

# Prior Knowledge-Guided Transformer for Remote Sensing Image Captioning

Lingwu Meng, Jing Wang, Yang Yang<sup>ID</sup>, and Liang Xiao<sup>ID</sup>, *Member, IEEE*

**Abstract**—Remote sensing image (RSI) captioning aims to generate meaningful and grammatically accurate sentences for RSIs. However, in comparison to natural image captioning, RSI captioning encounters additional challenges due to the unique characteristics of RSIs. The first challenge arises from the abundance of objects present in these images. As the number of objects increases, it becomes increasingly difficult to determine the main focus of the description. Moreover, the objects in RSIs often share similar appearances, which further complicates the generation of accurate descriptions. To overcome these challenges, we propose a prior knowledge-guided transformer (PKG-Transformer) for RSI captioning. First, scene-level and object-level features are extracted in a multilevel feature extraction (MFE) module. To further refine and enhance the extracted multilevel features, we introduce a feature enhancement (FE) module. This module utilizes a combination of graph neural networks and attention mechanisms to capture the correlation and difference between different objects or scene regions. Moreover, we propose a prior knowledge augmented attention (PKA) mechanism to select the objects that are more relevant to the scene regions by establishing the relationships between them. This attention mechanism is seamlessly integrated into the transformer structure, providing valuable prior knowledge that promotes the caption generation process. Extensive experiments on three RSI captioning datasets verify the superiority of the proposed method. Compared with the baseline methods, the proposed method achieves more impressive performance. The code will be publicly available at <https://github.com/One-paper-luck/PKG-Transformer>.

**Index Terms**—Image captioning, prior knowledge, remote sensing, transformer.

## I. INTRODUCTION

IMAGE captioning is the task of generating a meaningful and syntactically correct sentence to describe an image in natural language [1], [2], [3], [4]. Although great improvement has been witnessed in recent years, the research mainly focuses

Manuscript received 20 March 2023; revised 28 August 2023 and 22 September 2023; accepted 16 October 2023. Date of publication 27 October 2023; date of current version 16 November 2023. This work was supported in part by the Jiangsu Geological Bureau Research Project under Grant 2023KY11, in part by the Open Research Fund in 2021 of Jiangsu Key Laboratory of Spectral Imaging and Intelligent Sense under Grant JSGP202101 and Grant JSGP202204, and in part by the China Postdoctoral Science Foundation under Grant 2023TQ0181. (Corresponding author: Liang Xiao.)

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

Jing Wang is with the Department of Automation, Tsinghua University, Beijing 100084, China (e-mail: wangjing-wj@mail.tsinghua.edu.cn).

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

Digital Object Identifier 10.1109/TGRS.2023.3328181

1558-0644 © 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.  
See <https://www.ieee.org/publications/rights/index.html> for more information.

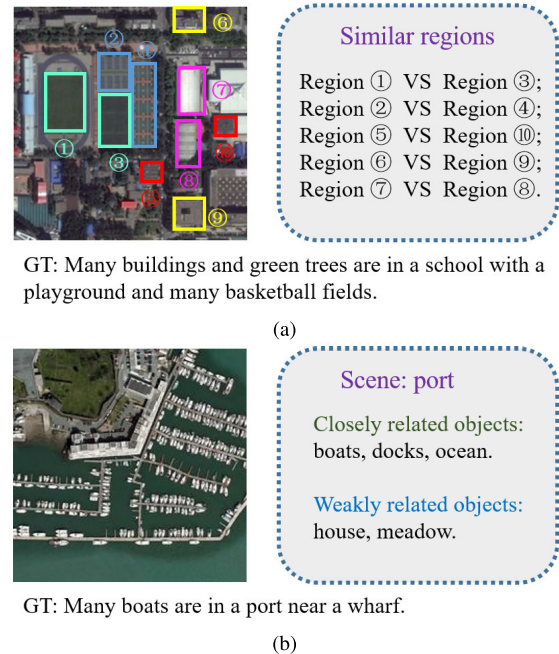


Fig. 1. Examples of (a) similar regions and (b) “port” scenes in RSIs. (a) Similar region pairs are listed on the right and the regions are highlighted in the image. (b) Objects that are closely or less related to the “port” scene are listed on the right. GT is the ground truth caption.

on images in the real world, leaving the description of remote sensing images (RSIs) still largely unexplored. In this article, we take one step further to study the RSI captioning (RSIC) task, which plays a critical role in the accurate and quick understanding of RSIs.

Compared to natural images, RSIs exhibit distinctive characteristics. First, a significant portion of the objects within RSIs possess a similar visual appearance. For instance, in Fig. 1(a), regions 2 and 4 represent “basketball fields” and “tennis courts,” respectively, despite their visual similarities. Second, RSIs typically contain a larger number of objects, which are more challenging to understand and describe.

These characteristics present great challenges for the captioning task in RSIs. 1) The first challenge stems from visually similar objects within RSIs. This similarity in appearance can result in potential misidentification by the model, leading to inaccuracies and inconsistencies in the generated descriptions. 2) The complexity of the captioning task is amplified by the presence of numerous objects, making it challenging to accurately capture and describe objects that are closely related to the scene. As shown in Fig. 1(b), objects such as “boat,”

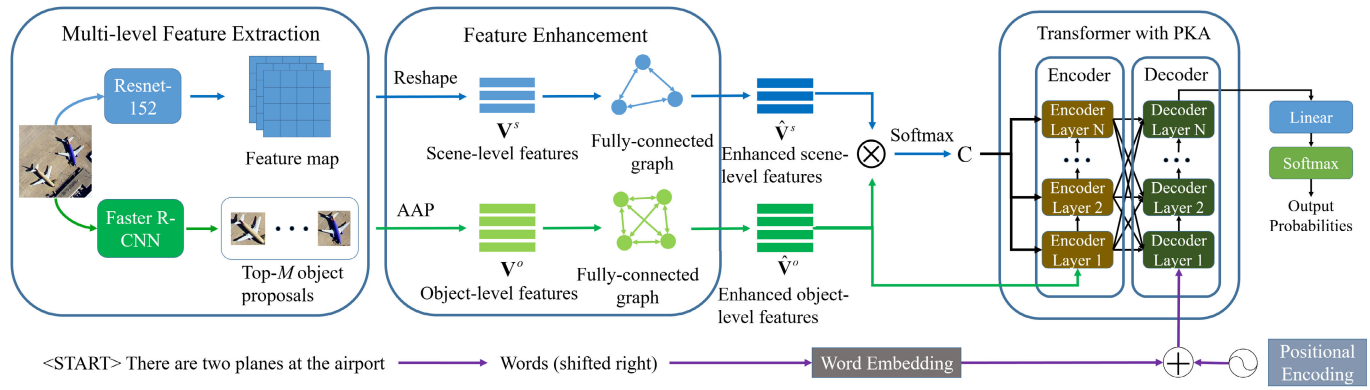


Fig. 2. Framework of the proposed method in this article. It consists of an MFE module, an FE module, and a transformer with PKA. The MFE module employs Faster R-CNN and ResNet-152 to extract object-level features  $V^o$  and scene-level features  $V^s$  from an input image, respectively. The FE module obtains enhanced object-level features  $\hat{V}^o$  and scene-level features  $\hat{V}^s$  by leveraging the object-object and scene-scene relationships. Then, we obtain a prior scene-object knowledge matrix  $C$  by calculating the dot product of the enhanced scene-level features  $\hat{V}^s$  and object-level features  $\hat{V}^o$ . Finally, a transformer with PKA is devised to generate descriptions based on the enhanced object-level features  $\hat{V}^o$  and text input (the sum of word embedding and positional encoding), guided by the prior knowledge. AAP,  $\oplus$ , and  $\otimes$  represent adaptive average pooling, addition operator, and matrix multiplication, respectively.

“dock,” and “ocean” are highly correlated with the concept of a “port” scene. Conversely, objects like “houses” and “meadows” are less related. When given such an RSI, existing methods [5], [6], [7], [8], [9], [10], [11], [12], [13], [14] may create generations like “some boats are sailing on the ocean” or “the water is dark green,” overlooking relevant objects or focusing on irrelevant content in the RSI. However, it is crucial to identify and describe the pertinent objects that contribute to the overall scene. To address the above challenges, some methods [15], [16], [17], [18] have attempted to extract multiscale or multilevel features from RSIs to obtain more detailed information. They employed various convolutional layers or neural network models to extract the features. However, these structures are not specifically designed to highlight the differences between objects, thereby limiting their effectiveness in comprehending RSIs.

In this article, we propose a solution to address these challenges by exploring two types of relationships in RSIs: intramodality (object-object/scene-scene) relationships and intermodality (scene-object) relationships. First, we study the object-object and scene-scene relationships within a feature enhancement (FE) module. This module refines the object-level and scene-level features based on the learned relationships using graph attention networks [19]. By incorporating the FE module, the correlation and difference between different objects or scene regions are injected into the object-level or scene-level features. This enables the model to distinguish between visually similar objects, resulting in more accurate descriptions for RSIs. Furthermore, we investigate the relationships between objects and scene regions in RSIs. The relationships are further integrated into the scene-object relationship into the transformer [5], [6], [7] structure as prior knowledge. By doing so, the model becomes aware of the degree of relevance between an object and a scene, which encourages the generation of more contextually relevant descriptions of the RSIs.

By consolidating the above ideas, we present a prior knowledge-guided transformer (PKG-Transformer) for RSIC, as shown in Fig. 2. The overall framework comprises a multilevel feature extraction (MFE) module, an FE module,

and a transformer with prior knowledge augmented attention (PKA). In the MFE module, object-level features and scene-level features are first extracted using faster R-CNN [20] and ResNet-152 [21], respectively. Then, the FE module combines graph attention networks and the multihead attention mechanism to enrich the features with more detailed information. Subsequently, the prior scene-object knowledge is acquired by establishing relationships between the enhanced object-level features and scene-level features. This acquired knowledge is then integrated into the transformer to facilitate the caption generation process.

The main contributions of this work can be summarized as follows.

