A Multiscale Grouping Transformer with CLIP Latents for Remote Sensing Image Captioning

Lingwu Meng, Jing Wang, Ran Meng, Yang Yang, and Liang Xiao, *Senior Member, IEEE*

***Abstract*—Recent progress has shown that integrating multi-scale visual features with advanced Transformer architectures is a promising approach for remote sensing image captioning (RSIC). However, the lack of local modeling ability in self-attention may potentially lead to inaccurate contextual information. Moreover, the scarcity of trainable image-caption pairs poses challenges in effectively harnessing the semantic alignment between images and texts. To mitigate these issues, we propose a Multiscale Grouping Transformer with Contrastive Language-Image Pre-training (CLIP) latents (MG-Transformer) for RSIC. First of all, a CLIP image embedding and a set of region features are extracted within a Multi-level Feature Extraction module. To achieve a comprehensive image representation, a Semantic Correlation module is designed to integrate the image embedding and region features with an attention gate. Subsequently, the integrated image features are fed into a Transformer model. The Transformer encoder utilizes dilated convolutions with different dilation rates to obtain multiscale visual features. To enhance the local modeling ability of the self-attention mechanism in the encoder, we introduce a Global Grouping Attention mechanism. This mechanism incorporates a grouping operation into self-attention, allowing each attention head to focus on different contextual information. The Transformer decoder then adopts the Meshed Cross-Attention mechanism to establish relationships between various scales of visual features and text features. This facilitates the generation of captions for images by the decoder. Experimental results on three RSIC datasets demonstrate the superiority of the proposed MG-Transformer. The code will be publicly available at https://github.com/One-paper-luck/MG-Transformer.**

***Index Terms*—Remote sensing image captioning, Transformer, CLIP, multiscale, Grouping.**

I. Introduction

REMOTE sensing image captioning (RSIC) aims to translate remote sensing images (RSIs) into natural language descriptions, enabling non-experts to intuitively understand the content of the images. Compared with traditional tasks such as object detection [1]–[4] and scene classification [5]–[8] for RSIs, the RSIC task further elucidates the rich information in RSIs and provides intuitive guidance for a wide range of application fields such as disaster monitoring, agricultural management, and urban planning.

Despite substantial advancements in natural image captioning (NIC), the investigation of RSIC remains largely unexplored. The main challenges in RSIC arise from two aspects: First, an RSI often comprises a vast number of objects with varying scales and different categories, making it difficult to achieve effective abstraction and representation of the RSI. Second, the limited quantity of image-text pairs in existing RSIC datasets can result in reduced model generalization and potentially inaccurate predictions.

To mitigate the first challenge, early RSIC methods [9], [10] utilized Convolutional Neural Networks (CNNs) [11] to extract region features from a specific layer as image representations. Despite exhibiting decent performance, they cannot effectively capture objects at different scales and may overlook important objects. To overcome this limitation, there has been an increasing emphasis among researchers on exploring the multiscale information included in RSIs. These methods [12]–[14] extracted multiscale features from different convolutional layers of CNNs and fed the features into a text decoder to generate sentences. Recent studies [15], [16] have also integrated attention mechanisms with multiscale features to discern the importance of features at different scales. Expanding on this approach, the research conducted by [17]–[19] has further employed the self-attention mechanism, renowned for its exceptional ability to capture global information. However, it is worth noting that the self-attention mechanism lacks the ability to effectively model local contextual information [20]. This limitation hinders its ability to capture fine-grained dependency relationships.

To address the second challenge, one straightforward way is to employ data augmentation techniques to expand the image-text pairs. However, these techniques may result in a semantic mismatch between the expanded images and the corresponding captions. For example, the rotation operation may alter the position of objects, while the cropping operation may result in the loss of important objects.

To tackle the challenges mentioned above, this paper proposes a solution that combines the multiscale features plus self-attention mechanism framework, along with the utilization of CLIP image embedding [21], [22]. In order to enhance the self-attention's ability to model local context, we introduce a novel Global Grouping Attention mechanism (GGA) by

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 & Intelligent Sense under Grant JSGP202101 and Grant JSGP202204, in part by the National Natural Science Foundation of China under Grant 62302255, and in part by the China Postdoctoral Science Foundation 2023TQ0181. (*Corresponding authors:* Liang Xiao; Jing Wang.)

Lingwu Meng, Ran 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; rmeng@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, Nanjing University of Science and Technology, Nanjing 210094, China, and also with 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).

incorporating a grouping operation. The integration of GGA facilitates improved interaction between the multiscale features and ensures a more comprehensive understanding of the image context. Furthermore, to overcome the data scarcity issue, we leverage the CLIP image embedding, which is known for its impressive zero-shot capabilities and robustness to variations in image distribution [23]. Specifically, we design a Semantic Correlation (SC) module to integrate the global CLIP image embedding and the local region features with an attention gate.

Based on the aforementioned idea of enhancing the self-attention's local modeling ability and leveraging the knowledge from the pre-trained vision-language CLIP model, we propose a Multiscale Grouping Transformer with CLIP latents (MG-Transformer) to improve RSIC, as shown in Figure. 1. The proposed framework consists of three main components: a Multi-level Feature Extraction (MFE) module, an SC module, and a Transformer that incorporates dilated convolution, GGA, and MCA. Firstly, the MFE module extracts global and region features using the pre-trained CLIP model and the ResNet-152 [24], respectively. Subsequently, the SC module combines these features to establish global-local semantic correlations, yielding rich visual features. The obtained features are then fed into the Transformer. The Transformer encoder employs dilated convolutions with different dilation rates to obtain multiscale visual features and further focuses on valuable local information using the GGA. The decoder utilizes the MCA to integrate and correlate multiscale visual features with text features. Finally, the output of the decoder is used to generate sentences for RSIs.

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

(1) To enhance the local modeling ability of the self-attention mechanism, we propose a GGA mechanism that utilizes a grouping mechanism to enable each attention head to focus on distinct localized information.

(2) We propose to address the issue of inadequate image-caption pairs by incorporating the CLIP image embedding as supplementary knowledge. This embedding provides the model with abundant pre-aligned image-text information, enabling more semantically accurate descriptions for RSIs. To seamlessly integrate this embedding into the captioning framework, we design an SC module that combines it with the region features using an attention gate.

(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 remote sensing image captioning.

A. Natural Image Captioning

Experts initially focused on the natural image captioning (NIC) task. Early template-based methods [25]–[27] and retrieval-based methods [28]–[30] heavily rely on manually designed features and retrieval results, resulting in less natural descriptions that are unable to adapt well to the content of the images. However, with the advent of deep learning-based methods [31]–[34], significant improvements have been achieved. Vinyals *et al.* [31] first proposed the vanilla encoder-decoder paradigm. In the encoder, a CNN was applied to extract high-level visual features from input images. In the decoder, a recurrent neural network (RNN) [35] was trained to generate sentences based on the visual features. Later, Xu *et al.* [32] made significant advancements by incorporating the attention mechanism into the CNN-RNN framework. Rennie *et al.* [36] introduced the self-critical training strategy based on reinforcement learning mechanisms to further improve the performance of image captioning. By combining the self-critical training method with the sampling operation, the discrepancy between training and inference can be mitigated. Unlike the above methods, Anderson *et al.* [33] extracted the more expressive object features with an object detection model Faster R-CNN [37]. Then, They encoded the relationships between the objects and employed a two-layer LSTM decoder to facilitate caption generation. In recent years, most state-of-the-art methods for image captioning have incorporated advanced Transformer frameworks. The Transformer is primarily composed of self-attention, which enables it to capture global dependencies. Huang *et al.* [38] proposed an Attention on Attention (AoA) that enhances self-attention by examining the correlation between attention outputs and queries. Cornia *et al.* [39] presented a Memory-Augmented Attention, which extended self-attention by modeling a priori knowledge on relationships between objects.

B. Remote Sensing Image Captioning

Inspired by NIC, Qu *et al.* [9] first proposed a deep multimodal neural network model based on the CNN-RNN framework for RSIC. Later, Lu *et al.* [10] demonstrated that attention-based methods can achieve superior performance compared to the multimodal model, which laid the foundation for RSIC. To produce more detailed and accurate captions, several attribute-based methods [15], [16] have been proposed. Zhang *et al.* [15] proposed an attribute attention to combine high-level features extracted from the relatively deep fully connected layer (or softmax layer) with low-level features extracted from the relatively shallow convolution layers. In the same year, Zhang *et al.* [16] proposed a label-attention mechanism to utilize label information to guide the computation of attention masks. Furthermore, Zhang *et al.* [17] 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. However, these methods do not express multiscale information about RSIs, which is crucial for the RSIC task. To address this issue, Wang *et al.* [12] collected features from conv4 and conv5. These features were concatenated as the image feature representation after a self-attention and a gated cross-attention. Afterward, Li *et al.* [13] adopted the efficient spatial pyramid (ESP) [40] to extract the multiscale visual features from the region features obtained by a pre-trained CNN, which are fed to an adaptive average pooling operation for a global representation.

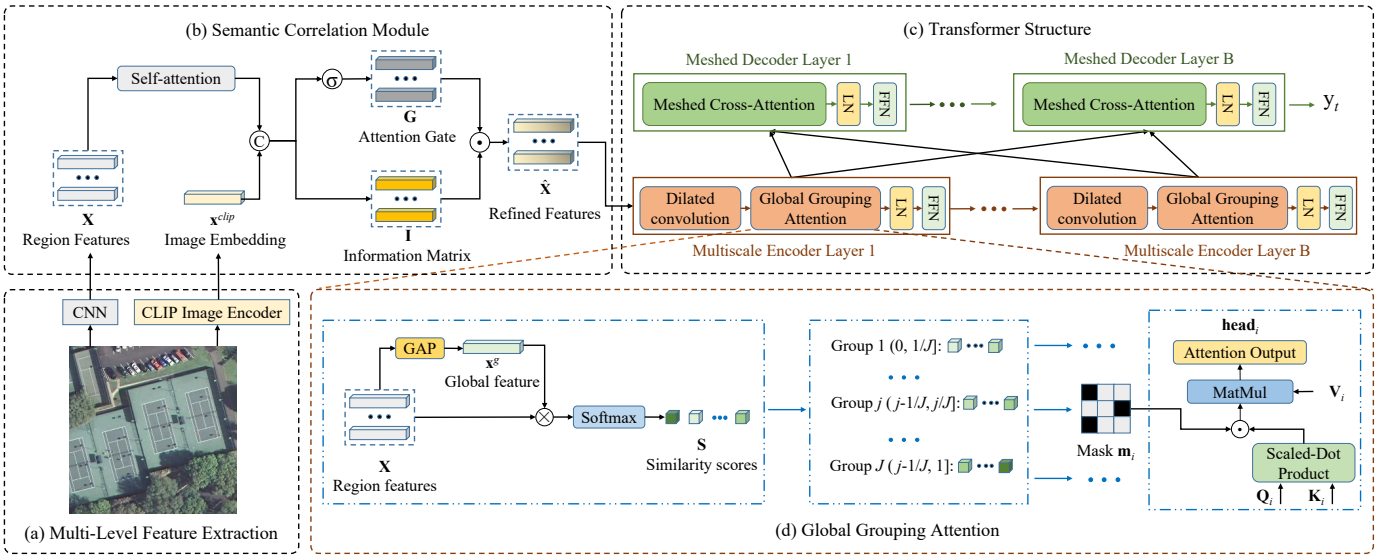


Fig. 1. Framework of the proposed MG-Transformer in this paper. The Multi-level Feature Extraction module employs a pre-trained CNN (ResNet-152) and CLIP image encoder to extract region features and the image embedding from an input image, respectively. The Semantic Correlation module integrates them to obtain refined features. Then, the refined features are fed into a Transformer. The Transformer encoder adopts dilated convolutions with different dilation rates to obtain multiscale visual features. The Global Grouping Attention learns more valuable information from multiscale visual features. The Transformer decoder establishes intermodal interactions between different scales of visual features and text features, and calculates the probability p_t over a vocabulary of possible words. For the sake of clarity, fully connected layers are not shown. Ⓒ, $\odot$, σ, $\otimes$, and GAP represent concatenation operation, Hadamard product, sigmoid activation function, matrix multiplication, and global average pooling operation, respectively.

