International Journal of Digital Earth

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tjde20

MC-Net: multi-scale contextual information aggregation network for image captioning on remote sensing images

Haiyan Huang, Zhenfeng Shao, Qimin Cheng, Xiao Huang, Xiaoping Wu, Guoming Li & Li Tan

To cite this article: Haiyan Huang, Zhenfeng Shao, Qimin Cheng, Xiao Huang, Xiaoping Wu, Guoming Li & Li Tan (2023) MC-Net: multi-scale contextual information aggregation network for image captioning on remote sensing images, International Journal of Digital Earth, 16:2, 4848-4866, DOI: 10.1080/17538947.2023.2283482

To link to this article: https://doi.org/10.1080/17538947.2023.2283482

INTERNATIONAL JOURNAL OF DIGITAL EARTH
2023, VOL. 16, NO. 2, 4848–4866
https://doi.org/10.1080/17538947.2023.2283482

MC-Net: multi-scale contextual information aggregation network for image captioning on remote sensing images

Haiyan Huang [a], Zhenfeng Shao [a], Qimin Cheng[b], Xiao Huang [c], Xiaoping Wu[d], Guoming Li[e] and Li Tan[f]

[a]State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, People's Republic of China; [b]School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, People's Republic of China; [c]Department of Geosciences, University of Arkansas, Fayetteville, USA; [d]School of Geography and Resources Science, Sichuan Normal University, Sichuan, People's Republic of China; [e]School of Resources and Environment, University of Electronic Science and Technology, Sichuan, People's Republic of China; [f]School of Geophysics, Chengdu University of Technology, Sichuan, People's Republic of China

ABSTRACT

Remote Sensing Image Captioning (RSIC) plays a crucial role in advancing semantic understanding and has increasingly become a focal point of research. Nevertheless, existing RSIC methods grapple with challenges due to the intricate multi-scale nature and multifaceted backgrounds inherent in Remote Sensing Images (RSIs). Compounding these challenges are the perceptible information disparities across diverse modalities. In response to these challenges, we propose a novel multi-scale contextual information aggregation image captioning network (MC-Net). This network incorporates an image encoder enhanced with a multi-scale feature extraction module, a feature fusion module, and a finely tuned adaptive decoder equipped with a visual-text alignment module. Notably, MC-Net possesses the capability to extract informative multiscale features, facilitated by the multilayer perceptron and transformer. We also introduce an adaptive gating mechanism during the decoding phase to ensure precise alignment between visual regions and their corresponding text descriptions. Empirical studies conducted on four publicly recognized cross-modal datasets unequivocally demonstrate the superior robustness and efficacy of MC-Net in comparison to contemporaneous RSIC methods.

ARTICLE HISTORY
Received 28 August 2023
Accepted 8 November 2023

KEYWORDS
Image captioning; deep learning; semantic understanding; visual-text alignment

1. Introduction

Recent remarkable advances in deep learning technology have facilitated a better comprehension of remote sensing images. Conventional visual tasks, such as classification, detection, and segmentation, have made significant achievements, whereas semantic-level visual tasks such as image captioning remain challenging problems. RSIC involves a concise and coherent description of a complex scene captured in a specific image using natural language. The goal is not only to recognize the information pertaining to target objects in an image but also to understand the relationship between objects and to generate syntactically accurate and semantically fluent descriptive sentences.

CONTACT Zhenfeng Shao shaozhenfeng@whu.edu.cn State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, 430079 People's Republic of China

In recent years, inspired by advances in machine translation and computer vision, researchers have explored various techniques for the task of RSIC, which holds promise for applications in remote sensing image retrieval (Cheng et al. 2021), object detection (Zhang et al. 2019), military intelligence generation (Shi and Zou 2017), and disaster assessment (Liu et al. 2018).

Traditional image captioning methods can be classified into template-based (Farhadi et al. 2010; Ordonez, Kulkarni, and Berg 2011) and retrieval-based (Kulkarni et al. 2013; Ushiku et al. 2015) methods. Template-based methods used a predefined template to generate image descriptions. Retrieval-based methods generate a description of the query image based on the description of the similarity image. The generating sentences of above methods rely on the performance of the template and retrieval.

The advent of deep learning has significantly advanced the captioning of remote sensing images, where the encoder-decoder network (Anderson et al. 2018; Lu et al. 2017; Vinyals et al. 2015; Xu et al. 2015) has been widely adopted, and it can overcome the limitations of traditional template and retrieval approaches. Natural image captioning has made remarkable progress with the rapid development of artificial intelligence (AI). Remote sensing images have unique characteristics such as large imaging ranges, the presence of objects of different scales, and complex scenes in the same image (Du et al. 2021; Feng et al. 2023). In addition, complex scenes often depict several types of feature targets, making it challenging for category labels to capture the entire image content. These characteristics significantly increase the difficulty of captioning RSIs.

