Visual Rotated Position Encoding Transformer for Remote Sensing Image Captioning

Anli Liu, Lingwu Meng, and Liang Xiao, *Senior Member, IEEE*

***Abstract*—Remote sensing image captioning (RSIC) is a crucial task in interpreting remote sensing images (RSIs), as it involves describing their content using clear and precise natural language. However, the RSIC encounters difficulties due to the intricate structure and distinctive features of the images, such as the issue of rotational ambiguity. The existence of visually alike objects or areas can result in misidentification. In addition, prioritizing groups of objects with strong relational ties during the captioning process poses a significant challenge. To address these challenges, we propose the visual rotated position encoding transformer for RSIC. First of all, rotation-invariant features and global features are extracted using a multilevel feature extraction (MFE) module. To focus on closely related rotated objects, we design a visual rotated position encoding module, which is incorporated into the transformer encoder to model directional relationships between objects. To distinguish similar features and guide caption generation, we propose a feature enhancement fusion module consisting of feature enhancement and feature fusion. The feature enhancement component adopts a self-attention mechanism to construct fully connected graphs for object features. The feature fusion component integrates global features and word vectors to guide the caption generation process. In addition, we construct an RSI rotated object detection dataset RSIC-ROD and pretrain a rotated object detector. The proposed method demonstrates significant performance improvements on four datasets, showcasing enhanced capabilities in preserving descriptive details, distinguishing similar objects, and accurately capturing object relationships.**

***Index Terms*—Image captioning, remote sensing, transformer, visual position encoding.**

I. Introduction

THE widespread availability of remote sensing technology has led to a growing focus on applications based on RSIs, including object segmentation [1], detection [2], and classification tasks [3], [4]. However, as the demand increases and new application scenarios are explored, these applications find it difficult to fully convey the semantic and content information of images [5].

In recent years, there has been significant progress in natural image captioning task, aiming to describe the visual content of images in natural language. This has inspired scholars to focus on RSIC task and achieved initial success.

Compared to natural images obtained from horizontal shooting, the RSIs obtained from top-down perspective have more complex scenes and objects. This poses the following challenges for RSIC.

1) The rotation ambiguity problem [13] in RSIs leads to the unclear of object orientation, making it difficult to obtain accurate semantic information.
2) The abundance of objects in RSIs poses a challenge in describing closely related groups of objects [38], [44]. For example, "ship," "port," and "ocean" have close conceptual and spatial relationships. Likewise, "building," "road," and "car" is also close. In contrast, objects like "ship" and "road" are not related.
3) The presence of visually similar objects or regions can potentially cause misidentification by the model, leading to inaccurate and inconsistent captions [44].

Moreover, detailed captions often depend on high-level vocabulary and structural information embedded in global image features. To mitigate these challenges, most methods [9], [10], [11], [12], [13], [14] extract image features from the last convolutional layer of a CNN [19]. However, these methods are unable to model fine-grained semantic relationships, resulting in low-quality sentences. Some methods [63] consider the multiscale problems, and there are also methods [64] that attempt to address these challenges by incorporating image segmentation. Recently, object-based methods [44] have demonstrated their potential on RSIC.

In this article, we propose a method to overcome these challenges by exploring the guiding role of global features in the process of generating descriptions from object features in RSIs, and relationships among rotated objects. First, we extract object-level rotation-invariant features with their corresponding position and orientation information and extract global features in multilevel feature extraction (MFE) module. Existing rotated object detection datasets have fewer classes, which is significantly different from the object classes in RSIC tasks. To accurately extract rotated object features for RSIC datasets, we propose a rotated object detection dataset RSIC-ROD based on NWPU-Captions [37] and pretrain a rotated object detector. Furthermore, we design a feature enhancement fusion

Received 6 May 2024; revised 2 September 2024; accepted 23 October 2024. Date of publication 29 October 2024; date of current version 15 November 2024. This work was supported in part by the Jiangsu Provincial Frontier Technology Research and Development Program under Grant BF2024070 and in part by the National Science Foundation of China under Grant 62471235. *(Corresponding author: Liang Xiao.)*

Anli Liu and Lingwu Meng are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: liuanli@njust.edu.cn; menglw815@njust.edu.cn).

Liang Xiao is with the Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education and the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: xiaoliang@mail.njust.edu.cn).

The code is available at https://github.com/AnliLiu/VRoPE.

Digital Object Identifier 10.1109/JSTARS.2024.3487846

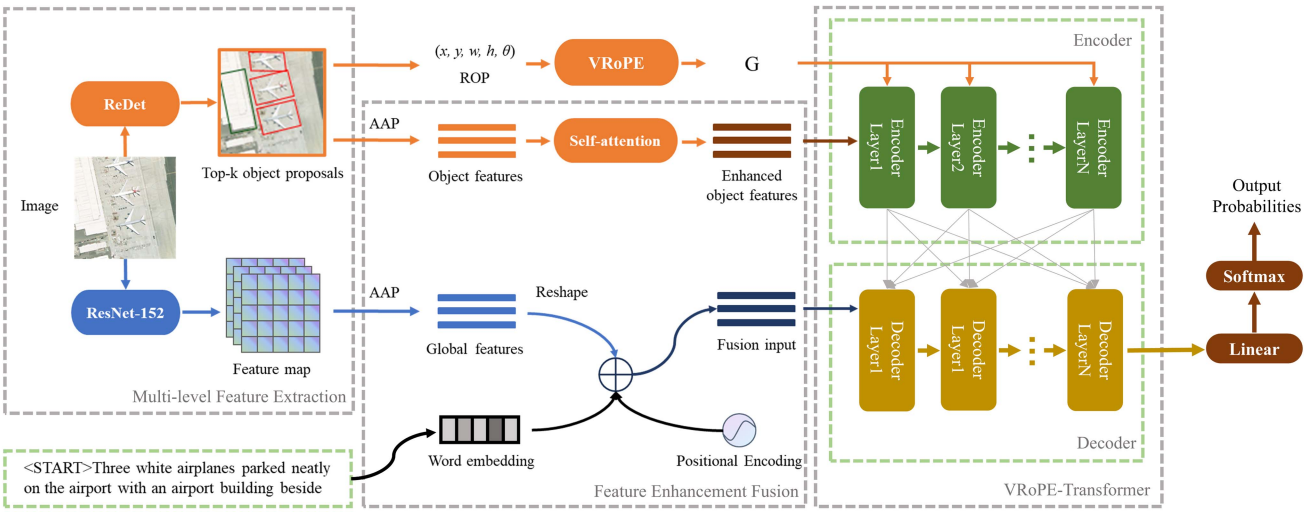

Fig. 1. Framework of the proposed method in this article. For image representation, the MFE module employs ResNet-152 and Redet to extract global features and object features from an input image. The FEF module uses self-attention to enhance the internal relationships between object features and fuses the global feature with word embeddings and positional encoding, which is employed as the input to the decoder. Then, we obtain a VRoPE module through operating of object features" rotated positions. The enhanced object features serve as the input to the first encoder layer. Finally, the VRoPEA incorporates the VRoPE $\mathbf{G}$ into self-attention, and it is hired in the transformer encoder. $\oplus$ and AAP, represent addition operator and adaptive average pooling.

(FEF) module that consists of feature enhancement and feature fusion components. Follow [28], feature enhancement component refines object–object relationships by a graph attention network [20]. The feature fusion component establishes the relationship between global features and text features to generate accurate captions.

In addition, we propose a visual rotated position encoding (VRoPE) module that encodes the spatial relationships of rotated objects. Then, the relationships are incorporated into self-attention of transformer [9], [10], [11] to regulate semantic closeness and encourage the generation of more reasonable and contextually relevant image captions.

Integrating the concepts outlined above, we propose a VRoPE-Transformer for RSIC. As shown in Fig. 1, VRoPE-Transformer comprises an MFE module, an FEF module, and a transformer equipped with a VRoPE module. In the MFE module, object-level rotation-invariant features and their corresponding position and orientation information are first extracted using a pretrained *REDET* [1]; global features are extracted using a pretrained *Resnet-152* [19]. The FEF module enhance rotated object features and text features. The enhanced rotated object features are fed into the transformer with VRoPE module to generate captions.

The main contributions of this article are as follows.

1) To better model the relationships among rotated objects, we propose the visual rotated position encoding (VRoPE) module, which integrates the position and geometric information of the objects into self-attention mechanisms.
2) To generate more comprehensive and detailed captions and to differentiate similar objects, we introduce the FEF module. This module integrates global features and textual information to guide the generation of object feature captions, while also utilizing self-attention mechanisms to enhance the relevance and distinctiveness of the objects.
3) To improve the identification of rotated objects in RSIC datasets, we have constructed the RSIC-ROD dataset, which includes 15 scenes with a diverse range of rotated objects, and have annotated 32 object categories.
4) Extensive experiments on four RSIC datasets have been conducted to validate the superior performance of our proposed method.

II. Related Work

In this section, we delve into the related research in two key areas: RSIC and visual position encoding.

A. Remote Sensing Image Captioning

RSIC has witnessed prominent evolution in technology. Initial attempts at image captioning primarily relied on template-based [21], [22], [23] and retrieval-based [24], [25], [26] methods. Retrieval-based approaches involved selecting the most similar sentence or a collection of highly related sentences from a predefined pool based on the given image, whereas template-based approaches generated sentences by populating predefined slots with detected visual elements. However, these methods often suffered from limited adaptability and heavy reliance on manually crafted features, which hindered the generation of high-quality captions.

With the advent of deep learning, substantial progress has been made in neural network-based methods [8], [9], [11]. The majority of these methods have adopted the encoder–decoder framework [15], [17]. Vinyals et al. [27] were the pioneers in introducing this framework, utilizing a CNN [28] for encoding to abstract high-level visual features, while an RNN [29] is used for decoding to produce captions. Following this, Xu et al. [30] subsequently enhanced an decoder based on LSTM by

incorporating hard and soft attention mechanisms which allowed it to concentrate on various areas of the image.