Then, the global representation was concatenated with region features and then fed into a two-layer LSTM decoder to generate captions.

Besides the CNN-RNN framework, some methods leverage the Transformer framework to enhance RSIC. Wang *et al.* [41] designed a two-stage Word–Sentence framework. They extracted valuable words through a classification task and utilized the Transformer framework to organize these words into coherent sentences. Liu *et al.* [14] proposed a multilayer aggregated Transformer (MLAT). They fused multiscale visual features extracted from a CNN and then fed these features into a Transformer to generate sentences. Chen *et al.* [15] proposed a pure Transformer architecture with caption-type controller. They adopted a multiscale Vision Transformer (ViT) for image representation and introduced a standard Transformer decoder to generate sentences. Recently, a series of works [42]–[46] have been proposed for remote sensing image change captioning (RSICC), which aims to describe the differences between bitemporal images by natural language. Liu *et al.* [42] achieved impressive results by leveraging pre-trained large language models. To address the multiscale problems in RSICC, Liu *et al.* [46] focused on changing regions and sufficiently extracted multiscale visual features from different layers.

Despite the significant progress achieved by multiscale methods, they often fail to adequately consider the interactions between visual and textual features at different scales, leading to inaccurate descriptions. In this paper, we propose a novel MG-Transformer framework to address this issue by integrating the contributions of these interactions through MCA. Furthermore, we incorporate a grouping mechanism into self-attention to enhance the local modeling ability of the self-attention mechanism.

III. APPROACH

A. Overall Framework

The overall framework of the proposed method, MG-Transformer, is depicted in Figure. 1. It consists of an MFE module, an SC module, and a Transformer with dilated convolution, GGA and MCA. Specifically, an image embedding and a set of region features are first extracted from the MFE module. Then, the SC module integrates the image embedding and region features with an attention gate to obtain refined features. The refined features are subsequently fed into the Transformer to promote the captioning process. The Transformer encoder employs dilated convolutions with different dilation rates to capture multiscale visual features. Furthermore, each encoder layer employs a GGA to learn the fine-grained relationships among feature groups from different regions. Subsequently, the output of the GGA is fed into a position-wise fully connected feed-forward network to obtain the output of the encoder layer. The Transformer decoder leverages the MCA to facilitate intermodal interactions between text features and multiscale visual features from the encoder, and finally generates a sentence for the input RSI.

B. Multi-level Feature Extraction

We extract region features and the CLIP image embedding to obtain a comprehensive semantic representation for an RSI. The region features capture the local information of the RSI, while the image embedding aggregates the information of the entire RSI.

Following a standard practice in RSIC [9], [10], we adopt the ResNet-152 pre-trained on ImageNet [47] to extract N region features $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\} \in \mathbf{R}^{N \times d_1}$.

Then, we adopt the pre-trained CLIP image encoder with ViT-B/32 backbone to extract the representation of the whole image, denoted as $\mathbf{x}^{clip} \in \mathbf{R}^{d_2}$. By using the ViT encoder, the CLIP image embedding is able to encode visual clues from all image patches. Moreover, the integration between vision and text enriches the embedding of information from both modalities.

By employing these multi-level feature representations, we acquire rich local and global semantic information from RSIs.

C. Semantic Correlation Module

Drawing inspiration from the AoA mechanism [38], we devise the SC module to integrate the CLIP image embedding and the region features. This integration is achieved through an attention gate, which serves to strengthen the correlation between the attention queries and their corresponding attention results. First of all, the image embedding $\mathbf{x}^{clip}$ is aligned with the number of region features by a copy operation to obtain the image embedding matrix $\mathbf{X}^{clip} \in \mathbf{R}^{N\times d_2}$. Both the region features $\mathbf{X}$ and the image embedding matrix $\mathbf{X}^{clip}$ are mapped to the same d-dimensional space through linear projections:

$$\mathbf{X}^r = \mathbf{X}\mathbf{W}_1 + \mathbf{b}_1, \tag{1}$$

$$\mathbf{X}^c = \mathbf{X}^{clip}\mathbf{W}_2 + \mathbf{b}_2, \tag{2}$$

where $\mathbf{W}_1 \in \mathbf{R}^{d_1\times d}$ and $\mathbf{W}_2 \in \mathbf{R}^{d_2\times d}$ are learnable parameter matrices; $\mathbf{b}_1 \in \mathbf{R}^d$ and $\mathbf{b}_2 \in \mathbf{R}^d$ are biases.

Subsequently, the region features $\mathbf{X}^r \in \mathbf{R}^{N\times d}$ are processed through a self-attention mechanism, which integrates the interrelations among regions into $\mathbf{X}^r$ and obtains the feature $\mathbf{Z} \in \mathbf{R}^{N\times d}$:

$$\begin{aligned}\mathbf{Z} &= \mathrm{Att}(\mathbf{X}^r\mathbf{W}_q, \mathbf{X}^r\mathbf{W}_k, \mathbf{X}^r\mathbf{W}_v) \\ &= \mathrm{softmax}\left(\frac{(\mathbf{X}^r\mathbf{W}_q)(\mathbf{X}^r\mathbf{W}_k)^{\mathbf{T}}}{\sqrt{d}}\right) \times (\mathbf{X}^r\mathbf{W}_v),\end{aligned} \tag{3}$$

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

Finally, the feature $\mathbf{Z}$ is concatenated with the image embedding matrix $\mathbf{X}^c$ to form an information matrix $\mathbf{I}$, which integrates the local and global semantic information, resulting in a more comprehensive and enriched feature representation. The computation of the information matrix $\mathbf{I} \in \mathbf{R}^{N\times d}$ can be expressed as follows:

$$\mathbf{I} = \mathrm{Concat}(\mathbf{Z},\mathbf{X}^c)\mathbf{W}_\mathrm{I} + \mathbf{b}_\mathrm{I}, \tag{4}$$

where $\mathbf{W}_\mathrm{I} \in \mathbf{R}^{2d\times d}$ is the learnable parameter matrix and $\mathbf{b}_\mathrm{I} \in \mathbf{R}^d$ is the bias; Concat means concatenation operator.

To filter irrelevant or weakly correlated content in the information matrix $\mathbf{I}$, we design an attention gate $\mathbf{G} \in \mathbf{R}^{N\times d}$ based on the sigmoid activation function:

$$\mathbf{G} = \sigma(\mathrm{Concat}(\mathbf{Z},\mathbf{X}^c)\mathbf{W}_\mathrm{G} + \mathbf{b}_\mathrm{G}), \tag{5}$$

where $\mathbf{W}_\mathrm{G} \in \mathbf{R}^{2d\times d}$ is the learnable parameter matrix and $\mathbf{b}_\mathrm{G} \in \mathbf{R}^d$ is the bias; σ is the sigmoid activation function. The attention gate $\mathbf{G}$ selectively emphasizes significant information within the information matrix, providing a more discriminative image representation. Based on this gate, the refined features $\hat{\mathbf{X}} \in \mathbf{R}^{N\times d}$ can be calculated as:

$$\hat{\mathbf{X}} = \mathbf{G} \odot \mathbf{I}, \tag{6}$$

where $\odot$ denotes Hadamard product. Afterwards, the refined features $\hat{\mathbf{X}}$ are fed into the Transformer structure.

D. Multiscale Grouping Transformer

Our Transformer is composed of a multiscale encoder and a meshed decoder, both of which consist of stacked attention layers. In the multiscale encoder, we first extract multiscale visual features using dilated convolutions with different dilation rates. Then, the GGA transforms these features of each scale into a series of intermediate states, which are enhanced with the contextual information in between. The meshed decoder adopts the MCA to encourage intermodal interactions between different scales of visual features and text features, subsequently generating captions word by word. We will start by introducing GGA, followed by a multiscale encoder based on GGA, and then the meshed decoder based on MCA. The details are explained below.

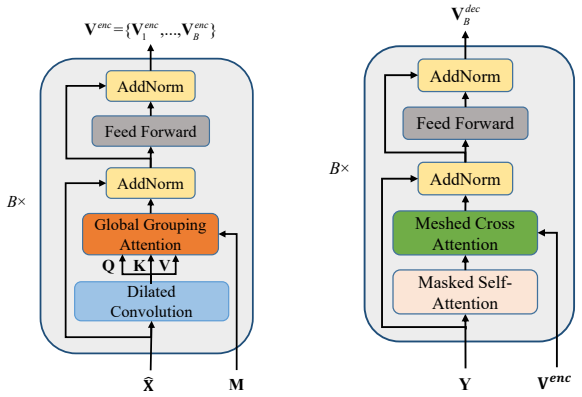
Global Grouping Attention: Given a set of image regions (In our context, the region features correspond to the output feature $\tilde{\mathbf{X}} \in \mathbf{R}^{\mathbf{N}\times\mathbf{d}}$ of the dilated convolution), we introduce the GGA that incorporates a grouping mechanism into self-attention to enhance its local modeling ability.

To expedite training, we initially categorize the regions into different groups based on the original region features $\mathbf{X}$ in an offline manner. Technically, a global feature $\mathbf{x}^g \in \mathbf{R}^{d_1}$ is first obtained by performing an average pooling operation on region features $\mathbf{X}$ as shown in Figure. 1(d). We then adopt the dot product similarity between $\mathbf{x}^g$ and $\mathbf{X}$ to measure the correlation between the global feature and region features in the same feature space. Subsequently, their similarity scores $\mathbf{S} = \{\mathbf{s}_1, ..., \mathbf{s}_N\} \in \mathbf{R}^N$ are normalized using a softmax function. The i-th region is assigned to Group j if $\mathbf{s}_i$ within the interval $(j - 1/J, j/J]$. As a result, we can get J groups.