Researchers have conducted studies in the field of remote sensing. Some benchmark datasets (Lu et al. 2017; Qu et al. 2016) have been established to meet the data dependence of deep learning methods, including the largest cross-modal dataset established in an earlier work (Cheng et al. 2022). Researchers have used multilevel attention (Yuan, Li, and Wang 2019), denoising (Huang, Wang, and Li 2020), label information (Zhang et al. 2019), and attribute information (Zhang et al. 2019) to obtain richer image features. Wang et al. (2022) combined multi-label semantic information as a priori information and designed two methods for fusing semantic attributes and image features. On the decoder side, researchers have optimized the generated statements by extending long short-term memory (LSTM) Fu et al. (2020), using support vector machines (SVM) Hoxha and Melgani (2020), and optimizing loss functions during the training process (Li et al. 2020). Some innovative approaches have been introduced to model architectures that incorporate image retrieval (Wang et al. 2020), sound information (Lu, Wang, and Zheng 2019), and interpretability enhancement (Wang et al. 2020). Sumbul, Nayak, and Demir (2020) developed a summary-driven model, whereas (Hoxha, Melgani, and Demir 2020) combined retrieval and captioning methods to generate multiple captions and then compared them to reference headings to determine the final sentence captioning. Aiming at the problem of label scarcity in remote sensing image captioning, Yang, Ni, and Ren (2022) combined meta-learning into a remote sensing image captioning framework, which extracts meta features from two support tasks, including natural image classification and remote sensing image classification, and transfers the meta features to RSIC. To address the multi-scale problem, Zhang et al. (2019) introduced multiscale cropping to enhance data. Li et al. (2021) incorporated static and multiscale features by using recurrent attention. Zhang et al. (2021) reduced the burden of hidden states because the proposed language state (LS) provides text features exclusively so that the hidden states only guide the visual-textual attention process. Furthermore, Cheng et al. (2022) fused specific spatial regions and image scales. Wang et al. (2022) proposed a multiscale feature representation structure based on the two-stage training strategy. Existing remote sensing image captioning methods neglect to filter the redundant features of remote sensing images and do not consider the effective utilization of multi-scale image features, which hinder the development of remote sensing image description, therefore, how to obtain significant visual features is a problem that needs to be solved.

To address these problems, we propose a multi-scale contextual information aggregation network, including multi-scale feature image encoder and adaptive visual-text alignment decoder. The image encoder uses a multiscale feature extraction module (MS) to extract image features,

for which channel relation modeling is proposed to filter redundant features. Furthermore, two feature fusion methods were proposed using MLP and a transformer to model the local and global context information of the images. Finally, a visual-text alignment mechanism is used to generate descriptions that are both syntactically accurate and semantically fluent. Based on the above design, the three proposed modules can work together to generate accurate and informative descriptions. The contributions of this paper can be summarized as follows:

(1) We propose a multi-scale contextual information aggregation network for RSIC that mainly extracts deep multi-scale features and multifaceted backgrounds inherent in remote sensing images (RSIs) by expoloiting multi-scale feature image encoder and adaptive visual-text alignment decoder.
(2) We propose multi-scale feature extraction and feature aggregation modules in the image encoder. The former module extracts deep features from remote sensing images at various scales and filter redundant features. The latter module can adaptively fuse cross scale feature information with the aim of improving the ability to understand the contextual information of RSIs. A visual-text alignment mechanism is introduced to the decoder to generate accurate descriptive sentences.
(3) We demonstrate the effectiveness and robustness of MC-Net on four cross-modal benchmark datasets. The results show that the proposed MC-Net achieves state-of -the-art performance.

2. Method

The architecture of the MC-Net model, which is based on an encoder-decoder structure, is shown in Figure 1. We provide a detailed overview of the MC-Net model, which includes the encoder based on multi-scale feature extraction and fusion, and the decoder, which uses the visual-text alignment mechanism. Specifically, the encoder leverages the MS to extract features at various scales and then combines this information to obtain optimized image features. On the decoding side, the fused multiscale feature information is input to an LSTM decoder, which generates the captioning of the remote sensing image. In the training process, cross-entropy loss is used to train MC-Net.

2.1. Multi-scale feature extraction module

Remote sensing image captioning aims to understand the information of an image at a fine-grained level and describe it in a fluent sentence. In this work, we design a MS module to extract visual features at different scales and filter redundant features of RSIs.

A pretrained VGG-16 was used to extract the visual features. Specifically, we divide the input features into four groups along the channel dimension, with each group denoted as

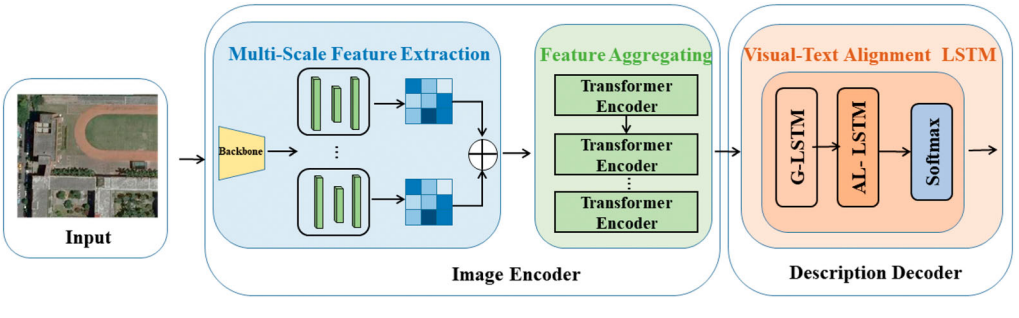

Figure 1. The overall architecture of the MC-Net. The image encoder with multi-scale feature extraction and feature aggregating module is used to extract visual features, and the description decoder with visual-text alignment LSTM is used to generate description sentences.

$X_i \in^{H\times W\times C_i}, i \in \{1, 2, 3, 4\}$, where H, W and C_i denote the different dimensions of image, respectively indicates height, width, and channels in each group, respectively. For each group of vectors, the convolution is calculated. Specifically, for the first group of input feature vectors, the output features are obtained directly by a convolution of 1×1. The second, third, and fourth groups of input feature vectors were generated with the output of the previous group. For which, the output features were obtained using a 3×3 convolution block after each remaining group of inputs can obtain a larger sensory field (1×1, 3×3, 5×5, 7×7). Furthermore, the channel relationship modeling module adopts the previous set of feature vectors output by the spatial relationship modeling module and the current feature vectors as the input. This design achieves filtering of redundant information with an optimized image feature representation. Figure 2 shows the structure of the MS module based on a multiscale convolution layer. The right panel of Figure 2 shows an expansion diagram of the channel relationship modeling module.