In the context of RSIC, Shiet al. [31] highlighted the challenge of capturing a range of objects across various scales, and conveying their characteristics and relational statuses. Lu et al. [13] raised issues such as scale, category, rotation ambiguities in RSIC and contributed the largest RSICD dataset to the field. This underscored the importance of accurate feature extraction and fusion for this task. Early attempts at RSIC relied on pretrained CNNs to extract pixel-level features. However, these methods were limited in their ability to capture semantic information effectively. Anderson et al. [9] made a significant advancement by introducing bottom-up and top-down attention mechanisms, leveraging Faster R-CNN [32] to extract features at the object level. This approach elevated the visual features in attention from the levels of pixels and grid to the levels of prominent region and object, greatly enhancing the descriptive power of captions. To fuse features from different levels. Zhang et al. [33] introduced attribute attention, which combines high-level features derived from deeper fully connected (FC) layers with low-level features from shallower convolutional or softmax layers. This approach enables the model to encapsulate both detailed and abstract information, leading to more accurate and informative captions. Other researchers have explored different strategies for feature fusion. Ma et al. [34] extracted object-level features from convolutional layers of *VGG-16* [35] and scene-level features using the *ResNet-50*. By combining these features, they achieved better performance in RSIC. Wang et al. [36] proposed a two-phase multiscale structure representation approach, collecting features from the *conv4* and *conv5* layers and using self-attention and gated cross-attention for multilevel CNN feature interaction. This approach enhanced the object discrimination capability of the model and improved the accuracy of descriptions. Cheng et al. [37] introduced MLCA-Net, which dynamically fuses image features from distinct spatial locations and across various scales. This network enabled the model to focus on relevant regions and scales, leading to more accurate and detailed captions. Zhang et al. [38] furthered this research by introducing GVFGA and LSGA mechanisms. These mechanisms allowed the model to attend to both semantic and visual information, improving the quality of captions.

In recent years, the attention-based methods have experienced rapid progress, with the transformer model emerging as a popular choice for RSIC. Chen et al. [39] utilized a multiscale vision transformer (ViT) for encoding images, while introducing a decoder of transformer for sentence generation. Liu et al. [40] further advanced this line of research by introducing the MLAT, which integrated the strengths of both transformer and LSTM [41]. In addition, Herdade et al. [42] innovated by incorporating geometric relationships among regional features into the transformer framework for generating captions. More recently, Cornia et al. [43] introduced trainable prior information to enhance the attention mechanism in the encoder of transformer, establishing comprehensive connections between each decoder layer and encoder layer through a grid structure. Building upon these advancements, Meng et al. [44] proposed a prior knowledge enhanced attention module that establishes relationships between objects to select those more relevant to the scene area.

While the previously mentioned methods can produce grammatically descriptions, they usually overlook the consequences of rotation ambiguity and may struggle with creating more nuanced captions. To preserve the positional and directional data of features as thoroughly as possible for later stages of processing, we incorporate rotation-invariant attributes, along with their respective rotated coordinates as input.

B. Visual Position Encoding

As deep learning technology has progressed, a growing body of research has concentrated on empowering models to effectively harness position information within images. Initial approaches primarily involved embedding position information into network architectures through hard-coding. CoordConv [45] is a technique that infuses position information into CNNs by representing absolute or relative coordinates via supplementary channels in the input feature maps. Wang et al. [46] developed Axial-DeepLab, a CNN model for semantic segmentation that leverages an axial attention mechanism in conjunction with axial absolute position encoding. This enables the network to gain a more nuanced understanding of semantic information at various positions within an image.

Then, Dosovitskiy et al. [47] introduced visual position encoding and proposed the ViT model, which divides images into fixed-size patches and uses trainable position embeddings to encode the position information of each image patch, embedding this information into the image's feature representation. Wu et al. [48] studied the forms of visual position encoding and how they are embedded into the image's feature representation. Graham et al. introduced LeViT [49], a vision model that uses a trainable position encoding method to improve inference speed on image classification tasks. Wang et al. [46] introduced the Axial-DeepLab model, which uses an axial attention mechanism and axial position encoding to better handle the semantic content at various locations within the image. Guo et al. [50] proposed a geometry-aware self-attention mechanism capable of capturing the relative geometry relationships among objects, guiding the development of the RSTNet [8].

However, most of the aforementioned position encodings have been applied to grid features rather than objects, lacking consideration for rotation and neglecting the importance of direction and positional relationships between objects. Therefore, we propose a novel rotated position encoding that introduces directional information and acts on rotated objects, integrating the orientation relationships between objects into the Transformer architecture to direct the model in focusing on content with higher contextual relevance, thereby producing more contextually appropriate captions.

III. Proposed Methodology

A. Overall Framework

The overall architecture of the proposed VRoPE-Transformer approach is shown in Fig. 1. It includes an MFE module, an FEF module, as well as a transformer equipped with a VRoPE module. We first present the MFE module in Section III-B, which extracts global features $\mathbf{V}^g$ and object features $\mathbf{V}^o$ from

the RSI $\mathbf{I}$, along with their corresponding rotated positions $\mathbf{P}$. Then, in Section III-C, we demonstrate the feature fusion enhancement (FFE) module. It enriches object features $\mathbf{V}^o$ by leveraging relationships between objects to obtain enhanced object features $\mathbf{V}^e$, and fuses the global feature $\mathbf{V}^g$ with word embeddings $\mathbf{W}$ to create an enhanced decoder input $\mathbf{Y}$. Next in Section III-D, we describe the VRoPE $\mathbf{G}$, which is obtained through the operation of object features" rotated positions $\mathbf{P}$. In Section III-E, we introduce the implementation of the VRoPE-based transformer designed by us, which uses the VRoPE and enhanced object features $\mathbf{V}^e$, combined with the global feature, to generate captions. In particular, every encoding layer within the transformer incorporates a visual rotated position encoding attention (VRoPEA) module that merges the VRoPE $\mathbf{G}$ into the multihead self-attention mechanism. Then, the decoder forecasts new words by utilizing the output of the encoder along with previously produced words. Finally, we established a dataset, RSIC-ROD.

B. Multilevel Feature Extraction

To enhance the semantic representation of RSI, feature extraction is divided into two components: global feature extraction and object feature extraction. Global features are crafted to encapsulate the overall structural and high-level semantic content of the imagery, providing a comprehensive contextual overview of the scene. In contrast, object features focus on the distinct details of each entity within the RSI, capturing the fine-grained information that is critical for identifying each object's characteristics and attributes.

1) Global Feature Extraction: To obtain a global feature representation that includes more advanced semantic information and capture relationships between objects and the overall structure of the scene, we leverage the powerful feature extraction capabilities of CNNs. We use the deep features extracted by CNNs as the depiction of RSIs, serving for the subsequent generation of captions. Among them, *ResNet-152* [20] is a typical CNN architecture that excels at extracting multiscale features and is, thus, well-suited for RSIs. Therefore, we choose the backbone of pretrained *ResNet-152*, denoted as $\mathbf{ResNet}_{\text{avgpool}}$, where *avgpool* is the last adaptive average pooling layer, to serve as the CNN feature extractor. The result after the *AAP* operation is used as the global feature. This yields the feature representation $\mathbf{V}^g = \{\mathbf{v}_1^g, \ldots, \mathbf{v}_N^g\}$, with the shape of $\mathbf{R}^{N\times D_1}$. This process can be represented as

$$\mathbf{V}^g = \mathbf{ResNet}_{\text{avgpool}}(\mathbf{I}) \tag{1}$$

where $\mathbf{I} \in \mathbf{R}^{H_i\times W_i\times C_i}$ is the input image.

2) Object Feature Extraction: To ensure consistent categorization across varying orientations, we utilize the backbone of *ReDet* [1] to extract rotation-invariant features along with their corresponding rotated positions $\mathbf{P}$. Trained on our RSIC-ROD dataset, we employ RRoI features derived from the Rotation-invariant RoI Align operation as the representative object features. We select the top $\mathbf{M}$ objects, and their feature vectors are denoted as $\mathbf{V}^o = \{\mathbf{v}_1^o, \ldots, \mathbf{v}_M^o\}$, forming a matrix of size $\mathbf{R}^{M\times D_2}$. The extraction process is formulated as follows:

$$\mathbf{V}^o = \mathbf{Redet}_{\text{RROIAlign}}(\mathbf{I}). \tag{2}$$

For each object, the result after performing bounding box regression is used as the rotated position $\mathbf{P}$, represented as $\mathbf{P} = \{\mathbf{p}_1(x_1, y_1, w_1, h_1, \theta_1), \ldots, \mathbf{p}_M(x_M, y_M, w_M, h_M, \theta_M)\}$, with a shape of $\mathbf{R}^{M\times 5}$. This process can be represented as

$$\mathbf{P} = \mathbf{Redet}_{\text{bboxRegression}}(\mathbf{I}). \tag{3}$$

C. Feature Enhancement Fusion

To enhance the distinguish ability of similar features and guide caption generation using global features rich in advanced semantic information, we propose the FFE module. It includes two parts: feature enhancement and feature fusion.

1) Feature Enhancement: Since similar object features may be extracted for different objects, we need to enhance the differences between similar features. The self-attention mechanism can adjust each feature during the learning process, ensuring the model's heightened focus on distinguishing subtle differences between similar features. By adding a self-attention module to enhance the input, we improve the distinguish ability of different object features in the encoder.

First, we apply dot product to each pair of object features $\mathbf{v}_i^o$ and $\mathbf{v}_j^o$ to represent the relationship between object features, resulting in attention weights $\mathbf{e}_{ij}^o$

$$e_{ij}^o = \text{softmax}_j\left(\mathbf{v}_i^o\mathbf{W}_q^o\left(\mathbf{v}_j^o\mathbf{W}_k^o\right)^{\mathbf{T}}/\sqrt{d}\right) \tag{4}$$

where $\mathbf{W}_k^o \in \mathbf{R}^{d\times d}$ and $\mathbf{W}_q^o \in \mathbf{R}^{d\times d}$ refer to the trainable parameter matrices; d refers to the scaling factor.

