, and comparable results in UCM-captions. Our source code is available at https://github.com/hw2hwei/GLCM.

II. Related Work

A. Remote Sensing Image Captioning (RSIC)

In the field of RSIC, the existing methods obey the general architecture of encoder-decoder framework [20] composing of two components of encoder and decoder. For the encoder, it is used to detect and recognize the visual features and objects from remote sensing images and store them in a feature vector. The encoder consists of hand-crafted feature extractor such as *Fisher Vector* (FV) [21], *Scale-Invariant Feature Transform* (SIFT) [22], *Bag Of Words* (BOW) [23] and *Vector of Locally Aggregated Descriptors* (VLAD) [24], and deep feature extractor that usually refers to CNN backbones like GoogLeNet [25], VGG [26], ResNet [27]. For the decoder, it needs to generate semantically relevant sentences to describe the images according to the features extracted by the encoder. To achieve this goal, the decoder is realized by the sequential model of RNN or LSTM in the existing methods of RSIC.

Researchers have proposed many methods to improve the quality of RSIC from different views. Inspired by *Natural Image Captioning* (NIC) [28]–[35], Qu *et al.* [14] first introduced CNN + RNN/LSTM mechanism in RSIC, publishing two data sets of UCM-captions and Sydney-captions. Lu *et al.* released the largest data set of RSICD in [12] and attempted different visual features including hand-crafted features and deep features using soft- and hard-attention mechanisms to further improve the caption quality. Following that, Zhang *et al.* [15] introduced a multi-scale image cropping and training mechanism to extract multi-scale visual features, which is a kind of data augmentation strategy. Besides, Zhang *et al.* [36] embedded an attribute attention mechanism into an encoder-decoder model to explore the impact of the attributes of remote sensing images. Recently, Zhang *et al.* [37] utilized visual aligning attention to enhance location accuracy of the interesting regions of remote sensing images. To handle the inconsistency of descriptions from different observers, Lu *et al.* [38] presented a sound activation attention framework that generates captions according to the interest of the observer. In order to deal with the dramatic scale variation of objects and explore the visual relationship of remote sensing images, Yuan *et al.* [16] proposed a multi-level attention and multi-label attribute graph convolution based framework.

B. Region-Based Image Captioning

In natural image captioning (NIC) [39], [40], region-based methods are continuously studied for more interpretable and accurate descriptions. Karpathy *et al.* [41] proposed an alignment model to generate some sentence snippets for the corresponding regions extracted by Region Convolutional Neural Network (RCNN) from a natural image. To both localize and describe regions in natural images in natural language, Justin *et al.* [42] presented a Fully Convolutional Localization Network (FCLN), which requires no external regions proposals, to process an image with a single forward pass. Ting *et al.* [32] utilized the weakly-supervised approach of Multiple Instance Learning (MIL) to detect attributes and integrate these attributes into CNN+RNN architecture in an end-to-end manner. And Li *et al.* [43] used Faster RCNN to extract region features as local features and take them as part of the input of LSTM. Besides, Marcella *et al.* [44] treated a sequence or set of image regions as a control signal to predict textual chunks with the explicit ground regions. How the features and objects of remote sensing images, such as *river* and *building*, usually are irregular and massively dispersed. Thus, in this paper, we propose an attention method to aggregate similar local features instead of region location methods.

The idea of combining global and local features has been also proposed for image captioning in global-local attention (GLA) [43]. However, there are two main differences between [43] and the proposed GLCM. Firstly, GLA uses LSTM as the decoder to generate the caption which can only absorb information from the single previous word and hidden state while the proposed GLCM use self-attention that can utilize

all the generated words simultaneously beyond the limitation of time steps. Therefore, the proposed GLCM can achieve more flexible and explicable word prediction at each time step in such a one-to-many manner. Secondly, GLA has two visual feature extractors of image feature extractor (VGG16) and object feature extractor (Faster R-CNN) to extract global and local features respectively, while the proposed GLCM use the shared CNN backbone and two heads of global and local attention modules to extract global and local features simultaneously. In view of this, compared with GLA, the proposed GLCM has three points of advantages for RSIC:

(a) The object feature extractor of GLA needs to be pre-trained on an object detection data set of remote sensing images. Nevertheless, there are not as sufficient and complete object annotations in remote sensing images as in natural images. GLCM has no demand for it.
(b) Different from natural images, remote sensing images cover a large number of multi-scale irregular and discrete objects that belong to scene-level features. It is not suitable to use an object detector to detect all of these objects for RSIC. Instead, GLCM utilizes multiple local attention modules to gather similar feature vectors so as to form multiple local visual features from bottom to top, without the need for remote sensing object annotation.
(c) Two CNN streams of GLA have much more parameters and higher computation complexity than the shared CNN backbone of GLCM.

C. Attention Mechanism

The attention mechanism is an effective strategy in both Computer Vision (CV) and Natural Language Processing (NLP). As a cross-modal task between CV and NLP, image captioning would be beneficial from the progress of attention technologies in the fields of both CV and NLP.

In CV, attention mechanism is used to select the most valuable or relevant parts of an image and aggregate their features for more accurate classification or location. According to the receptive field it covers, it could be divided into soft-attention [28], [45] and hard-attention [28], [46]. Soft-attention considers the features of all the parts of the image and gives them different weights, while hard-attention only utilizes the feature of the most relevant part. Besides spatial parts, visual attention can also operate in feature channels [47]–[49] and time sequence [43], [50].

Attention mechanism is also widely used in NLP [51], [52], which is utilized to match the target words or documents with the closed source words or documents to pick up the most effective information. Similar to soft- and hard-attention in CV, attention mechanism in NLP can be divided into global-attention and local-attention [53]. To further achieve multi-level language understanding, researchers proposed all kinds of attention variants such as hierarchical attention [54], attention over attention [55] and multi-step attention [56]. There is a popular attention variant of self-attention [19], [57] with achieving outstanding results. Different from the other types of attention whose key and query are from different sources, key and query of self-attention are from the same source. Based on self-attention, Transformer has became popular with the sequential generation tasks [19], [58].

For the task of video captioning [58], its input of temporal images and its output of shifted sentences are both sequential data that are born for the application of Transformer. Different from video captioning, the visual input of image captioning is usually a feature map or vector, directly applying Transformer to RSIC is not quite reasonable because of the large number of multi-scale and dispersed objects in remote sensing images. In view of this, the proposed GLCM focuses on integrating the related feature vectors from the visual feature map to multiple relatively independent local visual features, which can enhance the discrimination of each word in the caption.

III. Global-Local Captioning Model

The workflow of the proposed Global-Local Captioning Model (GLCM) is shown in Fig. 2. The whole model consists of two sub-networks of a global-local feature encoding network and an attention-based decoding network, which conforms to the encoder-decoder framework. In the global-local feature encoding network, high-level visual semantic feature maps are first extracted from remote sensing images. Then a global feature vector and $N\times$ local feature vectors are further obtained from the feature map. In the attention-based decoding network, different combinations of Self-Attention and Co-Attention layers followed by a word classifier are used to explore the roles of global and local features in caption generation. In this section, global-local feature encoding network and attention-based decoding network are introduced in detail respectively.