To enhance the robustness and generalization ability of the captioning model, we randomly assign λ different groups to each $\mathbf{head}_i$ of self-attention, which can be achieved through building a grouping mask matrix $\mathbf{m}_i \in \mathbf{R}^{N\times N}$. The $\mathbf{head}_i$ can be calculated as follows:

$$\mathbf{head}_i = \left(\mathrm{softmax}(\frac{\mathbf{Q}_i\mathbf{K}_i^{\mathbf{T}}}{\sqrt{D}}) \odot \mathbf{m}_i\right) \times \mathbf{V}_i, \tag{7}$$

where $\mathbf{Q}_i \in \mathbf{R}^{N\times D}$, $\mathbf{K}_i \in \mathbf{R}^{N\times D}$ and $\mathbf{V}_i \in \mathbf{R}^{N\times D}$ are obtained by linearly projecting input region features $\tilde{\mathbf{X}}$; h is the number of heads; D is the scaling factor, and $D = d/h$; $\odot$ denotes the Hadamard product. While the original self-attention mechanism lacks effective modeling of local contextual information by having every attention head focus on all regions [20], our approach encourages each attention head to concentrate on semantically relevant regions. During the attention operation, only objects within those groups contribute to enhancing the target object feature, thereby reinforcing the contextual relationships among local regions.



(a) (b)

Fig. 2. (a) Multiscale encoder and (b) Meshed decoder for MG-Transformer.

Finally, we concatenate all h heads and obtain the output feature $\mathbf{V}^{GGA}$:

$$\mathbf{V}^{GGA} = \text{Concat}(\mathbf{head}_1, ..., \mathbf{head}_h) \times \mathbf{W}_h, \tag{8}$$

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

Multiscale Encoder: As can be seen from Figure. 2(a), the encoder is composed of a stack of B identical layers. Each encoding layer stacks a dilated convolution, a GGA, and a position-wise fully connected feed-forward network (FFN) in turn. We incorporate a residual connection and a layer normalization around the 2nd and 3rd sublayers, respectively. The inputs to each encoding layer include the output of the previous encoding layer (we use the output $\hat{\mathbf{X}}$ of the SC module for the first layer) and a set of grouping mask matrices $\mathbf{M} = \{\mathbf{m}_1, ..., \mathbf{m}_h\} \in \mathbf{R}^{h\times N\times N}$.

Let's consider the first encoder layer as an example. The output $\hat{\mathbf{X}}$ of the SC module is fed into a dilated convolution to obtain new scale features $\tilde{\mathbf{X}}_\mathbf{1} \in \mathbf{R}^{\mathbf{N}\times\mathbf{d}}$. Then, $\tilde{\mathbf{X}}_\mathbf{1}$ and the grouping mask matrices $\mathbf{M}$ are fed into a GGA. The obtained features $\mathbf{V}^{GGA}$ are then passed through an FFN layer to acquire the output $\mathbf{V}_1^{enc}$ of the first encoder layer. Consequently, we can obtain the outputs $\mathbf{V}^{enc} = \{\mathbf{V}_1^{enc}, ..., \mathbf{V}_B^{enc}\}$ of all encoder layers as the output of the multiscale encoder.

Meshed Decoder: To construct relationships between different scale visual features and text features, we adopt the meshed decoder architecture from the M^2 Transformer [39]. As can be seen from Figure. 2(b), the decoder is built with a stack of B identical decoding layers. Each decoder layer is composed of an MCA and an FFN layer. We use a residual connection and layer normalization around each sublayer. Similar to the encoder, the inputs to each decoding layer include the output of the previous decoding layer and the outputs $\mathbf{V}^{enc} = \{\mathbf{V}_1^{enc}, ..., \mathbf{V}_B^{enc}\}$ of the encoder. For the first layer, we input the text sequence $\mathbf{Y} = \{\mathbf{y}_1, ..., \mathbf{y}_T\} \in \mathbf{R}^{T\times d}$ instead of the previous layer output, where T denotes the length of the text sequence.

In the case of the first decoding layer, we initially perform a masked attention mechanism to focus on the left subsequence $\mathbf{Y}_{mask}$. Subsequently, the cross-attention is employed to enable interactions between text features and multiscale visual features from all encoder layers:

$$\mathbf{S}^i = \text{Att}(\mathbf{Y}_{mask}\mathbf{W}_q^{d_i}, \mathbf{V}_i^{enc}\mathbf{W}_k^{d_i}, \mathbf{V}_i^{enc}\mathbf{W}_v^{d_i}), \tag{9}$$

where $\mathbf{W}_q^{d_i} \in \mathbf{R}^{d\times d}$, $\mathbf{W}_k^{d_i} \in \mathbf{R}^{d\times d}$, and $\mathbf{W}_v^{d_i} \in \mathbf{R}^{d\times d}$ are learning parameter matrices; $\mathbf{V}_i^{enc}$ is the output of the i-th encoder layer. Then, we measure the correlation between the attention results and the left subsequence to obtain a weight matrix:

$$\alpha_i = \sigma\left(\text{Concat}(\mathbf{Y}_{mask}, \mathbf{S}^i)\mathbf{W}_i^\alpha + \mathbf{b}_i^\alpha\right), \tag{10}$$

where σ 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. The contributions of all encoder layers and their respective weight matrix are weighted in MCA as follows:

$$\text{MCA}(\mathbf{V}^{enc}, \mathbf{Y}_{mask}) = \sum_{i=1}^{B} \alpha_i \odot \mathbf{S}^i. \tag{11}$$

After an FFN layer, we acquire the output $\mathbf{V}_1^{dec}$ of the first decoding layer.

The output $\mathbf{V}_B^{dec}$ of the last decoding layer is regarded as the decoder output. Subsequently, $\mathbf{V}_B^{dec}$ is sequentially fed into a fully connected layer and a softmax layer to calculate the probability over a vocabulary of possible words.

E. Training and Objectives

Training with Cross Entropy Loss: Following a standard practice in image captioning [25], [38], [39], we first train our model with a word-level cross entropy loss (XE):

$$L_{XE}(\theta) = -\sum_{t=1}^{T} \log(p_\theta(y_t^*|y_{1:t-1}^*)), \tag{12}$$

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

CIDEr-D Score Optimization: Recent studies [36], [48], [49] have demonstrated that the XE training may lead to the exposure bias problem, where the model always makes predictions on the true labels during training. Consequently, we employ the Self-Critical Sequence Training method [36] to further train our model. During the training process, we employ a beam search strategy to select the top-k sentences with the highest probability. Then we compute their CIDEr scores as rewards. To update the model parameters, the gradient expression is formulated as follows:

$$\nabla_\theta L_{RL}(\theta) = -\frac{1}{k}\sum_{i=1}^{k}((r(w^i) - b)\nabla_\theta \log p_\theta(w^i)), \tag{13}$$

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

IV. EXPERIMENT RESULTS

A. Datasets

This paper evaluates the proposed MG-Transformer on three RSIC datasets: Sydney-Captions [9], UCM-Captions [10], and RSICD [10].

Sydney-Captions: This dataset is a collection of 600 images carefully selected from the Sydney Dataset [50] and contains 7 land cover categories. The size of each image within the dataset is consistently maintained at a resolution of 500 × 500 pixels, ensuring uniform scale and quality throughout. Each image is associated with five human-annotated captions.

UCM-Captions: This dataset is a collection of 2100 images carefully selected from the UC Merced Land Use [51] and contains 21 land cover categories. The size of each image within the dataset is consistently maintained at a resolution of 256 × 256 pixels, ensuring uniform scale and quality throughout. Each image is associated with five human-annotated captions.

RSICD: It consists of 10921 RSIs selected from the AID dataset [52] and other platforms, such as Baidu Map, MapABC. In addition, it contains 30 land cover categories. The size of each image within the dataset is consistently maintained at a resolution of 224 × 224 pixels, ensuring uniform scale and quality throughout. Each image is associated with five human-annotated captions.

B. Evaluation Metrics

To evaluate the quality of the generated captions, we illustrate the RSIC model performance by ten metrics: BLEU-n [53], METEOR [54], ROUGE$_{\mathrm{L}}$ [55], CIDEr [56], SPICE [57], and S_m^* [17]. They can be calculated by the COCO caption evaluation tool[1].

BLEU-n: Bilingual evaluation understudy (BLEU) is a metric used to measure the quality of machine-generated translations. It was developed by IBM Research. BLEU's main strength lies in its simplicity and efficiency, which has made it one of the most commonly used metrics in the field of machine translation. The "n" in BLEU-n refers to the maximum length of the n-grams (contiguous sequence of "n" words) that the metric will consider when comparing a candidate translation against one or more reference translations. For example, BLEU-1 only looks at unigrams (single words), BLEU-2 considers up to bigrams (two-word phrases), BLEU-3 up to trigrams (three-word phrases), and so on. 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: Metric for Evaluation of Translation with Explicit ORdering (METEOR) is another metric used to evaluate the quality of machine-generated translations. Developed by researchers at the Language Technologies Institute at Carnegie Mellon University, METEOR aims to address some of the shortcomings found in earlier metrics such as BLEU. It computes a harmonic mean of the precision and recall values, with recall being weighted higher to prioritize the "completeness" of the translated information. METEOR often correlates better with human judgement compared to BLEU.

ROUGE$_{\mathrm{L}}$: Recall-Oriented Understudy for Gisting Evaluation (ROUGE) is a widely used set of metrics for evaluating automatic summarization and machine translation. It was originally developed by Chin-Yew Lin at the University of Southern California Information Sciences Institute. ROUGE$_{\mathrm{L}}$ specifically measures the longest common subsequence between the candidate sentences and reference sentences.

CIDEr: Consensus-based Image Description Evaluation (CIDEr) is a metric used to evaluate the quality of image descriptions. The key idea behind CIDEr is that human consensus can serve as a proxy for image description quality. First, the term frequency inverse document frequency (TF-IDF) weighting is used to give more importance to words that are discriminative. Then, the cosine similarity between the n-grams of the candidate sentences and the reference sentences is calculated. Next, the n-gram similarities for different n are combined into a single score using a geometric mean. A final CIDEr score is computed by averaging the CIDEr scores for each reference sentence. An important aspect of CIDEr is that it attempts to capture both the relevance and the saliency of the words in the generated caption.

SPICE: Semantic Propositional Image Caption Evaluation (SPICE) is an automated metric proposed for evaluating image captioning. It is a relatively new evaluation method, aiming to better capture the semantic information in image captioning. Unlike traditional n-gram-based evaluation methods like BLEU, ROUGE, METEOR, and CIDEr, which mainly evaluate through calculating the lexical overlap between reference annotations and generated annotations, SPICE analyzes at a higher level (i.e., the semantic level). SPICE first converts captions into a series of semantic tuples, each representing a specific semantic concept, such as objects, attributes, relations, etc. Then, an F-score is used to measure the degree of match between the generated annotation and the reference annotation.