- 1) We propose to enrich the object-level and scene-level features by leveraging the object-object and scene-scene relationship within the FE module. This effectively incorporates correlation and difference information between different objects or scene regions into their feature embeddings, resulting in a more precise understanding of RSIs.
- 2) We introduce a PKA mechanism and explore the relationships between the objects and scene regions in RSIs, which promotes the generation of contextually relevant descriptions of the images.
- 3) Extensive experiments are conducted on three RSIC datasets, demonstrating the superior performance of the proposed method.

## II. RELATED WORK

In this section, we will review the related works from two aspects: natural image captioning and RSI captioning.

### A. Natural Image Captioning

In the early stages, many template-based methods [22], [23], [24] and retrieval-based methods [25], [26], [27] were proposed for the image captioning task. However, these methods face challenges such as limited flexibility and reliance on hand-designed features, making it difficult to obtain high-quality captions. With the success of neural networks for

various computer vision and natural language processing tasks, deep-learning-based image captioning methods [4], [5], [7], [15], [18], [28] have made significant progress in recent years. Vinyals et al. [15] first introduced the encoder–decoder framework, where the encoder utilized a convolutional neural network (CNN) [16] to extract high-level visual features, and the decoder employed a recurrent neural network (RNN) [17] to generate captions. Furthermore, Xu et al. [18] incorporated the attention mechanism into the framework in order to enhance the focus on distinct regions of the feature maps within the image. Later, Rennie et al. [29] proposed a self-critical sequence training method, which adopted a better reward signal normalization method to achieve a more stable training process. Unlike the above methods that extract region-level features with pretrained CNNs [30], [31], [32]. Anderson et al. [5] extracted object-level features with the faster R-CNN and presented a new bottom–up–top–down combined attention. Since then, object-level features have been widely adopted in the NIC task. Nowadays, transformer-based methods [4], [6], [7], [28] have achieved state-of-the-art performance in the NIC task. Cornia et al. [28] incorporated prior information about the relationships between objects into self-attention to generate more precise captions. Later, Yan et al. [33] embedded a learnable memory vector inside the self-attention to obtain prior language knowledge.

### B. RSI Captioning

Inspired by the NIC methods, most RSIC methods adopt the CNN-RNN framework. Qu et al. [8] first proposed a deep multimodal neural network to select the best combination of CNN and RNN for RSIC. Later, Lu et al. [9] explored the performance of multimodal and attention-based methods, showcasing that attention-based methods can achieve the best performance. To accurately describe the semantic content of RSIs, Zhang et al. [34] proposed attribute attention to combine high-level features extracted from the relatively deep fully connected (FC) layer with low-level features extracted from the relatively shallow convolution layers or softmax layer. Furthermore, Zhang et al. [35] proposed a label-attention mechanism to utilize label information to guide the computation of attention masks. However, these methods ignore the relationship between objects and scene regions. To describe the regions related to the scene, Ma et al. [36] extracted scene-level semantic features from ResNet-50 and extracted target-level semantic features by using the convolution layers of VGG-16. Then, they fed the multilevel features and the previously hidden state  $h_{t-1}$  of the decoder output into the multihead self-attention mechanism [6] to obtain the context vectors at time step  $t$ . Later, Zhang et al. [37] proposed a global visual feature-guided attention mechanism to filter out redundant feature components in the fused local features and global features through an attention gate. To take full advantage of multiscale features, Wang et al. [38] collected features from conv4 and conv5 of ResNet. These features were concatenated as the image feature representation after self-attention and gated cross-attention. Soon after, Li et al. [39] adopted the efficient spatial pyramid [40] to extract multi-scale features from the spatial features obtained by a pretrained

CNN and fed them into an adaptive average pooling operation for a global representation. Then, the global representation is concatenated with the original CNN features. Finally, they designed a recurrent attention and semantic gate framework to facilitate the RSIC task.

Recently, the advanced transformer framework has been investigated for the RSIC task. Chen et al. [41] adopted a multiscale vision transformer for image representation and introduced a transformer decoder to generate sentences. In the same year, Liu et al. [42] proposed a multilayer aggregated transformer by combining transformer and LSTM [43]. First, they fused multiscale features extracted from different layers of ResNet-50. Then, the fused multiscale features were fed into the transformer encoder, and LSTM was used to aggregate the features from different encoding layers. Finally, the aggregated features were input into the transformer decoder to generate captions. Liu et al. [44] utilized pretrained large language models based on the vision transformer to describe the differences between bitemporal images by natural language.

While the aforementioned approaches may generate grammatically correct captions, they often fail to accurately capture the focus or subject of the image content. In this article, we propose a novel approach to address this issue by leveraging scene–object relationships. By incorporating this prior knowledge into the transformer structure, we guide the model to describe objects that are closely related to the scene, resulting in a more contextually appropriate description.

## III. APPROACH

### A. Overall Framework

The overall framework of the proposed method, PKG-Transformer, is shown in Fig. 2. It consists of an MFE module, an FE module, and a transformer with PKA. In the MFE module, a set of scene-level features  $\mathbf{V}^s$  and a set of object-level features  $\mathbf{V}^o$  are first extracted from an RSI  $\mathbf{I}$ . The FE module then enriches these obtained features by leveraging the object–object and scene–scene relationships, resulting in enhanced object-level features  $\hat{\mathbf{V}}^o$  and scene-level features  $\hat{\mathbf{V}}^s$ . Subsequently, the prior scene–object knowledge  $\mathbf{C}$  is acquired by establishing the relationships between the enhanced features. Finally, a transformer with PKA is devised to generate descriptions based on the enhanced features, guided by prior knowledge. Specifically, each encoding layer of the transformer employs a PKA that incorporates the prior knowledge  $\mathbf{C}$  into the self-attention mechanism. The transformer decoder then predicts new words based on the outputs from the encoder and the previously generated words.

### B. Multilevel Feature Representation

*Multilevel Feature Extraction:* To obtain a more detailed and comprehensive semantic representation for RSIs, scene-level features and object-level features are extracted, respectively. The scene-level feature captures the overall scene information of the entire image, while the object-level features correspond to specific objects within RSIs.

To extract the object-level features, faster R-CNN is used to produce a set of object regions, and the features for the  $M$  objects are denoted as  $\mathbf{V}^o = \{\mathbf{v}_1^o, \dots, \mathbf{v}_M^o\} \in \mathbf{R}^{M \times D_1}$ .

Following [36], we adopt a pretrained CNN (ResNet-152 in this article) to extract the scene-level features. The last FC layer is removed and the scene-level features  $\mathbf{V}^s = \{\mathbf{v}_1^s, \dots, \mathbf{v}_N^s\}$  are extracted after the adaptive average pooling operation, with the shape of  $\mathbf{R}^{N \times D_2}$ .

Finally, the object-level features  $\mathbf{V}^o$  and scene-level features  $\mathbf{V}^s$  are projected into a  $d$ -dimensional space.

*Feature Enhancement:* Inspired by the concept of leveraging adjacent nodes to strengthen the current node in graph structures, we propose a fusion of graph attention networks [19] and the multihead attention mechanism [6] to refine the object and scene features.

*Object-Level FE:* Given the object-level features  $\mathbf{V}^o$ , we build a fully-connected graph  $\mathbf{G}^o = (\mathbf{V}^o, \mathbf{E}^o)$ , where  $\mathbf{V}^o = \{\mathbf{v}_1^o, \dots, \mathbf{v}_M^o\}$  is the node set and  $\mathbf{E}^o$  is the edge set. The edge  $\mathbf{e}_{ij}^o$  between  $\mathbf{v}_i^o$  and  $\mathbf{v}_j^o$  is defined as the attention weight, which is obtained through dot-product attention

$$\begin{aligned} \mathbf{e}_{ij} &= \mathbf{v}_i^o \mathbf{W}_q^o (\mathbf{v}_j^o \mathbf{W}_k^o)^T / \sqrt{d} \\ \mathbf{e}_{ij}^o &= \text{softmax}(e_{ij}) = \frac{\exp(e_{ij})}{\sum_{k \in N_i} \exp(e_{ik})} \end{aligned} \quad (1)$$

where  $\mathbf{W}_q^o \in \mathbf{R}^{d \times d}$  and  $\mathbf{W}_k^o \in \mathbf{R}^{d \times d}$  are the learnable parameter matrices;  $d$  is the scaling factor; and  $N_i$  is the neighborhood of node  $i$  in the graph.

The refined object-level feature  $\hat{\mathbf{v}}_i^o$  is then computed based on the other nodes and the edges connecting them

$$\hat{\mathbf{v}}_i^o = \sum_{j \in N_i} \mathbf{e}_{ij}^o (\mathbf{v}_j^o \mathbf{W}_v^o) \quad (2)$$

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

Given the proven effectiveness of the multihead attention mechanism in various domains [6], [19], [45], we extend this concept to our refinement module. Consequently, the final object-level features, denoted as  $\hat{\mathbf{V}}^o \in \mathbf{R}^{M \times d}$ , can be expressed as follows:

$$\hat{\mathbf{V}}^o = \text{Concat}(\hat{\mathbf{v}}_{\text{head}_1}^o, \dots, \hat{\mathbf{v}}_{\text{head}_p}^o) \mathbf{W}^r \quad (3)$$

where  $\mathbf{W}^r \in \mathbf{R}^{d \times d}$  is the learnable parameter matrix, each head  $i$  refers to a round of feature refinement, and produces the refined  $\hat{\mathbf{v}}_{\text{head}_i}^o$ .

*Scene-Level FE:* In a similar manner, we build a fully-connected graph  $\mathbf{G}^s = (\mathbf{V}^s, \mathbf{E}^s)$  to refine the scene-level features, which produces the enhanced scene-level features  $\hat{\mathbf{V}}^s \in \mathbf{R}^{N \times d}$ .

