From Plane to Hierarchy: Deformable Transformer for Remote Sensing Image Captioning

Runyan Du, Wei Cao, Wenkai Zhang, *Member, IEEE*, Guo Zhi, Xian Sun, *Senior Member, IEEE*, Shuoke Li, and Jihao Li

***Abstract*—With the growth of remote sensing images, understanding image content automatically has attracted many researchers' interests in deep learning for remote sensing image. Inspired from the natural image captioning, the model with convolutional neural network (CNN)-Recurrent neural network (RNN) as the backbone and supplemented by attention has been widely used in remote sensing image captioning. However, it is inefficient for the current attention layer to simultaneously mine hidden foreground from the background of remote sensing image and perform feature interactive learning. Meanwhile, the new mainstream language model has recently surpassed the traditional long short-term memory (LSTM) in sentence generation. For solving the above problems, in this article, we proposed a novel thought to make the flat remote sensing images stereoscopic by separating the foreground and background. Based on hierarchical image information, we designed a novel Deformable Transformer equipped with deformable scaled dot-product attention to learn multiscale feature from foreground and background through the powerful interactive learning ability. Evaluations are conducted on four classic remote sensing image captioning datasets. Compared with the state-of-the-art methods, our Transformer variant achieves higher captioning accuracy.**

***Index Terms*—Attention, remote sensing image captioning (RSIC), transformer.**

I. Introduction

REMOTE sensing image captioning (RSIC) [1], [9] is a challenging task for replacing human to automatically understand the growing mass of high-resolution remote sensing images. It is a translation task from visual modality to text modality. Even now, how to bridge the semantic gap between two modalities has always been a difficult problem for researchers to overcome. With the gradual maturity of deep learning technology, the main framework for processing RSIC is inspired from the natural image captioning (NIC) [4], [12] in natural scene, which is called "encoder–decoder framework." The encoder is responsible for turning each image into a numerical vector. Meanwhile, the decoder generates sentences by turning each numerical vector into a specific textual content. The largest difference between image captioning in natural scenes and remote sensing scenes lies in the image understanding module of the encoder part.

This difference stems from the fact that natural images and remote sensing images are worlds apart in terms of no matter view or content. The natural image is recorded in a horizontal view and its foreground and background are clear at a glance from human's sight. The protagonist in the image often occupies most of the image area and is easy to identify. Humans can naturally regard it as the subject and construct sentences. On the other hand, remote sensing image is recorded in a vertical view (God's view) [17], which has a large number of objects and more complex case for discriminating the foreground and background. The main character in the image is similar in size to the background, or is directly hidden in it, which is difficult to be found even for human. All these factors make it harder to generate decent descriptions under the same framework for RSIC than NIC. Since foreground elements are difficult to identify in remote sensing image, if we first extract the relevant prior knowledge from an upstream task and then directly tell the caption model the position of these foregrounds, the difficulty of RSIC will degenerate to be comparable to NIC. At the same time, fine foreground and background separation can also benefit the model to better understand remote sensing images.

In natural scenarios, using the object sequence extracted by an object detection network [18] as the input of the model has become a mainstream method. Due to the difference in perspective between remote sensing images and natural images, the objects in remote sensing images have smaller size, denser distribution, more diverse aspect ratios, and unfixed directions compared to natural images. The remote sensing images contain complex background information, and much foreground is hidden in it. These make ordinary object detector, which is very effective in natural image caption, defuncts in remote scene. Instead of spending much time on searching another better designed object detector, we plan to design a unified and better Transformer framework, which is customized for RSIC based on a widely known feature extractor.

First, the attention in the current CNN-RNN framework is weak in mining hidden foreground in remote sensing images. In order to perceive the hidden foreground information from the complex background, a pixel-by-pixel analysis

Manuscript received 12 July 2023; revised 22 July 2023; accepted 10 August 2023. Date of publication 16 August 2023; date of current version 25 August 2023. This work was supported by the National Key R&D Program of China under Grant 2022ZD0118402. *(Corresponding author: Wenkai Zhang.)*

The authors are with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China, also with the School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100190, China, and also with the Key Laboratory of Network Information System Technology (NIST), Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China (e-mail: durunyan19@mails.ucas.ac.cn; caowei@aircas.ac.cn; zhangwk@aircas.ac.cn; guozhi@mail.ie.ac.cn; sunxian@aircas.ac.cn; lisk@aircas.ac.cn; lijihao17@mails.ucas.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3305889

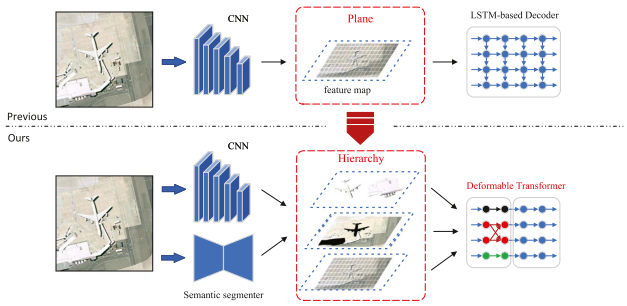

Fig. 1. Intuitive comparison diagram between the traditional RSIC framework and our framework. The main difference lies in the hierarchical strategy at the image understanding stage and the uniquely designed deformable Transformer.

method is adopted with a general pixel-level semantic segmentation pretrained model to distinguish the foreground from the background. In the previous methods, the foreground and background information in the grid-like feature map are entangled, which makes it more difficult for the attention mechanism to understand the visual content. Combining the above two points, we establish a conceptual hierarchical structure with three floors based on the foreground and background extracted by semantic segmentation [31]. Fig. 1 shows its concept diagram. The bottom floor is the original grid-like feature map, the second floor is the background, and the top floor is the foreground. Each floor has its own corresponding encoding method.

Second, the feature interactive learning ability is derived from the global weighted summation of single-layer attention, which does not consider the local interactive learning for each object and has poor module depth. For solving this problem, Transformer has become our backbone with its powerful interactive learning and better sentence generation abilities. Note that, the shape of the foreground is variable. For adapting to such variability, we propose a deformable scaled dot-product attention to learn the interaction between the subregions within the shape to reduce the interference from other background information. Based on the feature map, the deformable attention learns the interaction within one object at pixel level, while the subsequent self-attention learns the interaction between objects at object level. Our model has both intra- and interclass feature representation learning, to build a powerful Transformer [19] variant to generate better sentences. Overall, the main contributions of this article can be summarized as follows.

Contributions:

1) From the view of overall framework, we designed a novel Transformer framework specifically for RSIC, which is equipped with the ability of understanding pixel semantics. It is different from the commonly used Transformer for understanding object categories in natural scenes. Functionally, our encoder has the power of fine-grained aggregation dealing with information within the object's structure.
2) The central idea of our framework is to divide the multiscale information of the image into three floors: a) foreground, b) background, and c) raw information. The hierarchical encoding strategy we proposed has three corresponding perception functions for each floor to mine its content.
3) The perception functions for foreground and background are dedicated to learning the feature interaction between pixel and pixel inside the object's shape. We propose a novel attention operator: deformable scaled dot-product attention to implement this function. It understands the foreground and background content region by region.

The rest of this article is organized as follows. Section II introduces the related works of RSIC and Transformer. In Section III, we describe the proposed deformable Transformer in detail. The experimental results and analysis on four datasets are introduced in Section IV. Finally, Section V concludes the article.

II. Related Works

In this section, the relevant works for RSIC and Transformer will be briefly introduced.

A. Remote Sensing Image Captioning

In recent years, the mainstream frameworks for RSIC were divided into three types: 1) template-based methods; 2) retrieval-based methods; 3) encoder–decoder methods. The central idea of retrieval-based methods is to search relevant sample in a large database and take its caption as the result of current input image. The model performance is highly depended on the retrieval results and their match degree with the input. Wang et al. [1] has presented a collective semantic metric learning architecture which constructed a shared space to compute the distance for both image representation and caption representation. Given a test image, the caption with smallest distance will be taken as the final output. However, retrieval-based methods are suffered from the irrelevance of caption to image content, which affects the caption's accuracy. Another type is template-based methods, which generates independent words and fills them into an artificially set template by filling in the blanks. Shi and Zou [2] have employed a fully convolutional network (FCN)-based method [3] to capture the key words about objects, attributes, and relationship content. But the caption generated by the template-based method is too rigid in form and poor in readability.