A. Global-Local Feature Encoding Network

Image representation, which encodes the valuable features and objects of a given remote sensing image into a high-level multi-channel feature map, is crucial for understanding images. High-quality image representation is beneficial for extracting more effective global and local features and further improving caption performance.

*1) **CNN-based Feature Representation**:* Due to the powerful feature extraction ability, deep feature extracted by CNN is used as the representation of remote sensing images for the following caption generation. However, there are all kinds of CNNs proposed for different computer vision tasks. Among them, GoogLeNet [25] is a typical CNN architecture and it is good at extracting multi-scale features which is quite suitable for remote sensing images. Thus, the backbone of GoogLeNet denoted as $GooLeNet_{inception5b}$ where $inception5b$ is the last convolutional layer, is chosen as the CNN feature extractor in this paper. A remote sensing image is encoded into a high-dimension semantic feature map with the size of $H_g\times W_g\times C_g$ denoted as $\bar{F}\in\mathbb{R}^{H_g\times W_g\times C_g}$ by $GooLeNet_{inception5b}$, which is formulated as:

$$\bar{F}=GooLeNet_{inception5b}(I), \tag{1}$$

where $I\in\mathbb{R}^{H_i\times W_i\times C_i}$ refers to the input remote sensing image with the size of $H_i\times W_i\times C_i$.

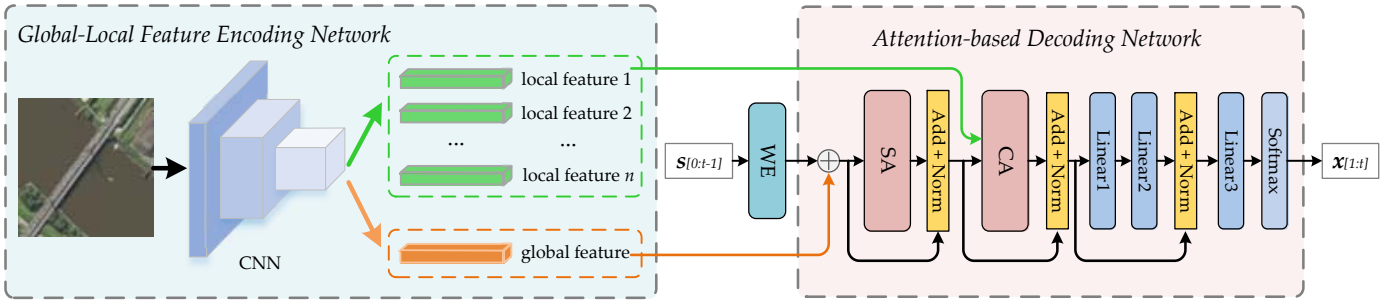


Fig. 2. The workflow of the proposed Global-Local Captioning Model (GLCM) for RSIC. WE: Word Embedding; SA: Self-Attention; CA: Co-Attention.

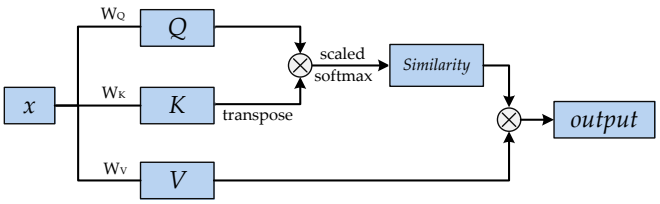


Fig. 3. Illustration of Self-Attention (Co-Attention).

To reduce model parameters, the channel dimension of $\bar{F}$ is reduced from C_g (C_g=1024 in GoogLeNet) to C, which is realized by:

$$F = \bar{F}W^F, \tag{2}$$

where $W^F \in \mathbb{R}^{C_G \times C}$ is the feature projection weight.

*2) **Global and Local Feature Extraction**:* Based on F_M, global feature vector denoted as $\boldsymbol{g} \in \mathbb{R}^C$ and multiple local feature vectors denoted as $\boldsymbol{l} \in \mathbb{R}^{N \times C}$ can be extracted respectively. N is the number of local feature vectors.

For global features, they should have the characteristics of invariant translation and rotation. To achieve it, Global Average Pooling (GAP) [59] layer is used to encode the feature map F into the global feature vector $\boldsymbol{g}$. The k-th element of $\boldsymbol{g}$ is calculated by:

$$\boldsymbol{g}(k) = \frac{1}{H \times W} \sum_{i=0}^{H} \sum_{j=0}^{W} F_{[i,j,k]}. \tag{3}$$

For local features, different local features need to concentrate on different parts of remote sensing images. To achieve it, $N\times$ soft attention modules are operated on F to generate multiple local feature vectors $\boldsymbol{l}$. It is the prerequisite to build $N\times$ attention weighting maps with the spatial size of $H \times W$, which is denoted as $\alpha \in \mathbb{R}^{N \times H \times W}$. The weighting coefficient at the location of (i, j) in the n-th attention map is calculated by:

$$\bar{\alpha}_{[n,i,j]} = \frac{1}{C} \sum_{k=0}^{C} F_{[i,j,k]} * w_{[n,k]}, \tag{4}$$

$$\alpha_{[n,i,j]} = \frac{exp(\bar{\alpha}_{[n,i,j]})}{\sum_{i=0}^{H} \sum_{j=0}^{W} exp(\bar{\alpha}_{[n,i,j]})}, \tag{5}$$

where $\bar{\alpha} \in \mathbb{R}^{N \times H \times W}$ and $\alpha \in \mathbb{R}^{N \times H \times W}$ are the intermediate and scaled weighting maps, respectively. In Eqn. (4), $w \in \mathbb{R}^{N \times C}$ projects $C\times$ feature channels into $N\times$ attention channels. And then $\bar{\alpha}_{[n,i,j]}$ is scaled by Softmax operation as in Eqn. (5).

Similar to global feature extraction, the GAP operation is also used to aggregate the local feature maps into multiple local feature vectors. The k-th element in the n-th vector of $\boldsymbol{l}$ is calculated by:

$$\boldsymbol{l}_{[n,k]} = \frac{1}{H \times W} \sum_{i=0}^{H} \sum_{j=0}^{W} F_{[i,j,k]} * \alpha_{[n,i,j]}. \tag{6}$$

B. *Attention-based Decoding Network*

After extracting global and local features from images, it needs to decode these visual features into a content-relevant sentence using the sequential model. Traditionally, RNN and LSTM are used as the sequential model. When predicting the next word, however, RNN and LSTM rely on the single previous word and the hidden state, which means that they cannot consider all the generated words simultaneously. To solve this problem, the attention mechanism is applied in NLP as described in Section II-C. Following [19], [60], Self-Attention is used to build information passageway between words while Co-Attention is used to connect words and local features. Before applying Self-Attention and Co-Attention, there is some preparation to do in advance.

Given the generated sentence by time step t denoted as $\boldsymbol{s} = [s_0, ..., s_{t-1}]$, the next word s_t would be predicted according to $\boldsymbol{s}$. Here i-th word of $s_i \in \mathbb{R}^1$ refers to the word number in the vocabulary and s_0 denotes the start symbol of the sentence. In order to adapt to network structure, the word numbers of the generated sentence $\boldsymbol{s}$ need to be embedded into sentence vectors $\boldsymbol{x} \in \mathbb{R}^{t \times C} = [x_0, ..., x_{t-1}]$ by word2vec technique [61], which is formulated as:

$$\boldsymbol{x} = Embedding(\boldsymbol{s}), \tag{7}$$

here it is worth mentioning that all the lengths of $\boldsymbol{x}$ of different time steps are expanded to the max sequence length T by

adding $(T-t)$ ending characters to it for parallel training, and therefore the dimension of the expanded $\boldsymbol{x}$ becomes $R^{T\times C}$.