### C. Prior Scene–Object Knowledge

The prior scene–object knowledge refers to the relationship between each object and the scene. Specifically, we compute the relationship between objects and all scene regions, rather than considering the entire scene as a whole. Given  $M$  objects and  $N$  scene regions, we begin by performing a dot-product operation between the refined object-level and scene-level features. Subsequently, we apply the softmax operation to each column, resulting in the knowledge matrix  $\mathbf{C} \in \mathbf{R}^{M \times N}$

$$\mathbf{C} = \text{softmax}\left(\frac{\hat{\mathbf{V}}^o \hat{\mathbf{V}}^s T}{\sqrt{d}}\right). \quad (4)$$

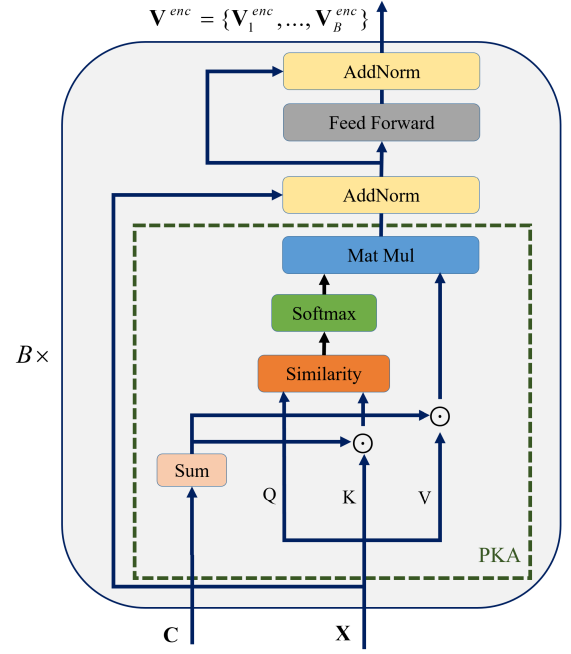


Fig. 3. Encoder for our transformer. The encoder is built with a stack of  $B$  identical encoding layers. Each encoding layer is mainly composed of the PKA. For the sake of clarity, the MHA module is not shown.  $\odot$  represents Hadamard product.

Each element in column  $j$  of  $\mathbf{C}$  represents the degree of relevance between the  $j$ th scene region and the objects, while the row  $i$  of  $\mathbf{C}$  denotes the relationship between the  $i$ th object and all scene regions. The acquired prior knowledge will be utilized to guide the process of generating captions in Section III-D.

### D. Transformer With PKA

Our transformer is composed of an encoder and a decoder module, both of which consist of stacked attention layers. The encoder module enriches the object-level features by incorporating the prior scene–object knowledge into the self-attention mechanism. The decoder module utilizes the output of each encoding layer to generate captions word by word. The specific details are explained below.

*Encoder:* The encoder of our transformer is based on the standard transformer architecture [6]. As shown in Fig. 3, the encoder is composed of a stack of  $B$  identical encoding layers. Each encoding layer consists of a PKA and a position-wise FC feed-forward network (FFN). A residual connection is employed around PKA or FFN, followed by layer normalization. The inputs to each encoding layer include the output of the previous encoding layer (object-level feature  $\hat{\mathbf{V}}^o$  for the first layer) and the prior scene–object knowledge matrix  $\mathbf{C}$ .

Let's consider the first encoding layer as an example. The PKA takes the object-level features  $\hat{\mathbf{V}}^o$  and the prior scene–object knowledge matrix  $\mathbf{C}$  as inputs. The queries  $\mathbf{Q}$  corresponds to the object-level features  $\hat{\mathbf{V}}^o$ , while the keys  $\mathbf{K}$  and values  $\mathbf{V}$  are the object-level features  $\hat{\mathbf{V}}^o$  filtered by the prior scene–object knowledge matrix  $\mathbf{C}$ . PKA can be

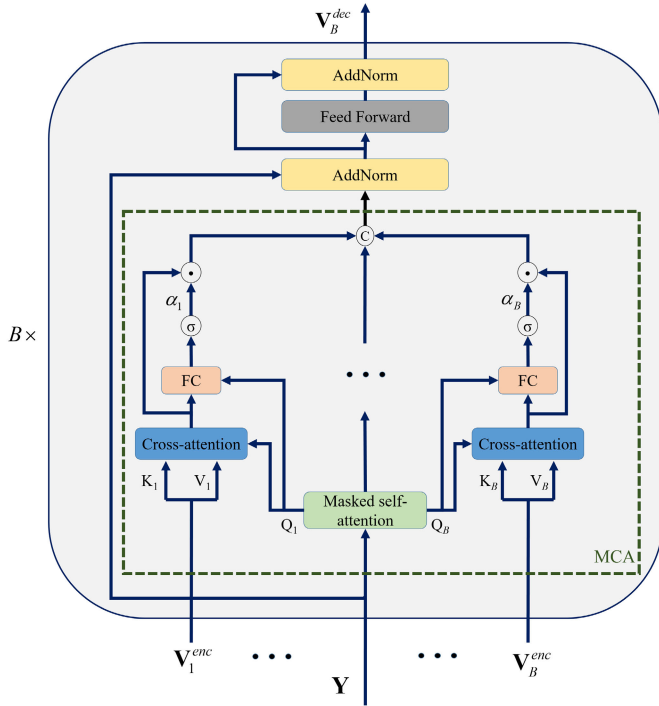


Fig. 4. Decoder for our transformer. The decoder is built with a stack of  $B$  identical decoding layers. Each decoding layer is mainly composed of MCA. For the sake of clarity, the MHA module, the final FC layer, and the softmax layer are not shown.  $\odot$ ,  $\otimes$ , and  $\oplus$  represent concatenation operation, Hadamard product, and sigmoid activation function, respectively.

represented mathematically as

$$\begin{aligned} \text{PKA}(\mathbf{X}, \mathbf{C}) &= \text{softmax}\left(\mathbf{Q}\mathbf{K}^T/\sqrt{d}\right)\mathbf{V} \\ \text{and } \mathbf{Q} &= \hat{\mathbf{V}}^o\mathbf{W}_q \\ \mathbf{K} &= (\hat{\mathbf{V}}^o\mathbf{W}_k) \odot (\text{sum}(\mathbf{C})\mathbf{W}_c) \\ \mathbf{V} &= (\hat{\mathbf{V}}^o\mathbf{W}_v) \odot (\text{sum}(\mathbf{C})\mathbf{W}_c) \end{aligned} \quad (5)$$

where  $\mathbf{W}_q \in \mathbf{R}^{d \times d}$ ,  $\mathbf{W}_k \in \mathbf{R}^{d \times d}$ ,  $\mathbf{W}_v \in \mathbf{R}^{d \times d}$ , and  $\mathbf{W}_c \in \mathbf{R}^{1 \times d}$  are learnable matrices;  $\text{sum}(\cdot)$  is utilized to calculate the row-wise sum of a matrix; the symbol  $\odot$  denotes the Hadamard product.

The key difference between PKA and self-attention is its ability to filter the keys and values based on the prior scene-object knowledge. This enables the model to incorporate scene-object knowledge into the self-attention mechanism, leading to improved inference of objects that are closely related to scenes.

The obtained features are then passed through a FFN layer to acquire the output  $\mathbf{V}_1^{\text{enc}}$  of the first encoding layer. Similarly, we can obtain the outputs of the encoder as  $\mathbf{V}^{\text{enc}} = \{\mathbf{V}_1^{\text{enc}}, \dots, \mathbf{V}_B^{\text{enc}}\}$ , where  $\mathbf{V}_i^{\text{enc}} \in \mathbf{R}^{M \times d}$  is from the  $i$ th layer.

**Decoder:** To fully utilize the output information from all the encoding layers, we adopt the decoder architecture from the  $M^2$  transformer [28]. As shown in Fig. 4, the decoder is composed of a series of  $B$  identical decoding layers. Each decoding layer consists of a meshed cross-attention (MCA) [28] and an FFN. We also adopt a residual connection around MCA or FFN, followed by layer normalization.

In the initial layer, we first apply the masked self-attention operation to the text sequence  $\mathbf{Y}$  to obtain  $\mathbf{Y}_{\text{mask}}$ . Then,

we apply cross-attention to each encoder layer output and the masked text sequence for  $B$  iterations. Given the  $i$ th encoder layer output  $\mathbf{V}_i^{\text{enc}}$  and the masked text sequence  $\mathbf{Y}_{\text{mask}}$ , the joint representation  $\mathbf{S}^i$  is calculated as follows:

$$\mathbf{S}^i = \text{softmax}\left(\left(\mathbf{Y}_{\text{mask}}\mathbf{W}_q^i\right)\left(\mathbf{V}_i^{\text{enc}}\mathbf{W}_k^i\right)^T/\sqrt{d}\right)\left(\mathbf{V}_i^{\text{enc}}\mathbf{W}_v^i\right) \quad (6)$$

where  $\mathbf{W}_q^i \in \mathbf{R}^{d \times d}$ ,  $\mathbf{W}_k^i \in \mathbf{R}^{d \times d}$ , and  $\mathbf{W}_v^i \in \mathbf{R}^{d \times d}$  are learnable matrices. Subsequently, we measure the relevance  $\alpha_i$  between  $\mathbf{S}^i$  and  $\mathbf{Y}_{\text{mask}}$  as follows:

$$\alpha_i = \sigma\left(\left[\mathbf{Y}_{\text{mask}}, \mathbf{S}^i\right]\mathbf{W}_i^\alpha + \mathbf{b}_i^\alpha\right) \quad (7)$$

where  $\sigma$  is the sigmoid activation function;  $\mathbf{W}_i^\alpha \in \mathbf{R}^{2d \times d}$  is the learnable parameter matrix;  $\mathbf{b}_i^\alpha \in \mathbf{R}^d$  is the bias; and  $[\cdot, \cdot]$  indicates concatenation operation. The preliminary hybrid feature  $\mathbf{Z}_1 \in \mathbf{R}^{M \times d}$  is obtained by aggregating all cross-attention results  $\mathbf{S}^1, \dots, \mathbf{S}^B$  according to their weights:  $\mathbf{Z}_1 = \sum_{i=1}^B \alpha_i \odot \mathbf{S}^i$ . This feature  $\mathbf{Z}_1$  is then further processed through an FFN. After the residual connection and layer normalization, we acquire the output  $\mathbf{V}_1^{\text{dec}}$  of the first decoding layer.

In the subsequent layers, the cross-attention operation is applied to the output of the previous decoding layer (rather than the text sequence  $\mathbf{Y}$  in the first layer) and the encoder outputs. The resulting output  $\mathbf{V}_B^{\text{dec}}$  from the last decoding layer is considered the final output of the decoder. Subsequently, the decoder output  $\mathbf{V}_B^{\text{dec}}$  is sequentially fed into an FC layer and a softmax layer to calculate the probability distribution over a vocabulary of possible words.