The above workflow can be summarized using the following equations:

$$X = VGG(I), \tag{1}$$

$$F_i = K_i^s(X_i), \quad i = 1, \tag{2}$$

$$F_i = K_i^s(X_i + F_{i-1}), \quad i \in 2, 3, 4, \tag{3}$$

where I denotes the input of the encoder, K_i^s denotes the convolution operation corresponding to each set of feature vectors, and F_i denotes the corresponding output.

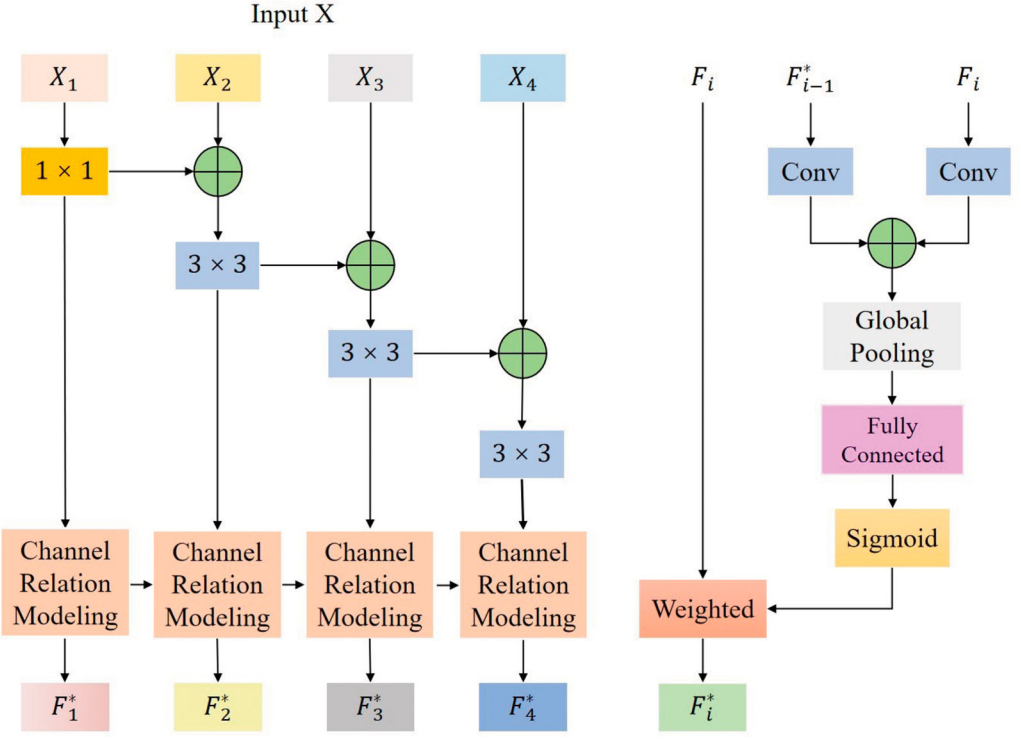

Figure 2. Outline of multi-scale visual feature extraction module.

Next, the channel-level information is obtained by using global average pooling:

$$U_i = W_{SS}^* F_{i-1}^* + W_{SS} F_i, \tag{4}$$

where , W_{SS}, respectively, denote the parameters of the two convolutions in module.

Next, the channel-level information is obtained by using global average pooling:

$$z_{ic} = F_{gp}(u_{ic}) = \frac{1}{H*W} \sum_{m=1}^{H} \sum_{n=1}^{W} u_{ic}(m, n). \tag{5}$$

We used a fully connected layer and an activation function to extract useful channel information. The output of the ith attention group is denoted by :

$$a_i = (z_i, W) = \sigma(W z_i). \tag{6}$$

The final output of spatial multi-scale features is obtained:

$$F_{ic}^* = F_{weighted}(F_{ic}, a_{isc}) = a_{isc} F_{ic}. \tag{7}$$

We further concatenated each group of optimized features along the channel dimensions to obtain the final feature.

$$F^* = concat(F_1^*, F_2^*, F_3^*, F_4^*), \tag{8}$$

where $F_i^* = F_{weighted}[F_{i1}^*, F_{i2}^*, F_{i3}^*, F_{i4}^*]$.

2.2. Feature aggregation module

We propose two feature aggregation methods to adaptively fuse cross-scale feature information with the aim of improving the ability to understand the contextual information of remote sensing images. Our strategy includes local modeling and global modeling.

2.2.1. Local modeling of images

The extracted multi-scale image features are fed into the MLP for feature learning, and the features on the four scales are attention-weighted using the sigmoid activation function. Figure 3 shows the local image modeling process. First, we obtained feature S by concatenating the extracted multiscale remote sensing image features:

$$S = concat(S_1^*, S_2^*, S_3^*, S_4^*). \tag{9}$$

Next, the concatenated image features were downsampled through the FC layer, and the correlation between the multiscale features was learned by the MLP. We calculate the weight matrix W using the sigmoid activation function for downsampling image features.

$$W = Sigmoid(MLP(S)). \tag{10}$$

Subsequently, the features of different scales are multiplied by the score weight matrix to obtain the weighted image features:

$$S_{final} = S + S*W, \tag{11}$$

where W is the weighting coefficient, and represents the image features obtained by post-attentional weighting.