Inspired from the natural image captioning, the encoder–decoder methods [4] are widely used in RSIC. The convolutional neural network (CNN) [5], [6] is responsible for image understanding, while the recurrent neural network (RNN) [7], [8] is responsible for sentence generation. During the same period, Qu et al. [9] and Zhang et al. [10] have successively proposed the encoder–decoder framework in RSIC and achieved good results. Lu et al. [11] added additional "soft" and "hard" attention mechanisms [12] to the CNN–RNN architecture, which allows the decoder to narrow the receptive field to one specific region when generating word. The efficiency of the attention mechanism has made it widely used in the field of RSIC. Researchers have proposed many excellent attention mechanism variants to insert between CNN and RNN.

Zhang et al. [13] integrated attribute information reasoned from the high-level image feature to assign higher attention weight for the salient object in image. Sumbul et al. [14] proposed a summarization-driven model, which used the summarization of ground truth captions to combine the standard captions. Li et al. [15] proposed a recurrent attention and semantic gate framework to enhance the visual feature integration and context vector generation. Zhang et al. [16] continued to polish the attention mechanism, and used global visual feature and linguistic state as the guiding factors to control the attention mechanism. Zhao et al. [17] proposed a structure-aware feature pooling method to change the grid-like feature map [12] according to the object's shape. This is an excellent work and takes into account the same utilization of object's structure as we do. But there is a fundamental difference with our method. Zhao et al. [17] used the CNN–RNN architecture and changed the RoI pooling [18] into a structure-aware pooling operator. On our side, we focus on the Transformer architecture [19] and change the interactive learning into a more fine-grained structure-aware interactive learning among pixels.

Except researches on attention mechanisms, there are many excellent works in RSIC. Lu et al. [20] introduced the sound information into the model, which introduced the active attention to generate sentence based on the sound-guided image feature. On the decoder side, Hoxha and Melgani [21] abandoned the RNN and proposed a novel SVM-based [22] decoder for better solving the long-term dependency problem. Wang et al. [23] concentrated on the textual modality and proposed a word-to-sentence framework for avoiding the unexplainable of encoder–decoder architecture. On the training side, Li et al. [24] proposed a novel truncation cross-entropy loss, which truncated a sample's loss value when it was well learned by model. Cheng et al. [25] proposed a grand new dataset NWPU-captions, which was elaborate and data-heavy. They also proposed a new network MLCA-Net for updating the attention into multilevel attention. To our best knowledge, there are few effective Transformer-based methods existing in the current stage of RSIC. Thus, in this article, we specifically designed a Transformer-based method for RSIC to understand multiscale information in remote sensing images.

B. Transformer

Transformer was first proposed by Vaswani et al. [19], and its novel full-attention network structure successfully achieved breakthrough results in all text domain tasks [27], [28]. In the field of multimodality [29], [30], Transformer has also achieved great success. The scaled dot-product attention in Transformer is a lightweight version of the attention embedded in CNN–RNN [12]. The generation of its attention weight is no longer through the fully connected network but through the vector dot product. Zhu et al. [26] firstly introduced the Transformer into image captioning, which took the object's features reasoned by the object detection network [18] as the input to generate sentences. Many subsequent Transformer variants also continue this framework. In this article, our method is not only to simply carry Transformer into RSIC but also to construct a customized interactive learning strategy for RSIC. Specifically, the Transformer framework we designed takes the interactive learning of pixel features within the object's shape as the core to fully understand the small-scale foreground content in remote sensing images.

III. Methodology

In this section, a brief introduction about the backbone Transformer and its functioning operator is firstly given, and then we illustrate in detail each part of our model built on it.

A. Overview of the Method

The classical captioning model with Transformer involves three parts: 1) an encoder (feature extractor), 2) a refiner (refining visual feature), and 3) a decoder (generating sentence). The overview of our method is shown in Fig. 2. Given a remote sensing image, it will be feed into our feature extractor, which charges the generation of its semantic map and feature map [12]. We use the 101-layer deep residual network (ResNet-101) [6] to generate feature map produced by the last convolution block. On the other side, the raw image will be first split into several subparts and each part will be transferred to its corresponding semantic map through the DeepLab v3+ [31]. Second, stitching semantic maps of all subparts into one and scale it to be the same size as the feature map. Inspired from zhao et al. [17], we additionally employ the selective search [32] to enhance the semantic map in the end.

Stepping in the feature refining stage, the foreground and background are split with the semantic map. The three visual contents, foreground, background, and raw feature map, are hierarchically encoded. We propose deformable scaled dot-product attention to constrain the interaction learning process in object's structure. The deformable scaled dot-product attention is a variant of conventional scaled dot-product attention [19]. Different from Transformer, our refiner combines deformable scaled dot-product attention and conventional self-attention [19] for learning the interaction in structure-level and object-level. We make a specifical design for it and this part will be described in detail in the subsequent section. We first introduce some preliminary knowledge about Transformer [19] and its scaled dot-product attention. The conventional Transformer employed an encoder–decoder framework. Both encoder and decoder are stacked by several attention layers following layer normalization [33] and residual connection [6]. In each attention layer, it uses the scaled dot-product attention, which replaces the traditional learnable attention weight with the similarity between two vectors. The general scaled dot-product attention (Att) is formulated as follows:

$$Att\left(\mathbf{Q},\mathbf{K},\mathbf{V}\right)=\text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{T}}{\sqrt{d_k}}\right)\mathbf{V} \tag{1}$$

$$MAtt\left(\mathbf{Q},\mathbf{K},\mathbf{V}\right)=Concat_{i=1:h}\times\left(Att\left(\mathbf{W}_q\mathbf{Q}_i,\mathbf{W}_k\mathbf{K}_i,\mathbf{W}_v\mathbf{V}_i\right)\right). \tag{2}$$

$MAtt$ is the multihead attention [19], which concatenates h attention result after synchronously processing scaled dot-product

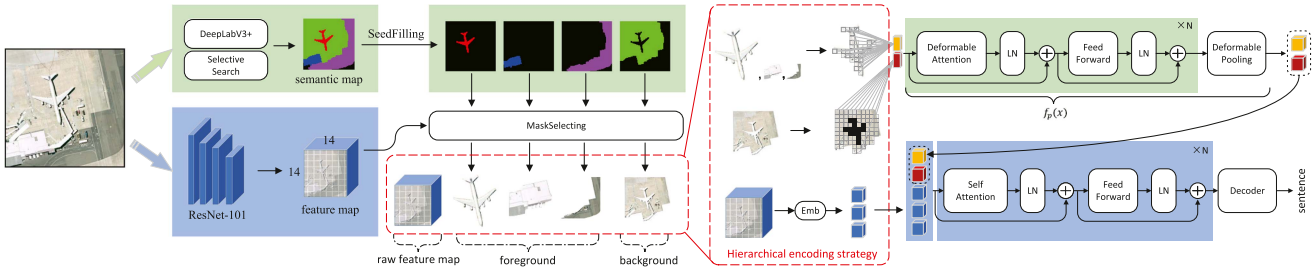

Fig. 2. Overall flowchart of our framework. Remote sensing images pass through ResNet-101, DeeplabV3+, and selective search method to obtain feature maps and semantic maps. After SeedFilling, different objects are divided according to the foreground and background and input into the perception function (functioned by deformable attention) to perform intraclass interactive learning. In the end, the output of deformable pooling will concatenate with the raw feature map and learn interclass interactive among its elements, and then generate sentences through the Transformer decoder.

attention in h feature subspaces. $\mathbf{W}_*$ are the learnable parameters. $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are the query, key, and value matrices. The computation of each scaled dot-product attention is more like a feature aggregation process. This makes the features of each position in the sequence continuously strengthen the representation ability of its own characteristics as the number of layers goes deeper and deeper, thereby, improving the performance of model.

Finally, in the sentence generating stage, the decoder also inherits a similar attention pattern and aligns the visual feature and textual feature in each decoder layer. Different from the sequential generation in the recurrent network, the words are synchronously generated in each layer with consideration of all previous context and encoder's output. After going through several layers, the final sequence is further fed into a fully connected layer with a softmax to produce tth word probabilities $\mathbf{p}_t$

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

B. Hierarchical Encoding Strategy

In the previous section, we mentioned about our feature extraction strategy. Through it, we obtain an additional semantic map and a conventional feature map. If following the traditional framework, the boundary between foreground and background information is difficult to be recognized by the model without any additional guidance. The small-scale foreground information is easily disturbed by surrounding large-scale background information. However, the existence of semantic map makes us having the ability to leverage a good prior knowledge to separate small-scale foreground objects from their surrounding background.