Thus, each feature is enhanced based on its relationships with other features to obtain the enhanced feature $\mathbf{v}_i^e$

$$\mathbf{v}_i^e = \sum_{j\in N_i} e_{ij}^o\left(\mathbf{v}_j^o\mathbf{W}_v^o\right) \tag{5}$$

where $\mathbf{W}_v^o \in \mathbf{R}^{d\times d}$ is the trainable parameter matrix.

We utilize multihead attention to further optimize the relationship between object features, obtaining the final enhanced object features $\mathbf{V}^e \in \mathbf{R}^{M\times d}$, represented as

$$\mathbf{V}^e = \text{Concat}(\mathbf{V}_{\text{head}_1}^e, \ldots, \mathbf{V}_{\text{head}_p}^e)\mathbf{W}^r \tag{6}$$

where $\mathbf{W}^r \in \mathbf{R}^{d\times d}$ refers to the trainable parameter matrix, and each head i denotes a round of dot-product attention that generates the refined $\mathbf{V}_{\text{head}_i}^e$.

2) Feature Fusion: Inspired by GLCM [51], we apply global features to the decoder to directly guide image caption generation through cross-attention. Before that, it is necessary to fuse the text word with the global feature.

First, the text word are processed. Given a sentence $\mathbf{L}$ generated at time step $\mathbf{t}$, represented as $\mathbf{L} = \{\mathbf{l}_0, \ldots, \mathbf{l}_{t-1}\}$, the subsequent word $\mathbf{l}_t$ is forecasted based on $\mathbf{l}$. the ith word $\mathbf{l}_i \in \mathbf{R}^1$ corresponds to an index in the vocabulary, with $\mathbf{l}_0$ indicating the sentence's initial token. To align with the network structure, the words of the sentence $\mathbf{L}$ must be incorporated into the sentence vector $\mathbf{W} \in \mathbf{R}^{t\times C} = \{\mathbf{w}_0, \ldots, \mathbf{w}_{t-1}\}$. Through the word2vec

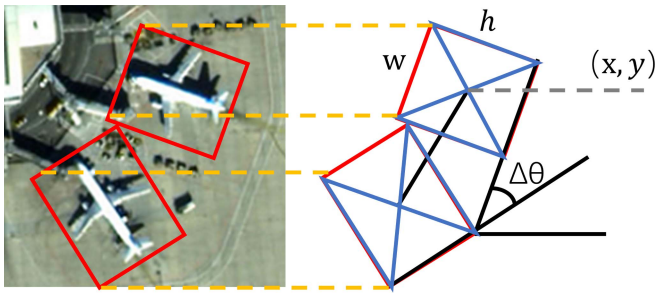

Fig. 2. Principle of VRoPE. Extracting the rotated object box of the object. $(x_i, y_i), w_i, h_i, \theta_i$ are the center coordinates, width, height, and rotation angle of box i.

technology [61], it is represented as

$$\mathbf{W} = \mathbf{Embedding}(\mathbf{L}). \tag{7}$$

After adaptive average pooling and linear projection, the global feature $\mathbf{V}^g$ is projected to the same dimension as $\mathbf{W}$

$$\hat{\mathbf{V}}^g = \mathrm{FC}(\mathrm{AAP}(\mathbf{V}^g)) \tag{8}$$

$$\mathbf{Y} = \mathbf{W} + \hat{\mathbf{V}}^g \tag{9}$$

where the dimension of $\mathbf{Y}$ is $\mathbf{R}^{T\times C}$. Then, a masked self-attention operation is utilized on the fused input $\mathbf{Y}$ to derive $\mathbf{Y}_{\text{mask}}$, which is used for final cross-attention.

D. Visual Rotated Position Embedding

To represent the spatial relationships between described objects and to calculate their degrees of closeness, we propose the visual rotated position embedding (VRoPE) to encode the position and geometric relationships of objects, which is integrated into self-attention mechanism. Initially, we obtain the rotated positions $\mathbf{P} = \{\mathbf{p}_1(x_1, y_1, w_1, h_1, \theta_1), \ldots, \mathbf{p}_M(x_M, y_M, w_M, h_M, \theta_M)\}$, where we consider both the positional and geometric characteristics of the object bounding boxes, as shown in Fig. 2. For each pair of objects $\mathbf{i}$ and $\mathbf{j}$, their positional relationship is represented by the ratio of the difference in center coordinates to the geometric size, their geometric relationship is represented by the aspect ratio, and their directional relationship is represented by the difference in rotation angles. These are concatenated into a 5-dimensional vector λ_{ij} representing the rotated relative geometric relationship:

$$\lambda_{ij} = \begin{bmatrix} \log(|x_i - x_j|/w_i) \\ \log(|y_i - y_j|/h_i) \\ \log(w_i/w_j) \\ \log(h_i/h_j) \\ \theta_i - \theta_j \end{bmatrix} \tag{10}$$

where $(x_i, y_i), w_i, h_i, \theta_i$ are the center coordinates, width, height, and rotation angle of box i, respectively. Since we need to assess the influence of each piece of information segment on attention, we weight the components of λ_{ij}. Then, to unify the dimensions and integrate with the attention mechanism, we transform it into a high-dimensional representation $\mathbf{G}_{ij}$ using an FC layer and subsequently apply an ReLU activation function activation as

$$\mathbf{G}_{ij} = \mathrm{ReLU}(\mathrm{FC}(\lambda_{ij})\mathbf{W}_g^T) \tag{11}$$

where W_g are the trainable weight parameters, and $\mathbf{G}_{ij} \in R^{N\times N}$ is the rotated relative geometric relationship, where $N = h \times w$. The obtained G is used as the VRoPE to improve the encoder module in the next section, optimizing the final caption generation.

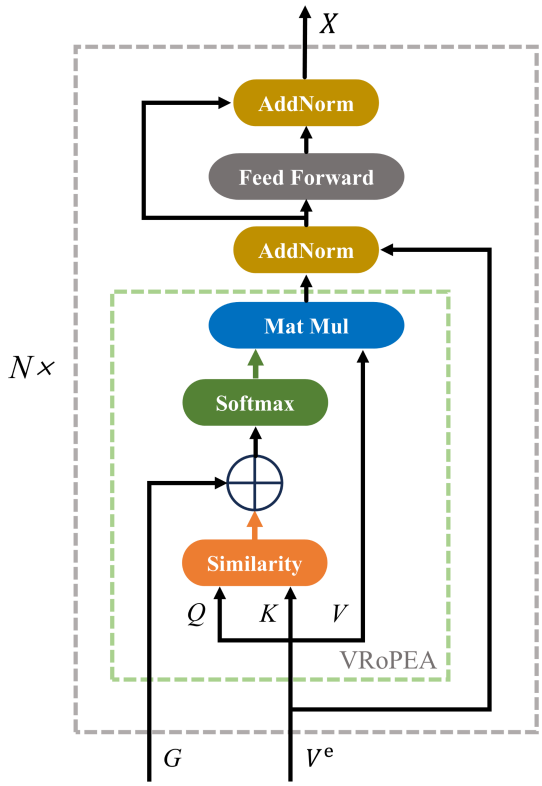

Fig. 3. Encoder of the VRoPE-Transformer. The encoder consists of a series of N uniform encoding layers. Each encoding Layer primarily consists of the VRoPEA. $\oplus$ denotes the addition operation.

E. Vrope-Transformer

The VRoPE-Transformer employs an encoder–decoder framework, with both components comprising layered multi-head attention mechanisms. The encoder enhances self-attention by integrating rotated position encoding for object pairs, thereby encoding directional relationships within the object features. The decoder leverages the outputs from these encoding layers to produce captions word by word incrementally. The architecture is detailed as follows.

1) Encoder: Drawing from the conventional transformer design [10]. As shown in Fig. 3, the encoder consists of *N* identical layers stacked atop one another. Each encoding layer integrates a VRoPEA module alongside a position-aware FC feedforward network (FFN). The input to the first layer is the enhanced object feature $\mathbf{V}^e$ combined with the VRoPE $\mathbf{G}$, and the input to each subsequent encoding layer includes the output from the preceding encoding layer combined with $\mathbf{G}$. Taking the first

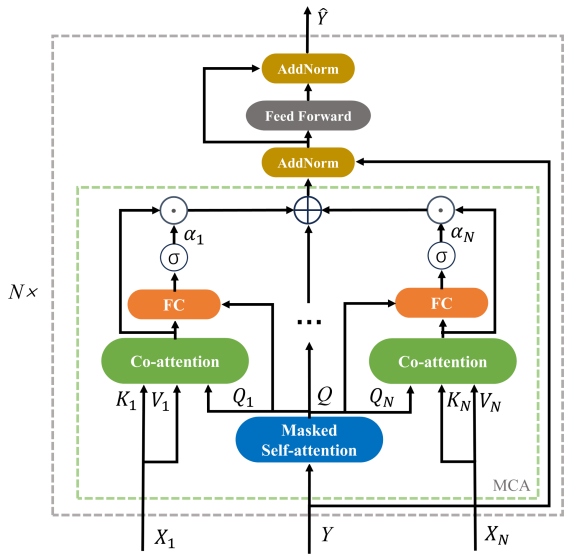

Fig. 4. Decoder of the VRoPE-Transformer. The decoder is constructed from a series of N uniform decoding layers. Each decoding Layer is primarily made up of meshed cross-attention (MCA). σ, $\odot$, and $\oplus$, denote the sigmoid activation function, the concatenation operation, and the Hadamard product.

encoding layer as an example, the queries $\mathbf{Q}$, keys $\mathbf{K}$, and values $\mathbf{V}$ are all enhanced object features $\mathbf{V}^e$. After matrix operations, $\mathbf{Q}$ and $\mathbf{K}$ are first fused with $\mathbf{G}$ and then with $\mathbf{V}$ for computation, as shown in the following equation:

$$\mathbf{Q} = \mathbf{V}^e\mathbf{W}_q, \mathbf{K} = \mathbf{V}^e\mathbf{W}_k, \mathbf{V} = \mathbf{V}^e\mathbf{W}_v$$
$$\mathbf{V}^r = \text{softmax}(\mathbf{Q}\mathbf{K}^T/\sqrt{d} + \mathbf{G})\mathbf{V} \tag{12}$$

where $\mathbf{W}_v \in \mathbf{R}^{d\times d}$, $\mathbf{W}_k \in \mathbf{R}^{d\times d}$, $\mathbf{W}_q \in \mathbf{R}^{d\times d}$ are trainable matrices.