### E. Training and Objectives

**Training With Cross Entropy Loss:** Following a widely adopted approach in image captioning [5], [28], we start by training our model using the cross-entropy loss (XE):

$$L_{\text{XE}}(\theta) = -\sum_{t=1}^T \log(p_\theta(y_t^* | y_{1:t-1}^*)) \quad (8)$$

where  $y_{1:t-1}^* = [y_1^*, \dots, y_{t-1}^*]$  represents a portion of the ground truth (GT) sequence, specifically the sequence of words from the start till the  $(t-1)$ th time step;  $T$  is the maximum time step;  $\theta$  denotes the variable parameter of the model.

**CIDEr-D Score Optimization:** Then, we employ a modified version of the self-critical sequence training method [29]. During training, at each time step, we sample the top- $k$  words based on the decoder probability distribution and keep track of the top- $k$  sentences with the highest probability. Then, we compute the CIDEr scores for these  $k$  sentences, which serve as the rewards. To update the model parameters, we define the gradient expression based on the average reward of the sentences

$$\nabla_\theta L_{RL}(\theta) = -\frac{1}{k} \sum_{i=1}^k ((r(w^i) - b) \nabla_\theta \log p_\theta(w^i)) \quad (9)$$

where  $k$  is the number of samples;  $w^i$  denotes the  $i$ th sentence sampled in the beam;  $r(\cdot)$  represents the reward function;  $b = 1/k \sum_{i=1}^k r(w^i)$  is the baseline reward, which is introduced to ensure training stability.

## IV. EXPERIMENT RESULTS

### A. Datasets

This article evaluates the proposed PKG-Transformer on three RSIC datasets: Sydney-Captions [8], UCM-Captions [9], and RSICD [9].

*Sydney-Captions*: It consists of 613 RSIs selected from the Sydney Dataset [46]. This dataset contains seven land cover categories, the size of each image is  $500 \times 500$  pixels, and each image is associated with five human-annotated captions.

*UCM-Captions*: It consists of 2100 RSIs selected from UC Merced Land Use [47], and the target is distributed in 21 classes. Each class contains 100 images with a size of  $256 \times 256$  pixels. Each image is associated with five human-annotated captions.

*RSICD*: RSICD is the largest RSIC dataset among three datasets, which consists of 10921 RSIs selected from the AID dataset [48] and other platforms, such as Baidu Map and MapABC. This dataset contains 30 land cover categories and the image size is  $224 \times 224$  pixels. Unlike the above two datasets, each image in this dataset is associated with no more than five human-annotated captions.

### B. Evaluation Metrics

To evaluate the quality of the generated captions, we illustrate the RSIC performance by ten metrics: BLEU-n [49], METEOR [50], ROUGE<sub>L</sub> [51], CIDE<sub>r</sub> [52], SPICE [53],  $S_m^*$  [37], and  $S_m$  [37]. BLEU-n, METEOR, ROUGE<sub>L</sub>, CIDE<sub>r</sub>, and SPICE. They can be calculated by the COCO caption evaluation tool<sup>1</sup>. BLEU-n, METEOR, and ROUGE<sub>L</sub> are proposed for machine translation, which are more concerned with the accuracy of captions. CIDE<sub>r</sub> and SPICE are proposed for image captioning, focusing on whether captions are more human-like.  $S_m^*$  and  $S_m$  are overall metrics, which evaluate the two aspects in balance.

*BLEU-n*: BLEU-n is first applied to the machine translation task and later to the image captioning task. Generally, it divides candidate sentences and reference sentences into phrases (from 1- to 4-gram). Then, it calculates the proportion of phrase overlap between the candidate sentence and reference sentence to measure quality. The value of this evaluation metric is between 0 and 1. The closer the score is to 1, the higher the quality of the translation.

*METEOR*: This evaluation metric is also proposed for machine translation tasks. It refers to both precision and recall (harmonic average), where recall is more important than precision. It is designed to correct certain shortcomings of BLEU, which is closer to human translation.

*ROUGE<sub>L</sub>*: ROUGE is a set of metrics which is another metric to evaluate the quality of machine translation/abstracts. ROUGE<sub>L</sub> is commonly employed as an evaluation metric for image captioning. It measures the similarity of candidate sentences and reference sentences by calculating the longest common subsequence among them. Different from BLEU, it is more concerned with recalls.

*CIDE<sub>r</sub>*: It is specially designed for image captioning. It treats each sentence as a document and then computes a

term frequency-inverse document frequency (TF-IDF) vector. By calculating the weight of each n-gram, cosine similarity between the reference caption and the candidate one is further used to measure the consistency of captions.

*SPICE*: It is also specially designed for image captioning. It encodes the objects, attributes, and relations for the given sentence based on the graph-method. First, it parses the candidate caption and references caption into syntactic dependency trees through a probabilistic context-free grammar dependency parser. Then, a rule-based method is employed to map the dependency trees into a scene graph. Finally, it calculates the F-score of the objects, attributes, and relations by leveraging the parsed tree.

$S_m^*$  and  $S_m$ : They are averages of part of the above metrics. Their detailed definitions are as follows:

$$S_m^* = \frac{1}{4}(B4 + M + R + C) \quad (10)$$

$$S_m = \frac{1}{5}(B4 + M + R + C + S). \quad (11)$$

### C. Experimental Settings and Training Details

*Dataset Splitting*: To ensure a fair comparison with the compared methods [8], [9], [28], [37], [42], [56], [57], [58], [59], each dataset is randomly shuffled and divided into three parts for training, validation, and testing by the ratio of 80%, 10%, and 10%, respectively. To reduce the impact of randomness, we conducted five experiments on each of the three RSIC datasets. The best and worst results are excluded, and the remaining outcomes are averaged to obtain solid results.

*Multilevel Feature Extraction*: For **object-level** features, we fine-tune the Faster R-CNN model with a ResNet-50 backbone pretrained on the ImageNet dataset [54] on each RSIC dataset. During the fine-tuning process, we freeze the batch normalization layers for the first 200 epochs and then unfreeze them for the remaining 1000 epochs. The initial learning rate is set to  $1e - 5$ , and the IoU threshold is set to 0.3. For an RSI, we select the top 50 object proposals as object-level features (i.e.,  $M = 50$ ), with each object having a dimension of 1024 (i.e.,  $D_1 = 1024$ ). For **scene-level** features, we remove the FC layer of ResNet-152 pretrained on the ImageNet dataset. The scene-level features are extracted after the adaptive average pooling operation. This process results in a feature map with dimensions of  $14 \times 14 \times 2048$ . We then flatten this feature map into a matrix of size  $196 \times 2048$  (i.e.,  $N = 196$  and  $D_2 = 2048$ ). The projection space dimensionality is set to  $d = 512$  for both the object-level and scene-level features.

*Model Setting and Training*: Following [28], we use  $B = 3$  identical layers in a sequential transformer framework and employ  $h = 8$  parallel attention layers in multihead attention. The word embedding dimension is set to 512. The positional encoding is applied to sequences with a maximum length of 128, and the output word sequence has a maximum length constraint of 20. For model training, we set the batch size to 50 and patience to 5. Dropout keeps a probability of 0.9 after each attention and FFN layer. We employ the Adam optimizer with  $\beta_1 = 0.9$ ,  $\beta_2 = 0.98$  to optimize the model parameters. And the learning rate is set to 1.

<sup>1</sup><https://github.com/tylin/coco-caption>

TABLE I

COMPARISON RESULTS ON SYDNEY-CAPTIONS. ALL RESULTS ARE REPORTED AS PERCENTAGE (%). BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, ROUGE<sub>L</sub>, CIDER, AND SPICE ARE DENOTED AS B1, B2, B3, B4, M, R, C, AND S, RESPECTIVELY. THE SYMBOL “-” INDICATES THAT THE RESULT IS NOT REPORTED BY THE ARTICLE

Methods	B1	B2	B3	B4	M	R	C	S	$S_m^*$	$S_m$
mRNN [8]	51.30	37.50	20.40	19.30	18.50	-	32.20	-	-	-
mLSTM [8]	54.60	39.50	22.30	21.20	20.50	-	37.20	-	-	-
Soft-attention [9]	73.22	66.74	62.23	58.20	39.42	71.27	249.93	-	104.71	-
Hard-attention [9]	75.91	66.10	58.89	52.58	38.98	71.89	218.19	-	95.41	-
SD-RSIC [57]	73.3	61.9	51.7	42.5	31.8	62.0	114.6	-	62.73	-
Word-Sentence Framework [58]	78.91	70.94	63.17	56.25	41.81	69.22	204.11	-	92.85	-
AoANet [59]	75.20	66.20	58.85	52.30	37.92	69.31	228.99	42.09	97.13	86.12
GVFGA+LSGA [37]	76.81	68.46	61.45	55.04	38.66	70.30	245.22	45.32	102.31	90.91
MLAT [42]	82.96±2.15	77.77±2.27	72.14±2.72	67.97±3.37	44.24±3.25	75.88±4.14	278.97±11.23	-	116.77±2.27	-
M <sup>2</sup> 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
Baseline	80.17±2.53	73.38±2.55	67.73±1.91	65.53±1.46	41.62±1.18	72.40±1.90	275.29±11.89	42.06±1.28	112.96±2.42	98.78±2.19
PKG-Transformer	<b>83.17±1.02</b>	<b>77.83±2.31</b>	<b>72.84±1.88</b>	<b>68.24±1.39</b>	<b>45.28±0.86</b>	<b>77.06±1.54</b>	<b>284.76±10.84</b>	<b>44.05±1.05</b>	<b>118.83±2.51</b>	<b>103.88±2.15</b>

All experiments are conducted on an NVIDIA GeForce RTX 3090 with PyTorch version 1.10.0.