From this perspective, we propose a tower-like hierarchical encoding strategy as shown in the middle of Fig. 2. For the foreground, background, and feature map, we design corresponding perception functions for encoding them. From the process, the extracted feature map $\mathbf{X}_{raw}$ first passes through an embedding layer to map each subregion' feature to the model's feature space. Subsequently, the foreground and background are divided according to the label value $\mathbf{S}_i$ of each pixel belonging to ith object in the semantic map $\mathbf{S}$. The above three kinds of information are encoded by the following formulas:

$$\mathbf{E}_f = f_p\left((\mathbf{S} == \mathbf{S}_i) \odot \mathbf{X}_{raw}\right) \tag{4}$$

$$\mathbf{E}_b = f_p\left(\left(\prod_{i}^{K} \mathbf{S}! = \mathbf{S}_i\right) \odot \mathbf{X}_{raw}\right) \tag{5}$$

$$\mathbf{E}_{raw} = emb\left(\mathbf{X}_{raw}\right). \tag{6}$$

where K is the total number of detected objects and emb is the embedding layer for raw feature map. $\mathbf{E}_f$, $\mathbf{E}_b$, and $\mathbf{E}_{raw}$ are the encoded foreground, background, and raw feature map, respectively. The f_p is the perception function, which is composed by several modified Transformer encoder layers. In each layer, the scaled dot-product attention is replaced by our deformable scaled dot-product attention. With the mask of foreground and background, the interaction learning process of deformable scaled dot-product attention is completely constrained within the object's shape. Theoretically, f_p charges the Internal feature aggregation of a specifical object's category and enhance the representation ability of its feature.

The last layer of f_p is a deformable max/average pooling, which is proposed for pooling features within the object's shape and summarize the statistical properties of an object from a variable-length feature sequence. After passing through the f_p, the $\mathbf{E}_f$, $\mathbf{E}_b$, and $\mathbf{E}_{raw}$ are concatenated together as a final hierarchical feature sequence for further feeding into the conventional Transformer encoder layers. From the view of overall encoder, the deformable scaled dot-product attention equipped by f_p learns the interaction of internal object's category. The scaled dot-product attention in the subsequent conventional Transformer learns the interaction between different object's categories.

C. Encoder of Deformable Transformer

Different from the encoder of conventional Transformer, our encoder is composed by the following two parts: 1) the perception function f_p (main operator is deformable scaled dot-product attention) and 2) self-attention layers (main operator is scaled dot-product attention).

1) Deformable Scaled Dot-Product Attention: In each layer of the perception function, the input of deformable scaled

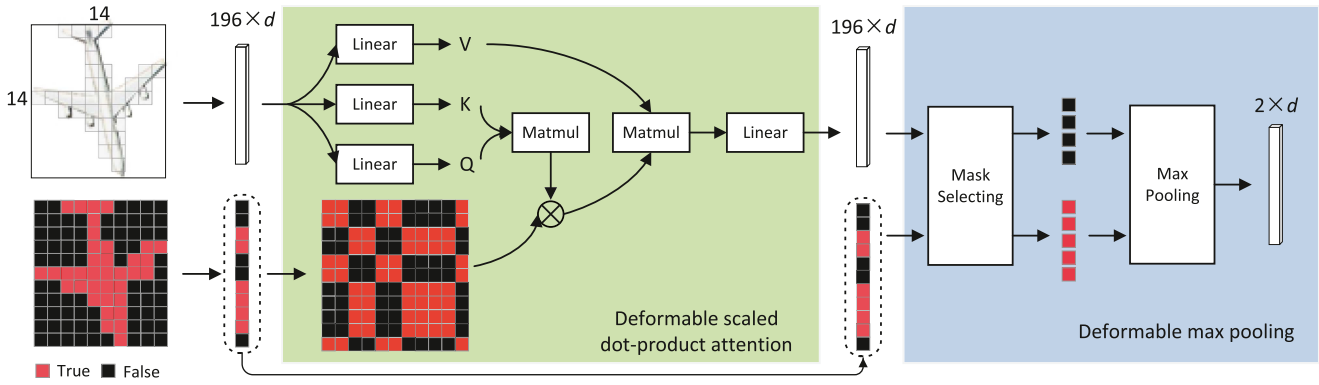

Fig. 3. Schematic diagram of deformable scaled dot-product attention and deformable pooling. The former calculates the intraclass interaction of all foreground and background at one time according to the connectivity relationship in the semantic map. The latter is to statistically process the features of the pixels within one object's shape.

dot-product attention is a flattened sequence: $X = \{x_1, x_2, \ldots, x_N\}$. Different from the conventional scaled dot-product attention, when one element x_i is taken as a query vector to compute the similarity with its key sequence $\mathbf{K} = X$, the elements in $\mathbf{K}$ will be verified with considering their connectivity in the semantic map. Specifically, only the elements of $\mathbf{K}$ which share the same connected domain R with the query x_i in the semantic map will be preserved. Fig. 3 shows the architecture of our attention. However, the scaled dot-product attention has no such restriction.

This advantage makes the feature aggregation only consider the elements within the object's shape and will not be disturbed by external noise. The formula of our deformable scaled dot-product attention is shown as follows:

$$DAtt\left(\mathbf{x_i}, \mathbf{K}, \mathbf{V}\right) = \text{softmax}\left(\frac{\mathbf{x_i}\left(\mathbf{K} \odot \mathbb{I}\right)^T}{\sqrt{d_k}}\right)\left(\mathbf{V} \odot \mathbb{I}\right) \quad (7)$$

$$\mathbb{I}_j = \begin{cases} 1, & P_{K_j} \in R_{x_i} \\ 0, & P_{K_j} \notin R_{x_i} \end{cases} \quad (8)$$

where $\mathbb{I}$ is the indicator function and its jth element determines whether jth element of K needs to be contained. P_{K_j} means the pixel of K_j in the semantic map. R_{x_i} represents the connected domain where x_i is located. Colloquially, when the pixel of K_j is in the same connected domain as x_i, the K_j will be put into key sequence K. The connected domain of x_i is detected with Seed-Filling method [34]. For enhancing the feature representation ability, the multihead mechanism [19] is also exploited which replaces the Att with $DAtt$ in (2). After the attention, the layer normalization [33] and residual connection [6] are followed. The deformable scaled dot-product attention layer can also be stacked multiple times for enhancing the feature representation ability.

2) Deformable Max/Average Pooling: The deformable scaled dot-product attention learns the interaction within one object's shape. Specifically, there are N pixels existing in the shape of one object and the output will be an N-length sequence $\mathbf{X}^L$ through L deformable scaled dot-product attention layers. Inspired from roi pooling [18] and structure attention [17], an appropriate way needs to be found to summarize the features of each pixel in the shape, and get the feature of the foreground object. The architecture is shown in Fig. 3.

We perform max pooling on the pixel features within the shape according to the connected area information provided in the semantic map. The deformable average pooling is also implemented. Both the calculation formulas are shown as follows:

$$DMP(\mathbf{X}^L) = \underset{i=1:C}{Concat} \max(\mathbf{X}_i^L), \text{i: feature dimension} \quad (9)$$

$$DAP(\mathbf{X}^L) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i^L, \text{i: index in sequence } \mathbf{X}. \quad (10)$$

The deformable pooling is effective to understand the objects in the remote sensing image. Due to the particularity of God's perspective, the foreground objects concerned in remote sensing images usually have a fixed shape. For example, whether a plane is taking off or landing, the shape is fixed when viewed from above. In contrast to natural scenes, the same person may have wildly different shapes due to his different poses. Therefore, if a human observes two remote sensing images, the objects of the same category can be easily located and identified according to their similar shapes. The deformable pooling and deformable attention actually use the semantic map to make the positioning work more detailed, and then use the excellent intraclass feature aggregation capabilities to recognize the object's category.

In summary, our encoder takes the split three visual information as the input: 1) foreground, 2) background and 3) raw feature map. The foreground and background information go through the perception function, which contains L deformable attention layers following one deformable pooling layer. According to our hierarchical encoding strategy, through the perception function, the encoded foreground and background information $\mathbf{E}_f$ and $\mathbf{E}_b$ are obtained. The raw feature map goes through the embedding layer to obtain $\mathbf{E}_{raw}$. Concatenating above three kinds of features, the hierarchical sequence has contained the large-scale information $\mathbf{E}_{raw}$, the middle-scale background $\mathbf{E}_b$, and small-scale foreground $\mathbf{E}_f$. They will be further fed into

the second part of our encoder: conventional Transformer encoder [19] to further learn the interclass interactions. In the following Transformer encoder, the information of all scales will be concatenated together as a visual sequence.