S_m^*: It is an average of BLEU-4, METEOR, ROUGE$_{\mathrm{L}}$, and CIDEr. It is defined as follows:

$$S_m^* = \frac{1}{4}(\text{BLEU-4} + \text{METEOR} + \text{ROUGE}_{\mathrm{L}} + \text{CIDEr}). \quad (14)$$

C. Experimental Settings and Training Details

Dataset Splitting: To ensure a fair comparison with compared methods [9], [10], [14]–[17], [40], [41], each dataset is also randomly shuffled and divided into three parts for training, validation, and testing by the ratio of 80%, 10%, and 10%, respectively. To minimize the impact of random splits, we conduct five experiments on each RSIC dataset. For each dataset, we perform five experiments with five different random splits. The best and worst results are excluded, and the remaining outcomes are averaged to obtain more reliable and solid results.

Feature Extraction: The size of each input image is set to 224 × 224 pixels. For the **region features**, we employ the ResNet-152 pre-trained on ImageNet whose last fully-connect layer is removed. This process results in a feature map with dimensions of 7×7×512. Then we flatten the feature map into a matrix of size 49×512 (i.e., N = 49 and d_1 = 512). For the

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

CLIP image embedding, we utilize the pre-trained CLIP with ViT-B/32 backbone to extract the Class token with dimensions of 768 (i.e., d_2 = 768).

Model Setting and Training: Following [14], [39], we employ a sequential Transformer framework with B = 3 identical layers and use h = 8 parallel attention layers in the multi-head attention mechanism. According to prior research experience [13], [40], the dilation rates of the dilated convolutions in the encoder are set as 2, 4, and 8 in sequence. For the region features, the number of groups J is set to 8 and each head focuses on λ = 6 groups. In terms of text representation, a start token <bos> and an end token <eos> are added before and after each ground truth sequence. During the training phase, the input of the decoder is the ground truth sequence. However, during the inference period, only the start token <bos> needs to be input into the decoder. The word embedding dimension is configured to be 512. The maximum sequence length for positional encoding is 128, and there is a constraint of a maximum output sequence length of 20. We set the batch size to 50 and use the Adam Optimizer with β_1 = 0.9, β_2 = 0.98 to optimize the model parameters. The initial learning rate is set to 1. To prevent overfitting, dropout is applied with a probability of 0.1 after each attention and FFN layer. The hidden dimension of each FFN layer is set to 2048. Furthermore, we define patience to monitor the model's performance on the validation set and decide whether to halt the training prematurely. All experiments are conducted on an NVIDIA GeForce RTX 3090 with PyTorch version 1.10.0.

During the cross entropy training process, we adopt the learning rate scheduling strategy [19], which incorporates a warm-up operation comprising 10,000 iterations. We save the model with the highest CIDEr score on the validation set as the initialization for the subsequent training phase or inference. When the patience reaches to 5, the focus of training shifts towards CIDEr optimization and the learning rate is fixed at $5e-6$. In both the optimization and decoding phases, we employ a beam size of 5 to generate candidate sequences.

Compared Methods: To evaluate the effectiveness of the proposed MG-Transformer, we compare the proposed MG-Transformer with several state-of-the-art methods as below:

(1) mRNN [9] and mLSTM [9] adopt the vanilla CNN-RNN framework, with different CNNs as their encoders and different RNNs (naive RNN, LSTM) as the decoders.

(2) Soft-attention [10] and Hard-attention [10] introduce hard attention and soft attention [32] into the CNN-RNN framework, respectively.

(3) FC-Att+LSTM [15] and SM-Att+LSTM [15] integrate low-level features and high-level attribute features based on the attribute attention mechanism. The results are from [17].

(4) SAT(LAM) [16] and Adaptive (LAM) [16] adopt predicted categories' word embedding vectors to guide the calculation of attention masks, which helps filter out redundant image features.

(5) Word–Sentence framework [41] consists of a word extractor and a sentence generator. The word extractor employs various CNNs and loss functions to extract as many words as possible from RSIs. And the sentence generator utilizes different Transformer structures to generate sentences. Here, we present the best results.

(6) GVFGA+LSGA [17] introduces a Global Visual Feature-Guided Attention to filter out redundant visual information and designs a Linguistic State-Guided Attention to optimize the fusion of visual and text features.

(7) RASG [40] fuses multiscale visual features extracted by the ESP module. It designs a recurrent attention mechanism to capture high-level attentive maps and devises a semantic gate to merge the semantic information from two LSTMs.

(8) MLAT [14] integrates multiscale visual features from different convolutional layers, which are fed into a Transformer. It employs LSTMs to aggregate information from all encoder layers and feeds it into the Transformer decoder to generate sentences.

(9) M^2 Transformer [39] embeds the prior knowledge of relationships between objects into self-attention in Transformer encoder.

(10) PKG-Transformer [58] first enriches the object and scene features by leveraging the object-object and scene-scene relationship. Then, it integrates the scene-object relationship into the Transformer encoder as prior knowledge.

To validate the effectiveness of each component and facilitate comparison with other RSIC methods, **we establish a strong baseline with a standard Transformer encoder and the meshed decoder**. The encoder and decoder of the baseline also consist of three identical layers. In particular, we will place emphasis on comparing with state-of-the-art multiscale methods (i.e., RASG and MLAT). They also employ ResNet-152 as the CNN backbone. Their patiences are also set to 5.

D. Comparison With Other Methods

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

As can be seen from the three tables, our method achieves superior performance over other comparison methods on almost all metrics for the three datasets. For RSICD, our method is only slightly lower in terms of BLEU-1 and ROUGE_L compared to the best results. Specifically, it attains the highest scores on image captioning metrics (i.e., CIDEr and SPICE) and the overall metric (i.e., S_m^*). Next, we provide a detailed analysis of the experimental results on each dataset.

Results on Sydney-Captions: The mRNN and mLSTM adopt the vanilla CNN-RNN framework to achieve inferior CIDEr scores (i.e., 32.20% and 37.20%). Benefiting from the attention mechanism, Soft-attention and Hard-attention achieve significant performance. FC-Att+LSTM, SM-Att+LSTM, SAT(LAM), Adaptive(LAM), Word–Sentence, and GVFGA+LSGA further improve the performance by incorporating attribute information (such as label information and global information) into the attention mechanism. Further, RASG and MLAT have achieved better performance by fusing multiscale visual information. Compared to the better-performing MLAT, the strong baseline integrates the contributions of all encoder layers to improve CIDEr by

TABLE I

COMPARISON RESULTS ON SYDNEY-CAPTIONS. ALL RESULTS ARE REPORTED AS PERCENTAGE (%). BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, ROUGE_L, 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 PAPER.

Methods	B1	B2	B3	B4	M	R	C	S	S_m^*
mRNN [9]	51.30	37.50	20.40	19.30	18.50	-	32.20	-	-
mLSTM [9]	54.60	39.50	22.30	21.20	20.50	-	37.20	-	-
Soft-attention [10]	73.22	66.74	62.23	58.20	39.42	71.27	249.93	-	104.71
Hard-attention [10]	75.91	66.10	58.89	52.58	38.98	71.89	218.19	-	95.41
FC-Att+LSTM [15]	73.83	64.40	57.01	50.85	36.38	66.89	224.15	39.51	94.57
SM-Att+LSTM [15]	74.30	65.35	58.59	51.81	36.41	67.72	234.02	39.76	97.49
SAT(LAM) [16]	74.05	65.50	59.04	53.04	36.89	68.14	235.19	40.38	98.32
Adaptive(LAM) [16]	73.23	63.16	56.29	50.74	36.13	67.75	234.55	42.43	97.29
Word–Sentence [41]	78.91	70.94	63.17	56.25	41.81	69.22	204.11	-	92.85 -
GVFGA+LSGA [17]	76.81	68.46	61.45	55.04	38.66	70.30	245.22	45.32	102.31
RASG [40]	79.83±2.34	73.71±3.10	68.70±3.89	64.21±4.51	41.29±1.64	72.03±2.57	250.18±11.50	-	106.93±4.99
MLAT [14]	83.23±1.35	77.88±3.99	72.98±4.11	68.29±3.93	43.81±2.11	75.99±2.90	277.24±6.06	-	116.32±2.96
M^2 Transformer [39]	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
PKG-Transformer [58]	83.17±1.02	77.83±2.31	72.84±1.88	68.24±1.39	45.28±0.86	77.06±1.54	284.76±10.84	44.05±1.05	118.83±2.51
Baseline	83.11±0.89	77.57±2.14	72.78±2.95	68.38±3.46	44.58±2.29	76.11±2.03	289.79±19.78	43.93±1.68	120.23±5.46
MG-Transformer	**86.68±1.07**	**80.87±2.47**	**75.88±3.54**	**71.22±4.36**	**47.70±3.11**	**79.70±3.86**	**327.76±14.49**	**48.09±1.93**	**131.37±5.19**

TABLE II

COMPARISON RESULTS ON UCM-CAPTIONS. ALL RESULTS ARE REPORTED AS PERCENTAGE (%). BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, ROUGE_L, 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 PAPER.

Methods	B1	B2	B3	B4	M	R	C	S	S_m^*
mRNN [9]	60.10	50.70	32.80	20.80	19.30	-	42.80	-	-
mLSTM [9]	63.50	53.20	37.50	21.30	20.30	-	44.50	-	-
Soft-attention [10]	74.54	65.45	58.55	52.50	38.86	72.37	261.24	-	106.24
Hard-attention [10]	81.57	73.12	67.02	61.82	42.63	76.98	299.47	-	120.23
FC-Att+LSTM [15]	81.02	73.30	67.27	61.88	42.80	76.67	337.00	48.67	129.59
SM-Att+LSTM [15]	81.15	74.18	68.14	62.96	43.54	77.93	338.60	48.75	130.76
SAT(LAM) [16]	81.95	77.64	74.85	71.61	48.37	79.08	361.71	50.24	140.19
Adaptive(LAM) [16]	81.7	75.1	69.9	65.4	44.8	78.7	328.0	50.3	129.23
Word–Sentence [41]	79.31	72.37	66.71	62.02	43.95	71.32	278.71	-	114.00
GVFGA+LSGA [17]	83.19	76.57	71.03	65.96	44.36	78.45	332.70	48.53	130.37
RASG [40]	82.05±1.38	77.47±1.30	73.86±1.32	70.85±1.26	47.40±0.36	78.49±1.02	326.01±3.73	-	130.69±1.48
MLAT [14]	90.35±1.04	86.86±1.31	83.68±1.65	80.77±2.11	54.57±1.61	87.67±1.46	383.15±17.52	-	151.54±5.60
M^2 Transformer [39]	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
PKG-Transformer [58]	90.48±1.15	87.04±1.09	84.10±1.33	81.39±1.67	54.66±2.05	86.57±2.00	427.49±12.37	57.01±2.40	162.53±3.52
Baseline	86.65±1.27	82.54±1.63	79.27±1.85	76.40±2.02	50.55±1.58	82.74±1.57	401.54±5.25	52.87±1.70	152.81±2.60
MG-Transformer	**91.28±1.23**	**88.11±2.29**	**85.37±2.17**	**82.95±2.08**	**56.01±1.66**	**88.77±2.06**	**448.39±4.08**	**58.39±1.32**	**169.03±2.05**

TABLE III

COMPARISON RESULTS ON RSICD. ALL RESULTS ARE REPORTED AS PERCENTAGE (%). BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, ROUGE_L, 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 PAPER.