We employ a two-stage training approach to improve the performance of the model. In the first stage, we optimize the model using the cross entropy loss and follow the learning rate scheduling strategy from [6], which incorporates a warm-up operation comprising 10 000 iterations. After completing the first stage of training, we assess the model’s performance on the validation set at each epoch and select the model with the highest CIDEr score as the basis for the subsequent training phase. The second stage of training focuses on CIDEr optimization with a fixed learning rate of  $5e - 6$ . During the optimization and decoding process, a beam size of 5 is employed to generate multiple candidate sequences.

*Compared Methods:* To evaluate the effectiveness of the proposed PKG-Transformer, we compare our method with several state-of-the-art methods as below.

a) mRNN [8] and mLSTM [8] use VGG-16 as their encoders but employ different RNNs (naive RNN, LSTM, and GRU [55]) as their decoders.

b) Soft-attention [9] and Hard-attention [9] are based on soft attention [5] and hard attention [5], respectively. They utilize VGG-16 as the encoders and LSTM as the decoders.

c) RTRMN (semantic) [56] and RTRMN (statistical) [56] select semantic topics and statistical topics from RSIC datasets. Then, they incorporate the retrieved topic words into a recurrent memory network to guide the sentence generation process. The input image is represented by ResNet-101.

d) SD-RSIC [57] summarizes the GT captions into a single caption for an RSI. Then, it integrates the summarized captions with standard captions using an adaptive weighting strategy. It employs ResNet-152 as the encoder and LSTM as the decoder.

e) Word-Sentence Framework [58] is a two-stage method, consisting of a word extractor and a sentence generator. It employs ResNet-18 as the word extractor and the transformer as the sentence generator.

f) AoANet [59] designs an attention-on-attention (AoA) mechanism to establish the relationship between the image features. The decoder contains an LSTM and an AoA module. The results are from [37].

g) GVFGA + LSGA [37] introduces a global visual feature-guided attention based on the AoA mechanism to filter

out redundant information in the image features. It designs a linguistic state-guided attention to further refine the fusion of visual features and textual features. It also adopts the CNN-RNN framework.

h) MLAT [42] is a multiscale RSIC method. It takes fused multiscale features as input to the transformer encoder. Then, it aggregates the encoding layers with LSTMs. Finally, it inputs the aggregated representation to the decoder to generate sentences. The encoder and decoder are both composed of 3 identical layers.

i) M<sup>2</sup> transformer [28] embeds the prior knowledge of relationships between objects into self-attention in the transformer encoder. This model adopts three identical layers in both the encoder and decoder modules.

j) To evaluate the individual contributions of each component and facilitate comparison with the M<sup>2</sup> transformer, we establish our baseline as a combination of the encoder from the standard Transformer and the decoder from the M<sup>2</sup> transformer. This baseline architecture also utilizes three identical layers in both the encoder and decoder.

#### D. Comparison With Other Methods

We present the comparison results with the other methods in Tables I–III, which correspond to Sydney-Captions, UCM-Captions, and RSICD datasets, respectively. All the results are reported as percentages (%).

Table I shows the comparison results of our method with other methods on Sydney-Captions. Overall, our method achieves the best performance, outperforming the current state-of-the-art method MLAT in terms of METEOR, ROUGE<sub>L</sub>, CIDEr, and  $S_m^*$  by 2.35%, 1.56%, 2.08%, and 1.76% (relative improvements, same below), respectively. The methods mRNN and mLSTM, which are solely based on the CNN-RNN framework exhibit the least favorable performance, with CIDEr scores of 32.20% and 37.20%, respectively. This is because they do not consider the spatial relationships between regions. Subsequently, the incorporation of attention mechanisms results in substantial performance gains. Notably, Soft-attention significantly surpasses mLSTM, particularly on CIDEr with an impressive gain of 249.93%. Moreover, the employment of the transformer architecture in our designed baseline further enhances the CIDEr score by 12.26% com-

TABLE II

COMPARISON RESULTS ON UCM-CAPTIONS. ALL RESULTS ARE REPORTED AS PERCENTAGE (%). BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, ROUGE<sub>L</sub>, CIDER, AND SPICE ARE DENOTED AS B1, B2, B3, B4, M, R, C, AND S, RESPECTIVELY. THE SYMBOL “-” INDICATES THAT THE RESULT IS NOT REPORTED BY THE ARTICLE

Methods	B1	B2	B3	B4	M	R	C	S	$S_m^*$	$S_m$
mRNN [8]	60.10	50.70	32.80	20.80	19.30	-	42.80	-	-	-
mLSTM [8]	63.50	53.20	37.50	21.30	20.30	-	44.50	-	-	-
Soft-attention [9]	74.54	65.45	58.55	52.50	38.86	72.37	261.24	-	106.24	-
Hard-attention [9]	81.57	73.12	67.02	61.82	42.63	76.98	299.47	-	120.23	-
RTRMN (semantic) [56]	55.26	45.15	39.62	35.87	25.98	55.38	180.25	26.57	74.37	64.81
RTRMN (statistical) [56]	80.28	73.22	68.21	63.93	42.58	77.26	312.70	45.35	124.12	108.36
SD-RSIC [57]	71.4	62.5	55.3	49.2	36.3	65.8	197.8	-	87.28	-
Word-Sentence Framework [58]	79.31	72.37	66.71	62.02	43.95	71.32	278.71	-	114.00	-
AoANet [59]	81.85	74.73	68.80	63.27	41.30	75.43	308.73	43.96	122.18	106.54
GVFGA+LSGA [37]	83.19	76.57	71.03	65.96	44.36	78.45	332.70	48.53	130.37	114.00
MLAT [42]	86.42±1.37	82.28±1.29	78.51±1.35	77.98±1.28	50.45±2.77	83.75±2.05	382.50±16.72	-	152.05±5.01	-
M <sup>2</sup> 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
Baseline	84.47±1.19	80.16±1.02	76.66±1.21	73.57±1.34	48.28±2.00	79.77±2.46	363.82±17.39	50.61±2.35	141.36±5.80	123.21±5.11
PKG-Transformer	<b>90.48±1.15</b>	<b>87.04±1.09</b>	<b>84.10±1.33</b>	<b>81.39±1.67</b>	<b>54.66±2.05</b>	<b>86.57±2.00</b>	<b>427.49±12.37</b>	<b>57.01±2.40</b>	<b>162.53±3.52</b>	<b>141.42±3.09</b>

TABLE III

COMPARISON RESULTS ON RSICD. ALL RESULTS ARE REPORTED AS PERCENTAGE (%). BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, ROUGE<sub>L</sub>, CIDER, AND SPICE ARE DENOTED AS B1, B2, B3, B4, M, R, C, AND S, RESPECTIVELY. THE SYMBOL “-” INDICATES THAT THE RESULT IS NOT REPORTED BY THE ARTICLE

Methods	B1	B2	B3	B4	M	R	C	S	$S_m^*$	$S_m$
mRNN [8]	45.58	28.25	18.09	12.13	15.69	31.26	19.15	-	19.56	-
mLSTM [8]	50.57	32.42	23.19	17.46	17.84	35.02	31.61	-	25.48	-
Soft-attention [9]	67.53	53.08	43.33	36.17	32.55	61.09	196.43	-	81.56	-
Hard-attention [9]	66.69	51.82	41.64	34.07	32.01	60.84	179.25	-	76.54	-
RTRMN (semantic) [56]	62.01	46.23	36.44	29.71	28.29	55.39	151.46	33.22	66.21	59.61
RTRMN (statistical) [56]	61.02	45.14	35.35	28.59	27.51	54.52	148.20	32.36	64.71	58.24
SD-RSIC [57]	64.40	47.40	36.90	30.0	24.90	52.30	79.40	-	46.65	-
Word-Sentence Framework [58]	<b>72.40</b>	58.61	49.33	42.50	31.97	<b>62.60</b>	206.29	-	85.84	-
AoANet [59]	67.18	55.52	47.35	41.01	32.51	58.52	256.47	46.12	97.13	86.93
GVFGA+LSGA [37]	67.79	56.00	47.81	41.65	32.85	59.29	260.12	46.83	98.48	88.15
MLAT [42]	68.20±1.45	57.18±1.24	49.43±1.30	42.98±1.27	32.22±1.28	59.60±1.42	265.58±5.53	-	98.80±1.55	-
M <sup>2</sup> 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
Baseline	66.86±1.79	55.94±0.83	48.09±0.94	41.99±1.03	31.74±0.47	58.79±0.40	255.84±3.39	44.02±0.71	97.09±1.05	86.48±0.99
PKG-Transformer	69.67±1.74	<b>58.30±1.39</b>	<b>50.45±1.13</b>	<b>44.31±1.24</b>	<b>33.32±1.31</b>	60.78±1.30	<b>274.01±2.88</b>	<b>46.91±0.79</b>	<b>103.11±1.31</b>	<b>91.87±1.20</b>

pared to GVFGA + LSGA. Finally, by exploring two types of relationships, namely, object-object/scene-scene relationship and scene-object relationship in RSIs, our method achieves a 3.44% higher CIDEr score compared to the baseline.

The comparative results on UCM-Captions and RSICD are shown in Tables II and III, respectively. Consistent with the findings from Sydney-Captions, the experimental results indicate that transformer-based methods outperform both the CNN-RNN-based methods and attention-based methods. In Table II, our method demonstrates improvements across all metrics when compared to the best performing M<sup>2</sup> transformer for UCM-Captions, with the CIDEr score showing the most significant increase (+2.17%). Furthermore, in Table III, our method surpasses the best-performing MLAT on most metrics, especially the CIDEr score (+3.17%). Only metrics BLEU-1 and ROUGE<sub>L</sub> are slightly lower than Word-Sentence Framework. Experimental results across the three RSIC datasets confirm the generality of our method.

### E. Ablation Study

To verify the effectiveness of each component, we conducted ablation experiments by removing the FE module, the

PKA, and both components individually on the three RSIC datasets. The experimental results are shown in Tables IV-VI.

The experimental results for the three datasets have similar trends. Taking Table V as an example, it presents the experimental results on UCM-Captions. The FE module alone leads to improvements of 14.24% on CIDEr and 11.77% on  $S_m$ , affirming its efficacy. This improvement is a result of the FE module facilitating the generation of stronger image representations. Specifically, the FE module establishes a fully-connected graph to enhance the relationships between each feature node and its neighbors, thereby producing more effective scene-level features and object-level features.