2.2.2. Global modeling of images

Image global context information is important to the image description results. We use the advantage of transformer to fuse the image features of different scales to obtain the global information.

The process of global modeling of images is shown in Figure 4. After the extraction of multiscale features, they are fed into the transformer encoder, where the position information of each element in the sequence is considered by positional encoding. Further, the encoded information is fed into the encoder that consists of four identical encoding modules, each containing two sublayers, that is, a multi-head attention and feed forward network. The forward network extracts the features of the input sequence, and each sublayer reduces the loss of information and ensures training stability by connecting the residuals and regularizing the layers. The specific workflow can be summarized as follows:

$$Z_0 = (x_{class}; x_p^1 E; x_p^2 E \cdots x_p^N E) + Epos, \tag{12}$$

$$Z_l' = MSA(LN(Z_{l-1})) + Z_{l-1}, \quad l = 1, \ldots L, \tag{13}$$

$$Z_l = MLP(LN(Z_l')) + Z_l', \quad l = 1, \ldots L, \tag{14}$$

$$Y = LN(Z_L^o), \tag{15}$$

where E is the input image block feature, E_{pos} is the location encoding, MSA denotes the multi-head self-attention, MLP denotes the forward network, and LN denotes the layer regularization.

The Transformer Encoder module employs multi-head self-attention to identify the interrelationships among objects in remote sensing images. The encoder module does not alter the dimensionality of the image features.

2.3. Visual-text alignment decoder

We employed an adaptive gating mechanism to achieve the adaptive selection of image information and language during the decoding phase. The Visual-Text alignment module consists of gated attention LSTM (G-LSTM) and adaptive language LSTM (AL-LSTM). The multiscale contextual features of the image extracted by the encoding network are fed to the decoding LSTM to generate descriptive statements of the image. The input vectors for the gate attention LSTM at each time step serve as the embedding vector of the current word, average pooling features of the image, and previously hidden state of the adaptive language LSTM. Then, the specific positions of the LSTM multiscale features are guided according to the attention mechanism, whereas the attention vector is optimized by the gating mechanism. Furthermore, the semantic gating vector facilitates the adaptive alignment of the visual features of the decoding process and textual information of the description statement. Finally, the context vector and Attention LSTM hidden state generated by gating

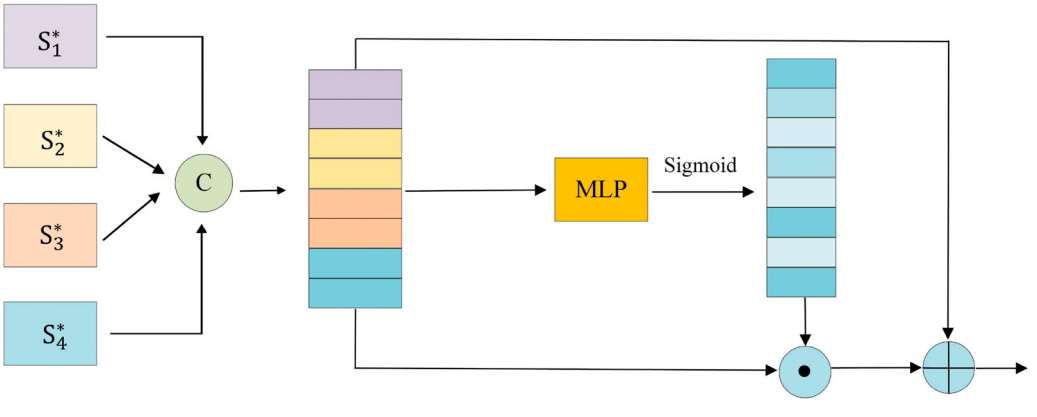

Figure 3. The process of local modeling of images.

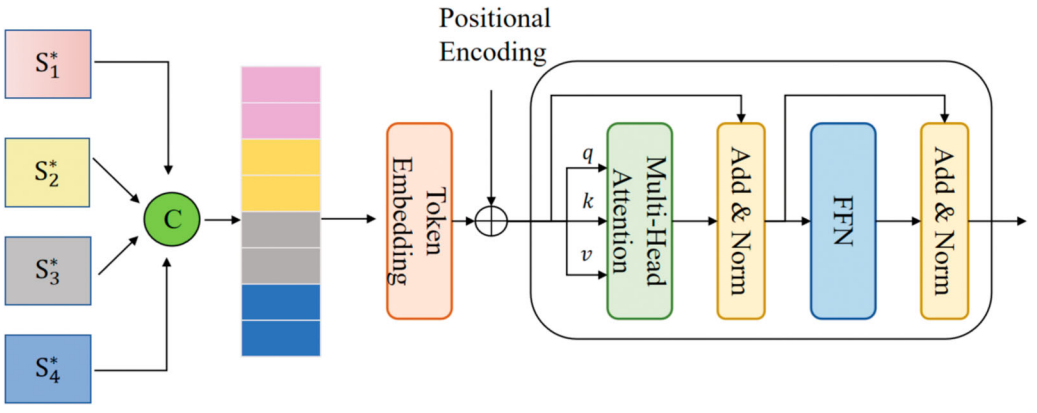

Figure 4. The process of global modeling of images.

attention are fed to the language LSTM. The decoding process of the adaptive visual alignment of the two-layer LSTM is as follows.

$$h_t^1 = LSTM_1(concat[h_{t-1}^2, \bar{V}, W_e x_t], h_t^1), \tag{16}$$

$$a_{j,t} = W_a^T \tanh(W_{va} V_i + W_{ha} h_t^1), \tag{17}$$

$$\alpha_t = softmax(a_t), \tag{18}$$

$$\hat{V} = \sum_{j=1}^{H \times W} \alpha_{j,t} V_i, \tag{19}$$

where $\bar{V}$ denotes the global average feature; W_e is the word embedding matrix; W_{va}, W_{ha} and W_a^T are the learnable parameters; and α_t refers to the vector composed of the attention weights corresponding to each of the regional feature vectors. $\hat{V}$ denotes the visual attention vector.