D. Decoder

For decoder, we apply the standard Transformer decoder [19]. Each decoder layer consists of two attention layers and one feedforward layer. The attention layers are self-attention and cross-attention, respectively. Self-attention operator is responsible for encoding the textual information and cross-attention reweights the internal features generated by encoder according to the similarity between both modality. Following the standard procedure, we employ the subsequent mask to prevent the word from foreseeing subsequent words in ground true sentence. Formally, given a textual input $\mathbf{T}^{l-1} = \{\mathbf{t}_1^{l-1}, \mathbf{t}_2^{l-1}, \ldots, \mathbf{t}_N^{l-1}\}$ from the previous decoder layer and a encoder output $\mathbf{E}_{out}$, the output of lth decoder layer is calculated as follows:

$$\mathbf{D}_{\text{self}}^{l} = LN\left(MAtt(\mathbf{T}^{l-1}, \mathbf{T}^{l-1}, \mathbf{T}^{l-1}) + \mathbf{T}^{l-1}\right) \tag{11}$$

$$\mathbf{D}_{\text{cross}}^{l} = LN\left(MAtt(\mathbf{D}_{\text{self}}^{l}, \mathbf{E}_{\text{out}}, \mathbf{E}_{\text{out}}) + \mathbf{D}_{\text{self}}^{l}\right). \tag{12}$$

The $\mathbf{D}_{\text{self}}^{l}$ and $\mathbf{D}_{\text{cross}}^{l}$ are output of self-attention layer and cross-attention layer in the lth decoder layer. LN is the LayerNorm [33] and $MAtt$ is the multihead attention shown in 2. Each attention layer is followed with the LayerNorm and residual connection [6]. After stacking several decoder layers, the $\mathbf{D}^{L}$ (last decoder layer) will be used as the final textual features for sentence generation.

From the view of overall framework, a raw remote image becomes the feature map $\mathbf{X}_{raw} \in \mathbb{R}^{N_r \times N_r \times d}$ with a pretrained ResNet [6] and the segmentation map $S \in \mathbb{R}^{N_r \times N_r \times C}$ with a pretrained DeepLabv3+ [31]. N_r is the weight or height of the feature map, C is the number of detected instances, and d is the model dimension. As shown in Fig. 2, with the segmentation map S, the different foreground and background objects are split and fed into the perception function f_p to learn the intraobject interaction with the deformable attention. The output of f_p is the visual sequence $\mathbb{E} = [\mathbb{E}_f, \mathbb{E}_b, \mathbb{E}_{raw}]$ which contains the embedding of foreground 4, background 5, and feature map 6. The $\mathbb{E}$ will be further fed into the conventional self-attention to generate the encoder output $\mathbb{E}_{\text{out}} \in \mathbb{R}^{(N_r * N_r + N_f + N_b) \times d}$, where N_f and N_b are the number of foreground and background, respectively. Then, $\mathbb{E}_{\text{out}}$ is passed to each decoder layer for performing cross-attention. After stacking several decoder layers, the $\mathbf{D}^{L}$ (last decoder layer) will further feed into an FC following a $softmax$ to generate logit for each word from the vocabulary.

E. Loss Functions

The output of encoder will be fed into decoder for generating words. Our decoder is the classic Transformer decoder [19]. It is worth mentioning that the encoder framework we proposed can also be easily spliced on any decoder based on the recurrent network. Following the previous works, the cross-entropy loss [35] has been used for training model and ensures fair performance comparisons. Assuming that the generated words sequence of final decoder layer is $\hat{\mathbf{y}} = \{\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, \ldots, \hat{\mathbf{y}}_N\}$, the ground truth sentence is $\mathbf{y}$. The loss calculation formula is as follows:

$$\mathcal{L}_{CE} = -\sum_{i=1}^{N} y^i * log\hat{y}^i. \tag{13}$$

In order to alleviate the model overfitting problem to a certain extent, the label smoothing [36] is employed on the ground-truth label. For ensuring the fairness of model performance comparison, we do not use reinforcement learning [37] to further train the model as the natural image captioning task.

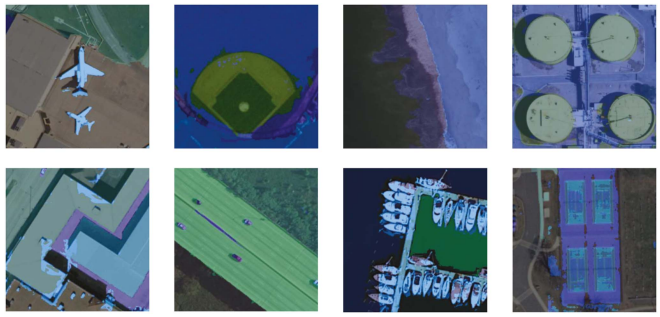

Fig. 4. Some samples of our semantic map. The foreground (for example: aircraft, etc.) is extracted by DeepLabV3+, while the background is mainly enhanced by selective search.

F. Implementation Details

1) Training Details: In the training phase, the Adam [38] is used as the optimizer with its regularization coefficients α and β is set to be 0.9 and 0.98, respectively. Our framework is designed for end-to-end training stage and the learning rate is set to be 1×10^{-8}. The batch size of images is set to 5 and each image contains five samples with different ground-truth caption. The model is trained for 100 epochs. In our feature extractor, the ResNet-101 [6] is pretrained on the ImageNet [39] dataset and the DeepLab v3+ [31] is pretrained on the iSAID [40] dataset. Both feature extractors are set to be untrainable. The initialization of our model employs the xavier normalization [6]. In the testing phase, we use the beam search [4] method to generate more stable captions. The beam size is set to 5.

2) Model Details: In our feature extraction stage, we divide each input image into 16 subimages with the same size and inference the semantic map on each subimage through a DeepLab v3+ [31]. Finally, the submaps are spliced together to obtain the final total semantic map. For detecting the connected domain of each object in the semantic map, we use the seed-filling [34] method. Fig. 4 shows some samples of our semantic map. The encoder of our model contains three deformable attention layers and six conventional Transformer encoder layers. On the decoder side, our model contains six decoder layers. The drop rate of the dropout layer in all attention layers is set to be 0.1. In the deformable pooling layer, we use Max operator as the pooling core. Besides, the d_{model} and d_{ff} of Transformer is set to 512, 1024. The setting of the selective search [32] is followed as Zhao et al. [17]

IV. Experiments

In this section, we introduce our experimental datasets, evaluation metrics, and the comparative results with other state-of-the-art methods in detail. The ablation experiments, parameter analysis, and attention visualization are provided to qualitatively and quantitatively verify the effectiveness of our model.

A. Datasets

In the RSIC, there are three widely used datasets, including Sydney-captions [9], UCM-captions [9], and RSICD [11]. Meanwhile, we note a grand new fascinating dataset proposed in recent years. In order to verify the generalization ability of our model, we also conduct experiments on NWPU-Captions [25].

1) Sydney-Captions: The Sydney-captions dataset was proposed by Qu et al. [9], which is based on the images provided in Sydney [41]. There are a total of 613 images with seven different scenes, including industrial, rivers, residential, meadow, runway, airport, and ocean. All of them are collected from Google Earth of Sydney, Australia. The resolution of each image is 0.5 m. For each image, five reference sentences are given to abstract its content from different observers. There are 80% images in Sydney-captions used for training, 10% for validation, and the rest 10% for testing.

2) UCM-Captions: The UCM-captions dataset was also proposed by Qu et al. [9], which was based on the UC Merced (UCM) land-use dataset [42]. There are a total of 2100 high-resolution remote sensing images. The UCM-captions dataset contains 21 scene categories, including building, beach, airplane, chaparral, forest, harbor, freeway, overpass, intersection, runway, river, agricultural, dense residential, tennis court, sparse residential, golf course, baseball diamond, medium residential, parking lot, mobile home park, and storage tank. All the images measure 256×256 pixels with a resolution of 0.3048 m. For each image, five descriptions are also given by different observers. For training, there are 80% images. The rest is split equally for validation and testing.

3) RSICD: The RSICD dataset was proposed by Lu et al. [11], which contained 10 921 images measuring 224×224 pixels. All the images are collected from MapABC, BaiduMap, Google Earth, and Tianditu with different resolutions. Similar to the previous datasets, five reference sentences are provided for each image. Its splitting ratio is the same as Sydney-captions and UCM-captions.

4) NWPU-Captions: The NWPU-captions dataset was proposed by Cheng et al. [25], which contains 31 500 images and 157 500 sentences. It is constructed based on NWPU-RESISC45 [43], which contains 45 scene categories. The images are collected from Google Earth and each one is annotated manually in five different sentences. It is a balanced dataset and each scene category contains 700 images (its size is 256×256). In NWPU-captions dataset, there are 25 200 images used for training. The validation set and test set has 3150 images for each.

B. Evaluation Metrics

In RSIC, five evaluation metrics proposed in the text field are commonly used, including BLEU [44], METEOR [45], ROUGE-L [46], CIDEr [47], and SPICE [48]. With higher score of the above evaluation indicators, the generated sentence is closer to the reference sentence. The value ranges of BLEU, METEOR, ROUGE-L, SPICE is from 0 to 1. The value ranges of CIDEr is from 0 to 5.