If global features $\boldsymbol{g}$ are used, they would be incorporated into $\boldsymbol{x}$ by:

$$\boldsymbol{x} = \boldsymbol{x} + \boldsymbol{g}. \tag{8}$$

*1) **Self-Attention**:* After finishing the preparation of Eqn. (7)-(8) , Self-Attention [19] is used to pass information among the word vectors $\boldsymbol{x}$ as shown in Fig. 3, which is formulated as:

$$Q_S = \boldsymbol{x}W_S^Q, K_S = \boldsymbol{x}W_S^K, V_S = \boldsymbol{x}W_S^V, \tag{9}$$

$$\boldsymbol{x} = \boldsymbol{x} + softmax(\frac{Q_S {K_S}^T}{\sqrt{C}})V_S, \tag{10}$$

$$\boldsymbol{x} = LayerNorm(\boldsymbol{x}), \tag{11}$$

where Q_S, K_S and V_S (all $\in \mathbb{R}^{T\times C}$) refer to queries, keys and values and all of them are vectors. W_S^Q, W_S^K and W_S^V (all $\in \mathbb{R}^{C\times C}$) are the projection matrices. In Eqn. (10), scaled dot-product is used to calculate attention coefficient distribution among word vectors $\boldsymbol{x}$, and skip-connection is applied to improve the stability of the network. At time step $t-1$, information of all the generated word vectors [x_1,..., x_{t-1}] would be passed to the current word vector x_{t-1} in a proportion of their similarity coefficients. $LayerNorm$ in Eqn. (11) is a normalization layer. It is worth mentioning that Self-Attention is not limited by the length of $\boldsymbol{x}$ thanks to the dot-product operation.

*2) **Co-Attention**:* Following Self-Attention, Co-Attention [19], [60] is used to build a bridge between local features $\boldsymbol{l}$ and word vectors $\boldsymbol{x}$. As shown in Fig. 3, Co-Attention has the same architecture as Self-Attention but their inputs are different. Here, Co-Attention is formulated as:

$$Q_C = \boldsymbol{x}W_C^Q, K_C = \boldsymbol{l}W_C^K, V_C = \boldsymbol{l}W_C^V, \tag{12}$$

$$\boldsymbol{x} = \boldsymbol{x} + softmax(\frac{Q_C {K_C}^T}{\sqrt{C}})V_C,, \tag{13}$$

$$\boldsymbol{x} = LayerNorm(\boldsymbol{x}), \tag{14}$$

where Q_C and K_C (both $\in \mathbb{R}^{N\times C}$) are key-value pairs projected from $\boldsymbol{l}$ while $V_C \in \mathbb{R}^{T\times C}$ is values projected from $\boldsymbol{x}$. In Eqn. (13), information of all the local features $\boldsymbol{l}$ are sent to each word vector in $\boldsymbol{x}$ according to their similarity coefficients given by scaled dot-product operation.

*3) **Word Prediction**:* At time step t, information of [x_0, ..., x_{t-1}] are aggregated into x_{t-1} and x_{t-1} is further enhanced by global and local features, which are realized by Self-Attention and Co-Attention as described above. In image captioning, the words of sentences are predicted step by step. In this paper, it is to use the enhanced x_{t-1} to predict the next word x_t. Following [19], the word prediction is realized by:

$$\boldsymbol{x} = \boldsymbol{x} + ReLU(\boldsymbol{x}W_F^1)W_F^2, \tag{15}$$

$$\boldsymbol{x} = LayerNorm(\boldsymbol{x}), \tag{16}$$

$$\boldsymbol{x} = softmax(\boldsymbol{x}W_F^3), \tag{17}$$

where $W_F^1 \in \mathbb{R}^{C\times 2C}$ projects $\boldsymbol{x}$ into the feature space of $2\times$ dimension, and $W_F^1 \in \mathbb{R}^{2C\times C}$ restores its dimension. ReLU is a non-linear activation function. $W_F^3 \in \mathbb{R}^{C\times K}$ plays the role of classifier predicting the word number of x_t according to x_{t-1}, where K is the vocabulary size of data set. Eqn. (15)-(17) output the word probability distribution of $\boldsymbol{x} \in \mathbb{R}^{T\times K}$.

Finally, cross-entropy loss is used to optimize the whole model, which is defined as:

$$L(\theta) = -\sum_{t=1}^{T} log(p_\theta(x_t|s_0^*, ..., s_{t-1}^*)), \tag{18}$$

where θ denotes the parameters of the whole model and [s_0^*, ..., s_{t-1}^*] is the label sequence.

*4) **Ablation Combination of Self-Attention and Co-Attention**:* To independently explore the effect of global features $\boldsymbol{g}$ and local features $\boldsymbol{l}$, as shown in Fig. V, four different combinations of Self-Attention and Co-Attention are designed for ablation study: *(1) Local*. To study the effect of local features $\boldsymbol{l}$, global features $\boldsymbol{g}$ is not added to word vectors $\boldsymbol{x}$ as in Eqn. (8) while only $\boldsymbol{l}$ is utilized by Co-Attention; *(2) Global_A*. To study the effect of $\boldsymbol{g}$, only $\boldsymbol{g}$ is added to $\boldsymbol{x}$ while l is not utilized; *(3) Global_B*. To study the impact of stacking Self-Attention, one more Self-Attention layer is added to the first one in *Global_A*; *(4) Global_Local*. In order to study the influence of both $\boldsymbol{g}$ and $\boldsymbol{l}$, $\boldsymbol{g}$ is added to $\boldsymbol{x}$ in Eqn. (8) and $\boldsymbol{l}$ is incorporated into $\boldsymbol{x}$ by Co-Attention. These four combinations can be regarded as variations of GLCM.

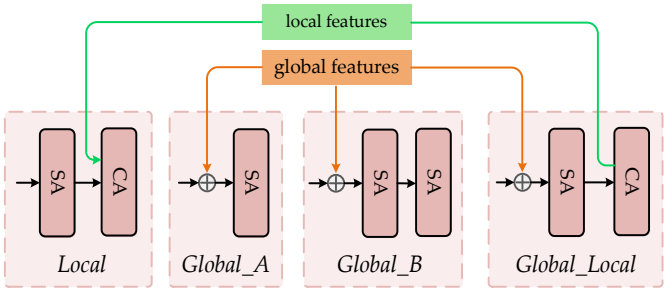


Fig. 4. Illustration of four variations of GLCM. In the figure, only Self-Attention and Co-Attention are shown and the other components are ignored. There are two attention layers of Self-Attention and Self-Attention in *Local* and *Global_Local*, and there is only a single layer in *Global_A*, and there are two Self-Attention layers of the same architecture in *Global_B*. SA: Self-Attention; CA: Co-Attention.