The core mechanism of the VRoPEA module is to use the geometric and positional relationships between object pairs to adjust the correlation between queries and keys, that is, feature relevance. This allows object pairs with higher geometric closeness to be given more attention and integrated into the self-attention mechanism, thereby improving the ability to infer closely related object pairs. Then, the resulting features are added to the original features and pass through an FFN layer to obtain the output of the first encoding layer, $\mathbf{X}_1$. In this way, we can obtain the encoder's output $\mathbf{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$, where $\mathbf{X}_i \in R^{M\times D}$ from the ith layer, represented as follows:

$$X = \text{AddNorm}(\text{FFN}(\text{AddNorm}(V_r))). \tag{13}$$

This procedure is iterated for every following encoding layer, with the output of each layer serving as the input to the decoder.

2) Decoder: We adopt the meshed cross-attention from M^2 transformer [43] to design our decoder architecture. Instead of only focusing on the last encoding layer, the meshed attention operator performs cross-attention on all encoding layers. This effectively utilizes the multilevel information from the encoding layers. Specifically, the decoder comprises N uniform decoder layers as shown in Fig. 4. For each decoder layer, the output from the preceding decoder layer is employed as the query, and the output of each encoding layer is used as the value and key. This cross-attention operation allows the decoder to consider the context from both the previously produced words and the encoded image features. We perform cross-attention between the masked input $\mathbf{Y}_{\text{mask}}$ and the output of each encoding layer

$$\mathbf{Q} = \mathbf{Y}_{\text{mask}}\mathbf{W}_q^i, \mathbf{K} = \mathbf{V}_i\mathbf{W}_k^i, \mathbf{V} = \mathbf{V}_i\mathbf{W}_v^i$$
$$\mathbf{C}_i = \text{softmax}(\mathbf{Q}\mathbf{K}^T/\sqrt{d})\mathbf{V} \tag{14}$$

where $\mathbf{W}_k^i \in \mathbf{R}^{d\times d}$, $\mathbf{W}_q^i \in \mathbf{R}^{d\times d}$, and $\mathbf{W}_v^i \in \mathbf{R}^{d\times d}$ are trainable matrices. After that, in order to produce the output of this layer, we must mix the features. First, we calculate the weighted matrix α_i, which represents the contribution of each encoding layer, calculated from the masked input $\mathbf{Y}_{\text{mask}}$ and the output of each encoder layer. Then, we weight $\mathbf{C}_i$ to obtain the mixed features $Z_1 \in \mathbf{R}^{M\times d}$, followed by *AddNorm* and FFN operations to obtain the output $\mathbf{Y_1}$ of the first decoder, as expressed in the following equation:

$$\alpha_i = \text{softmax}\left([\mathbf{Y}_{\text{mask}}, \mathbf{C}_i]\mathbf{W}_i^\alpha + \mathbf{b}_i^\alpha\right) \tag{15}$$

$$Z_1 = \sum_{i=1}^{N} \alpha_i \odot \mathbf{C}_i \tag{16}$$

$$Y_1 = \text{AddNorm}(\text{FFN}(\text{AddNorm}(Z_1))) \tag{17}$$

where $\mathbf{W}_i^\alpha \in \mathbf{R}^{2d\times d}$ represents the trainable parameter matrix, $\odot$ refers to the Hadamard product, $\mathbf{b}_i^\alpha \in \mathbf{R}^d$ denotes the bias term, and $[\cdot,\cdot]$ signifies the concatenation operation.

In the following layers, we sequentially perform cross-attention operations between the output of the previous decoder layer and the output of each encoding layer, and use the output from the final decoder layer as the ultimate output of the decoder. The final output $\hat{\mathbf{Y}}$ is then passed through a mapping head, which consists of an *FC* layer that projects the decoder's output space into the vocabulary space, preceded by a softmax layer that converts the output into a probability distribution across the vocabulary. This cycle is iterated for each successive decoder layer, with the final output of the last decoder layer being the completed caption or a chain of words which represents the caption for the input image.

F. Training and Objectives

Inspired by prevalent practices in image captioning [9], we initiate the training process by employing the cross-entropy loss (XE). This loss function is formulated as follows:

$$L_{\text{XE}}(\theta) = -\sum_{t=1}^{T} \log(p_\theta(s_t^*|s_{1:t-1}^*)) \tag{18}$$

where T refers to the maximum time step; $s_{1:t-1}^* = [s_1^*, \ldots, s_{t-1}^*]$ denotes the ground truth sequence up to the $(t-1)$th time step; θ represents the model's variable parameter. The objective is to maximize the logarithm of the likelihood for the true word (s_t^*) based on the preceding words in the sequence.

G. RSIC-ROD

To train the backbone of *REDET* for extracting object features, we constructed the rotated object detection dataset RSIC-ROD.

This dataset encompasses the object categories of objects described in current mainstream RSIC datasets. Therefore, it can be used for the rotated feature extraction part of object-level RSIC models.

Data source: We built the RSIC-ROD dataset based on the NWPU-Captions [37] RSIC dataset.

Data processing: We selected images rich in objects, clear, and representative of each scene, and manually annotated rotated object bounding boxes on them by three volunteers with extensive experience.

Dataset scale: The dataset consists of 10 500 256 × 256 RSIs and their corresponding rotated object annotations. The dataset contains 15 scene categories, which are as follows: *Wharf, ground track, field, basketball court, airport, baseball diamond, tennis court, storage tank, ship, stadium, roundabout, railway, intersection, overpass, and harbor* The dataset contains 32 object categories, which are as follows: *Airport, baseball diamond, ground track, basketball court, field, harbor, wharf, intersection, overpass, railway, roundabout, ship, storage tank, stadium, tennis court, airplane, bridge, river, railway, boat, farmland, crosswalk, runway, freeway, highway, car, truck, road, meadow, trees, buildings, parking lot, and bare land.*

Previous rotated object detection datasets had too large a gap in object categories compared to existing RSIC datasets, which could lead to object-level feature-based RSIC models having more difficulty with category ambiguity. RSIC-ROD is large in scale, has many categories, and is more compatible with RSIC datasets, making it significant for future object feature RSIC models.

IV. Experiment Results

A. Datasets

We evaluate the proposed method on four RSIC datasets: RSICD [13], Sydney-Captions [12], UC Merced (UCM)-Captions [12], and NWPU-Captions [37].

Sydney-Captions: Sydney-Captions is a derivative of the Sydney Dataset [52]. The Sydney Dataset comprises 613 images categorized into seven classes: airport, residential, runway, industrial, ocean, rivers, and meadow. These images, obtained from Google Earth and manually cropped, are all 500 × 500 pixels at a resolution of 0.5 m.

UCM-Captions: UCM-Captions is an extension based on the UCM Land Use Dataset [53]. The UCM dataset consists of 2100 images covering 21 scene classes, including baseball diamond, harbor, dense residential, intersection, beach, chaparral, river, golf course, forest, mobile home park, tennis courts, freeway, parking lot, overpass, medium-density residential, airplane, storage tanks, buildings, sparse residential, and agricultural. Every category in the dataset consists of 100 images, each measuring 256 by 256 pixels, and each pixel corresponds to a ground distance of 0.3048 m.

RSICD: RSICD comprises 10 921 RSIs covering various geographical areas and scenes. The dataset includes 30 scene types such as open land, industrial area, beach, medium residential, stadium, playground, mountain, bridge, farmland, church, school, desert, meadow, pond, sparse residential area, commercial center, port, river, dense residential, viaduct, railway station, forest, park, airport, baseball field, parking area, resort, square, and storage tanks. The images in RSICD, obtained from Google Earth, are sized 224 × 224 pixels with varying pixel resolutions.

NWPU-Captions: NWPU-Captions is the largest RSIC dataset. NWPU offers a wealth of image diversity, encompassing images from various satellites and aerial sensors. The dataset contains abundant scenes and objects such as airports, building complexes, and ports. It comprises 331 500 aerial RSIs, each equipped with RGB three-band data, and ground sampling distances varying from 0.2 to 30 m.

B. Evaluation Metrics

To assess the quality of produced captions for RSIs, we employ ten metrics: BLEU-n [54], METEOR [55], ROUGE_L [56], CIDEr [57], SPICE [58], S_m^*, and S_m [38].The metrics SPICE, CIDEr, ROUGE_L, METEOR, and BLEU-n can be computed using the COCO caption evaluation toolkit. METEOR, BLEU-n, and ROUGE_L were originally devised for machine translation and emphasize caption accuracy. CIDEr and SPICE, were specifically crafted for image captioning, with a focus on human-likeness. S_m^* and S_m serve as overall metrics, balancing both aspects.

BLEU: Originally designed for machine translation, BLEU measures the overlap of phrases between candidate and reference sentences based on the match of n consecutive words, represented as n-grams. In our study, we use four variations of the BLEU metric, namely BLEU-1, BLEU-2, BLEU-3, and BLEU-4, with n set to 1, 2, 3, and 4. These scores are normalized on a scale between 0 and 1, indicating the quality of the generated captions.

ROUGE_L: ROUGE_L is part of the Recall-oriented understudy for gisting evaluation (ROUGE), it evaluates similarity between reference sentences and candidate by determining the longest common subsequence, prioritizing recalls over precision. Commonly used in image captioning evaluations.