1) BLEU: BiLingual Evaluation Understudy (BLEU) is a commonly used evaluation for sentence generation tasks. By calculating the precision of n-gram of different lengths between generated and reference sentences, BLEU focuses on measuring the n-gram coincidence. In the captioning task, the value of n is set to be 1, 2, 3, and 4 corresponding to BLEU1–4.

2) METEOR: Metric for Evaluation of Translation with Explicit Ordering (METEOR) computes the single-precision weighted harmonic mean and single-word recall rate when comparing the generated and referenced sentences. It measures the degree of the alignment between them. Comparing with the BLEU, METEOR considers both precision and recall rate, which can solve some of the defects inherent to the BLEU.

3) ROUGE-L: Recall-Oriented Understudy for Gisting Evaluation (ROUGE) is introduced from the field of text summarization. Similar with BLEU, the ROUGE-L concentrates on calculating the recall rate. The F-measure based on the longest common subsequence (LCS) is adopted in the ROUGE-L metric to measure the similarity of the reference and generated sentences.

4) CIDEr: Consensus-based Image Description Evaluation (CIDEr) is a specialized evaluation metric for image captioning, which is one of the most important indicators for comparing model performance. In CIDEr, each sentence is regarded as a "document" and transformed into a term frequency inverse document frequency (TF-TDF) vector. The cosine similarity between the reference and generated sentences will be calculated by a specialized evaluation model. Compared with the aforementioned evaluation metrics, the CIDEr considers the semantic correlation between the generated sentences and reference sentences to some extent. This makes the CIDEr evaluation indicator more convincing.

5) SPICE: Semantic Propositional Image Caption Evaluation (SPICE) is another specialized evaluation metric for image captioning, which measures how effectively image captions recover objects, attributes, and the relationships between them. In SPICE, the reference and generated sentences are represented with a syntactic dependency tree through a Probabilistic Context-Free Grammar (PCFG) dependency parser. According to the dependency tree, the sentence will be transformed into a scene graph and F-score will be measured between the objects, attributes, and the relationships contained in caption.

C. Ablation Studies

The ablation studies are designed for verifying the effectiveness of three modules: 1) hierarchical encoding strategy; 2) deformable scaled dot-product attention; 3) deformable pooling. The ablation studies are performed on all four datasets: 1) the UCM-Captions dataset, 2) the Sydney-Captions dataset, 3) the

TABLE I
COMPARISON OF MODEL PERFORMANCE UNDER DIFFERENT BASIC FRAMEWORKS

Frameworks	BLEU-4	METEOR	ROUGE-L	CIDEr
Transformer	0.620	0.430	0.750	2.701
obj+Transformer	0.616	0.437	0.746	2.689
seg+Transformer	**0.665**	**0.454**	**0.786**	**3.036**

* *obj* refers to the object detector as the feature extractor, and *seg* refers to the semantic segmentation model as the feature extractor.

TABLE II
COMPARISON OF MODEL PERFORMANCE UNDER THE SEMANTIC MAP WITH DIFFERENT DIALATION TIMES

Dilation	BLEU-4	METEOR	ROUGE-L	CIDEr
3	**0.615**	**0.443**	**0.767**	**2.769**
5	0.584	0.410	0.714	2.465
7	0.524	0.411	0.743	2.529
9	0.574	0.419	0.732	2.585
∞	0.553	0.386	0.707	2.566

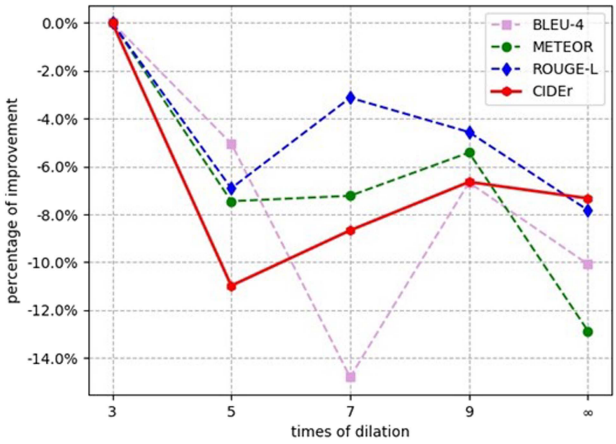

Fig. 5. Performance curve of the model under the semantic map with different dialation times.

RSICD dataset, and 4) the NWPU-Captions. For brevity, we only report results on the Sydney-Captions, because our method behaves similarly on all four datasets. For the simplicity of the discussion, we only compare BLEU-4, METEOR, ROUGE-L, and CIDEr evaluation indicators in the ablation experiments.

1) Hierarchical Encoding Strategy: We remove the hierarchical encoding strategy and design two experiments to demonstrate the effectiveness of our hierarchical idea. First, the first experiment is to directly replace the original CNN–RNN frame with Transformer. The input is the feature map of 14×14, which is flattened into a sequence of 196 and directly input into the Transformer. This experiment shows the effect that if we do not separate the foreground from the background, but directly use the Transformer, so as to reflect the necessity of separating the foreground and background. The second experiment is to directly relocate the framework in natural image captioning and input the object's features extracted by a pretrained object detector into Transformer. This experiment is to illustrate the necessity of using image segmentation in combination with the specificity of remote sensing images. Table I shows the improvement, when we use our hierarchical strategy. In summary, the method of directly relocating image captioning framework in natural scenes is not ideal. The object detector: Faster RCNN we used is pretrained on the MSCOCO dataset [49]. Before passing through nms [18], we saved the feature of each region proposals after RoI pooling as the object's feature. Similar to using the CNN pretrained on ImageNet, this method can retain as much image information as possible. We do not recommend using the result after nms, which leads to poor model performance because most of the information is abandoned. A single sample has an average of about 150 region proposals. Directly transforming the object detector framework does not bring much improvement to the model. Compared with using Transformer directly on the feature map, the improvement is insignificant (BLEU-4: -0.4%, METEOR: $+0.7\%$, ROUGE-L: -0.4%, CIDEr: -1.2%). Therefore, it is necessary to redesign the model considering the characteristics of remote sensing images. Our hierarchical encoding strategy performs better. The hierarchical encoding strategy improves the accuracy with a large margin in terms of all evaluation metrics (BLEU-4: $+4.5\%$, METEOR: $+2.4\%$, ROUGE-L: $+3.6\%$, CIDEr: $+33.5\%$).

2) Deformable Scaled Dot-Product Attention: Deformable attention is designed for constraining the feature interaction among pixels within the object's shape. For verifying the effectiveness of this operator, we design experiments to gradually remove the limitation, allowing attention to be computed beyond the structure of object. Specifically, our method is to use the morphological processing to expand the accurate mask, so that it can gradually cover the surrounding irrelevant environmental elements. Our dilation operator is a cross-shaped structure with the size of 3×3, and the original mask is expanded 3, 5, 7, and 9 times to observe the impact on model performance. The extreme case is to degenerate deformable attention into self-attention to consider the whole picture. Table II shows the results. Fig. 5 shows the curve of the evaluation indicators as the times of expansion increases. In summary, after the dilation operation, as the noise continues to increase, the ability of the foreground and background perception function to understand image gradually deteriorates, which in turn leads to a decrease in the accuracy of the generated sentence. When the dilation reaches a certain level, thanks to the robustness of the dot-product attention, the model performance gradually reaches a stable lower limit.

In summary, the purpose of deformable attention is to learn the pixel-to-pixel interactions inside the object's structure, while the pixels outside the object naturally become noise. As shown in Table II, after the segmentation mask is removed, the deformable attention degenerates into a traditional self-attention model, and the performance of the model drops significantly. The reason is that after the segmentation mask disappears, the interactive learning within the object introduces external noise.