Similarly, the PKA alone yields notable improvements in CIDEr and  $S_m$  by 15.79% and 12.77%, which highlights the effectiveness of the PKA. This improvement can be attributed to the encoding of prior knowledge regarding the relationships between objects and scene regions in PKA, thereby enhancing the model’s ability to generate scene-related captions.

When both the PKA and FE modules are employed, the CIDEr and  $S_m$  achieve superior performance. Compared to using the FE module alone, the joint utilization of the PKA and FE module leads to further enhancements on CIDEr (+2.86%) and  $S_m$  (+2.69%). Similarly, compared to solely utilizing the PKA, incorporating both the PKA and FE

TABLE IV

ABLATION STUDIES ON SYDNEY-CAPTIONS. ALL RESULTS ARE REPORTED AS PERCENTAGE (%). BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, ROUGE<sub>L</sub>, CIDER, AND SPICE ARE DENOTED AS B1, B2, B3, B4, M, R, C, AND S, RESPECTIVELY

FE	PKA	B1	B2	B3	B4	M	R	C	S	$S_m^*$	$S_m$
✗	✗	80.17±2.53	73.38±2.55	67.73±1.91	65.53±1.46	41.62±1.18	72.40±1.90	275.29±11.89	42.06±1.28	112.96±2.42	98.78±2.19
✓	✗	82.52±1.33	76.95±2.93	71.85±1.66	67.01±1.44	44.81±0.93	76.13±1.66	278.71±10.25	43.50±1.19	117.42±2.48	102.68±2.53
✗	✓	82.82±1.06	77.19±2.24	72.18±1.63	67.65±1.30	44.72±0.94	76.89±1.51	280.45±11.21	43.73±1.21	118.18±2.19	103.24±2.59
✓	✓	<b>83.17±1.02</b>	<b>77.83±2.31</b>	<b>72.84±1.88</b>	<b>68.24±1.39</b>	<b>45.28±0.86</b>	<b>77.06±1.54</b>	<b>284.76±10.84</b>	<b>44.05±1.05</b>	<b>118.83±2.51</b>	<b>103.88±2.15</b>

TABLE V

ABLATION STUDIES ON UCM-CAPTIONS. ALL RESULTS ARE REPORTED AS PERCENTAGE (%). BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, ROUGE<sub>L</sub>, CIDER, AND SPICE ARE DENOTED AS B1, B2, B3, B4, M, R, C, AND S, RESPECTIVELY

FE	PKA	B1	B2	B3	B4	M	R	C	S	$S_m^*$	$S_m$
✗	✗	84.47±1.19	80.16±1.02	76.66±1.21	73.57±1.34	48.28±2.00	79.77±2.46	363.82±17.39	50.61±2.35	141.36±5.80	123.21±5.11
✓	✗	88.96±1.77	85.33±1.60	82.39±1.14	79.67±1.78	51.75±1.30	84.13±1.55	415.62±13.94	54.53±1.13	157.79±2.78	137.71±2.45
✗	✓	89.24±1.05	85.78±1.17	82.97±1.64	80.40±1.43	52.81±1.98	85.22±1.18	421.25±9.51	53.99±3.62	159.92±2.49	138.94±3.62
✓	✓	<b>90.48±1.15</b>	<b>87.04±1.09</b>	<b>84.10±1.33</b>	<b>81.39±1.67</b>	<b>54.66±2.05</b>	<b>86.57±2.00</b>	<b>427.49±12.37</b>	<b>57.01±2.40</b>	<b>162.53±3.52</b>	<b>141.42±3.09</b>

TABLE VI

ABLATION STUDIES ON RSICD. ALL RESULTS ARE REPORTED AS PERCENTAGE (%). BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, ROUGE<sub>L</sub>, CIDER, AND SPICE ARE DENOTED AS B1, B2, B3, B4, M, R, C, AND S, RESPECTIVELY

FE	PKA	B1	B2	B3	B4	M	R	C	S	$S_m^*$	$S_m$
✗	✗	66.86±1.79	55.94±0.83	48.09±0.94	41.99±1.03	31.74±0.47	58.79±0.40	255.84±3.39	44.02±0.71	97.09±1.05	86.48±0.99
✓	✗	68.77±0.95	57.77±1.25	49.77±1.28	43.62±1.26	32.85±0.52	60.09±0.89	267.37±3.12	45.97±0.73	100.98±1.18	89.98±0.92
✗	✓	68.72±1.40	58.03±1.23	50.17±1.36	43.99±1.53	32.97±0.58	60.50±0.78	269.41±2.57	46.17±0.55	101.72±1.32	90.61±0.75
✓	✓	<b>69.67±1.74</b>	<b>58.30±1.39</b>	<b>50.45±1.13</b>	<b>44.31±1.24</b>	<b>33.32±1.31</b>	<b>60.78±1.30</b>	<b>274.01±2.88</b>	<b>46.91±0.79</b>	<b>103.11±1.31</b>	<b>91.87±1.20</b>

TABLE VII

EXPERIMENTAL RESULTS OF IMAGE DEGRADATION ON SYDNEY-CAPTIONS. ALL RESULTS ARE REPORTED AS PERCENTAGE (%). BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, ROUGE<sub>L</sub>, CIDER, AND SPICE ARE DENOTED AS B1, B2, B3, B4, M, R, C, AND S, RESPECTIVELY. THE SYMBOL "w/o" MEANS NO NOISE ADDED

Noise	Method	B1	B2	B3	B4	M	R	C	S	$S_m^*$	$S_m$
Case1	Baseline	80.94±2.78	74.94±3.10	69.62±3.64	64.85±3.99	<b>44.06±2.43</b>	75.11±2.45	262.61±7.82	42.42±1.02	111.66±3.40	97.81±2.86
	PKG-Transformer	<b>81.21±0.89</b>	<b>75.53±0.28</b>	<b>70.26±1.15</b>	<b>65.37±1.77</b>	42.98±0.52	<b>75.37±0.79</b>	<b>269.82±31.09</b>	<b>44.10±1.18</b>	<b>113.38±8.54</b>	<b>99.53±7.07</b>
Case2	Baseline	<b>82.83±3.36</b>	<b>76.88±4.59</b>	<b>71.55±5.66</b>	<b>66.78±6.44</b>	<b>43.52±2.04</b>	<b>75.50±2.15</b>	264.33±8.54	<b>43.09±1.17</b>	112.53±2.80	98.64±2.47
	PKG-Transformer	81.54±0.45	75.87±1.56	70.76±2.72	66.09±3.55	42.78±1.14	74.77±1.40	<b>268.02±18.52</b>	42.64±1.38	<b>112.91±6.15</b>	<b>98.86±5.20</b>
Case3	Baseline	80.18±7.26	<b>74.40±7.81</b>	<b>68.96±8.32</b>	<b>64.08±8.59</b>	41.53±4.13	72.40±6.15	237.93±27.17	39.18±4.54	103.98±11.51	91.02±10.12
	PKG-Transformer	<b>80.36±0.32</b>	74.19±0.65	68.67±1.28	63.71±1.99	<b>42.70±1.24</b>	<b>74.44±0.91</b>	<b>257.84±14.91</b>	<b>43.17±0.50</b>	<b>109.67±3.00</b>	<b>96.37±2.50</b>
Case4	Baseline	82.11±3.78	75.95±5.53	70.62±6.46	65.87±7.12	43.61±2.94	75.06±2.80	267.21±4.93	43.97±0.05	112.94±4.45	99.14±3.57
	PKG-Transformer	<b>83.32±2.86</b>	<b>77.91±3.31</b>	<b>73.05±3.51</b>	<b>68.58±3.57</b>	<b>45.49±1.66</b>	<b>76.95±2.13</b>	<b>279.74±17.51</b>	<b>44.71±0.56</b>	<b>117.69±6.03</b>	<b>103.09±4.94</b>
Case5	Baseline	<b>81.44±3.49</b>	<b>75.49±4.95</b>	<b>70.27±5.88</b>	<b>65.56±6.45</b>	<b>44.36±2.74</b>	<b>75.34±3.16</b>	265.75±2.23	<b>44.22±0.29</b>	112.75±3.33	<b>99.05±2.72</b>
	PKG-Transformer	80.94±2.12	75.03±2.35	69.82±2.70	65.13±3.04	43.51±0.54	75.31±1.41	<b>268.23±4.73</b>	42.00±2.55	<b>113.04±1.19</b>	98.84±0.94
Case6	Baseline	<b>82.24±1.06</b>	75.61±1.77	69.52±2.46	64.24±2.94	42.90±2.91	74.28±2.86	244.76±6.63	41.48±0.30	106.55±1.03	93.53±0.88
	PKG-Transformer	81.21±0.83	<b>75.64±0.72</b>	<b>70.49±1.06</b>	<b>65.75±1.70</b>	<b>43.41±1.26</b>	<b>75.65±2.45</b>	<b>267.72±14.07</b>	<b>44.06±1.98</b>	<b>113.13±4.47</b>	<b>99.32±3.97</b>
w/o	Baseline	80.17±2.53	73.38±2.55	67.73±1.91	65.53±1.46	41.62±1.18	72.40±1.90	275.29±11.89	42.06±1.28	112.96±2.42	98.78±2.19
	PKG-Transformer	<b>83.17±1.02</b>	<b>77.83±2.31</b>	<b>72.84±1.88</b>	<b>68.24±1.39</b>	<b>45.28±0.86</b>	<b>77.06±1.54</b>	<b>284.76±10.84</b>	<b>44.05±1.05</b>	<b>118.83±2.51</b>	<b>103.88±2.15</b>

module results in notable improvements on CIDER (+1.48%) and  $S_m$  (+1.78%). This observation indicates that the PKA and the FE module exhibit complementary effects. The FE module enhances the internal relationships among features, thereby improving the quality of prior scene-object knowledge. Subsequently, the PKA assimilates this prior knowledge, aiding the model in selecting objects closely related to scenes.

In conclusion, both the FE module and PKA contribute to improved metric accuracy, with the best performance achieved when both modules are utilized in combination rather than independently.