CIDEr: Specifically designed for image captioning, CIDEr considers each sentence as an individual document. It then calculates a TF-IDF vector for each sentence, which is used to measure the similarity between reference captions and candidate. CIDEr focuses on the accuracy of n-grams and considers their frequency across the entire dataset through weighting operations. It measures caption consistency by computing the cosine similarity.

SPICE: Also designed for image captioning, it encodes relations, objects, and attributes using a graph-based approach. It parses sentences into syntactic dependency trees, maps them to scene graphs, and calculates F-scores for objects, attributes, and relations.

S_m^* and S_m: Averages of the aforementioned metrics, offering comprehensive evaluation of caption quality. Calculation processes are provided as follows:

$$S_m^* = \frac{1}{4}(\text{C} + \text{R} + \text{M} + \text{BELU4}) \tag{19}$$

$$S_m = \frac{1}{5}(\text{C} + \text{R} + \text{M} + \text{S} + \text{BELU4}). \tag{20}$$

C. Experimental Settings and Training Details

1) Dataset Splitting: To facilitate an equitable and objective comparison with state-of-the-art methods and to analyze the effectiveness of the proposed method, we randomly shuffle the first three datasets and split them into 10% for testing, 80% for training, and 10% for validation. To eliminate the impact of random splitting, we conduct five experiments on each dataset. After discard the best and worst results, we calculate the average and standard deviation of the remaining results as the final result. For NWPU-Captions, we use its original split [37].

2) Feature Extraction: For global features, we employ the *ResNet-152* architecture pretrained on ImageNet [59]. We extract the feature map from the final adaptive average pooling layer, which is of size 14 $\times$14$\times$ 2048. This is then flattened into a matrix with dimensions 196 $\times$ 2048 ($D1 = 2048$ and $N = 196$), serving as our final global feature.

For object features, we use the backbone of *Redet* pretrained on the ImageNet and fine-tune it on our proposed dataset RSIC-ROD. Specifically, we freeze the *ReResNet* modules pretrained on ImageNet and train the subsequent *ReFPN* modules on RSIC-ROD for 200 epochs with the learning rate of $1e-5$. The intersection over union threshold is set at 0.1. For each RSI, we choose the highest-scoring 50 object proposals to serve as object features ($M = 50$), each with a size of 7 $\times$7$\times$ 256. After that, we perform an AAP operation followed by a linear mapping operation to reduce each object's dimension to 1024 ($D2 = 1024$), which serves as our final object feature.

3) Model Configuration and Training: For all our transformer-based models, we configure the encoder and decoder with $N = 3$ layers each, and we utilize 8 parallel attention heads within the multihead attention mechanism. The dimension for word embeddings is set at $C = 512$. We apply positional encoding to sequences with a maximum length of 128, and we limit the maximum length of the output sequences to 20. During training, we set the batch size to 50 and apply Dropout at a rate of 0.9 following each feedforward and attention layer. Adam is utilized as the optimizer with parameters $\beta1 = 0.9$ and $\beta2 = 0.98$. The model is trained using XE, and we employ a learning rate scheduling strategy that includes a warmup period of 10 000 iterations. Our experiments are conducted on an NVIDIA GeForce RTX 3080Ti, using PyTorch 1.10.0 and CUDA 10.2 as our software environment.

D. Comparisons With the State-of-the-Art

To objectively validate the performance of the proposed VRoPE-Transformer on RSIC, we selected several state-of-the-art methods and some representative methods to generate captions for RSIs from different viewpoints. These methods were then compared through experimental analysis. They are as follows.

1) *mRNN [12] and* mLSTM *[12]:* Classical encoder–decoder architectures, using VGG-16 as their encoder and employing different RNNs as their decoders.
2) *Attention-based (hard/soft) method:* Building upon the CNN-LSTM framework, hard attention and soft attention [9] utilize VGG-16 for encoding and LSTM for decoding, pioneering the investigation into the application of attention mechanisms for the generation of captions from RSIs.
3) *SD-RSIC [60]:* A method for RSIC. The method first summarizes multiple ground-truth captions (GT) for RSIs into a single caption. Then, it utilizes an adaptive weighting strategy that merges the condensed caption with typical captions to produce more precise and informative captions. It utilizes *ResNet-152* for encoding and LSTM for decoding purposes.
4) *WS-RSIC [61]:* It decomposes the image caption tas into two distinct phases: extracting relevant words and crafting sentences. In the first stage, the method uses *ResNet-18* as a word extractor to extract words that represent different visual features from the image. In the second stage, a transformer is employed as the sentence generator to combine these words into coherent sentence captions.
5) *AoANet [62]:* It introduces an attention mechanism called "attention on attention (AoA)." This attention mechanism enables the model to concentrate on image features at multiple levels during the decoding process, allowing the model to interpret the image content more accurately and generate corresponding captions. The encoder part commonly employs CNN for image feature extraction, whereas the decoder part utilizes an LSTM combined with the AoA module to generate captions.
6) *GVFGA+LSGA [38]:* Using a CNN-RNN framework and combines two attention mechanisms to better utilize image and text information. First, through GVFGA, the model filters out redundant information and places greater emphasis on the most important regions within the image. Second, through LSGA, the model further adjusts the fusion of image and text information based on the current language state to produce more precise and consistent captions.
7) *MLCA-Net [37]:* Integrating a multilevel attention mechanism to dynamically combine image features from targeted spatial areas and scales. In addition, a contextual attention module is used to reveal the hidden context in RSIs. This approach improves the flexibility and detail of captions, ensuring they remain accurate and concise.
8) *RSGPT [67]:* A large visual language model tailored for the remote sensing field, trained on the high-quality RSIC dataset RSICap.
9) *RS-LLAVA [68]:* An improved version of the large language and vision assistant model (LLAVA), specifically adapted for RSIs through a low-rank adaptation approach.
10) M^2 *Transformer [43]:* Embedding prior knowledge about object relationships into the self-attention mechanism of the transformer encoder. The encoder's role is to encoding image features into semantic representations while embedding object relationships within the self-attention mechanism. The decoder then creates captions from the encoder's output, using multilayer self-attention and position encoding to enhance the interpretation of the input image and produce corresponding captions.

TABLE I
QUANTITATIVE COMPARISON RESULTS ON SYDNEY-CAPTIONS

Methods	BELU-1	BELU-2	BELU-3	BELU-4	Meteor	ROUGE_L	CIDEr	SPICE	S_m^*	S_m
mRNN [12]	51.30	37.50	20.40	19.30	18.50	-	32.20	-	-	-
mLSTM [12]	54.60	39.50	22.30	21.20	20.50	-	37.20	-	-	-
Hard-attention [13]	75.91	66.10	58.89	52.58	38.98	71.89	218.19	-	95.41	-
Soft-attention [13]	73.22	66.74	62.23	58.20	39.42	71.27	249.93	-	104.71	-
WS-RSIC [61]	78.91	70.94	63.17	56.25	41.81	69.22	204.11	-	92.85	-
GVFGA+LSGA [38]	76.81	68.46	61.45	55.04	38.66	70.30	245.22	45.32	102.31	90.91
SD-RSIC [60]	73.3	61.9	51.7	42.5	31.8	62.0	114.6	-	62.73	-
AoANet [62]	75.20	66.20	58.85	52.30	37.92	69.31	228.99	42.09	97.13	86.12
RSGPT [67]	82.26	75.28	68.57	62.23	41.37	74.77	273.08	-	-	-
RS-LLAVA(7B) [68]	-	-	-	-	-	-	-	-	-	-
M^2 Transformer [28]	82.25±1.65	76.19±1.12	71.04±1.83	66.30±1.83	44.20±1.49	75.21±1.98	275.44±13.65	43.64±1.74	115.29±2.69	101.16±2.10
PKG-Transformer [44]	84.00±1.48	78.65±2.04	73.78±2.42	69.01±2.11	46.49±1.82	78.25±1.56	303.22±6.03	46.77±1.04	124.24±0.85	108.75±0.81
Baseline	82.60±2.75	76.72±2.92	71.75±3.37	67.14±3.61	45.53±1.35	76.93±1.14	285.22±14.47	45.35±1.71	118.70±2.22	104.03±1.91
VRoPE-Transformer	**87.30±1.45**	**82.37±1.78**	**78.31±2.29**	**74.38±2.65**	**49.79±1.78**	**81.54±1.25**	**343.88±14.70**	**48.22±1.08**	**137.40±4.99**	**119.56±4.21**

The "-" is used to signify that the results were not disclosed in the publication.
The bold values represents the optimal results.

TABLE II
QUANTITATIVE COMPARISON RESULTS ON UCM-CAPTIONS

Methods	BELU-1	BELU-2	BELU-3	BELU-4	Meteor	ROUGE_L	CIDEr	SPICE	S_m^*	S_m
mRNN [12]	60.10	50.70	32.80	20.80	19.30	-	42.80	-	-	-
mLSTM [12]	63.50	53.20	37.50	21.30	20.30	-	44.50	-	-	-
Hard-attention [13]	81.57	73.12	67.02	61.82	42.63	76.98	299.47	-	120.23	-
Soft-attention [13]	74.54	65.45	58.55	52.50	38.86	72.37	261.24	-	106.24	-
WS-RSIC [61]	79.31	72.37	66.71	62.02	43.95	71.32	278.71	-	114.00	-
GVFGA+LSGA [38]	83.19	76.57	71.03	65.96	44.36	78.45	332.70	48.53	130.37	114.00
SD-RSIC [60]	71.4	62.5	55.3	49.2	36.3	65.8	197.8	-	87.28	-
AoANet [62]	81.85	74.73	68.80	63.27	41.30	75.43	308.73	43.96	122.18	106.54
RSGPT [67]	86.12	79.14	72.31	65.74	42.21	78.34	333.23	-	-	-
RS-LLAVA(7B) [68]	88.70	82.88	77.70	72.84	47.98	85.17	349.43	-	-	-
M^2 Transformer [28]	88.90±1.32	85.98±1.49	82.74±1.51	80.89±1.36	51.13±2.21	84.45±2.61	418.41±17.27	53.43±2.43	155.97±4.36	138.66±2.98
PKG-Transformer [44]	89.72±0.43	86.08±0.71	82.94±0.79	80.08±0.78	**53.62±0.25**	**86.66±0.54**	423.18±7.71	**56.80±0.35**	160.88±1.95	140.07±1.59
Baseline	86.73±1.21	82.79±1.37	79.55±1.08	76.62±1.16	50.64±2.03	83.48±2.38	382.93±8.36	51.84±1.38	148.41±2.67	129.10±2.04
VRoPE-Transformer	**89.81±0.22**	**86.45±0.42**	**83.56±0.53**	**80.98±0.63**	53.36±1.12	85.93±0.74	**434.63±2.74**	55.38±0.26	**163.73±0.38**	**142.06±0.36**

The "-" is used to signify that the results were not disclosed in the publication.
The bold values represents the optimal results.