3) Deformable Max/Average Pooling: Pooling is to aggregate the features of each pixel in the object's shape after the deformable attention feature interaction and turn it into one feature vector representing a foreground target. For this module, we mainly make a comparison between max pooling and average pooling to find which operator performs better under our framework. Table III shows the results. In terms of results,

TABLE III
COMPARISON OF MODEL PERFORMANCE UNDER THE SEMANTIC MAP WITH DIFFERENT DIALATION TIMES

Methods	BLEU-4	METEOR	ROUGE-L	CIDEr
Avg pooling	0.611	0.441	0.776	2.887
Max pooling	**0.665**	**0.454**	**0.786**	**3.036**

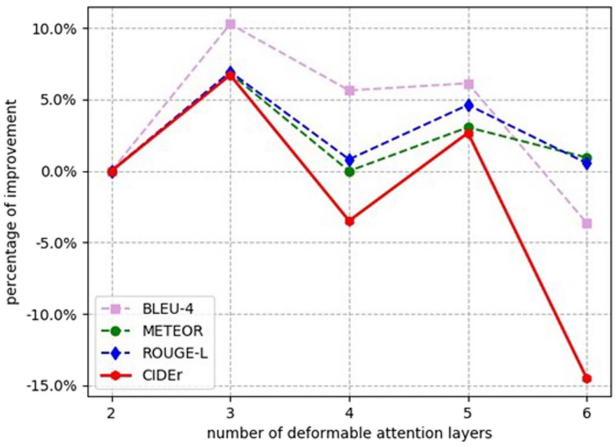

Fig. 6. Performance curve of the model under different number of deformable attention layers.

the max pooling has improvement compared to average pooling (BLEU-4: +5.4%, METEOR: +1.3%, ROUGE-L: +1%, CIDEr: +14.9%). We conjecture the reason is due to the particularity of our framework, the role of deformable attention layers is similar to the multilayer convolution in CNN, which needs to understand the image. Compared with avg pooling, max pooling is stronger for highlighting the difference in attribute of one object.

D. Parameter Analysis

Both our deformable attention and the original self-attention can be stacked with multiple times to enhance model's learning ability. However, there is a valuable problem we want to explore in this section: How many layers would we stacked are appropriate to benefit model performance?

Therefore, we first fix the number of self-attention layer, and constantly increase the number of deformable attention layer to observe the result. Specifically, we set the number of deformation attention layers to 2, 3, 4, 5, 6 coupled with six self-attention layers. We show the curve of the evaluation index on Fig. 6. With the increase of deformable attention layers, the CIDEr score first increases and achieves the highest point when we use three layers. When the number of layers exceeds three, the model appears overfitting problem but still has strong general ability.

After discussing the change in the number of deformable attention layers, we continue to adjust the number of self-attention layers to observe its impact. We fixed the deformable attention layers to 3 and set the layers of self-attention to 2, 3, 4, 6, 8, respectively. We also show the curve of the evaluation indicators in Fig. 7. In comparison, when the self-attention layers are set to 6, the model performance is best. There exists an interesting case: When the number of self-attention layers is a multiple of

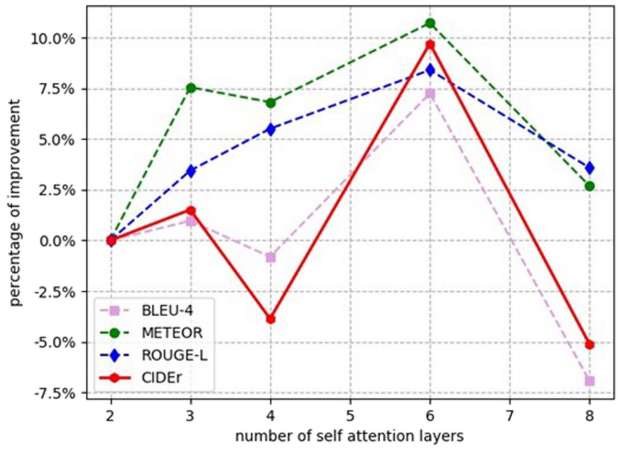

Fig. 7. Performance curve of the model under different number of self-attention layers.

3, its performance is often better than a multiple of 2. Besides, we suppose that self-attention is more important to improve the model ability than our deformable attention. However, it is not true. In the case of underfitting, when we upgrade the layer number from 2 to 3, its CIDEr improvement is only 4% points (deformable attention increased by 23% points). In summary, we believe that the number of self-attention layers does not have a decisive impact on the performance of the model. Meanwhile, in order to ensure the versatility of the model on different-scale datasets, we finally adopt a 6-layer self-attention.

E. Comparison With Other Methods

In this section, we evaluate our method on four datasets and compared our best model performance with a variety of recent captioning methods. The comparison methods include the CNN + RNN [9], [20], [50], ConvCap [51], Soft-attention [11], Hard-attention [11], CSMLF [1], RTRMN [52], structure attention [17], GVFGA+LSGA [16], SVM-D CONC [21], Word _ sentence [23], MLCA-Net [25], RASG [15], CNN-T [54], SCAMET [53]. Since the reinforcement learning of image captioning in natural scenes will greatly improve the CIDEr score (verified in natural scenes). Therefore, for the fairness of the comparison, we only use CE loss to train our framework, and do not use reinforcement learning. In this comparison, we also do not compare against any method that uses reinforcement learning.

1) CNN+RNN: CNN+RNN uses the VGG-16 [5] as the encoder and RNN-based model as the decoder. In this series, naïve RNN [9], LSTM [9], GRU [50], and GRU-embedword [20] are used, respectively. We report the highest score given by GRU-embedword for comparison.

2) ConvCap: The ConvCap [51] employs the VGG-16 as the encoder and also equips a CNN-based decoder.

3) Soft-Attention and Hard-Attention: They use the VGG-16 as the encoder and integrate the "soft" and "hard" attention mechanism [11] with the LSTM [8] decoder. We choose the "soft" attention method with better performance for reporting.

TABLE IV
COMPARISON OF MODEL PERFORMANCE WITH OTHER STATE-OF-THE-ARTS IN SYDNEY-CAPTIONS DATASET

Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
CNN+RNN [20]	0.6885	0.6003	0.5181	0.4429	0.3036	0.5747	1.6894	0.3701
ConvCap [51]	0.7472	0.6512	0.5725	0.5012	0.3476	0.6674	2.1484	0.3917
Soft-attention [11]	0.7322	0.6674	0.6223	0.5820	0.3942	0.7127	2.4993	-
CSMLF [1]	0.5998	0.4583	0.3869	0.3433	0.2475	0.5018	0.7555	
Structure attention [17]	0.7795	0.7019	0.6392	0.5861	0.3954	0.7299	2.3791	-
Word_Sentence [23]	0.7891	0.7094	0.6317	0.5625	0.4181	0.6922	2.0411	-
SVM-D CONC [21]	0.7547	0.6711	0.5970	0.5308	0.3643	0.6746	2.2222	-
GVFGA+LSGA [16]	0.7681	0.6846	0.6145	0.5504	0.3866	0.7030	2.4522	0.4532
MLCA-Net [25]	0.8310	0.7420	0.6590	0.5800	0.3900	0.7110	2.3240	0.4090
RASG [15]	0.8000	0.7217	0.6531	0.5909	0.3908	0.7218	2.6311	0.4301
CNN-T [54]	0.8220	0.7410	0.6620	0.5940	0.3970	-	2.7050	-
SCAMET [53]	0.8072	0.7136	0.6431	0.5846	**0.4614**	0.7218	2.3570	-
Ours	**0.8373**	**0.7771**	**0.7198**	**0.6659**	0.4548	**0.7860**	**3.0369**	**0.4839**

TABLE V
COMPARISON OF MODEL PERFORMANCE WITH OTHER STATE-OF-THE-ARTS IN UCM-CAPTIONS DATASET

Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
CNN+RNN [20]	0.7574	0.6983	0.6451	0.5998	0.3685	0.6674	2.7924	0.4147
ConvCap [51]	0.7034	0.5647	0.4624	0.3857	0.2831	0.5962	1.9015	0.2948
Soft-attention [11]	0.7454	0.6545	0.5855	0.5250	0.3886	0.7237	2.6124	-
CSMLF [1]	0.3671	0.1485	0.0763	0.0505	0.0944	0.2986	0.1351	-
RTRMN [52]	0.8028	0.7322	0.6821	0.6393	0.4258	0.7726	3.1270	0.4535
Structure attention [17]	**0.8538**	**0.8035**	**0.7572**	**0.7149**	0.4632	0.8141	3.3489	-
Word_Sentence [23]	0.7931	0.7237	0.6671	0.6202	0.4395	0.7132	2.7871	-
SVM-D CONC [21]	0.7653	0.6947	0.6417	0.5942	0.3702	0.6877	2.9228	-
GVFGA+LSGA [16]	0.8319	0.7657	0.7103	0.6596	0.4436	0.7845	3.3270	0.4853
MLCA-Net [25]	0.8260	0.7700	0.7170	0.6680	0.4350	0.7720	3.2400	0.4730
RASG [15]	0.8518	0.7925	0.7432	0.6976	0.4571	0.8072	3.3887	**0.4891**
CNN-T [54]	0.8390	0.7690	0.7150	0.6750	0.4460	-	3.2310	-
SCAMET [53]	0.8460	0.7772	0.7262	0.6812	**0.5257**	**0.8166**	3.3772	-
Ours	0.8230	0.7700	0.7228	0.6792	0.4439	0.7839	**3.4629**	0.4825

4) CSMLF: It is a retrieval-based method that uses semantic embedding to measure the similarity between input image representation and the candidate sentence representation.