*Noise Analysis:* In order to evaluate the robustness of the proposed method, we also analyze the effects of Gaussian noise and Gaussian blur on RSIs. To accomplish this, we assess the algorithm's performance under six distinct



**GT:** There is a small tennis court next to a basketball court surrounded by some plants and some cars parked beside.  
**Baseline:** Four tennis courts are surrounded by some buildings and plants and some cars parked beside.  
**M<sup>2</sup> Transformer:** There are two tennis courts next to a basketball court next to buildings.  
**PKG-Transformer:** There is a small tennis court next to a basketball court surrounded by some trees and some cars parked beside.



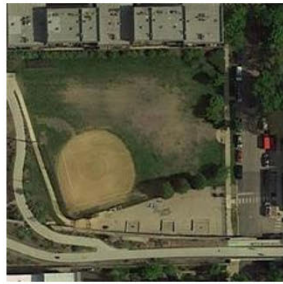
**GT:** Three tennis courts arranged neatly with some plants surrounded.  
**Baseline:** Four tennis courts are surrounded by some plants.  
**M<sup>2</sup> Transformer:** There are Four tennis courts on the basketball ground surrounded by some plants.  
**PKG-Transformer:** There are three tennis courts arranged neatly and surrounded by some plants.



**GT:** Lots of mobile homes are closed to each other and some roads go through this area.  
**Baseline:** Many mobile homes parked at the mobile home park.  
**M<sup>2</sup> Transformer:** Lots of mobile homes are closed to each other in the mobile home park.  
**PKG-Transformer:** Many mobile homes arranged haphazardly in the mobile home park and some roads go through this area.



**GT:** Some green trees are around two basketball fields and three tennis courts.  
**Baseline:** A playground is surrounded by many green trees and buildings.  
**M<sup>2</sup> Transformer:** A playground with a basketball field in it is surrounded by many green trees.  
**PKG-Transformer:** Three tennis courts and two basketball fields are surrounded by many green trees.



**GT:** Some buildings and green trees are around a baseball field.  
**Baseline:** A baseball field is next to a crossroads.  
**M<sup>2</sup> Transformer:** A baseball field is surrounded by many buildings.  
**PKG-Transformer:** A baseball field is next to a crossroads and some buildings.



**GT:** Four baseball fields in different sizes are near several buildings and green trees.  
**Baseline:** Four baseball fields are surrounded by some green trees.  
**M<sup>2</sup> Transformer:** Four baseball fields are near some green trees.  
**PKG-Transformer:** Four baseball fields are near several green trees and buildings.



**GT:** Many green trees and a building are around three tennis courts and a baseball field.  
**Baseline:** A baseball field is surrounded by many green trees.  
**M<sup>2</sup> Transformer:** A baseball field is surrounded by many green trees.  
**PKG-Transformer:** A baseball field with three tennis courts in it is next to a road.



**GT:** A small tennis court with some plants and buildings beside.  
**Baseline:** A medium residential area with a road goes through this area.  
**M<sup>2</sup> Transformer:** There is a small tennis court with a swimming pool beside.  
**PKG-Transformer:** A tennis court is surrounded by some trees while a swimming pool beside.

Fig. 5. Examples of captions generated by the baseline, M<sup>2</sup> transformer, and the proposed PKG-transformer, as well as the corresponding GT captions. Some detailed and accurate words are highlighted in blue. The words that are inconsistent with the image content are highlighted in red.

degradation cases. This analysis allows us to understand how the proposed method performs under various levels of noise and blur. The six cases are described as follows:

- Case1:* Gaussian noise is added to each image with a mean of 0 and a variance of 0.01.
- Case2:* Gaussian noise is added to each image with a mean of 0 and a variance of 0.1.
- Case3:* Gaussian noise is added to each image with a mean of 3 and a variance of 0.01.
- Case4:* Gaussian blur is added to each image with a kernel size of 3 and a standard deviation of 0.5.
- Case5:* Gaussian blur is added to each image with a kernel size of 5 and a standard deviation of 0.5.
- Case6:* Gaussian blur is added to each image with a kernel size of 5 and a standard deviation of 1.5.

Following the above feature extraction process, we acquire object-level features and scene-level features with different noises. We report the experimental results of the proposed PKG-Transformer and the baseline on Sydney-Captions in Table VII. First, the PKG-Transformer outperforms the baseline in different cases. Second, upon comparing the results of each metric, it is evident that most metrics exhibit only minor variations. Compared with Gaussian blur, Gaussian noise has

a greater impact on both the proposed PKG-Transformer and baseline.

$S_m$  represents the average of BLEU-4, METEOR, ROUGE<sub>L</sub>, CIDEr, and SPICE. Therefore, to assess the robustness of the algorithm, we can measure the maximum difference between the mean values of  $S_m$  before and after degradation. The maximum difference for PKG-Transformer is 7.51%, whereas for baseline, it is 7.76%. This indicates that the proposed PKG-Transformer exhibits better robustness compared to the baseline.

#### F. Qualitative Analysis

To visually illustrate the effectiveness of the proposed method, qualitative results comparing the baseline, M<sup>2</sup> transformer, with the proposed PKG-Transformer are presented in Fig. 5. These results are accompanied by human-annotated GT captions, providing an intuitive demonstration of the performance of the proposed method. Obviously, the captions generated by the proposed PKG-Transformer are more accurate and comprehensive compared to the baseline.

For instance, considering the first image in Fig. 5, the baseline not only inaccurately depicts “four tennis courts” but



Fig. 6. Example of captions generated by baseline with/without the FE module, as well as the GT, closely/weakly related objects, similar objects. Some detailed and accurate words are highlighted in blue. The words that are inconsistent with the image content are highlighted in red.

also misses an important object “a basketball court”. Although the  $M^2$  transformer portrays the important object “basketball court,” it inaccurately describes “two tennis courts.” In contrast, our method accurately describes “a tennis court” and “a basketball court.” This demonstrates that our approach is capable of generating captions that better align with the context of the scene.

Furthermore, to evaluate the effectiveness of each component, we visualize the results of the baseline with/without the FE module and with/without the PKA separately, as shown in Figs. 6 and 7, respectively. Taking the first image in Fig. 6 as an example, the baseline describes “four tennis courts,” while the baseline with the FE module accurately describes “six tennis courts.” This demonstrates that the graph structure facilitates the differentiation of similar objects, thus aiding the model in generating accurate descriptions.

Similarly, considering the first image in Fig. 7, the baseline describes “a dense residential area” and “houses” but misses the closely related object “swimming pool.” On the contrary, the baseline with the PKA describes “houses,” “swimming pool,” and “plants” in a way consistent with GT captions. This demonstrates that incorporating prior scene-object knowledge enables the model to describe objects that are closely related to the scene accurately.

Overall, the above qualitative results provide strong evidence of the effectiveness and improved performance of the proposed PKG-Transformer in generating more accurate and contextually aligned captions compared to the baseline.

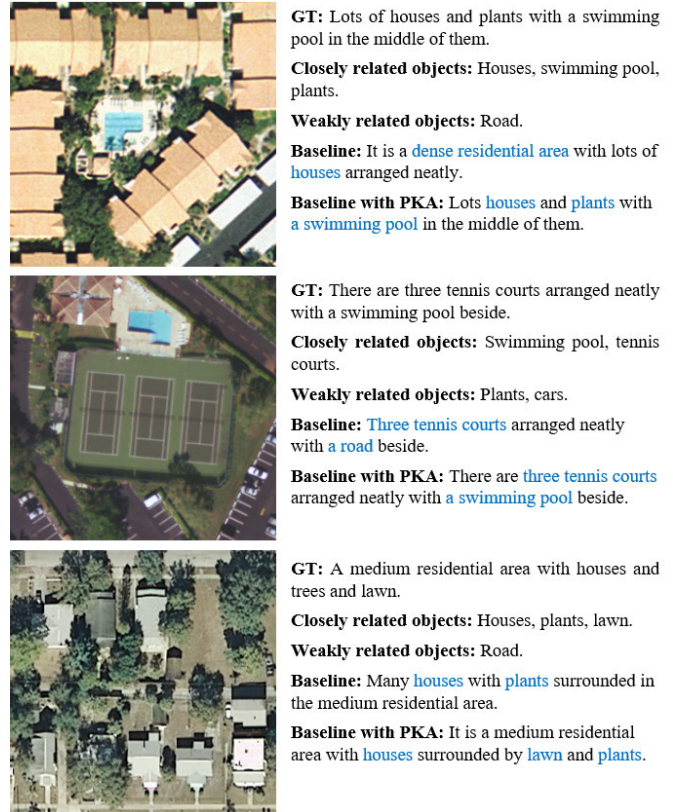


Fig. 7. Example of captions generated by baseline with/without the PKA, as well as the GT, closely/weakly related objects, similar objects. Some detailed and accurate words are highlighted in blue.

TABLE VIII  
 COMPARISON BETWEEN THE BASELINE,  $M^2$  TRANSFORMER, AND THE PROPOSED PKG-TRANSFORMER ON THE NUMBER OF PARAMETERS AND FLOPS

Method	Parameters	FLOPs
Baseline	27.67M	1.12G
$M^2$ Transformer [24]	28.85M	1.32G
PKG-Transformer	31.94M	1.58G

### G. Complexity Analysis

This section provides a brief analysis of the computational complexity of the proposed PKG-Transformer. We can see that the PKA incurs the majority of the computational cost. First, computing the new prior scene-object knowledge matrix  $C$  according to (4) requires  $O(MNd + MN + Md)$  cost. Then, computing  $PKA(X, C)$  via (5) costs  $O(2M^2d + 3Md^2 + 6Md + 2MN)$ . Therefore, the computational cost of the PKA is  $O(2M^2d + 3Md^2 + 7Md + 3MN + MNd)$ . The computational cost of the proposed PKG-Transformer is  $O(M^2)$ , which is identical to that of the  $M^2$  transformer and baseline. Table VIII shows the number of parameters and the number of floating-point operations (FLOPs) of baseline,  $M^2$  transformer and the proposed PKG-Transformer. While they consume comparable resources, the proposed PKG-Transformer achieved the best results.