11) *PKG-Transformer [44]:* Extracting object-level and scene-level features in the MFE module. In addition, a feature enhancement module employs a combination of attention mechanisms and graph neural networks to encapsulate correlations and differences among diverse objects or distinct scene areas. Finally, the prior knowledge-enhanced attention module establishes relationships between objects to select more relevant objects to the scene area. The decoder part is consistent with M^2 transformer [43].
12) *Baseline*: To assess the independent contributions of each component and promote comparison with PKG-Transformer, we created a baseline model that merges the standard transformer encoder with object features as input with the decoder of PKG-Transformer [44].

Tables I–IV display the comparative outcomes against other approaches, which correspond to UCM-Captions, RSICD datasets, Sydney-Captions, and NWPU-Captions. The presented results are all in the form of percentages (%).

1) Sydney-Captions: Table I presents the comparative outcomes for the Sydney-Captions dataset. The proposed method exhibits the best performance across all metrics, with relative improvements of 3.93%, 4.73%, 6.14%, 7.78%, 7.10%, 4.20%, 13.41%, 3.10%, 10.59%, and 9.94%. Notably, the proposed method shows a significant increase in CIDEr, indicating that the generated descriptions are more natural and precise in language expression and semantics. In addition, the BLEU-3 and BLEU-4 scores, which have stricter requirements for generating multiple consecutive words, suggest that the proposed method can produce more precise phrases.

2) UCM-Captions: Table II presents the comparative outcomes for the UCM-Captions dataset. The proposed method outperforms PKG-Transformer in BLEU, CIDEr, S_m^* and S_m, with relative improvements of 0.10%, 0.42%, 0.74%, 1.12%, 2.70%, 1.77%, and 1.42%, respectively. The other evaluation metrics also surpass those of the comparative methods. On the METEOR and SPICE metrics, PKG-Transformer performs better, indicating a higher recall rate due to its exploration of scene-scene/object-object and object-scene relationships in RSIs, resulting in more accurate objects, attributes, and relationships. However, the S_m^* and S_m metrics indicate that the proposed method still has a better overall performance.

3) RSICD: Table III lists the comparison results for the RSICD dataset. Due to the lack of descriptions of relative spatial relationships in the RSICD dataset, the proposed method does not achieve the best results on all metrics. However, the results of the proposed method are very close to the best results, and it achieves the best result on the most critical metrics CIDEr, S_m^* and S_m, fully proving the competitiveness of the proposed method.

TABLE III
QUANTITATIVE COMPARISON RESULTS ON RSICD

Methods	BELU-1	BELU-2	BELU-3	BELU-4	Meteor	ROUGE_L	CIDEr	SPICE	S_m^*	S_m
mRNN [12]	45.58	28.25	18.09	12.13	15.69	31.26	19.15	-	19.56	-
mLSTM [12]	50.57	32.42	23.19	17.46	17.84	35.02	31.61	-	25.48	-
Hard-attention [13]	66.69	51.82	41.64	34.07	32.01	60.84	179.25	-	76.54	-
Soft-attention [13]	67.53	53.08	43.33	36.17	32.55	61.09	196.43	-	81.56	
WS-RSIC [61]	**72.40**	**58.61**	**49.33**	42.50	31.97	**62.60**	206.29	-	85.84	-
GVFGA+LSGA [38]	67.79	56.00	47.81	41.65	**32.85**	59.29	260.12	**46.83**	98.48	88.15
SD-RSIC [60]	64.40	47.40	36.90	30.0	24.90	52.30	79.40	-	46.65	-
AoANet [62]	67.18	55.52	47.35	41.01	32.51	58.52	256.47	46.12	97.13	86.93
RSGPT [67]	70.32	54.23	44.02	36.83	30.10	53.34	102.94	-	-	-
RS-LLAVA(7B) [68]	-	-	-	-	-	-	-	-	-	-
M^2 Transformer [28]	68.44±1.20	56.57±1.25	48.10±1.25	41.56±1.26	32.69±1.01	59.12±0.93	258.58±3.68	45.43±0.92	97.99±1.67	87.48±1.52
PKG-Transformer [44]	67.99±0.36	56.62±0.53	48.49±0.62	42.27±0.79	32.27±0.51	58.90±0.63	255.57±9.58	45.14±1.25	97.25±2.73	86.83±2.43
Baseline	67.90±0.08	55.50±0.40	46.65±0.64	39.86±0.90	31.79±0.29	57.39±0.74	235.97±8.42	43.55±0.82	91.25±2.56	81.71±2.17
VRoPE-Transformer	68.90±0.84	57.38±0.95	49.03±0.86	**42.56±0.74**	32.73±0.62	59.68±0.68	**260.28±6.48**	45.63±1.05	**98.82±2.03**	**88.18±1.83**

The "-" is used to signify that the results were not disclosed in the publication.
The bold values represents the optimal results.

TABLE IV
QUANTITATIVE COMPARISON RESULTS ON NWPU-CAPTIONS

Methods	BELU-1	BELU-2	BELU-3	BELU-4	Meteor	ROUGE_L	CIDEr	SPICE	S_m^*	S_m
Hard-attention [13]	73.3	61.0	52.7	46.4	34.0	60.0	110.3	28.4	-	-
Soft-attention [13]	73.1	60.9	52.5	46.2	33.9	59.9	113.6	28.5	-	-
MLCA-Net [37]	74.5	62.4	54.1	47.8	33.7	60.1	126.4	28.5	-	-
PKG-Transformer [44]	88.3	80.4	74.0	68.9	44.4	77.3	197.7	**28.9**	97.1	83.5
Baseline	87.0	78.7	72.2	67.0	43.8	76.0	193.3	28.5	95.0	81.7
VRoPE-Transformer	**90.5**	**83.5**	**77.6**	**72.7**	**44.6**	**78.6**	**210.8**	28.5	**101.7**	**87.0**

The "-" is used to signify that the results were not disclosed in the publication.
The bold values represents the optimal results.

TABLE V
ABLATION STUDIES ON SYDNEY-CAPTIONS

VRoPE	FEF	BELU-1	BELU-2	BELU-3	BELU-4	Meteor	ROUGE_L	CIDEr	SPICE	S_m^*	S_m
✗	✗	80.38	74.74	69.39	64.32	43.63	75.03	286.19	44.60	117.29	102.76
✓	✗	84.54	79.48	75.23	71.21	47.44	79.08	320.36	48.58	129.52	113.33
✗	✓	83.79	78.10	73.22	68.72	45.12	76.63	299.76	47.07	122.56	107.46
✓	✓	**85.46**	**80.35**	**76.59**	**73.06**	**47.57**	**78.13**	**335.31**	**45.48**	**133.52**	**115.91**

The "✓" indicates the use of the module, while the "✗" indicates non-use.
The bold values represents the optimal results.

4) NWPU-Captions: Table IV lists the comparison results for the NWPU-Captions dataset. The results of the proposed method are comparable to those of the comparative methods on METEOR and SPICE metrics, and it achieves significant improvements in BLEU, CIDEr, S_m^* and S_m, with relative improvements of 2.49%, 3.86%, 4.86%, 5.52%, 6.63%, 4.74%, and 4.19%, respectively. This confirms the robustness and universality of the proposed method on larger and more detailed datasets, offering strong evidence of its performance in practical application scenarios.

In summary, compared to the state-of-the-art methods, the proposed method achieves excellent results on the Sydney-Captions, UCM-Captions and NWPU-Captions datasets, and comparable results on the RSICD dataset. This indicates that the proposed method has greater potential in more detailed scene applications and also demonstrates good competitiveness in large datasets. These results validate the universality of the method we propose.

E. Ablation Study

To validate the impact of each individual component, we performed ablation experiments by eliminating the FEF module, the VRoPE module, and both modules on three RSIC datasets. For each dataset, we conduct experiments on the same split. The results of these experiments are presented in Tables V–VII.

The outcomes from the experiments on the three datasets show consistent patterns. When using the FEF module alone, CIDEr increased by 4.74% on Sydney-Captions, 4.43% on UCM-Captions, and 6.59% on RSICD-Captions, while Sm increased by 4.57% on Sydney-Captions, 3.75% on UCM-Captions, and 5.05% on RSICD-Captions. This verifies the effectiveness of the FEF module. The improvement is due to the FEF module's ability to enhance image representations. In particular, the FEF module enhances the independence of object features and fuses the global features to obtain more effective input features for guiding image caption generation.

Similarly, when using the VRoPE module alone, CIDEr increased by 11.9% on Sydney-Captions, 6.68% on UCM-Captions, and 6.59% on RSICD-Captions, while Sm increased by 10.2% on Sydney-Captions, 4.91% on UCM-Captions, and 5.23% on RSICD-Captions, verifying the effectiveness of the VRoPE module. The enhancement stems from the VRoPE module leveraging the positional and directional relationships between object groups to augment the model's capability to generate strong related object group captions. Specifically, by encoding position and direction, the caption weights of closely related object groups are increased, and the directional relationships are used to guide caption generation.