5) RTRMN: RTRMN [52] uses ResNet-101 as the encoder and adds topic information to guide the caption generation.

6) Structure Attention: Structure attention [17] uses the CNN+RNN framework combined with the structure-aware pooling to improve the attention mechanism.

7) GVFGA+LSGA: It supplies the missed global visual feature in the encoder and adds the linguistic state to guide the attention process.

8) SVM-D CONC: It replaces the commonly used RNN decoder with multiple SVMs to generate sentence.

9) Word _ Sentence: It abandons the CNN–RNN framework and employs a two-step sentence generator. Firstly, the content in the image is changed into several independent words, and then the words are connected into sentences.

10) MLCA-Net: It uses a multilevel attention module to adaptively aggregate visual features on the encoder side. For decoder, it introduces a contextual attention module to explore latent context.

11) RASG: It introduces a recurrent attention mechanism to improve the context vector and uses a semantic gate for more precise semantic understanding.

12) CNN-T: CNN-T uses a multiscale feature extractor based on CNN and a transformer-based decoder for generating captionings.

13) SCAMET: SCAMET constructs a multiattention encoder from CNN visual features, which further proceeded to memory-guided Transformer model.

Tables IV–VII report the accuracy of our method and the above methods on the four different datasets. As shown in the table, our method outperforms the abovementioned well-known methods published in excellent journals on most remote sensing datasets. On Sydney-captions, our method achieves state of the art. On the UCM-captions, our model has the highest CIDEr score compared to previous models. On the RSICD datset, our model's performance is higher than most of the methods but sightly lower than RASG [15]. BLEU1-4, METEOR, ROUGE-L of our model outperforms the GVFGA+LSGA [16] method. On the NWPU-captions dataset, BLEU1-4 of our model is higher than MLCA-Net [25].

F. Qualitative Analysis

1) Caption Generation Results: As shown in Fig. 8, we extracted several caption results generated from our model in RSICD, CNN+RNN, and basic Transformer. For comparison, we show four caption results: 1) CNN + RNN (SA); 2) basic Transformer (Trans); 3) our model; 4) ground-truth (GT). According to Fig. 8, compared with the previous method, the caption generated by our method is more accurate in describing the object and has a lower error rate. Specifically, note the 2-th image in Fig. 8, the foreground (red line) is hidden in the

TABLE VI
COMPARISON OF MODEL PERFORMANCE WITH OTHER STATE-OF-THE-ARTS IN RSICD DATASET

Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
CNN+RNN [20]	0.6885	0.6003	0.5181	0.4429	0.3036	0.5747	1.6894	0.3477
ConvCap [51]	0.6336	0.5103	0.4174	0.3452	0.3325	0.5770	1.6648	0.3933
Soft-attention [11]	0.6753	0.5308	0.4333	0.3617	0.3255	0.6109	1.9643	-
CSMLF [1]	0.5106	0.2911	0.1903	0.1352	0.1693	0.3789	0.3388	-
RTRMN [52]	0.6201	0.4623	0.3644	0.2971	0.2829	0.5539	1.5146	0.3322
Structure attention [17]	0.7016	0.5614	0.4648	0.3934	0.3291	0.5706	1.7031	-
Word_Sentence [23]	0.7240	0.5861	0.4933	0.4250	0.3197	0.6260	2.0629	-
SVM-D CONC [21]	0.5999	0.4347	0.3355	0.2689	0.2299	0.4557	0.6854	-
GVFGA+LSGA [16]	0.6779	0.5600	0.4781	0.4165	0.3285	0.5929	**2.6012**	**0.4683**
MLCA-Net [25]	0.7570	0.6340	0.5390	0.4610	0.3510	0.6460	2.3560	0.4440
Ours	**0.7581**	**0.6416**	**0.5585**	**0.4923**	**0.3550**	**0.6523**	2.5814	0.4579
RASG [15]	0.7729	**0.6651**	**0.5782**	**0.5062**	0.3626	0.6691	**2.7549**	**0.4719**
CNN-T [54]	**0.7980**	0.6470	0.5690	0.4890	0.2850	-	2.4040	-
SCAMET [53]	0.7681	0.6309	0.5352	0.4611	**0.4572**	**0.6979**	2.4681	-

TABLE VII
COMPARISON OF MODEL PERFORMANCE WITH OTHER STATE-OF-THE-ARTS IN NWPU-CAPTIONS DATASET

Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
CNN+RNN [20]	0.7390	0.6170	0.5320	0.4680	0.3300	0.5930	1.2360	0.2760
Soft-attention [11]	0.7310	0.6090	0.5250	0.4620	0.3390	0.5990	1.1360	0.2850
CSMLF [1]	0.7170	0.5900	0.5040	0.4400	0.3200	0.5780	1.0650	0.2650
MLCA-Net [25]	0.7450	0.6240	0.5410	0.4780	**0.3370**	**0.6010**	**1.2640**	**0.2850**
Ours	**0.7515**	**0.6291**	**0.5457**	**0.4828**	0.3187	0.5858	1.2071	0.2678

background (ocean). In this case, it is very challenging and lacks interpretability for the SA to generate sentences containing red lines. In fact, the description generated by SA doesn't contain redline information. However, our model is able to effectively extract this hidden foreground and embody it in the generated sentences. Meanwhile, the sufficiency of the sentence also meets the requirements. These improvements in sentences all benefit from the fact that our deformable attention can reduce the interference of external noise in the process of feature interactive learning. The similar situations are present in 4-th and 5-th sample shown in Fig. 8. Specifically, the "runway" is in the corner of the image, which is difficult to be identified without prior. As a result, both SA and $Trans$ are missing this object in the generated sentence. In the 5-th sample, the "wide river" is missing because its a rare sample in the training set about "residential area". With the prior we extracted, our model has the stronger generalization ability. The other samples presented are also proved the superiority of our model. In the 1-th and 3-th samples, the missing "buildings" and "river" are detected by our method. When describing the frequent appeared case like "house" shown in 6-th sample of Fig. 8, our model also has the satisfactory result.

2) Visualization of the Attention: Fig. 9 shows the feature interaction learning process of a single aircraft object in the image when the model generates the word "airplane". The Soft attention (SA) method is to make a global feature weighted sum on raw feature map. Due to the shallow depth of the attention module and the large number of regions involved in the calculation, the actual allocation for all weights of aircraft regions is only 46%. The remaining 54% contains a large amount of irrelevant noise, which affects the performance to a certain extent for shallow depth modules.

In contrast, the interactive learning of our model is more effective and powerful. In the process of intra-class feature interaction, the weight allocation for one single connected domain of "airplane" achieves 100% due to the effect of deformable attention, which avoiding the influence of other irrelevant noise. In the subsequent process of inter-class feature interaction, the weight allocation of the relevant regions in the low-level layer reaches 68%(+22%). Compared with SA, our model is more effective on interactive learning.

The second sample shown in Fig. 9 also proves the effectiveness of our hierarchical encoding strategy and Deformable attention. The "river" in the image is an important object for describing and also present in the ground truth. However, the soft attention framework fails to attend river and can't generate the correct word (5-th sample in Fig. 8). The result is the lack of prior information. This problem is well offset with our framework by introducing the hierarchical encoding strategy with the segmentation prior. Meanwhile, the deformable attention further learns the interaction within the river's structure to provide the valid information.

V. CONCLUSION

A brand-new Deformable Transformer is proposed which is customized for remote sensing image captioning, instead of simply carrying the Transformer from the natural scene. Our framework doesn't require a very powerful pretrained feature extractor to support extremely accurate semantic map. A widely used semantic segmenter is satisfied to effectively drive our model. Relying on the powerful interactive learning and noise reduction capabilities of the deformable attention proposed in our paper, our model achieves better caption accuracy on most remote sensing datasets compared with previous methods. The visualization experiments also demonstrate the effectiveness of our framework. With the development of remote sensing technology, image quality becomes clearer and clearer is an

Fig. 8. Comparison of generated sentences among soft-attention (SA), Transformer (Trans), and our framework. Red words indicate the object that the rest of the baseline models ignore but our model pays attention to. GT is ground truth.