Methods	B1	B2	B3	B4	M	R	C	S	S_m^*
mRNN [9]	45.58	28.25	18.09	12.13	15.69	31.26	19.15	-	19.56
mLSTM [9]	50.57	32.42	23.19	17.46	17.84	35.02	31.61	-	25.48
Soft-attention [10]	67.53	53.08	43.33	36.17	32.55	61.09	196.43	-	
Hard-attention [10]	66.69	51.82	41.64	34.07	32.01	60.84	179.25	-	76.54
FC-Att+LSTM [15]	66.71	55.11	46.91	40.59	32.25	57.81	257.63	46.73	97.07
SM-Att+LSTM [15]	66.99	55.23	47.03	40.68	32.55	58.02	257.38	46.87	97.16
SAT(LAM) [16]	67.53	55.37	46.86	40.26	32.54	58.23	258.50	46.36	97.38
Adaptive(LAM) [16]	66.64	54.86	46.76	40.70	32.30	58.43	260.55	46.73	98.00
Word–Sentence [41]	**72.40**	58.61	49.33	42.50	31.97	62.60	206.29	-	85.84
GVFGA+LSGA [17]	67.79	56.00	47.81	41.65	32.85	59.29	260.12	46.83	98.48
RASG [40]	69.57±1.39	57.24±0.42	47.10±0.52	39.75±0.60	34.00±0.24	**64.48±0.42**	247.09±3.44	-	96.33±1.10
MLAT [14]	69.46±1.06	59.20±1.25	51.02±1.32	44.69±1.30	33.95±0.50	61.81±1.22	269.75±3.70	-	102.55±1.39
M^2 Transformer [39]	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
PKG-Transformer [58]	69.67±1.74	58.30±1.39	50.45±1.13	44.31±1.24	33.32±1.31	60.78±1.30	274.01±2.88	46.91±0.79	103.11±1.31
Baseline	66.42±0.84	55.10±1.29	46.98±1.57	40.71±1.85	31.29±0.50	58.07±0.85	246.12±9.18	43.88±0.57	94.05±3.10
MG-Transformer	70.27±1.54	**59.80±0.24**	**52.08±0.46**	**46.01±0.65**	**34.17±0.52**	62.27±0.93	**285.10±5.01**	**48.07±0.98**	**107.89±1.78**

4.53% and S_m^* by 3.36% (relative improvements, same below). The proposed MG-Transformer exceeds the MLAT by 18.22% on CIDEr and 12.94% on S_m^*, validating the effectiveness of our approach. In addition, object-based

TABLE IV
ABLATION STUDIES ON SYDNEY-CAPTIONS. ALL RESULTS ARE REPORTED AS PERCENTAGE (%). BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, ROUGE_L, CIDER, AND SPICE ARE DENOTED AS B1, B2, B3, B4, M, R, C AND S, RESPECTIVELY. DC, SC, AND GGA REPRESENT DILATED CONVOLUTION, SEMANTIC CORRELATION MODULE, AND GLOBAL GROUPING ATTENTION, RESPECTIVELY.

DC	SC	GGA	B1	B2	B3	B4	M	R	C	S	S_m^*
✗	✗	✗	83.11±0.89	77.57±2.14	72.78±2.95	68.38±3.46	44.58±2.29	76.11±2.03	289.79±19.78	43.93±1.68	120.23±5.46
✓	✗	✗	84.15±1.64	78.89±1.83	73.74±2.81	69.00±3.02	45.98±2.44	78.56±3.03	291.86±17.98	45.43±1.39	120.83±6.57
✓	✓	✗	86.06±1.46	80.29±2.77	75.18±3.61	70.33±4.10	46.40±2.74	78.69±2.37	307.59±10.88	45.85±1.52	125.97±5.85
✓	✓	✓	**86.68±1.07**	**80.87±2.47**	**75.88±3.54**	**71.22±4.36**	**47.70±3.11**	**79.70±3.86**	**327.76±14.49**	**48.09±1.93**	**131.37±5.19**

TABLE V
ABLATION STUDIES ON UCM-CAPTIONS. ALL RESULTS ARE REPORTED AS PERCENTAGE (%). BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, ROUGE_L, CIDER, AND SPICE ARE DENOTED AS B1, B2, B3, B4, M, R, C AND S, RESPECTIVELY. DC, SC, AND GGA REPRESENT DILATED CONVOLUTION, SEMANTIC CORRELATION MODULE, AND GLOBAL GROUPING ATTENTION, RESPECTIVELY.

DC	SC	GGA	B1	B2	B3	B4	M	R	C	S	S_m^*
✗	✗	✗	86.65±1.27	82.54±1.63	79.27±1.85	76.40±2.02	50.55±1.58	82.74±1.57	401.54±5.25	52.87±1.70	152.81±2.60
✓	✗	✗	87.00±1.51	82.85±1.64	79.50±1.66	76.58±1.60	50.95±0.52	82.94±1.54	402.38±6.09	53.91±0.25	153.21±2.44
✓	✓	✗	90.74±1.72	87.54±2.47	84.83±3.02	82.32±3.51	54.29±2.16	87.39±1.60	435.77±11.73	56.23±2.39	164.94±7.25
✓	✓	✓	**91.28±1.23**	**88.11±2.29**	**85.37±2.17**	**82.95±2.08**	**56.01±1.66**	**88.77±2.06**	**448.39±4.08**	**58.39± 1.32**	**169.03±2.05**

TABLE VI
ABLATION STUDIES ON RSICD. ALL RESULTS ARE REPORTED AS PERCENTAGE (%). BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, ROUGE_L, CIDER, AND SPICE ARE DENOTED AS B1, B2, B3, B4, M, R, C AND S, RESPECTIVELY. DC, SC, AND GGA REPRESENT DILATED CONVOLUTION, SEMANTIC CORRELATION MODULE, AND GLOBAL GROUPING ATTENTION, RESPECTIVELY.

DC	SC	GGA	B1	B2	B3	B4	M	R	C	S	S_m^*
✗	✗	✗	66.42±0.84	55.10±1.29	46.98±1.57	40.71±1.85	31.29±0.50	58.07±0.85	246.12±9.18	43.88±0.57	94.05±3.10
✓	✗	✗	66.53±0.80	55.14±0.61	47.02±1.41	40.78±1.83	31.74±0.40	58.62±0.65	252.90±6.32	44.41±0.77	96.01±2.20
✓	✓	✗	69.51±0.88	58.61±1.22	51.58±1.45	45.25±1.53	33.12±0.36	61.74±0.47	281.64±4.11	47.92±0.17	105.69±1.62
✓	✓	✓	**70.27±0.54**	**59.80±0.24**	**52.08±0.46**	**46.01±0.65**	**34.17±0.52**	**62.27±0.93**	**285.10±5.01**	**48.07±0.98**	**107.89±1.78**

methods (i.e. M^2 Transformer and PKG-Transformer) have also achieved significant performance. Compared to the state-of-the-art PKG-Transformer, CIDEr and S_m^* have increased by 15.10% and 10.55%, respectively. This demonstrates the superiority of the proposed MG-Transformer.

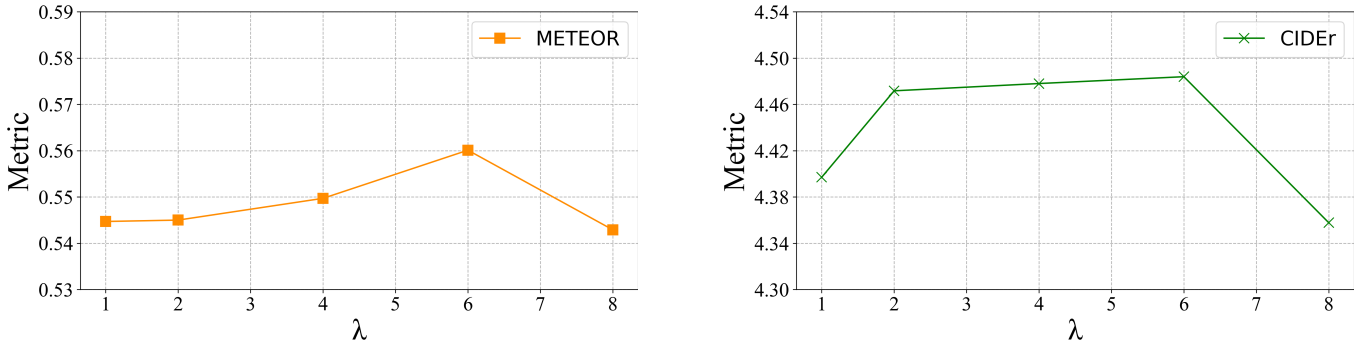
Results on UCM-Captions: Consistent with the results from Sydney-Captions, the mRNN and mLSTM achieve inferior CIDEr scores. With the incorporation of attention mechanisms and attribute information into the vanilla CNN-RNN framework, models such as Soft-attention, Hard-attention, FC-Att+LSTM, and SM-Att+LSTM achieve significant CIDEr and S_m^* scores. Our strong baseline outperforms the best multiscale method MLAT by 4.80% on CIDEr and 0.84% improvement on S_m^*. Additionally, the proposed MG-Transformer also improves 17.03% on CIDEr and 11.54% on S_m^*. And the proposed MG-Transformer exceeds the state-of-the-art PKG-Transformer by 4.89% on CIDEr and 4.00% on S_m^*.

Results on RSICD: From Table III, it is evident that among the compared methods, MLAT achieved the best results on most of the metrics, with scores of 269.75% and 102.55% on CIDEr and S_m^*, respectively. Unlike the results from the other two datasets, the performance of the strong baseline is inferior to that of MLAT but in close proximity to RASG. The proposed MG-Transformer exceeds MLAT by 5.69% on CIDEr and 5.21% on S_m^*. Furthermore, the proposed MG-Transformer outperforms the state-of-the-art PKG-Transformer by 4.05% in terms of CIDEr and 4.64% in terms of S_m^*.

E. Ablation Study

In this section, we further conduct extensive ablation experiments on the three RSIC datasets to evaluate the effectiveness of the dilated convolution (DC), SC module, and GGA in the proposed MG-Transformer. The results of these experiments are presented in Table IV - Table VI. Moreover, we validate the effectiveness of λ and CLIP, and their results are shown in Figure. 3 and Table VII, respectively.