TABLE VI
ABLATION STUDIES ON UCM-CAPTIONS

VRoPE	FEF	BELU-1	BELU-2	BELU-3	BELU-4	Meteor	ROUGE_L	CIDEr	SPICE	S_m^*	S_m
✗	✗	86.73	82.79	79.55	76.62	50.64	83.48	382.93	51.84	148.41	129.10
✓	✗	88.22	83.88	80.31	77.20	51.77	84.27	408.54	55.45	155.45	135.45
✗	✓	88.41	84.36	80.98	77.95	52.82	85.43	399.93	53.56	154.03	133.94
✓	✓	**90.06**	**86.84**	**84.03**	**81.61**	**52.08**	**85.08**	**437.52**	**55.59**	**164.07**	**142.38**

The "✓" indicates the use of the module, while the "✗" indicates non-use.
The bold values represents the optimal results.

TABLE VII
ABLATION STUDIES ON RSICD

VRoPE	FEF	BELU-1	BELU-2	BELU-3	BELU-4	Meteor	ROUGE_L	CIDEr	SPICE	S_m^*	S_m
✗	✗	67.67	56.12	47.87	41.57	32.00	58.50	247.32	44.18	94.85	84.72
✓	✗	68.08	57.29	49.29	42.93	32.73	60.71	263.01	46.37	99.84	89.15
✗	✓	67.99	57.00	48.82	42.68	32.66	60.29	263.63	45.74	99.82	89.00
✓	✓	**69.81**	**58.35**	**49.87**	**43.36**	**33.45**	**60.46**	**266.50**	**46.69**	**100.96**	**90.10**

The "✓" indicates the use of the module, while the "✗" indicates non-use.
The bold values represents the optimal results.

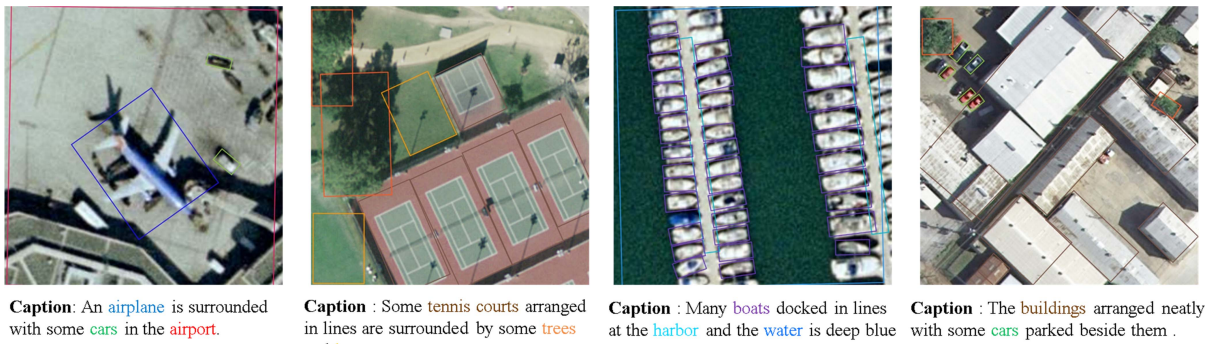

Fig. 5. Example of rotating object extraction and generated caption. The object words in the caption have the same color with the object box in the image.

Using both the VRoPE and FEF modules together further improves CIDEr and Sm scores. Specifically, CIDEr increased by 17.2% on Sydney-Captions, 14.26% on UCM-Captions, and 7.76% on RSICD-Captions, while Sm increased by 12.80% on Sydney-Captions, 10.29% on UCM-Captions, and 6.35% on RSICD-Captions, which is a significant improvement over using either module alone. The observation suggests that the VRoPE module and the FEF module not only have good effects when used independently but also exhibit a certain degree of complementary effect. In other words, while the VRoPE module uses positional and directional relationships between objects to guide caption generation, it does not incorporate global features. The FEF module enhances the independence of object features and integrates global features to compensate for this limitation. This makes the combined use of the two modules the best in terms of performance.

In conclusion, both the FEF and VRoPE modules significantly boost metric accuracy, with their combined use outperforming individual application.

F. Qualitative Analysis

To intuitively demonstrate the effectiveness of our VRoPE-Transformer, we provide qualitative comparisons in Fig. 5. We have conducted a detailed comparison between the outputs of the VRoPE-Transformer and the baseline model, incorporating manually annotated GT. This comparison encompasses multiple dimensions, including object identification, the use of adjectives, and the overall naturalness of the captions, providing a clear demonstration of the proposed method's capabilities. In addition, to evaluate the individual contributions of each component, we present a visual analysis contrasting the baseline results with those obtained from the FEF module and the VRoPE module when used in isolation.

Visually, the captions generated by VRoPE-Transformer are noticeably more precise and detailed than those produced by the baseline model. For example, in the fourth image of Fig. 6, the baseline lacks the important object "sand" and fails to describe its location. In the sixth image, it does not describe the more closely related "buildings" and "roads," instead describing the less closely related "lawn." In the fifth image, it even misidentifies the number of "tennis courts" and lacks adjectives like "arranged neatly." Conversely, the proposed method accurately describes high-relevance objects and provides precise captions for scenes and quantities. It also includes advanced semantic words, indicating a better ability to describe relevant objects and higher contextual match.

Examining each module separately, the FEF module successfully captures the adjective "small" in the first image but

Fig. 6. Example of rotated object extraction and caption generation. The object words in the caption correspond to the color of the target box in the image.

overlooks the closely related object "plants." In the third image, it captures the correct number of objects but describes "some plants" instead of the closely related object "road" In the seventh image, it distinguishes between "basketball courts" and "tennis courts" but overlooks surrounding elements. This demonstrates that the FEF module can extract advanced semantic information by combining global features, and the enhanced object features allow it to describe the correct number and distinguish similar objects. In contrast, the VRoPE module generally identifies correctly related objects, though it occasionally groups objects in ways that defy intuition. For example, in the last image, it judges "trees" rather than "road" as being more relevant, which may be due to the closer spatial geometric relationship of the "trees." Moreover, due to the lack of global feature guidance and enhanced input, it sometimes fails to capture adjectives, gets the number of objects wrong, and cannot distinguish between similar objects. This demonstrates the complementary nature of the two modules, as their combined use yields the most accurate and rich results, with more reasonable object descriptions.

We visualize the attention states for each word generated in the captions, highlight the regions that significantly contribute to the word generation. Unlike models that can easily extract attention maps with the entire image as input, the proposed method uses rotated regions as input, which requires us to analyze the contribution of different areas to the results. We mark the most attended regions. As shown in Fig. 7, in the result of baseline, the attended regions are more singular and the patterns and priorities are not obvious. The baseline model tends to focus on the most prominent areas during target detection. The PKG-Transformer can focus on more precise and diverse areas. It placing attention on parts related to the scene, such as "houses" and "lawn." Instead, our proposed method pays more attention to groups of objects with strong correlations, so it describes the "road" and elucidates its spatial relationship with the "residential area."

In summary, the qualitative results clearly show that the VRoPE-Transformer surpasses the baseline in accuracy and comprehensiveness, aligning better with human cognitive patterns in description generation.

Fig. 7. Visualization of attention states for Baseline, PKG-Transformer and VRoPE-Transformer. We show the attended image regions, outlining the region in high definition.

TABLE VIII
COMPARISON BETWEEN THE BASELINE, PKG-TRANSFORMER, AND THE PROPOSED VROPE-TRANSFORMER ON THE NUMBER OF PARAMETERS AND FLOPS

Method	Parameters	FLOPs
Baseline	28.61 M	1.28 G
PKG-Transformer [44]	31.94 M	1.58 G
VRoPE-Transformer*	30.70 M	1.36 G
VRoPE-Transformer	31.27 M	1.38 G

G. Complexity Analysis

This section provides a brief analysis of the computational complexity of the proposed VRoPE-Transformer. Table VIII details the number of model parameters and floating-point operations (FLOPs). VRoPE-Transformer* refers to the model without the FEF module. The FEF module introduces a negligible increase in the number of parameters and computational complexity, yet it confers a notable enhancement in performance. Although the VRoPE module modestly elevates the number of parameters and computational complexity, it markedly improves performance. Compared to the PKG-Transformer, the proposed VRoPE-Transformer achieves comparable results while consuming fewer resources.

V. CONCLUSION

The VRoPE-Transformer method introduced in this article represents a significant advancement in the RSIC domain. It adeptly extracts rotation-invariant features and incorporates dedicated modules—the FEF and VRoPE—to tackle the nuanced challenges of RSIC. By leveraging self-attention mechanisms, the FEF module effectively fuses multiscale features, augmenting their relevance and distinctiveness. Concurrently,

the VRoPE module captures positional and directional relationships among objects, integrating this information into the Transformer's attention mechanism. This synergy facilitates the generation of precise captions for objects with intricate interrelationships. Experimental outcomes across four datasets underscore the method's superior performance, affirming its efficacy. Moreover, we introduce a dataset comprising 10 500 256 $\times$ 256 RSIs, each with corresponding rotated object annotations. This dataset paves the way for a seamless fusion of object-level image caption and remote sensing object detection, thus broadening the scope of application and enhancing the depth of understanding in RSI. The VRoPE-Transformer ushers in a novel paradigm for comprehending RSIs, promising extensive utility in various domains. It offers a fresh viewpoint and a robust solution that advances the state of the art in RSI processing. Future research can explore more intricate modeling of object relationships to encapsulate even richer detail, and the application of this method within multimodal, large-scale models holds potential for extracting even more comprehensive information.

REFERENCES

[1] T. Zhang et al., "Semantic attention and scale complementary network for instance segmentation in remote sensing images," *IEEE Trans. Cybern.*, vol. 52, no. 10, pp. 10999–11013, Oct. 2022.

[2] J. Han et al., "ReDet: A rotation-equivariant detector for aerial object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2785–2794, doi: 10.1109/cvpr46437.2021.00281.

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

[4] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1155–1167, Feb. 2019.