The attention mechanism guides the decoding process to generate a weighted average feature vector at each time step, and the result of the image description generation greatly depends on the attention result. In this study, we optimize the attention vector by extending the existing attention mechanism and combining it with the gating mechanism, which leads to the retention of useful attention information during the decoding process. The optimized attention vector $\hat{V}'$ is derived as follows:

$$i = w_q^i h_t^1 + w_v^i \hat{V} + b^i, \tag{20}$$

$$g = \sigma(w_q^q h_t^1 + w_v^g \hat{V} + b^g), \tag{21}$$

$$\hat{V}' = g \odot i, \tag{22}$$

where w_q^i, w_v^i, b^i, b^g refer to the learnable parameters and $\odot$ denotes element-by-element multiplication.

To select visual or sentence context information to generate sentences, we introduce a semantic gate. The context vector trades off how much new information the network is considering from the

image with what it already knows in the decoder memory.

$$g_t = \sigma(W_x^1 x_t^1 + W_h^1 h_{t-1}^1), \tag{23}$$

$$S_t = g_t \odot \tanh(m_t^1), \tag{24}$$

$$c_t^{'} = \beta_t S_t + (1 - \beta_t)\hat{V}^{'}, \tag{25}$$

β_t produces a scalar in the range [0,1]. A value of 1 implies that long and short term visual and linguistic information of decoder memory is used and 0 means only spatial image information is used when generating the next word. To calculate β_t, we add an additional element S_t to the decoding model, which indicates the extent to which the model pays for the sentence context:

$$z_t = w_h \tanh(w_s V + W_h h_t^1), \tag{26}$$

$$\beta_t = softmax(concat(z_t, w_h \tanh(w_s S_t + W_h h_t^1))), \tag{27}$$

where $V = v_1, v_2, \ldots, v_L$ denotes an image feature vector. w_s and W_h represent learnable weight parameters. Furthermore, the context vector $c_t^{'}$, is input to the language LSTM to obtain the final output captioning after the softmax layer:

$$h_t^2 = LSTM_2(concat[h_t^1, c_t^{'}], h_{t-1}^2), \tag{28}$$

$$S_t = softmax(W_0 h_t^2 + b_0). \tag{29}$$

Our model is trained with maximum likelihood estimation of MLE loss, with the goal of minimizing MLE loss. The input x_t and previous hidden state h_{t-1} are combined to obtain the hidden state h_t in the training phase. Then, the softmax function is used to calculate the probability distribution of words during utterance generation, and the word with the highest probability is selected as the predicted word. The predicted word then served as the input for the next time step. The above steps were repeated until the network predicted the end vector. The loss function for model training is the sum of the negative log likelihoods that generate the correct description of words in each time step:

$$L(\theta) = -\sum_{t=1}^{T} \log(p_\theta(s_t^* | s_1^*, \ldots s_{t-1}^*)), \tag{30}$$

where θ is the parameter to be learned and $(s_1^*, \ldots s_t^*)$ represents the generated descriptive sentence.

3. Experiments

3.1. Dataset

(1) The UCM-Captions Dataset (Qu et al. 2016) was created by expanding of the UCMerced Land Use Dataset (Yang and Newsam 2010) originally designed for remote sensing image classification tasks. To enrich each remote sensing image, five distinct descriptive statements, each annotated by the same person, were incorporated. The annotated sentence descriptions consist of 21 categories, with each category containing 100 images, and the resolution of image is 256 × 256 pixels and the resolution size of a pixel is 0.3048 m. The dataset contains 2,100 images paired with 10,500 sentence descriptions.

(2) The Sydney-Captions Dataset (Qu et al. 2016) enriches each image by including five distinct sentences, each annotated by the same person, across seven categories. The number of each category is different. All images (500 × 500 pixels) in the dataset have a resolution of 0.5 m. It consists of 613 images and 3,065 sentence descriptions and has been frequently used

for comparative evaluations with the UCM-Captions Dataset in remote sensing description generation tasks.

(3) The RSICD (Lu et al. 2017) was developed in 2017, featuring 10,921 images distributed across 30 categories, each having a resolution of 224 × 224 pixels. Each image in the dataset is accompanied by 1-2 description statements, totaling 24,233 annotations. The description statements are sourced from multiple individuals, resulting in five sentence descriptions for each image. In instances where an image has fewer than five titles, existing sentence descriptions are extended through random copying. Consequently, the five descriptive statements for each image in this dataset may not be entirely distinct.

(4) The NWPU-Captions Dataset (Cheng et al. 2022) contains 31,500 images (256 × 256 pixels) across 45 categories, each having a resolution of 0.228 m. Unlike other datasets discussed previously, the NWPU-Captions Dataset comprises five unique descriptions labeled by different individuals, ensuring a diverse set of sentences. The dataset is also larger and more varied in its feature categories, which accurately reflects the intricate image variation in remote sensing images, including high intraclass diversity and interclass similarity. Students with a remote sensing background manually annotated the dataset using comprehensive expressions to ensure rich structural and lexical diversity in the description statements. The dataset contained a minimum of six words per description statement.