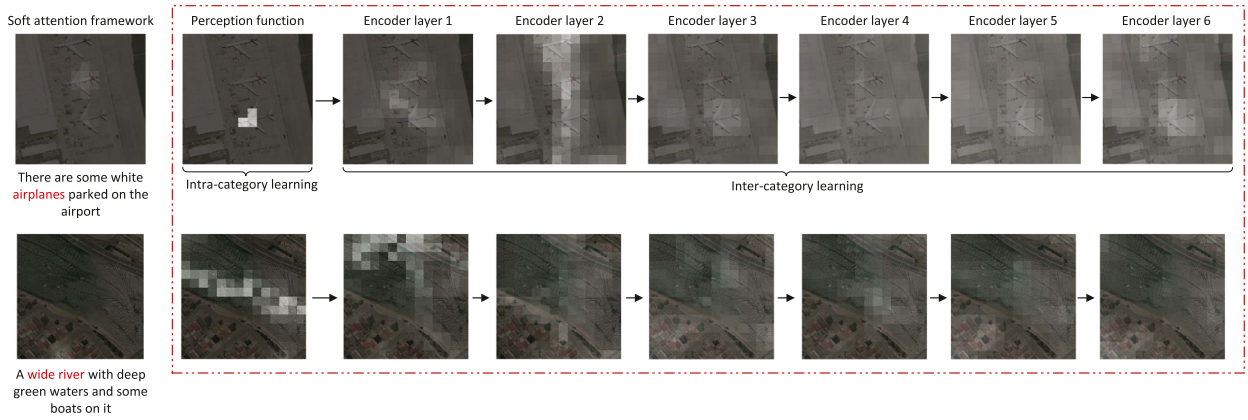

Fig. 9. Learning process of the specific object (airplane) by the attention mechanism of different models. The basic soft-attention only fuses several regions, and our framework can start from intraclass and interclass interactive learning to optimize the target features layer by layer.

inevitable trend. It is worth noting that higher quality images can further improve the performance of our model while bringing more accurate semantic maps.

The limitation of our model is mainly from the inaccuracy segmentation result and the limited representative ability of the image encoder. As a result, our future work is to combine our model with the full-transformer architecture. On the other hand, the fine-tuned segmentation model is also sufficient for improving current performance.

REFERENCES

[1] 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.

[2] 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.

[3] J. Long, E. Shelhamer, and T. Darrel, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3340.

[4] 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.*, 2015, pp. 3156–3164.

[5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

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

[7] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

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

[9] B. Qu, X. Li, D. Tao, and X. Lu, "Deep semantic understanding of high resolution remote sensing image," in *Proc. Int. Conf. Comput., Inf. Telecommun. Syst.*, 2016, pp. 1–5.

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

[11] 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.

[12] K. Xu et al., "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[13] 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, 2019, Art. no. 612.

[14] G. Sumbul, S. Nayak, and B. Demir, "SD-RSIC: Summarization-driven deep remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 1–13, Aug. 2021.

[15] Y. Li et al., "Recurrent attention and semantic gate for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022.

[16] Z. Zhang, W. Zhang, M. Yan, X. Gao, K. Fu, and X. Sun, "Global visual feature and linguistic state guided attention for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022.

[17] R. Zhao, Z. Shi, and Z. Zou, "High-resolution remote sensing image captioning based on structured attention," *IEEE Trans. Geosci.e Remote Sens.*, vol. 60, pp. 1–14, 2022, Art. no. 5603814, doi: 10.1109/TGRS.2021.3070383.

[18] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[19] A. Vaswani et al., "Attention is all you need," in *Neural Inf. Process. Syst.*, 2017.

[20] X. Lu, B. Wang, and X. Zheng, "Sound active attention framework for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 1985–2000, Mar. 2020.

[21] G. Hoxha and F. Melgani, "A novel SVM-Based decoder for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.

[22] V. Vapnik, *Statistical Learning Theory*. New York, NY, USA: Wiley, 1998, pp. 156–160.

[23] Q. Wang, W. Huang, X. Zhang, and X. Li, "Word–Sentence framework for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10532–10543, Dec. 2021.

[24] X. Li, X. Zhang, W. Huang, and Q. Wang, "Truncation cross entropy loss for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5246–5257, Jun. 2021.

[25] Q. Cheng, H. Huang, Y. Xu, Y. Zhou, H. Li, and Z. Wang, "NWPU-Captions dataset and MLCA-Net for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–19, 2022.

[26] X. Zhu, L. Li, J. Liu, H. Peng, and X. Niu, "Captioning transformer with stacked attention modules," *Appl. Sci.*, vol. 8, no. 5, 2018, Art. no. 739.

[27] X. Li et al., "Multilingual speech translation with efficient finetuning of pretrained models," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics, 11th Int. Joint Conf. Natural Lang. Process.*, vol. 1, pp. 827–838.

[28] J. Devlin, M. Chang, K. Kenton Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.

[29] J. Yu, J. Li, Z. Yu, and Q. Huang, "Multimodal transformer with multi-view visual representation for image captioning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 12, pp. 4467–4480, Dec. 2020.

[30] D. Lin, C. Kong, S. Fidler, and R. Urtasun, "Generating multisentence lingual descriptions of indoor scenes," *Computer Sci.*, pp. 2333–9721, 2015.

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

[32] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, Sep. 2013.

[33] J. Ba, J. Kiros, and G. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

[34] T. Pavlidis, "Filling algorithms for raster graphics," *Computer Graph. Image Process.*, vol. 10, no. 2, pp. 126–141, 1979.

[35] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Proc. 32 Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 8778–8788.

[36] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help ?," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 4694–4703.

[37] S. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1179–1195.

[38] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015.

[39] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[40] S. Z. Waqaset al., "iSAID: A large-scale dataset for instance segmentation in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 28–37.

[41] 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.

[42] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, 2010, pp. 270–279.

[43] 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.

[44] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2001, pp. 311–318.

[45] M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in *Proc. 9th Workshop Stat. Mach. Transl.*, 2014, pp. 376–380.

[46] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Proc. ACL Text Summarization Workshop*, 2004, pp. 74–81.

[47] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4566–4575.
[48] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: Semantic propositional image caption evaluation," in *Proc. Eur. Conf. Comput. Vis.*," 2016, pp. 382–398.
[49] X. Chen et al., "microsoft COCO captions: Data collection and evaluation Server," 2015, *arXiv:1504.00325*.
[50] X. Li, A. Yuan, and X. Lu, "Multi-modal gated recurrent units for image description," *Multimedia Tools Appl.*, vol. 77, no. 22, pp. 29847–29869, Nov. 2018.
[51] J. Aneja, A. Deshpande, and A. G. Schwing, "Convolutional image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5561–5570.
[52] B. Wang, X. Zheng, B. Qu, and X. Lu, "Retrieval topic recurrent memory network for remote sensing image captioning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 256–270, 2020.
[53] G. O. Gajbhiye and A. V. Nandedkar, "Generating the captions for remote sensing images: A spatial-channel attention based memoryguided transformer approach," *Eng. Appl. Artif. Intell.*, vol. 114, 2022, Art. no. 105076.
[54] U. Zia, M. M. Riaz, and A. Ghafoor, "Transforming remote sensing images to textual descriptions," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 108, 2022, Art. no. 102741.

Runyan Du received the bachelor's degree in electronic information engineering from Tianjin University, Tianjin, China, in 2019. He is currently working toward the Ph.D. degree in signal and information processing with the School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing, China.

His research interests include computer vision and multimodal learning.

Wei Cao received the master's degree in computer science and technology from the School of Computer and Cyber Sciences, Communication University of China, Beijing, China, in 2008. She is currently working toward the doctor's degree in signal and information processing with the School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing, China.

Her research interests include computer vision, deep learning, and image captioning.

Wenkai Zhang (Member, IEEE) received the B.Sc. degree from the China University of Petroleum, Qingdao, China, in 2013, and the Ph.D. degree from the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, in 2018.

He is an Assistant Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include multimodal signal processing, image segmentation, and pattern recognition.

Guo Zhi received the B.Sc. degree from Tsinghua University, Beijing, China, in 1998, and the M.Sc. and Ph.D. degrees from the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, in 2003.

He is a Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include computer vision, geospatial data mining, and remote sensing image understanding.

Xian Sun (Senior Member, IEEE) received the B.Sc. degree from the Beijing University of Aeronautics and Astronautics, Beijing, China, in 2004, and the M.Sc. and Ph.D. degrees from the Institute of Electronics, Chinese Academy of Sciences, Beijing, in 2009.

He is a Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include computer vision, geospatial data mining, and remote sensing image understanding.

Shuoke Li received the B.Sc. degree from Xi'an Jiaotong University, Xi'an, China, in 2017, and the M.Sc. degree from Peking University, Beijing, China, in 2020.

He is currently an Assistant Engineer with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research interests include multimodal signal processing, image segmentation, and pattern recognition.

Jihao Li received the B.Sc. degree from Xidian University, Xi'an China, in 2017, and the Ph.D. degree from Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China, in 2022.

He is currently an Assistant Researcher with the Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include computer vision and remote sensing image interpretation.