Effectiveness of DC, SC, and GGA: The experimental results for the three datasets have similar variations. Taking Table IV as an example, the baseline achieves an average score of 289.79% for CIDEr and 120.23% for S_m^*. Next, we add three additional modules one by one and conduct a performance evaluation. The first component is the dilated convolution, which is responsible for extracting multiscale visual features from the raw data. By adding this module, we find that the CIDEr score and S_m^* score are increased by 0.71% and 0.50%, respectively. The results demonstrate that the inclusion of the module slightly improves the performance compared to the baseline.



(a) (b)

Fig. 3. Variation on METEOR (a) and CIDEr (b) with different λ values.

TABLE VII
EXPERIMENTAL RESULTS OF MG-TRANSFORMER WITH VIT-IMAGENET (V-I) OR VIT-CLIP (V-C) AND RESNET-IMAGENET (R-I) OR RESNET-CLIP (R-C) ON SYDNEY-CAPTIONS, UCM-CAPTIONS, AND RSICD. ALL RESULTS ARE REPORTED AS PERCENTAGE (%). BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, ROUGE_L, CIDER, AND SPICE ARE DENOTED AS B1, B2, B3, B4, M, R, C AND S, RESPECTIVELY.

Datasets	Backbone	B1	B2	B3	B4	M	R	C	S
Sydney-Captions	R-I & V-I	84.80±1.90	78.88±2.68	73.78±3.32	69.16±3.68	45.60±1.95	77.31±1.80	310.90±19.14	44.93±3.11
	R-I & V-C	**86.68±1.07**	**80.87±2.47**	**75.88±3.54**	**71.22±4.36**	**47.70±3.11**	**79.70±3.86**	**327.76±14.49**	**48.09±1.93**
	R-C & V-C	85.64±2.56	80.35±3.16	75.60±3.92	71.06±4.39	47.34±2.34	79.68±2.81	326.04±13.57	47.41±3.68
UCM-Captions	R-I & V-I	88.68±2.10	84.83±2.22	81.69±2.34	78.92±2.36	52.57±0.62	84.68±1.55	417.66±5.52	55.26±1.42
	R-I & V-C	91.28±1.23	88.11±2.29	85.37±2.17	82.95±2.08	**56.01±1.66**	**88.77±2.06**	**448.39±4.08**	**58.39±1.32**
	R-C & V-C	**91.50±0.58**	**88.32±0.85**	**85.65±1.06**	**83.16±1.23**	55.67±0.54	88.51±0.65	446.28±3.84	57.78±0.45
RSICD	R-I & V-I	68.59±0.38	57.57±0.31	49.68±0.41	43.54±0.52	32.75±0.60	60.25±1.18	265.73±10.78	45.55±1.20
	R-I & V-C	**70.27±0.54**	**59.80±0.24**	**52.08±0.46**	**46.01±0.65**	**34.17±0.52**	**62.27±0.93**	**285.10±5.01**	**48.07±0.98**
	R-C & V-C	69.88±0.99	59.25±0.67	51.38±0.62	45.18±0.69	33.84±0.29	61.78±0.29	279.07±3.47	47.60±0.50

The second component is the SC module, which integrates global and local information to obtain refined features. The introduction of this module further enhances the performance of the model. The CIDEr score and the S_m^* score are further increased by 5.39% and 4.25%, respectively. The performance gain achieved by incorporating the module highlights its importance in our method. More detailed ablation studies of the universal adaptability of the semantic correlation module are provided in the supplementary material.

Finally, we add the GGA, which enhances the ability to capture local information by introducing a grouping mechanism within self-attention. This module also significantly contributed to the overall performance improvement. The incorporation of GGA enhances the CIDEr score by 6.56% and the S_m^* score by 4.29%. The ablation analysis reveals that the module contributes significantly to the improved results.

Based on the above analysis, we can draw the following conclusions: the dilated convolution, the SC module, and the GGA all contribute to the performance improvement of the baseline model.

Effectiveness of λ: In the proposed MG-transformer, there is a critical hyperparameter, λ, which corresponds to the number of groups each head focuses on in GGA. In all of our experiments, region features are divided into $L = 8$ groups. To observe the impact of different λ values in the captioning process, we conducted experiments with λ = 1, 2, 4, 6, and 8 on UCM-Captions. To visually illustrate the impact of λ, we present the variation curves of METEOR and CIDEr in Figure. 3.

In the Figure. 3, we can observe two key points: (1) The best results are obtained when λ = 6. Smaller values of λ lead to inadequate learning of contextual information by GGA, while larger values result in redundancy across heads. (2) The less favorable results are obtained when λ = 8 (each head in GGA focuses on all region features).

Effectiveness of CLIP: To validate the effectiveness of CLIP, we conduct experiments by replacing ResNet-ImageNet with ResNet-CLIP, and substituting ViT-CLIP with ViT-ImageNet, respectively. The experiments are shown in Table VII. We can observe that integrating ResNet-CLIP with ViT-CLIP outperforms the combination of ResNet-ImageNet and ViT-ImageNet, while achieving results comparable to merging ResNet-ImageNet and ViT-CLIP.

F. Qualitative Analysis and Attention Visualization

Qualitative Analysis: To intuitively demonstrate the performance of the proposed MG-Transformer, qualitative results of the baseline, the state-of-the-art PKG-Transformer, and the proposed MG-Transformer, coupled with human-annotated ground truth (GT) captions are presented in Figure. 4. Obviously, the captions generated by our MG-Transformer are

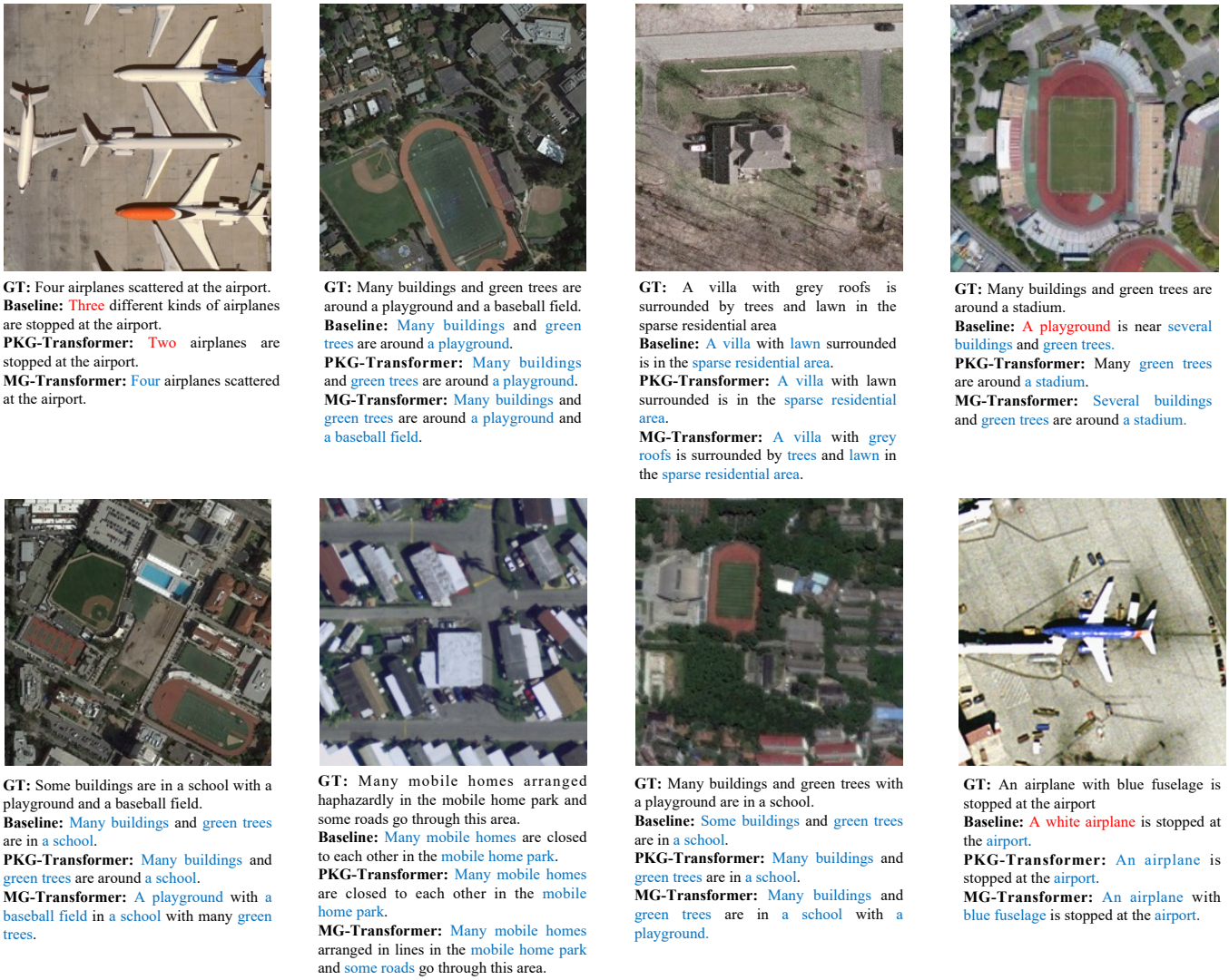


Fig. 4. Examples of captions generated by the baseline, PKG-Transformer, and the proposed MG-Transformer, as well as the corresponding ground truth captions. Some detailed and accurate words are marked in blue. And the words that are inconsistent with the image content are marked in red.

more accurate and comprehensive compared to the baseline and PKG-Transformer.

Taking the first subfigure with four different scale airplanes in Figure. 4 as an example, the baseline and PKG-Transformer inaccurately describe "three airplanes" and "two airplanes", while the proposed MG-Transformer accurately describe "Four airplanes". This fact provides evidence of the MG-Transformer's ability to address the multiscale problem. Further, taking the second subfigure with multiscale objects in Figure. 4 as an example, the baseline and PKG-Transformer describe "many buildings", "green trees" and "a playground", but miss the important object "a baseball field". In contrast, the proposed MG-Transformer provides a more comprehensive description, aligning effectively with GT captions. This indicates that the captions generated by our MG-Transformer are more accurate and comprehensive compared to the baseline and PKG-Transformer.

Attention Visualization: To better showcase the efficacy of our MG-Transformer in tackling multiscale challenges, we delve into the heatmaps of the encoding layers of the MG-Transformer, illustrated in Figure. 5, to identify the attended objects. We conduct comparisons by removing GGA (baseline+DC) and both GGA and DC (baseline) from the MG-Transformer to generate the heatmaps for comparison. Taking the first subfigure as an example, the RSI (top left) reveals a scene featuring a swimming pool, tennis courts, cars, buildings, and plants. In the baseline model, the three encoding layers primarily focuses on objects of similar scale, such as tennis courts and plants, while neglecting others. By utilizing dilated convolutions with diverse dilation rates, baseline+DC effectively captures objects across multiple scales, including tennis courts, the swimming pool, cars, and other objects. Through the integration of the grouping mechanism, the MG-Transformer, which builds upon baseline+DC, not only narrows down the focus of the model but also enhances its precision in attending to multiscale objects.