3.2. Evaluation metrics

Five commonly used evaluation metrics were employed to evaluate the performance of the proposed MC-Net for RSIC: BLEU (Papineni et al. 2002), ROGUE (Rouge 2004), METEOR (Banerjee and Lavie 2005), CIDEr (Vedantam, Zitnick, and Parikh 2015), and SPICE (Anderson et al. 2016). Evaluation metrics are utilized to objectively measure the correlation between the generated description sentences and the ground truth. A higher value for all the evaluation metrics indicates that the generated image description statement is closer to the ground truth. BLEU is a machine translation metric that analyzes the n-tuple correlation between generated description sentences and reference sentences. ROGUE, originally used for text summarization, calculates the longest common subsequence (LCS) and then obtains the F-measure. METEOR, another machine-translation metric, has a strong correlation with human judgment. CIDEr and SPICE are specifically designed for image captioning. CIDEr comprehensively evaluates the performance of the model, emphasizing the quality of image content. SPICE focuses on evaluating the sentence structure of a generated description sentence by utilizing a scene graph form to encode the targets, attributes, and their relationships within the sentence.

3.3. Implementation details

In our study, we evaluated the performance of the model using four datasets: 80% of the data reserved for model training, 10% for model validation, and 10% for model testing. Prior to training, the input model images were preprocessed to 224 × 224 pixels. The feature extraction network of all the models was standardized using VGG16 as the backbone model, with the encoder model fine-tuned. An Adam optimizer was used. The initialized learning rates of the encoder and decoder were set to 1e-4 and 5e-4, respectively. For every five epochs, the learning rate decreased to 0.8 times of the original during the training process. The batch size was 64 and the maximum number of epochs was 100. The word-embedding dimensionality was set to 512, while the number of encoding layers of the transformer was four. The model with the maximum value of CIDEr was selected for testing. During sentence generation, we utilized a beam search strategy set to five to generate candidate sentence fragments for an image. The maximum length of the generated sentences was set to 25.

3.4. Experimental results

3.4.1. Ablation experiments

A series of ablation experiments are desined to evaluate the effectiveness of different module of MC-Net. The baseline model serves as the control, while stacked convolution, Transformer-based feature aggregating, and feature aggregating method using MLP represent the different configurations tested. We named multi-scale feature extraction module as MSF, local modeling module as LM and global modeling module as GM.

Tables 1–4 display the captioning results of each submodule, demonstrating the impact of submodules on the captioning performance of the MC-Net model. The experimental results show that adding each submodule to the three different datasets outperformed the baseline model, and the best image description accuracy was achieved by adding global modeling module. Specifically, on the RSICD dataset, multi-scale feature module improved the Cider and Spice values by approximately 1%, METEOR and Rouge values by approximately 3%, and BLEU values by more than 5%, compared to the baseline model. Similarly, the global modeling module showed a significant improvement in each indicator, which was comparable to the performance of multiscale feature module, but the performance improvement varied slightly across different datasets. For instance, multi-scale feature module performed slightly better than global modeling module on the UCM-Captions and Sydney datasets, whereas global modeling module outperformed multi-scale feature module on the RSICD and NWPU-Captions datasets. Our findings demonstrate that incorporating multiscale feature extraction and aggregating modules can improve model performance. Furthermore, the Transformer-based global modeling approach outperformed the feature aggregation approach using MLP and Sigmoid in the fusion module. The highest performance was achieved by combining stacked-convolution multiscale feature extraction with transformer-based feature aggregation, highlighting the ability of MC-Net to generate more accurate and fluent sentences.

3.4.2. Comparative experiments

To evaluate the effectiveness of the MC-Net proposed in this study, we compared it with other comparative methods, including Attend (Xu et al. 2015), Convcap (Aneja, Deshpande, and Schwing 2018), CSMLF (Wang et al. 2019), Multimodal (Qu et al. 2016), Sound-a-a (Lu, Wang,

Table 1. Settings and results of ablation experiments on the UCM-Captions.

Method				Metrics							
Baseline	MSF	LM	GM	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE_L	CIDEr	SPICE
✓				0.806	0.724	0.656	0.592	0.403	0.744	3.009	0.462
✓	✓			0.841	0.771	0.709	0.651	0.418	0.777	3.202	0.482
✓		✓		0.823	0.754	0.694	0.637	0.412	0.753	3.154	0.471
✓			✓	0.840	0.771	0.713	0.659	0.435	0.784	3.329	0.503
✓	✓	✓		0.843	0.778	0.721	0.658	0.436	0.781	3.323	0.494
✓	✓		✓	**0.845**	**0.784**	**0.732**	**0.679**	**0.449**	**0.786**	**3.355**	**0.520**

Table 2. Settings and results of ablation experiments on the Sydney-Captions.

Method				Metrics							
Baseline	MSF	LM	GM	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE_L	CIDEr	SPICE
✓				0.789	0.711	0.642	0.579	0.388	0.708	2.261	0.391
✓	✓			0.816	0.733	0.655	0.581	0.381	0.728	2.290	0.407
✓		✓		0.819	0.728	0.648	0.580	0.339	0.724	2.278	0.396
✓			✓	0.824	0.740	0.658	0.581	0.395	0.734	2.336	0.415
✓	✓	✓		0.821	0.738	0.661	0.589	0.386	0.0.732	2.241	0.423
✓	✓		✓	**0.834**	**0.750**	**0.678**	**0.607**	**0.406**	**0.739**	**2.564**	**0.456**

Table 3. Settings and results of ablation experiments on the RSICD.