IV. Experiments

In this section, ablation and comparison experiments on three widely used data sets are performed to verify the effectiveness of the proposed GLCM for RSIC. First of all, data sets and evaluation metrics used in this paper are introduced. Secondly, experimental details and parameter settings are provided. Following that, ablation experiments based on four variations of GLCM are performed to study the impact of global and local features in RSIC. In the end, our method is compared with some existing state-of-the-art methods with detailed analysis.

A. Data Sets and Evaluation Metrics

1) Data Sets: In the field of remote sensing image captioning, three public data sets are widely used as follows:

TABLE I
RESULTS OF ABLATION STUDY OF DIFFERENT CNN BACKBONES COMBINED WITH $Global\_A$.

Dataset	Models	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE_L	CIDEr
Sydney-captions	ResNet18	78.61	71.60	65.33	59.72	41.23	67.66	209.90
	VGG16	79.98	72.67	66.15	61.31	42.89	70.27	228.92
	GoogLeNet	79.09	72.21	65.46	59.83	42.55	69.50	210.12
UCM-captions	ResNet18	79.94	72.64	66.92	61.83	43.36	70.86	270.08
	VGG16	77.71	70.11	63.84	58.23	43.24	70.80	264.43
	GoogLeNet	80.77	74.48	69.30	64.46	45.78	73.56	288.48
RSICD	ResNet18	75.04	62.51	53.44	46.50	34.32	65.50	238.75
	VGG16	75.67	62.55	53.03	45.77	34.39	65.48	233.51
	GoogLeNet	75.88	63.23	53.75	46.40	35.43	67.32	244.83

TABLE II
RESULTS ON THREE DATA SETS USING DIFFERENT VARIANTS OF GLCM BASED CNN BACKBONE OF GOOGLENET. THE HIGHER THESE SCORES ARE, THE BETTER THEIR PERFORMANCE IS.

Data Set	Model	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE_L	CIDEr
Sydney-captions	*Local*	77.83	72.12	**67.59**	**63.25**	42.91	67.96	223.14
	Global_A	79.09	72.21	65.46	59.83	42.55	69.50	210.12
	Global_B	**81.66**	**74.59**	67.23	60.84	42.05	**70.36**	208.50
	Global_Local	80.41	73.05	67.45	62.59	**44.21**	69.65	**243.37**
UCM-captions	*Local*	80.46	73.63	67.93	62.89	44.81	73.77	283.95
	Global_A	80.77	74.48	69.30	64.46	45.78	73.56	288.48
	Global_B	80.88	74.66	69.32	**64.83**	45.54	73.92	277.70
	Global_Local	**81.82**	**75.40**	**69.86**	64.68	**46.19**	**75.24**	**302.79**
RSICD	*Local*	77.52	64.67	55.56	48.76	35.83	67.27	247.38
	Global_A	75.88	63.23	53.75	46.40	35.43	67.32	244.83
	Global_B	76.74	64.84	56.13	49.34	35.25	67.16	252.95
	Global_Local	**77.67**	**64.92**	**56.42**	**49.37**	**36.27**	**67.79**	**254.91**

(a) ***Sydney-captions***. Sydney-Captions is developed in [14] based on Sydney Data set [65], which is used for remote sensing scene classification. Sydney Data set includes 613 images of seven classes composed of residential, airport, meadow, rivers, ocean, industrial, and runway. All the images are manually cropped from the image of Sydney that is downloaded on Google Earth, and they have the same size of 500 × 500 with a pixel resolution of 0.5 m.

(b) ***UCM-captions*.** UCM-captions [14] is also the secondary development of a remote sensing scene classification data set of UC Merced (UCM) Land Used data [66]. UCM data set has 2,100 images of 21 classical scene classes including agricultural, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium-density residential, mobile home park, overpass, parking lot, river, sparse residential, storage tanks, and tennis courts. Each class contains 100 images of the 256 × 256 size with a pixel resolution of 0.3048 m. And it is extracted from the United States Geological Survey, which is also downloaded on Google Earth.

(c) ***RSICD*.** RSICD is currently the largest RSIC data set provided by [12], containing a total of 10,921 remote sensing images of various geographical areas. There are 30 kinds of scenes in this data set including airport, bridge, beach, baseball field, open land, commercial, center, church, desert, dense residential, forest, farmland, industrial, mountain, medium residential, meadow, port, pond, parking, park, playground, river, railway station, resort, storage tanks, stadium, sparse residential, square, school, and viaduct. Similarly, images in this data set also come from Google Earth but have a size of 224 × 224 with various pixel resolutions.

For each image in all the above data sets, there are five sentences collected from several observers to describe it. In each of these three data sets, there are training, validation, and test sets, which are provided by their original literature. However, because their test sets are biased with unstable performance, their training and validation sets are used for training and test in this work, respectively.

2) Evaluation Metrics: Due to the huge flexibility of captions, it is important to comprehensively evaluate the caption

TABLE III
COMPARISON OF SOME STATE-OF-THE-ART METHODS ON SYDNEY-CAPTIONS.

Models	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE_L	CIDEr
FV-LSTM [12]	63.31	53.33	47.35	43.03	29.67	57.94	147.61
Attention-based (Hard) [12]	73.22	66.74	62.23	58.20	39.42	71.27	**249.93**
Sound-f-a [38]	71.55	63.23	54.69	46.60	31.32	60.35	180.27
VAA [37]	74.31	66.46	60.29	54.95	39.30	69.99	240.73
SM-ATT+LSTM [36]	81.43	73.51	65.86	58.06	41.11	71.95	230.21
ML_Attention+Attribute_GCN [16]	**82.33**	**75.48**	65.87	60.03	42.02	**72.37**	231.10
TCE [62]	79.37	73.04	67.17	61.93	**44.30**	71.30	240.42
SVM-D CONC [63]	75.47	67.11	59.70	53.08	36.43	67.46	222.22
The Proposed GLCM	80.41	73.05	**67.45**	**62.59**	44.21	69.65	243.37

TABLE IV
COMPARISON OF SOME STATE-OF-THE-ART METHODS ON UCM-CAPTIONS.

Models	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE_L	CIDEr
VLAD-LSTM [12]	70.16	60.85	54.96	50.30	34.64	65.20	231.31
Attention-based (Soft) [12]	81.57	73.12	67.02	61.82	42.63	76.98	299.47
Sound-a-f [38]	78.28	72.76	67.59	63.33	38.03	68.64	290.57
VAA [37]	81.92	75.11	69.27	63.87	43.80	**78.24**	**339.46**
SM-ATT+LSTM [36]	81.54	75.75	69.36	64.58	42.40	76.32	318.64
RTRMN [64]	80.28	73.22	68.21	63.93	42.58	77.26	312.70
ML_Attention+Attribute_GCN [16]	**83.30**	**77.12**	**71.54**	**66.23**	43.71	77.63	316.84
TCE [62]	82.10	76.22	71.40	67.00	47.75	75.67	285.47
SVM-D CONC [63]	76.53	69.47	64.17	59.42	37.02	68.77	292.28
The Proposed GLCM	81.82	75.40	69.86	64.68	**46.19**	75.24	302.79

TABLE V
COMPARISON OF SOME STATE-OF-THE-ART METHODS ON RSICD.