We also visualize the evolutions of attended image regions along the caption generation processes for the baseline and

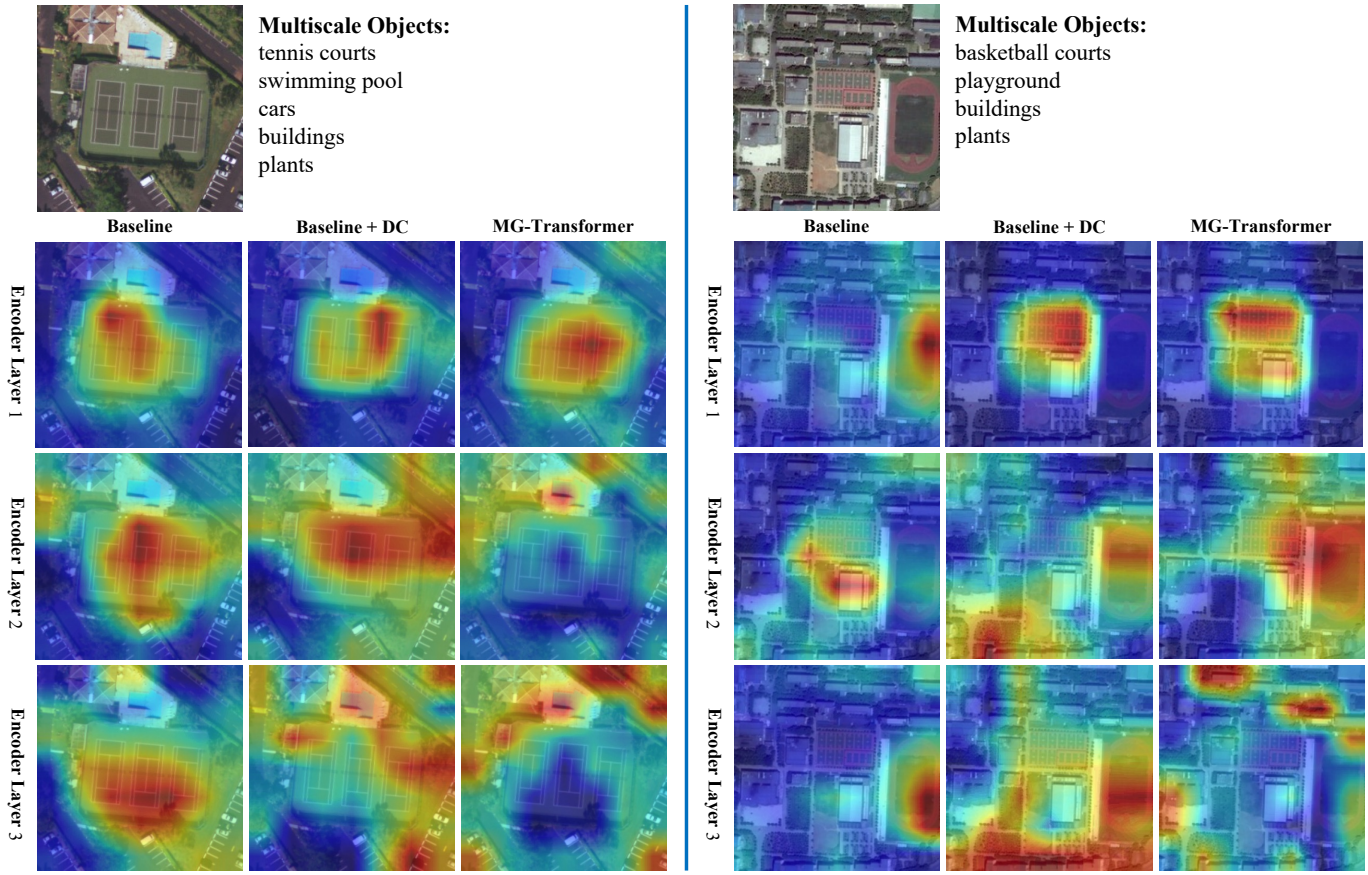


Fig. 5. Comparison of heatmaps generated by all encoding layers of baseline, baseline+DC, and MG-transformer. The degree of attention is indicated by the color intensity: red signifies a higher degree of attention, while blue represents a lower degree of attention.

MG-Transformer in Figure. 6. Each word pays more attention to the red pixels, which represent a larger attention weight.

Figure. 6 shows many "buildings", "trees" and a "playground" in a "school". The baseline not only inaccurately depicts the scene category as a "commercial area" but also misses an important object "playground". In contrast, our method provides a sentence that aligns perfectly with the GT captions. Moreover, both the baseline and our method generate the word "trees". However, our method demonstrates a more accurate correspondence between the word "trees" and the corresponding attention map. This once again demonstrates that our method can generate more accurate and comprehensive captions than the baseline. The results of qualitative analysis and attention visualization are included in the supplementary material.

G. Complexity Analysis

This section provides a brief analysis of the computational complexity of the proposed MG-Transformer. The computational complexity is mainly concentrated in the Transformer framework, with the most time-consuming operation being the GGA ($O(2N^2d+3Nd)$). First, computing the $\mathbf{head}_i$ according to Eq. (7) requires $O(2N^2D+3ND^2+3ND)$ cost. Then, computing $\mathrm{GGA}(\mathbf{X})$ via Eq. (8) costs $O(Nd^2+Nd)$. Therefore, the computational cost of the proposed MG-Transformer is $O(N^2)$, which is consistent with the computational complexity of the standard Transformer and the strong baseline. Baseline, the state-of-the-art multiscale method MLAT, and the proposed MG-Transformer all adopt the Transformer framework. Table VIII presents the number of model parameters and the number of floating-point operations (FLOPs). Among them, MLAT has the highest number of parameters. The Baseline and the proposed MG-Transformer exhibit a similar number of parameters and FLOPs. Compared to MLAT, our approach reduces the number of model parameters while achieving the best performance.

TABLE VIII
COMPARISON BETWEEN THE STRONG BASELINE, MLAT AND THE PROPOSED MG-TRANSFORMER ON THE NUMBER OF PARAMETERS AND FLOPS. THE SYMBOL "-" INDICATES THAT THE RESULT IS NOT REPORTED BY THE PAPER.

Method	Parameters	FLOPs
MLAT [14]	149.77M	-
Baseline	30.43M	1.34G
MG-Transformer	38.56M	1.60G

V. CONCLUSION

In this paper, we present a novel MG-Transformer framework that facilitates multiscale cross-modal interactions to

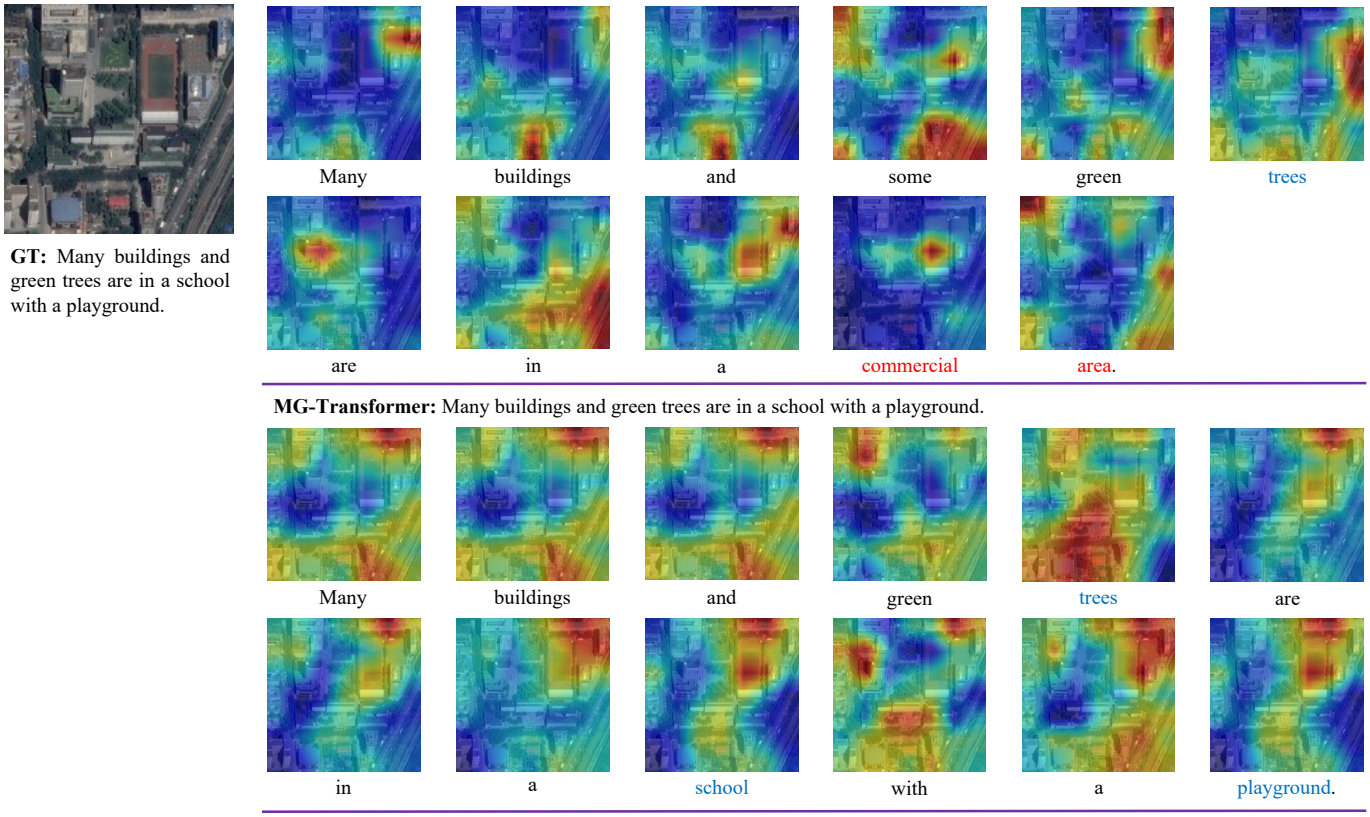


Fig. 6. The visualization of attended image regions along the caption generation processes for the baseline and our MG-Transformer. Pixels with larger attention weights are highlighted in red, while those with lower attention weights are highlighted in blue.

provide more precise captions for RSIC. We introduce the pre-trained CLIP model to extract the image embedding with pre-aligned image-text information as supplementary knowledge. An SC module is devised to integrate region features and the CLIP image embedding through an attention gate. Moreover, we design a GGA mechanism that incorporates a grouping mechanism into self-attention to enhance its local modeling ability. Experimental results and analysis demonstrate the effectiveness of the proposed MG-Transformer. In our future work, we will further explore the integration of multiscale features with Transformer.