Method				Metrics							
Baseline	MSF	LM	GM	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE_L	CIDEr	SPICE
✓				0.686	0.560	0.465	0.390	0.342	0.615	2.074	0.428
✓	✓			0.724	0.599	0.503	0.425	0.354	0.636	2.376	0.450
✓		✓		0.711	0.578	0.495	0.414	0.342	0.618	2.306	0.437
✓			✓	0.725	0.599	0.500	0.420	0.345	0.628	2.311	0.443
✓	✓	✓		0.725	0.603	0.508	0.428	0.357	0.639	2.412	0.458
✓	✓		✓	**0.728**	**0.606**	**0.511**	**0.433**	**0.360**	**0.641**	**2.454**	**0.463**

Table 4. Settings and results of ablation experiments on the NWPU-Captions.

Method				Metrics							
Baseline	MSF	LM	GM	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE_L	CIDEr	SPICE
✓				0.731	0.602	0.515	0.454	0.334	0.581	1.092	0.274
✓	✓			0.738	0.619	0.534	0.472	0.344	0.596	1.125	0.286
✓		✓		0.733	0.609	0.522	0.459	0.337	0.588	1.119	0.281
✓			✓	0.737	0.616	0.528	0.468	0.336	0.594	1.122	0.284
✓	✓	✓		0.739	0.623	0.538	0.475	0.346	0.604	1.135	0.288
✓	✓		✓	**0.741**	**0.626**	**0.544**	**0.478**	**0.347**	**0.611**	**1.159**	**0.289**

and Zheng 2019), FC-ATT/SM-ATT (Zhang et al. 2019), SAT(LAM) (Zhang et al. 2019), GVFGA +LSGA (Zhang et al. 2021), MLCA-Net(Cheng et al. 2022), SD-Net (Hoxha and Melgani 2022), PPIC-Net (Hoxha, Scuccato, and Melgani 2023).

Among the aforementioned methods, the CSMLF approach is a retrieval-based technique that utilizes metric learning to acquire semantic embeddings, project image features and sentence representations into the same embedding space, calculate the similarity of the input image and the description statement, and select the nearest neighboring sentence as the description statement for the test image. The Multimodal technique is a typical codec structure that employs a CNN encoder and an LSTM decoder to generate captioning. The FC-ATT/SM-ATT model is based on an attribute mechanism to guide the captioning model to focus on high-level features of images. The SAT(LAM)/adaptive (LAM) approach integrates label information and subsequently uses LSTM to maintain attention mapping for better sentence representation. The Sound-a-a method generates descriptions of remote sensing images by utilizing sound to guide the attention mechanism. The GVFGA+LSGA approach introduces the average pooled features in the encoding segment to guide the entire model. MLCA-Net uses multi-attention to strengthen the spatial and contextual information of an image. Among the compared computer vision domain techniques, Show, Attend and Tell is the first method to incorporate the attention mechanism into the codec network in the decoder to assign varying weights to different feature regions of the input image at different decoding steps, thereby guiding the dynamic concentration of the image region. The Convcap technique is employed to encode sentences and generate descriptive sentences using CNN as decoder.The SD-Net introduces a novel decoder that is based on support vector machines (SVMs). The proposed postprocessing strategies are based on hidden Markov models (HMMs) and Viterbi algorithm are introduced to PPIC-Net to rectify the errors and improve sentences quality.

Tables 5–8 show the performances of the different captioning methods on the four cross-model datasets. In general, the results indicate that MC-Net exhibits competitive results across all four datasets when compared to existing image description generation methods in both remote sensing and computer vision fields. Unlike natural images,remote sensing images possess multiscale characteristics and exhibit background complexity, rendering the direct application of natural domain methods to remote sensing images challenging. Compared with the latest SD-Net and PPIC-Net,

Table 5. Comparative results of MC-Net on the UCM-Captions.

Method	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE_L	CIDEr	SPICE
CSMLF	0.436	0.273	0.186	0.121	0.132	0.393	0.223	0.075
convcap	0.703	0.565	0.462	0.386	0.283	0.596	0.190	0.295
Multimodal	0.709	0.597	0.552	0.460	0.346	0.661	2.925	–
Sound-a-a	0.748	0.684	0.631	0.590	0.362	0.658	2.728	0.391
Attend	0.799	0.736	0.679	0.624	0.417	0.744	3.003	–
FC-ATT	0.810	0.733	0.673	0.619	0.428	0.767	3.370	0.487
SM-ATT	0.812	0.742	0.681	0.630	0.435	0.779	3.386	0.488
SAT(LAM)	0.820	0.776	0.749	0.716	0.484	0.791	3.617	0.502
GVFGA+LSGA	0.832	0.766	0.710	0.660	0.444	0.785	3.327	0.485
MLCA-Net	0.826	0.770	0.717	0.668	0.435	0.772	3.240	0.473
SD-Net	0.765	0.695	0.642	0.594	0.370	0.688	2.923	–
PPIC-Net	0.797	0.730	0.674	0.626	0.408	0.741	3.096	–
MC-Net	**0.845**	**0.784**	**0.732**	**0.679**	**0.449**	**0.786**	**3.355**	**0.520**

Table 6. Comparative results of MC-Net on the Sydney-Captions.