Models	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE_L	CIDEr
VLAD-LSTM [12]	50.04	31.95	23.19	17.78	20.46	43.34	118.01
Attention-based (Soft) [12]	67.53	53.08	43.33	36.17	32.55	61.09	196.43
Sound-f-a [38]	59.35	45.11	35.29	28.08	26.11	49.57	132.35
SM-ATT+LSTM [36]	75.71	63.36	53.85	46.12	35.13	64.58	235.63
ML_Attention+Attribute_GCN [16]	75.97	64.21	55.17	46.23	35.43	65.63	236.14
TCE [62]	76.08	63.58	54.71	47.91	34.25	66.87	246.65
SVM-D CONC [63]	59.99	43.47	33.55	26.89	22.99	45.57	68.54
The Proposed GLCM	**77.67**	**64.92**	**56.42**	**49.37**	**36.27**	**67.79**	**254.91**

quality. Therefore in this paper, the following classical metrics are utilized to evaluate the generated captions of remote sensing images from different views:

(a) ***BLEU***. Bilingual Evaluation Understudy (BLEU) is proposed in [67] to measure the matching degree of n consecutive words, denoted as n-grams, between the generated and reference texts. In this paper, n is set to 1, 2, 3, and 4, corresponding to BLEU1, BLEU2, BLEU3, and BLEU4, respectively. BLEU focuses on the accuracy of the words in the generated sentence.

(b) ***METEOR***. Metric for Evaluation of Translation with Explicit ORdering (METEOR) is proposed in [68] for machine translation evaluation between the machine-generated text and human-produced reference text. METEOR takes both accuracy and recall into consideration when evaluating the caption quality.

(c) ***ROUGE-L.*** Recall-Oriented Understudy for Gisting Evaluation (ROUGE) is proposed in [69] for text summary. There is a set of variations in ROUGE such as ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-W. In this paper, ROUGE-L is used for RSIC, which can calculate F-measure given the *Longest Common Subsequence* (LCS).

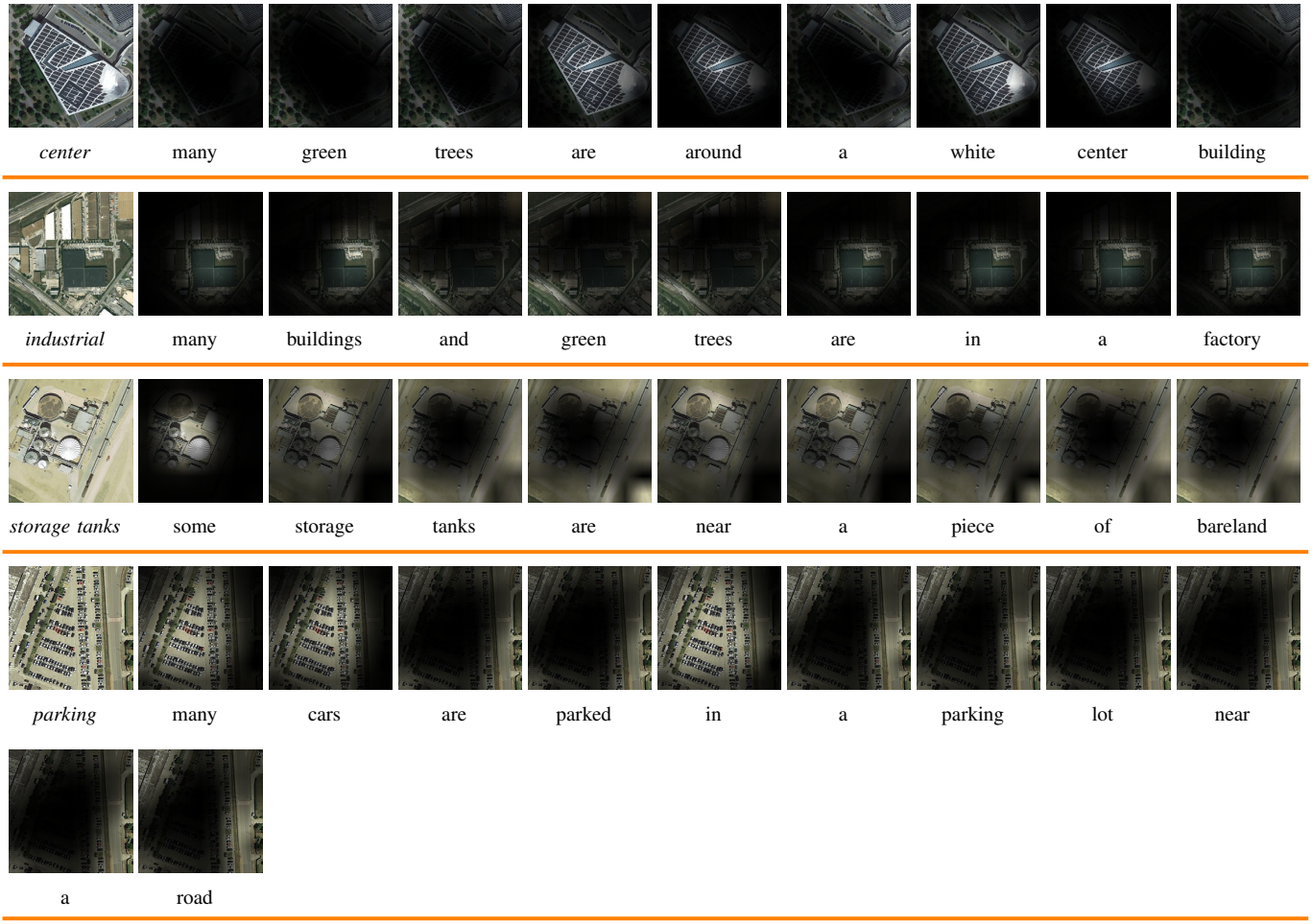


Fig. 5. Samples of four scenes of 'center', 'industrial', 'storage tanks', and 'parking' are provided. In the figure, only the local features which have the maximum dot-product response to the words are shown.

Different with BLEU, ROUGE-L pays attention to the recall degree of captions.

(d) ***CIDEr***. Consensus-based Image Description Evaluation (CIDEr) is proposed in [70] for image captioning. CIDEr concentrates on not only the accuracy of n-grams but also their frequency in the whole data set with the weighting operation.

B. Experimental Setup

Model Settings. The proposed GLCM obeys the encoder-decoder architecture, in which the encoder is used to extract global and local features while the decoder is used to fuse them and generate the corresponding sentence. For the encoder, we remove the last fully-connected layer (classifier) from GoogLeNet [25] pre-trained in ImageNet, and use the rest convolutional blocks as the CNN feature extractor. The number of local features of N is set to 64. For the decoder, four variations of global and local feature representation models, including $Local$, $Global\_A$, $Global\_B$ and $Global\_Local$, are combined with encoder and are performed on the three data sets to make ablation study.

Training details. In experiments, input images are resized to the same size of 224 $\times$ 224. During the training stage, images are randomly horizontally flipped with 50% probability. Adam is employed as the optimizer with the learning rate set from 5e-4 to 1e-4 due to the difference among data sets. The word vector dimension of C is set to 512. Under the condition of the mini-batch size of 64, all the models are trained for 30 epochs. In the Self-Attention layer, to make the sequence length of all the sentences consistent for parallel training, the max sequence length is set to 25. If the length of a training sentence in a mini-batch is not up to 25, it would be padded with zero. Besides, in this paper, all the experiments are conducted with PyTorch 1.3.0 in the computer of 64GB CPU and 1$\times$ 12GB GPU of NVIDIA GeForce GTX 1080Ti.