## V. CONCLUSION

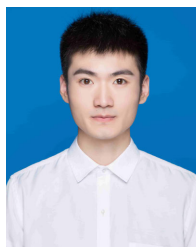
In this article, we propose a novel PKG-Transformer that generates contextually accurate and grammatically correct descriptions for RSIs. Our approach explores the

object–object/scene–scene relationships and scene–object relationships in RSIs through an FE module and a PKA, respectively. This addresses the limitations of previous methods and generates more accurate captions. The FE module, which is a fusion of graph attention networks and the multi-head attention mechanism, is designed to refine and enhance the object-level and scene-level features. In addition, we propose a PKA to select the objects that are more relevant to the scene regions by establishing the relationships between them. This attention mechanism is seamlessly integrated into the Transformer structure, providing valuable prior knowledge that promotes the caption generation process. Our extensive experimental results and analysis demonstrate the effectiveness of these proposed components. Compared with the state-of-the-art methods, the proposed PKG-Transformer achieves competitive performance in the RSIC task and generates more accurate and descriptive captions for RSIs.

## REFERENCES

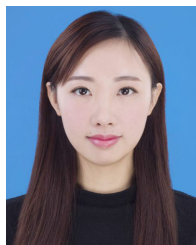
- [1] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, and R. Cucchiara, "From show to tell: A survey on deep learning-based image captioning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 539–559, Jan. 2023, doi: [10.1109/TPAMI.2022.3148210](https://doi.org/10.1109/TPAMI.2022.3148210).
- [2] 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](https://doi.org/10.1109/TIP.2022.3183434).
- [3] J. Ji, Y. Luo, and X. Sunet, "Improving image captioning by leveraging intra- and inter-layer global representation in transformer network," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 1655–1663.
- [4] X. Zhang et al., "RSTNet: Captioning with adaptive attention on visual and non-visual words," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15460–15469, doi: [10.1109/CVPR46437.2021.01521](https://doi.org/10.1109/CVPR46437.2021.01521).
- [5] P. Anderson et al., "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6077–6086.
- [6] A. Vaswani, N. Shazeer, and N. Parmar, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
- [7] Y. Pan, T. Yao, Y. Li, and T. Mei, "X-linear attention networks for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10968–10977, doi: [10.1109/CVPR42600.2020.01098](https://doi.org/10.1109/CVPR42600.2020.01098).
- [8] 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. (CITS)*, Jul. 2016, pp. 1–5, doi: [10.1109/CITS.2016.7546397](https://doi.org/10.1109/CITS.2016.7546397).
- [9] 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](https://doi.org/10.1109/TGRS.2017.2776321).
- [10] 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, doi: [10.1109/LGRS.2022.3198234](https://doi.org/10.1109/LGRS.2022.3198234).
- [11] 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.*, Sep. 2020, pp. 980–983, doi: [10.1109/IGARSS39084.2020.9323213](https://doi.org/10.1109/IGARSS39084.2020.9323213).
- [12] R. Zhao, Z. Shi, and Z. Zou, "High-resolution remote sensing image captioning based on structured attention," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5603814, doi: [10.1109/TGRS.2021.3070383](https://doi.org/10.1109/TGRS.2021.3070383).
- [13] 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. 1–5, 2022, doi: [10.1109/LGRS.2021.3135711](https://doi.org/10.1109/LGRS.2021.3135711).
- [14] 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, doi: [10.1109/LGRS.2019.2893772](https://doi.org/10.1109/LGRS.2019.2893772).
- [15] 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. (CVPR)*, Jun. 2015, pp. 3156–3164.
- [16] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biol. Cybern.*, vol. 36, no. 4, pp. 193–202, Apr. 1980, doi: [10.1007/BF00344251](https://doi.org/10.1007/BF00344251).
- [17] P. Razvan et al., "How to construct deep recurrent neural networks," in *Proc. Int. Conf. Learn. Represent.*, 2014. [Online]. Available: <https://arxiv.org/abs/1312.6026>
- [18] K. Xu, J. Ba, and R. Kiros, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.
- [19] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," 2017, *arXiv:1710.10903*.
- [20] S. Ren, K. He, and R. Girshick, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
- [21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [22] A. Gupta and P. Mannem, "From image annotation to image description," in *Proc. Int. Conf. Neural Inf. Process.*, 2012, pp. 196–204.
- [23] G. Kulkarni et al., "BabyTalk: Understanding and generating simple image descriptions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2891–2903, Dec. 2013.
- [24] S. Li, G. Kulkarni, T. Berg, A. Berg, and Y. Choi, "Composing simple image descriptions 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 descriptions," *Int. J. Comput. Vis.*, vol. 119, no. 1, pp. 46–59, Aug. 2016.
- [27] P. Kuznetsova, V. Ordonez, and A. Berg, "Collective generation of natural image descriptions," in *Proc. Annu. Meet. Assoc. Comput. Linguist.*, 2012, pp. 359–368.
- [28] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10575–10584, doi: [10.1109/CVPR42600.2020.01059](https://doi.org/10.1109/CVPR42600.2020.01059).
- [29] S. J. 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. (CVPR)*, Jul. 2017, pp. 1179–1195, doi: [10.1109/CVPR.2017.131](https://doi.org/10.1109/CVPR.2017.131).
- [30] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
- [31] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-RNN)," in *Proc. Int. Conf. Learn. Represent.*, 2015. [Online]. Available: <https://arxiv.org/abs/1412.6632>
- [32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
- [33] D. Yan, W. Yu, and Z. Zhang, "Transformer with prior language knowledge for image captioning," in *Proc. Int. Conf. Neural Inf. Process.*, 2021, pp. 40–51.
- [34] X. Zhang, X. Wang, X. Tang, H. Zhou, and C. Li, "Description generation for remote sensing images using attribute attention mechanism," *Remote Sens.*, vol. 11, no. 6, p. 612, Mar. 2019.
- [35] Z. Zhang, W. Diao, W. Zhang, M. Yan, X. Gao, and X. Sun, "LAM: Remote sensing image captioning with label-attention mechanism," *Remote Sens.*, vol. 11, no. 20, p. 2349, Oct. 2019.
- [36] 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](https://doi.org/10.1109/LGRS.2020.3009243).
- [37] 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, 2022, Art. no. 5615216, doi: [10.1109/TGRS.2021.3132095](https://doi.org/10.1109/TGRS.2021.3132095).
- [38] 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, doi: [10.1109/JSTARS.2022.3153636](https://doi.org/10.1109/JSTARS.2022.3153636).

- [39] Y. Li et al., "Recurrent attention and semantic gate for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5608816, doi: [10.1109/TGRS.2021.3102590](https://doi.org/10.1109/TGRS.2021.3102590).
- [40] S. Mehta, M. Rastegari, A. Caspi, L. Shapiro, and H. Hajishirzi, "ESPNet: Efficient spatial pyramid of dilated convolutions for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 552–568.
- [41] Z. Chen, J. Wang, A. Ma, and Y. Zhong, "TypeFormer: Multi-scale transformer with type controller for remote sensing image caption," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022, doi: [10.1109/LGRS.2022.3192062](https://doi.org/10.1109/LGRS.2022.3192062).
- [42] 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, doi: [10.1109/LGRS.2022.3150957](https://doi.org/10.1109/LGRS.2022.3150957).
- [43] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
- [44] C. Liu, R. Zhao, J. Chen, Z. Qi, Z. Zou, and Z. Shi, "A decoupling paradigm with prompt learning for remote sensing image change captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5622018, doi: [10.1109/TGRS.2023.3321752](https://doi.org/10.1109/TGRS.2023.3321752).
- [45] C. Liu, R. Zhao, H. Chen, Z. Zou, and Z. Shi, "Remote sensing image change captioning with dual-branch transformers: A new method and a large scale dataset," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5633520, doi: [10.1109/TGRS.2022.3218921](https://doi.org/10.1109/TGRS.2022.3218921).
- [46] 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, doi: [10.1109/TGRS.2014.2357078](https://doi.org/10.1109/TGRS.2014.2357078).
- [47] 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.*, Nov. 2010, pp. 2–5.
- [48] G.-S. Xia et al., "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017.
- [49] 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.
- [50] A. Lavie and A. Agarwal, "METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments," in *Proc. 2nd Workshop Stat. Mach. Transl.*, 2007, pp. 65–72.
- [51] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. Assoc. Comput. Linguistics Workshop*, 2004, pp. 74–81.
- [52] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4566–4575.
- [53] P. Anderson, B. Fernando, and M. Johnson, "SPICE: Semantic propositional image caption evaluation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 382–398.
- [54] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
- [55] K. Cho et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1724–1734.
- [56] 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, doi: [10.1109/JSTARS.2019.2959208](https://doi.org/10.1109/JSTARS.2019.2959208).
- [57] 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. 6922–6934, Aug. 2021, doi: [10.1109/TGRS.2020.3031111](https://doi.org/10.1109/TGRS.2020.3031111).
- [58] 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](https://doi.org/10.1109/TGRS.2020.3044054).
- [59] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4633–4642.



**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 pursuing the Ph.D. degree in computer science with the Nanjing University of Science and Technology (NJUST), Nanjing, China.

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



**Jing Wang** received the B.E. and Ph.D. degrees from the Nanjing University of Science and Technology, Nanjing, China, in 2015 and 2022, respectively.

From 2019 to 2020, she was a Visiting Scholar with the University of Rochester, Rochester, NY, USA. She is currently a Post-Doctoral Fellow with the Department of Automation, Tsinghua University, Beijing, China. Her research interests include computer vision and multimedia analysis, with a focus on vision and language.



**Yang Yang** received the Ph.D. degree in computer science from Nanjing University, Nanjing, China, in 2019.

In 2019, he became a Faculty Member at the Nanjing University of Science and Technology, Nanjing, where he is currently a Professor with the School of Computer Science and Engineering. He has published over ten papers in leading international journals/conferences. His research interests lie primarily in machine learning and data mining, including heterogeneous learning, model reuse, and

incremental mining.

Dr. Yang serves as a PC in leading conferences such as IJCAI, AAAI, ICML, and NIPS.



**Liang Xiao** (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 Post-Doctoral 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 the Ministry of Education, NJUST, where he is currently a Professor with the School of Computer Science. He has authored or coauthored more than 70 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.