REFERENCES

[1] D. Yu and S. Ji, "A New Spatial-Oriented Object Detection Framework for Remote Sensing Images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1-16, 2022, Art no. 4407416, doi: 10.1109/TGRS.2021.3127232.

[2] W. Ma, N. Li, H. Zhu *et al.*, "Feature Split–Merge–Enhancement Network for Remote Sensing Object Detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1-17, 2022, Art no. 5616217, doi: 10.1109/TGRS.2022.3140856.

[3] S. Tian, L. Kang, X. Xing *et al.*, "A Relation-Augmented Embedded Graph Attention Network for Remote Sensing Object Detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1-18, 2022, Art no. 1000718, doi: 10.1109/TGRS.2021.3073269.

[4] J. Han, J. Ding, N. Xue *et al.*, "ReDet: A Rotation-equivariant Detector for Aerial Object Detection," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2021, pp. 2785-2794, doi: 10.1109/CVPR46437.2021.00281.

[5] A. Ma, N. Yu, Z. Zheng *et al.*, "A Supervised Progressive Growing Generative Adversarial Network for Remote Sensing Image Scene Classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1-18, 2022, Art no. 5618818, doi: 10.1109/TGRS.2022.3151405.

[6] B. Zhang, S. Feng, X. Li *et al.*, "SGMNet: Scene Graph Matching Network for Few-Shot Remote Sensing Scene Classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1-15, 2022, Art no. 5628915, doi: 10.1109/TGRS.2022.3200056.

[7] Y. Li, Z. Zhu, J. -G. Yu *et al.*, "Learning Deep Cross-Modal Embedding Networks for Zero-Shot Remote Sensing Image Scene Classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10590-10603, Dec. 2021, doi: 10.1109/TGRS.2020.3047447.

[8] Y. Xu, B. Du and L. Zhang, "Assessing the Threat of Adversarial Examples on Deep Neural Networks for Remote Sensing Scene Classification: Attacks and Defenses," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1604-1617, Feb. 2021, doi: 10.1109/TGRS.2020.2999962.

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

[10] X. Lu, B. Wang, X. Zheng *et al.*, "Exploring Models and Data for Remote Sensing Image Caption Generation," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2183-2195, April 2018, doi: 10.1109/TGRS.2017.2776321.

[11] 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.

[12] Y. Wang, W. Zhang, Z. Zhang *et al.*, "Multiscale Multiinteraction Network for Remote Sensing Image Captioning," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 2154-2165, 2022, doi: 10.1109/JSTARS.2022.3153636.

[13] Y. Li, X. Zhang, J. Gu *et al.*, "Recurrent Attention and Semantic Gate for Remote Sensing Image Captioning," *IEEE Trans. Geosci. Remote Sens.* vol. 60, pp. 1-16, 2022, Art no. 5608816, doi: 10.1109/TGRS.2021.3102590.

[14] 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, no. 6506605, doi: 10.1109/LGRS.2022.3150957.

[15] X. Zhang, X. Wang, X. Tang *et al.*, "Description generation for remote sensing images using attribute attention mechanism," *Remote Sens.*, vol.

11, no. 6, Art. no. 612, Mar. 2019.
[16] Z. Zhang, W. Diao, W. Zhang *et al.*, "LAM: Remote sensing image captioning with label-attention mechanism," *Remote Sens.*, vol. 11, no. 20, Art. no. 2349, 2019.
[17] Z. Zhang, W. Zhang, M. Yan *et al.*, "Global Visual Feature and Linguistic State Guided Attention for Remote Sensing Image Captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1-16, 2022, no. 5615216, doi: 10.1109/TGRS.2021.3132095.
[18] C. Zihang, W. Junjue, M. Ailong *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, doi: 10.1109/LGRS.2022.3192062.
[19] A. Vaswani, N. Shazeer, N. Parmar *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998-6008.
[20] X. Wang, R. Girshick, A. Gupta *et al.*, "Non-local Neural Networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 7794-7803, doi: 10.1109/CVPR.2018.00813.
[21] A. Radford, J. W. Kim, C. Hallacy *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748-8763.
[22] A. Dosovitskiy, L. Beyer, A. Kolesnikov *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2021.
[23] A. Ramesh, P. Dhariwal, A. Nichol, *et al.*, "Hierarchical Text-Conditional Image Generation with CLIP Latents," 2022, [Online]. Available: https://arxiv.org/abs/2204.06125.
[24] K. He, X. Zhang, S. Ren *et al.*, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[25] A. Gupta and P. Mannem, "From image annotation to image description," in *Proc. Int. Conf. Neural Inf. Process.*, 2012, pp. 196–204.
[26] G. Kulkarni, V. Premraj, V. Ordonez *et al.*, "Babytalk: Understanding and generating simple image descriptions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2891–2903, Dec. 2013.
[27] S. Li, G. Kulkarni, T. L. Berg *et al.*, "Composing simple image descriptions using web-scale N-grams," in *Proc. Conf. Comput. Natural Lang. Learn.*, 2011, pp. 220–228.
[28] I. Sutskever, O. Vinyals, and Q. Le, "Sequence to sequence learning with neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1310–1318.
[29] V. Ordonez, X. Han, P. Kuznetsova *et al.*, "Large scale retrieval and generation of image descriptions," *Int. J. Comput. Vis.*, vol. 119, no. 1, pp. 46–59, 2016.
[30] P. Kuznetsova, V. Ordonez, A. Berg *et al.*, "Collective generation of natural image descriptions," in *Proc. Annu. Meet. Assoc. Comput. Linguist.*, 2012, pp. 359–368.
[31] O. Vinyals, A. Toshev, S. Bengio *et al.*, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 3156–3164.
[32] K. Xu, J. Ba, R. Kiros *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.
[33] P. Anderson, X. He, C. Buehler *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Jun. 2018, pp. 6077–6086.
[34] Y. Pan, T. Yao, Y. Li, *et al.*, "X-Linear Attention Networks for Image Captioning," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2020, pp. 10968-10977, doi: 10.1109/CVPR42600.2020.01098.
[35] P. Razvan, G. Caglar, K. Cho *et al.*, "How to construct deep recurrent neural networks," in *Proc. Int. Conf. Learn. Represent*, 2014.
[36] S. J. Rennie, E. Marcheret, Y. Mroueh *et al.*, "Self-Critical Sequence Training for Image Captioning," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 1179-1195, doi: 10.1109/CVPR.2017.131.
[37] S. Ren, K. He, R. Girshick *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.
[38] L. Huang, W. Wang, J. Chen *et al.* , "Attention on attention for image captioning," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 4634–4643.
[39] M. Cornia, M. Stefanini, L. Baraldi *et al.*, "Meshed-Memory Transformer for Image Captioning," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2020, pp. 10575-10584, doi: 10.1109/CVPR42600.2020.01059.
[40] S. Mehta, M. Rastegari, A. Caspi *et al.*, "ESPNet: Efficient spatial pyramid of dilated convolutions for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 552–568.
[41] Q. Wang, W. Huang, X. Zhang *et al.*, "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.
[42] C. Liu, R. Zhao, J. Chen *et al.*, "A Decoupling Paradigm with Prompt Learning for Remote Sensing Image Change Captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1-18, 2023, Art no. 5622018, doi: 10.1109/TGRS.2023.3321752.
[43] G. Hoxha, S. Chouaf, F. Melgani and Y. Smara, "Change Captioning: A New Paradigm for Multitemporal Remote Sensing Image Analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1-14, 2022, Art no. 5627414, doi: 10.1109/TGRS.2022.3195692.
[44] 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, pp. 1-20, 2022, Art no. 5633520, doi: 10.1109/TGRS.2022.3218921.
[45] S. Chang and P. Ghamisi, "Changes to Captions: An Attentive Network for Remote Sensing Change Captioning," *IEEE Trans. Image Process.*, vol. 32, pp. 6047-6060, 2023, doi: 10.1109/TIP.2023.3328224.
[46] C. Liu, J. Yang, Z. Qi, Z. Zou and Z. Shi, "Progressive Scale-Aware Network for Remote Sensing Image Change Captioning," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, 2023, pp. 6668-6671, doi: 10.1109/IGARSS52108.2023.10283451.
[47] J. Deng, W. Dong, R. Socher *et al.*, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[48] M. Ranzato, S. Chopra, M. Auli *et al.*, "Sequence level training with recurrent neural networks," 2015, [Online]. Available: http://arxiv.org/abs/1511.06732.
[49] S. Bengio, O. Vinyals, N. Jaitly *et al.*, "Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1171-1179.
[50] 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, April 2015, doi: 10.1109/TGRS.2014.2357078.
[51] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst.*, Nov. 2010, pp. 2–5.
[52] G. Xia, J. Hu, F. Hu *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.
[53] K. Papineni, S. Roukos, T. Ward *et al.*, "BLEU: A method for automatic evaluation of machine translation," in *Proc. Assoc. Comput. Linguist.*, 2002, pp. 311–318.
[54] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Worksh. Intrins. Extrins. Eval. Meas. Mach. Transl. Summar.*, vol. 29, 2005, pp. 65–72.
[55] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. Assoc. for Comput. Linguistics Workshop*, 2004, pp. 74–81.
[56] R. Vedantam, C. L. Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4566–4575.
[57] P. Anderson, B. Fernando, M. Johnson *et al.*, "Spice: Semantic propositional image caption evaluation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 382–398.
[58] 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, doi: 10.1109/TGRS.2023.3328181.

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 Shanghai University of Engineering Science, Shanghai, China, in 2021. He is pursuing the Ph.D. degree in computer science with the Nanjing University of Science and Technology (NJUST), Nanjing, China. His current research interests include pattern recognition, computer vision, and machine learning.

Jing Wang received the B.E. and Ph.D. degrees from Nanjing University of Science and Technology, Nanjing, China, in 2015 and 2022, respectively. From 2019 to 2020, she was a visiting scholar with University of Rochester, NY, USA. She is currently a Postdoctoral Fellow with the Department of Automation, Tsinghua University. Her research interests include computer vision and multimedia analysis, with a focus on vision and language.

Ran Meng received the B.A. degree from Central South University, Changsha, China, in 2017, and the MTI degree from Nanjing University of Science and Technology (NJUST), Nanjing, China, in 2021, where she is currently pursuing the Ph.D. degree in computer science. Her research interests include experimental phonetics, speech processing, pattern recognition, and machine learning.

Yang Yang received the Ph.D. degree in computer science, Nanjing University, China in 2019. At the same year, he became a faculty member at Nanjing University of Science and Technology, China. He is currently a Professor with the School of Computer Science and Engineering. His research interests lie primarily in machine learning and data mining, including heterogeneous learning, model reuse, and incremental mining. He has published over 10 papers in leading international journal/conferences. He serves as PC in leading conferences such as IJCAI, AAAI, ICML, NIPS, etc.

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 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.