C. Ablation Study of CNN Backbones

The visual feature extraction quality is important for the caption performance. To study the influence of different visual feature extractors and select a reasonable CNN backbone for the following ablation study and comparison experimental results, in this subsection, the ablation study of ResNet18 [27], VGG16 [26] and GoogLeNet [25] combined with $Global\_A$ are conducted on three RSIC data sets.

The experimental results are shown in Table I. It could be found that in general, GoogLeNet shows the best and the

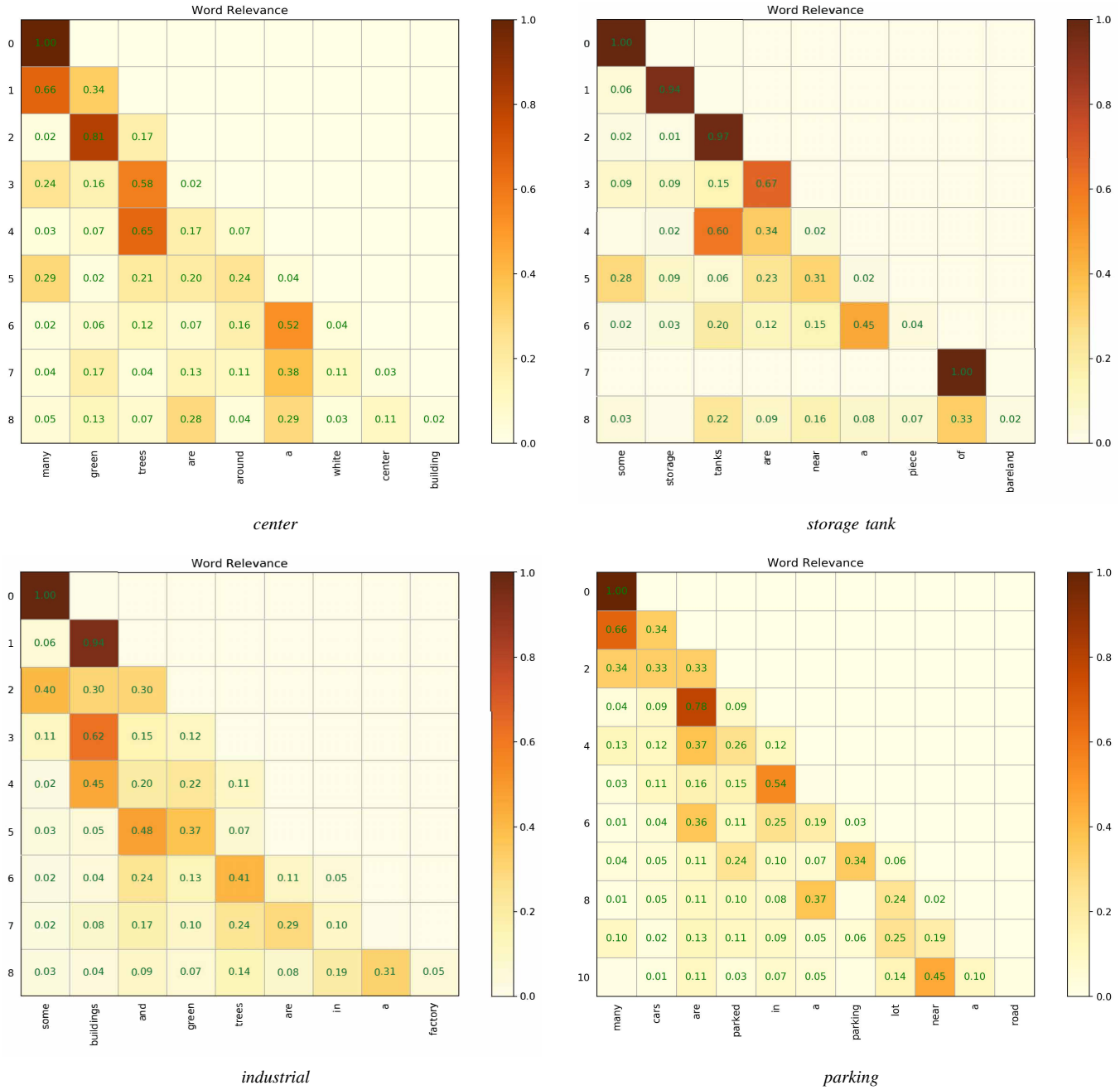


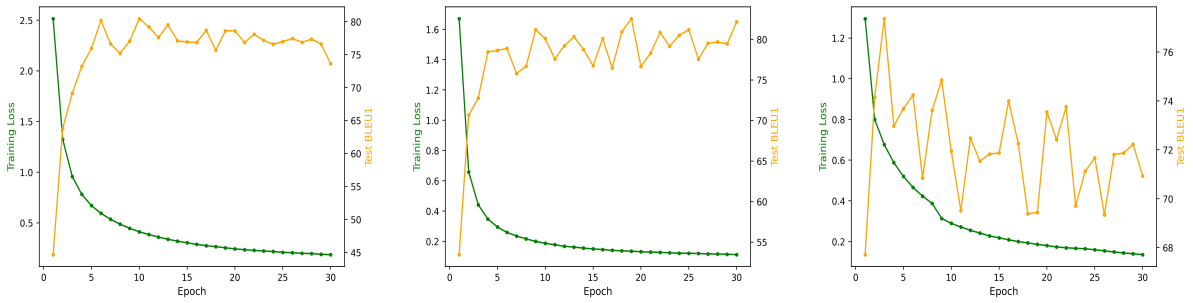
Fig. 6. Similarity among words of the generated sentences. For each word, it is only relevant to the generated words and itself. The grid on the diagonal represents the current word, and each grid to the left of the diagonal represents each generated word. The values of similarities among the current word and the generated words add up to 1 because of the Softmax function.

most stable performance in comparison with ResNet18 and VGG16. Besides, taking the model scale and inference speed into consideration, GoogLeNet has a significant advantage with much fewer model parameters compared with ResNet18 and especially VGG16. From this perspective, GoogLeNet is selected as the CNN extractor in the following experiments.

D. Ablation Study of GLCM Components

In this sub-section, to quantitatively explore the effect of global and local features, ablation experiments of *Local*, *Global_A*, *Global_B*, and *Global_Local* are conducted on three datasets. These four GLCM variations have different combination types of global and local features for RSIC. *Global_A* and *Global_B* take only global visual features as input, and *Local* takes only local visual features as input, and *Global_Local* takes both global and local visual features as input. Their results are provided in Table II.

Firstly, according to the comparison results between *Global_A* and *Global_B* of which the latter has one more Self-Attention layer than the former, it could be found that *Global_B* outperforms *Global_A* on all three datasets. Taking



Sydney-captions UCM-captions RSICD

Fig. 7. Training-test curves of three RSIC datasets based on our GLCM.

RSICD as an example, the average score of BLEU1-BLEU4 has improved from 59.82 of *Global_A* to 61.76 of *Global_B* with a gain of 1.95.