[5] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, and R. Cucchiara, "From show to tell: A survey on deep learning-based image captioning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 539–559, Jan. 2022, doi: 10.1109/TPAMI.2022.3148210.

[6] J. Ji, Y. Ma, X. Sun, Y. Zhou, Y. Wu, and R. Ji, "Knowing what to learn: A. metric-oriented focal mechanism for image captioning," *IEEE Trans. Image Process.*, vol. 31, pp. 4321–4335, 2022, doi: 10.1109/TIP.2022.3183434.

[7] J. Jiet al., "Improving image captioning by leveraging intra- and inter-layer global representation in transformer network," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 1655–1663.

[8] X. Zhang et al., "RSTNet: Captioning with adaptive attention on visual and non-visual words," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15460–15469, doi: 10.1109/CVPR46437.2021.01521.

[9] P. Anderson et al., "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6077–6086.

[10] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[11] Y. Pan, T. Yao, Y. Li, and T. Mei, "X-linear attention networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10968–10977, doi: 10.1109/CVPR42600.2020.01098.

[12] B. Qu et al., "Deep semantic understanding of high resolution remote sensing image," in *Proc. Int. Conf. Comput., Inf. Telecommun. Syst.*, 2016, pp. 1–5, doi: 10.1109/CITS.2016.7546397.

[13] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2183–2195, Apr. 2018, doi: 10.1109/TGRS.2017.2776321.

[14] H. Kandala, S. Saha, B. Banerjee, and X. X. Zhu, "Exploring transformer and multilabel classification for remote sensing image captioning," *IEEE Geosci. Remote. Sens. Lett.*, vol. 19, pp. 1–5, 2022, Art. no. 6514905, doi: 10.1109/LGRS.2022.3198234.

[15] C. Wang, Z. Jiang, and Y. Yuan, "Instance-aware remote sensing image captioning with cross-hierarchy attention," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2020, pp. 980–983, doi: 10.1109/IGARSS39084.2020.9323213.

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

[17] S. Zhuang, P. Wang, G. Wang, D. Wang, J. Chen, and F. Gao, "Improving remote sensing image captioning by combining grid features and transformer," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–15, 2022, Art. no. 6504905, doi: 10.1109/LGRS.2021.3135711.

[18] B. Wang, X. Lu, X. Zheng, and X. Li, "Semantic captions of high-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1274–1278, Aug. 2019, doi: 10.1109/LGRS.2019.2893772.

[19] 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.

[20] P. Veličković et al., "Graph attention networks," 2017. [Online]. Available: http://arxiv.org/abs/1710.10903

[21] P. Kuznetsova et al., "Collective generation of natural image captions," in *Proc. Annu. Meeting Assoc. Comput. Linguist.*, 2012, pp. 359–368.

[22] A. Gupta and P. Mannem, "From image annotation to image caption," in *Proc. Int. Conf. Neural Inf. Process.*, 2012, pp. 196–204.

[23] G. Kulkarni et al., "Babytalk: Understanding and generating simple image captions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2891–2903, Dec. 2013.

[24] S. Li et al., "Composing simple image captions using web-scale N-grams," in *Proc. Conf. Comput. Natural Lang. Learn.*, 2011, pp. 220–228.

[25] I. Sutskever, O. Vinyals, and Q. Le, "Sequence to sequence learning with neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1310–1318.

[26] V. Ordonez et al., "Large scale retrieval and generation of image captions," *Int. J. Comput. Vis.*, vol. 119, no. 1, pp. 46–59, 2016.

[27] O. Vinyals et al., "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3156–3164.

[28] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biol. Cybern.*, vol 36, pp. 193–202, 1980, doi: 10.1007/BF00344251.

[29] P. Razvan et al., "How to construct deep recurrent neural networks," in *Proc. Int. Conf. Learn. Representations*, 2014.

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

[31] Z. Shi and Z. Zou, "Can a machine generate humanlike language captions for a remote sensing image?," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 6, pp. 3623–3634, Jun. 2017.

[32] S. Ren et al., "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[33] X. Zhang et al., "Caption generation for remote sensing images using attribute attention mechanism," *Remote Sens.*, vol. 11, no. 6, Mar. 2019, Art. no. 612.

[34] X. Ma, R. Zhao, and Z. Shi, "Multiscale methods for optical remote-sensing image captioning," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 11, pp. 2001–2005, Nov. 2021, doi: 10.1109/LGRS.2020.3009243.

[35] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015.

[36] Y. Wang, W. Zhang, Z. Zhang, X. Gao, and X. Sun, "Multiscale multi-interaction network for remote sensing image captioning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 2154–2165, 2022, doi: 10.1109/JSTARS.2022.3153636.

[37] 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, Art. no. 5629419.

[38] 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, Art. no. 5615216, doi: 10.1109/TGRS.2021.3132095.

[39] C. Zihang et al., "TypeFormer: Multi-scale transformer with type controller for remote sensing image caption," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022, Art. no. 6514005, doi: 10.1109/LGRS.2022.3192062.

[40] C. Liu, R. Zhao, and Z. Shi, "Remote-sensing image captioning based on multilayer aggregated transformer," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022, Art. no. 6506605, doi: 10.1109/LGRS.2022.3150957.
[41] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[42] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, "Image captioning: Transforming objects into words," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 11137–11147.
[43] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10575–10584, doi: 10.1109/CVPR42600.2020.01059.
[44] L. Meng, J. Wang, Y. Yang, and L. Xiao, "Prior knowledge-guided transformer for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–13, 2023, Art. no. 4706213.
[45] R. Liu et al., "An intriguing failing of convolutional neural networks and the coordconv solution," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, vol. 31, pp. 9628–9639.
[46] H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, and L.-C. Chen, "Axial-deeplab: Stand-alone axial-attention for panoptic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 108–126.
[47] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021.
[48] K. Wu, H. Peng, M. Chen, J. Fu, and H. Chao, "Rethinking and improving relative position encoding for vision transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10033–10041.
[49] B. Graham et al., "LeViT: A vision transformer in convnet's clothing for faster inference," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 12259–12269.
[50] L. Guo, J. Liu, X. Zhu, P. Yao, S. Lu, and H. Lu, " Normalized and geometry-aware selfattention network for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10327–10336.
[51] Q. Wang, W. Huang, X. Zhang, and X. Li, "GLCM: Global–local captioning model for remote sensing image captioning," *IEEE Trans. Cybern.*, vol. 53, no. 11, pp. 6910–6922, Nov. 2023.
[52] 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.
[53] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst.*, 2010, pp. 270–279.
[54] K. Papineni et al., "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguist.*, 2002, pp. 311–318.
[55] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop Intrinsic Extrinsic Eval. Measures Mach. Transl. Summarization*, 2005, pp. 65–72.
[56] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. Assoc. Comput. Linguistics Workshop*, 2004, pp. 74–81.
[57] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image caption evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4566–4575.
[58] P. Anderson et al., "Spice: Semantic propositional image caption evaluation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 382–398.
[59] J. Deng et al., "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
[60] G. Sumbul, S. Nayak, and B. Demiret, "SD-RSIC: Summarization-driven deep remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 6922–6934, Aug. 2021, doi: 10.1109/TGRS.2020.3031111.
[61] 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, doi: 10.1109/TGRS.2020.3044054.
[62] L. Huang, W. Wang, J. Chen, and X. -Y. Wei, "Attention on attention for image captioning," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 4634–4643.
[63] Y. Wang, W. Zhang, Z. Zhang, X. Gao, and X. Sun, "Multiscale multiinteraction network for remote sensing image captioning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 2154–2165, 2022.
[64] R. Du et al., "From plane to hierarchy: Deformable transformer for remote sensing image captioning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 7704–7717, 2023.
[65] Z. Li et al., "STADE-CDNet: Spatial–temporal attention with difference enhancement-based network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–17, 2024, Art. no. 5611617, doi: 10.1109/TGRS.2024.3371463.
[66] R. Hang, P. Yang, F. Zhou, and Q. Liu, "Multiscale progressive segmentation network for high-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2022, Art. no. 5412012, doi: 10.1109/TGRS.2022.3207551.
[67] Y. Hu et al., "RSGPT: A remote sensing vision language model and benchmark," 2023, *arXiv:2307.15266*.
[68] Y. Bazi. et al., "RS-LLaVA: A large vision-language model for joint captioning and question answering in remote sensing imagery," *Remote Sens.*, vol. 16, no. 9, 2024, Art. no. 1477.
[69] T. Zhao et al., "Artificial intelligence for geoscience: Progress, challenges and perspectives," *Innovation*, vol. 5, 2024, Art. no. 100691.

Anli Liu received the B.E. degree in intelligence science and technology in 2022 from the Nanjing University of Science and Technology (NJUST), Nanjing, China, where he is currently working toward the M.S. degree in computer science.

His current research interests include pattern recognition, computer vision, and machine learning.

Lingwu Meng received the B.E. degree in electronics from Henan Agricultural University, Zhengzhou, China, in 2018 and the M.S. degree in mechatronic engineering from the Shanghai University of Engineering Science, Shanghai, China, in 2021. He is currently working toward the Ph.D. degree in computer science with the Nanjing University of Science and Technology, Nanjing, China.

His current research interests include pattern recognition, computer vision, and machine learning.

Liang Xiao (Senior Member, IEEE) received the B.S. degree in applied mathematics and the Ph.D. degree in computer science from the Nanjing University of Science and Technology (NJUST), Nanjing, China, in 1999 and 2004, respectively.

From 2009 to 2010, he was a Postdoctoral Fellow with the Rensselaer Polytechnic Institute, Troy, NY, USA. Since 2014, he has been the Deputy Director of the Jiangsu Key Laboratory of Spectral Imaging Intelligent Perception, Nanjing. He was the Second Director of the Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, NJUST, where he is currently a Professor with the School of Computer Science. He has authored or coauthored more than 100 international journal articles including IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING. His main research interests include inverse problems in image processing, computer vision and image understanding, pattern recognition, and remote sensing.