Secondly, when comparing the results of *Local* and *Global_B*, both of which have two attention layers, it could be found there is a trend that the performance of *Global_B* becomes better than *Local* when the data set gets larger. *Local* has superior performance than *Global_B* in the smallest data set of Sydney-captions, but the latter performs better in the larger data sets of UCM-captions and RSICD. It is probably because that the global features are mixed in the feature vector of the same size. It is difficult to learn the most effective part from the global features to predict the word at each time step especially when the training samples are limited. With the increase of training samples, the ability to extract valuable parts from global feature maps is promoted, which is beneficial for improving the caption quality.

Thirdly, it is meaningful to compare the results of *Global_B* and *Global_Local* which have the same number of learnable parameters but have different types of input. Specifically, *Global_B* can only get access to global features but *Global_Local* can absorb the information of both global and local features. Based on the same scale of learnable parameters, *Global_Local* obtains the better results on the two largest data sets of UCM-captions and RSICD which reveals that the introduction of local features can boost the caption performance of remote sensing images.

Finally, the number of the highest scores of *Global_Local* gets increases from 2 to 6 and then to 7, corresponding to Sydney-captions, UCM-captions, and RSICD. It could be observed that *Global_Local* almost gets the best results among the four GLCM variations and its advantage is more stable with the data set becoming larger. As described above, local features can provide more discriminative information while global features can provide complete information of features, objects, and their semantic relationship. Both global and local features are helpful for caption generation. As a model which can obtain joint global and local feature representation, it is reasonable for *Global_Local* to have the best performance.

In general, due to the ability of joint global and local feature representation, *Global_Local* is treated as the representation of GLCM in the next sub-section and its results are used to be compared with the existing state-of-the-art methods.

E. Comparison With State-of-the-Art Methods

To verify the performance of the proposed GLCM for RSIC objectively, some state-of-the-art methods generating captions for remote sensing images from different views are selected to make the analysis of comparison experiments. They are: (1) FV/VLAD-LSTM. SIFT, BOW, VLAD, and FV, which are the classical handcrafted features, are combined with LSTM for RSIC in [12]. Among these handcrafted features, FV-LSTM gets the best result of Sydney-captions and VLAD-LSTM gets the best results of UCM-captions and RSICD. (2) Attention-based (Soft/Hard) method. In [12], on the basis of CNN-LSTM architecture, two kinds of attention-based methods of soft and hard are further proposed to explore the role of attention mechanism in caption generation for remote sensing images for the first time. (3) Sound-f-a/Sound-a-f. Sound active attention is introduced into RSIC in [38] to capture the interest of an observer in images. Sound-f-a and Sound-a-f are different variations, and similarly, the best results of different data sets are compared with our method. (4) VAA. Visual Aligning Attention model (VAA) proposed in [37] aims to match the visual words with the corresponding image features. (5) SM-ATT+LSTM. In [36], to explore the affect of attributes hidden in high-level features, researchers present SM-ATT+LSTM based on the framework of encoder-decoder with an attribute attention module, which treats the softmax output of VGG16 as semantic attributes. (6) RTRMN. In order to take five sentences of a given image into consideration together, retrieval topic recurrent memory network (RTRMN) is proposed to incorporate the topic words among five sentences jointly to generate a determinate sentence. (7) ML_Attention+Attribute_GCN. To deal with the limitations of the various object scale and the underused visual relationship in remote sensing images, a framework based on multi-level attention (ML_Attention) and multi-label attribute graph convolutional network (Attribute_GCN) is specially designed in [16]. (8) TCE [62]. To preserve the possibility of other words with similar semantics at the same time steps, a truncation cross-entropy (TCE) loss is proposed for RSIC. (9) SVM-D CONC [63]. Instead of RNNs, a novel network of SVM is devised to decode the image information with word concatenation into a sentence description. Comparison results between these methods and the proposed GLCM on three data sets are provided in Table III, IV and V.

1) Sydney-captions: First of all, the comparison experiments are conduct on the smallest data set of Sydney-captions and the results are shown in Table III. For all the seven metric, ML_Attention+Attribute_GCN obtains the three highest scores of BLEU1, BLEU2, and ROUGE_L while the proposed GLCM takes the best results of BLEU3, BLEU4, and METEOR. Although both ML_Attention+Attribute_GCN and the proposed GLCM get the same number of the highest scores, the last CIDEr score of GLCM is higher than ML_Attention+Attribute_GCN. Besides, BLEU3 and BLEU4 scores have stricter requirements on generating multiple consecutive words than BLEU1 and BLEU2, which indicates that the proposed GLCM can generate the sentences which have more accurate phrases in Sydney-captions.

2) UCM-captions: Then, we make a comparison on UCM-captions, and the results are as shown in Table IV. Due to the limitation of handcrafted feature representation, VLAD-LSTM performs much worse than other deep learning feature-based methods. It is obvious that ML_Attention+Attribute_GCN takes the first place. Besides, the performance of the proposed GLCM is similar to Attention-based (Soft), VAA, and SM-ATT+LSTM. All of them are roughly in the second tier. In ML_Attention+Attribute_GCN, all the nouns and adjectives are manually selected from five sentences and used to guide the attribute extraction, which intrinsically is a second annotation for the data set. For GLCM, however, it is an end-to-end trainable framework without extra manual processing for the data set.

3) RSICD: In the end, the comparison experiments are made on RSICD and the results are listed in Table V. RSICD is the largest RSIC data set and therefore its results are more stable and convincing. According to the results, it could be found that our GLCM has an overall advantage and obtains all the highest scores of seven evaluation metrics. Especially for BLEU4, the proposed GLCM has an obvious score advantage of 3.14 over the second method of ML_Attention+Attribute_GCN. Such results suggest that the proposed GLCM can extract more discriminative and effective features for a large RSIC data set.

Overall, in comparison with the state-of-the-art methods, our GLCM obtains superior results of Sydney-captions and RSICD, and comparable results of UCM-captions. Such results verify the competitiveness of the proposed method.

F. Visualization

Benefiting from the attention layers in the decoder of GLCM, the most matched pairs of words and local visual features can be visualized from a Co-Attention layer while the similarity coefficients among words can be visualized from a Self-Attention layer. There are some visualization examples in Fig. 5, Fig. 6, and Fig. 7.

In Fig. 5, although not all the words are accurately located in the images, most of the activated areas are reasonable. Moreover, there is an interesting phenomenon that some consecutive words have the response to the same local features. For example, in the scene of *storage tank*, "many buildings", "and green trees", and "are in a factory" are respectively clustered into three corresponding local features. It reflects the hierarchical structure from word to phrase and to sentence in natural language.

In Fig. 6, the similarity among words in the same generated sentence is quantitatively visualized, which shows the contribution of all the previous words on predicting the next word. Different from LSTM which relies on the last word and the hidden state, Self-Attention can consider all the generated words simultaneously. It provides not only more interpretability of the caption generation but also more inspiration for further research on RSIC from the view of NLP.

Fig. 7 shows the training-test curves of our method on all three data sets, which reveals the fast convergence speed of our proposed method and proves its effectiveness.

V. Conclusion

In this paper, to make full use of the advantages of both global and local features of which the former can provide comprehensive information and the latter provide more discriminative information, a novel Global-Local Captioning Model (GLCM) is proposed for remote sensing image captioning (RSIC). The proposed GLCM utilizes a visual-linguistic attention mechanism with the corresponding attention-based visualization. To quantitatively explore the impact of global and local features in RSIC, ablation experiments of four variations based on GLCM are conducted. In comparison with some state-of-the-art methods, our GLCM obtains superior results in Sydney-captions and RSICD, and comparable results in UCM-captions.

In future work, we are going to improve the global-local mechanism for RSIC from the following points: (1) Current local visual features are extracted in a bottom-top manner, and it may be improved in combination with an up-bottom strategy to achieve more accurate local feature aggregation. (2) For the same remote sensing image, different observers can give different descriptions. The description uncertainty needs to be further considered and improved.

References

[1] W. Huang, Q. Wang, and X. Li, "Feature sparsity in convolutional neural networks for scene classification of remote sensing image," in *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium*, 2019.

[2] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of vhr remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, no. 99, pp. 1–13, 2018.

[3] Y. Li, Y. Zhang, and Z. Zhu, "Error-tolerant deep learning for remote sensing image scene classification," *IEEE Transactions on Cybernetics*, vol. 51, no. 4, pp. 1756–1768, 2021.

[4] Q. Wang, Z. Yuan, Q. Du, and X. Li, "Getnet: A general end-to-end 2-d cnn framework for hyperspectral image change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 1, pp. 3–13, 2018.

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

[6] X. Zhou, K. Shen, L. Weng, R. Cong, B. Zheng, J. Zhang, and C. Yan, "Edge-guided recurrent positioning network for salient object detection in optical remote sensing images," *IEEE Transactions on Cybernetics*, pp. 1–14, 2022.

[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] Y. Zhang, Y. Yuan, Y. Feng, and X. Lu, "Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 8, pp. 5535–5548, 2019.

[9] T. Zhang, X. Zhang, P. Zhu, X. Tang, C. Li, L. Jiao, and H. Zhou, "Semantic attention and scale complementary network for instance segmentation in remote sensing images," *IEEE Transactions on Cybernetics*, vol. 52, no. 10, pp. 10 999–11 013, 2022.

[10] X. Lu, X. Zheng, and X. Li, "Latent semantic minimal hashing for image retrieval," *IEEE Transactions on Image Processing*, vol. 26, no. 1, pp. 355–368, 2017.

[11] P. Li, L. Han, X. Tao, X. Zhang, C. Grecos, A. Plaza, and P. Ren, "Hashing nets for hashing: A quantized deep learning to hash framework for remote sensing image retrieval," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[12] 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, 2018.

[13] 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.

[14] 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.

[15] X. Zhang, Q. Wang, S. Chen, and X. Li, "Multi-scale cropping mechanism for remote sensing image captioning," in *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium*, 2019.

[16] Z. Yuan, X. Li, and Q. Wang, "Exploring multi-level attention and semantic relationship for remote sensing image captioning," *IEEE Access*, 2019.

[17] Q. Wang, W. Huang, X. Zhang, and X. Li, "Word–sentence framework for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 12, pp. 10 532–10 543, 2020.

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

[19] 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.

[20] 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.

[21] F. Perronnin, Y. Liu, J. Sánchez, and H. Poirier, "Large-scale image retrieval with compressed fisher vectors," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 3384–3391.

[22] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[23] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *null*. IEEE, 2003, p. 1470.

[24] 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.

[25] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

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

[27] 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.

[28] 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.

[29] 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.

[30] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-rnn)," *arXiv preprint arXiv:1412.6632*, 2014.

[31] W. Jiang, L. Ma, X. Chen, H. Zhang, and W. Liu, "Learning to guide decoding for image captioning," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[32] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4894–4902.

[33] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 4634–4643.

[34] L. Wang, Z. Bai, Y. Zhang, and H. Lu, "Show, recall, and tell: Image captioning with recall mechanism," *AAAI*, 2020.

[35] W. Zhang, Y. Ying, P. Lu, and H. Zha, "Learning long-and short-term user literal-preference with multimodal hierarchical transformer network for personalized image caption," *AAAI*, 2020.

[36] X. Zhang, X. Wang, X. Tang, H. Zhou, and C. Li, "Description generation for remote sensing images using attribute attention mechanism," *Remote Sensing*, vol. 11, no. 6, p. 612, 2019.

[37] Z. Zhang, W. Zhang, W. Diao, M. Yan, X. Gao, and X. Sun, "Vaa: Visual aligning attention model for remote sensing image captioning," *IEEE Access*, vol. 7, pp. 137 355–137 364, 2019.

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

[39] X. Li, A. Yuan, and X. Lu, "Vision-to-language tasks based on attributes and attention mechanism," *IEEE transactions on cybernetics*, 2019.

[40] Y. Bin, Y. Yang, F. Shen, N. Xie, H. T. Shen, and X. Li, "Describing video with attention-based bidirectional lstm," *IEEE transactions on cybernetics*, vol. 49, no. 7, pp. 2631–2641, 2018.

[41] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3128–3137.

[42] J. Johnson, A. Karpathy, and L. Fei-Fei, "Densecap: Fully convolutional localization networks for dense captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4565–4574.

[43] 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.

[44] M. Cornia, L. Baraldi, and R. Cucchiara, "Show, control and tell: a framework for generating controllable and grounded captions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8307–8316.

[45] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3156–3164.

[46] J. Ba, V. Mnih, and K. Kavukcuoglu, "Multiple object recognition with visual attention," *arXiv preprint arXiv:1412.7755*, 2014.

[47] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[48] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5659–5667.

[49] F. Wu, J. Cheng, X. Wang, L. Wang, and D. Tao, "Image hallucination from attribute pairs," *IEEE Transactions on Cybernetics*, 2020.

[50] L. Chen, M. Zhai, and G. Mori, "Attending to distinctive moments: Weakly-supervised attention models for action localization in video," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 328–336.

[51] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[52] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[53] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[54] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, 2016, pp. 1480–1489.

[55] Y. Cui, Z. Chen, S. Wei, S. Wang, T. Liu, and G. Hu, "Attention-over-attention neural networks for reading comprehension," *arXiv preprint arXiv:1607.04423*, 2016.

[56] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 1243–1252.
[57] Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A structured self-attentive sentence embedding," *arXiv preprint arXiv:1703.03130*, 2017.
[58] L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong, "End-to-end dense video captioning with masked transformer," pp. 8739–8748, 2018.
[59] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.
[60] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Advances in neural information processing systems*, 2016, pp. 289–297.
[61] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.
[62] 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*, vol. 59, no. 6, pp. 5246–5257, 2021.
[63] G. Hoxha and F. Melgani, "A novel svm-based decoder for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.
[64] 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.
[65] 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.
[66] 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.
[67] 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.
[68] 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.
[69] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," *Text Summarization Branches Out*, 2004.
[70] 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.

Qi Wang (M'15-SM'15) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and pattern recognition.

Wei Huang received the B.E. degree in control theory and engineering from the Northwestern Polytechnical University, Xi'an, China, in 2018. He is currently working toward the M.S. degree in computer science in the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include deep learning and remote sensing.

Xueting Zhang received the B.E. degree in control theory and engineering from the Northwestern Polytechnical University, Xi'an, China, in 2018. She is currently working toward the M.S. degree in computer science in the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. Her research mainly focuses remote sensing image processing.

Xuelong Li (M'02-SM'07-F'12) is currently a Professor with the School of Computer Science, with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China.