Method	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE_L	CIDEr	SPICE
CSMLF	0.600	0.458	0.387	0.343	0.248	0.502	0.756	0.263
convcap	0.747	0.651	0.573	0.501	0.348	0.667	0.215	0.392
Multimodal	0.697	0.612	0.543	0.504	0.359	0.635	2.202	–
Sound-a-a	0.709	0.623	0.539	0.460	0.312	0.597	1.748	0.384
Attend	0.739	0.640	0.562	0.525	0.349	0.672	2.201	0.395
FC-ATT	0.738	0.644	0.570	0.509	0.364	0.669	2.242	0.395
SM-ATT	0.743	0.634	0.586	0.518	0.364	0.677	2.340	0.398
SAT(LAM)	0.741	0.655	0.590	0.530	0.369	0.681	2.352	0.404
GVFGA+LSGA	0.768	0.685	0.615	0.550	0.387	0.703	2.452	0.453
MLCA-Net	0.831	0.742	0.659	0.580	0.390	0.711	2.324	0.409
SD-Net	0.755	0.671	0.597	0.531	0.364	0.675	2.222	–
PPIC-Net	0.784	0.699	0.632	0.572	0.395	0.711	2.555	–
MC-Net	**0.834**	**0.750**	**0.678**	**0.607**	**0.406**	**0.739**	**2.564**	**0.456**

Table 7. Comparative results of MC-Net on the RSICD.

Method	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE_L	CIDEr	SPICE
CSMLF	0.576	0.386	0.283	0.222	0.213	0.446	0.530	0.199
convcap	0.634	0.510	0.417	0.345	0.333	0.577	1.665	0.393
Multimodal	0.638	0.476	0.400	0.300	0.291	0.533	2.254	–
Sound-a-a	0.620	0.482	0.390	0.320	0.273	0.514	1.639	0.360
Attend	0.671	0.544	0.455	0.387	0.320	0.572	2.249	0.418
FC-ATT	0.667	0.551	0.469	0.406	0.323	0.578	2.576	0.467
SM-ATT	0.670	0.552	0.470	0.407	0.326	0.580	2.574	0.469
SAT(LAM)	0.675	0.554	0.469	0.403	0.325	0.582	2.585	0.464
GVFGA+LSGA	0.678	0.560	0.478	0.417	0.329	0.593	2.601	0.468
MLCA-Net	**0.757**	**0.634**	**0.539**	**0.461**	0.351	**0.646**	2.356	0.444
SD-Net	0.600	0.435	0.336	0.269	0.230	0.456	0.685	–
PPIC-Net	0.629	0.460	0.357	0.287	0.253	0.473	0.752	–
MC-Net	0.728	0.606	0.511	0.433	**0.360**	0.641	**2.454**	**0.463**

Table 8. Comparative results of MC-Net on the NWPU-Captions.

Method	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE_L	CIDEr	SPICE
CSMLF	0.715	0.588	0.502	0.438	0.318	0.576	1.063	0.263
Multimodal	0.634	0.604	0.514	0.458	0.339	0.589	1.078	0.281
Attend	0.734	0.612	0.528	0.469	0.337	0.601	1.109	0.284
FC-ATT	0.738	0.613	0.534	0.471	0.341	0.602	1.135	0.282
SM-ATT	0.739	0.616	0.534	0.469	0.338	0.595	1.137	0.279
MLCA-Net	**0.745**	0.624	0.541	0.478	0.337	0.601	**1.264**	0.285
MC-Net	0.741	**0.626**	**0.544**	**0.478**	**0.347**	**0.611**	1.159	**0.289**

our model has better performance in terms of semantic fluency and syntactic accuracy, for example, the cider evaluation metrics of our MC-Net has improved by 8%. The performance improvement is due to the use of a multi-scale feature extraction module, which can better deal with multi-scale targets. Notably, the GVFGA+LSGA remote sensing image captioning method exhibits the best performance among all the comparison models because of the introduction of global average pooling features as global image information at the encoding side and the proposal of a language state (LS) at the decoding side, reducing the burden of hidden states in providing exclusive text features. In comparison, the MC-Net model displays the highest performance, particularly demonstrating an improvement of 5% in BLEU4, METEOR, and CIDEr evaluation metrics on the Sydney Captions because of its effective extraction of multi-scale features and global contextual features, enabling a deeper comprehension of images. Additionally, the gated adaptive attention mechanism effectively guides sentence generation by considering both image information and sentence context information. On the UCM-Captions dataset, the MC-Net model outperforms the current best-performing model GVFGA+LSGA, exhibiting an improvement of approximately 2% across eight evaluation metrics, and a more significant improvement of approximately 4% in BLEU values and SPICE metrics, as presented in Tables 5–8. On the RSICD dataset, the overall improvement is more significant, especially with respect to the Meteor value, which exhibits an 8% improvement, albeit lower than the GVFGA+LSGA method in the ROUGH and SPICE metrics. This is related to the uneven data categories in the dataset, with relatively small-scale changes and many feature categories exhibiting similar scenes. Nonetheless, overall, the MC-Net model proposed in this paper displays apparent advantages, particularly exhibiting the best results on both the NWPU-Captions dataset, which has twice as many data categories as the UCM-Captions dataset, with richer intra-class diversity and interclass diversity. The CSMLF approach exhibited the lowest performance among all the compared models, indicating that the CNN + LSTM structure is better suited for remote sensing image captioning than the metric learning semantic embedding-based approach. The metric learning-based approach utilizes descriptions of similar images as descriptive information for the query image, resulting in generated statements with limited diversity. Additionally, on the UCM-Captions dataset, the LSTM-based decoder structure significantly outperforms the CNN-based decoder, highlighting the advantage of LSTM in sequence modeling tasks. The introduction of attention in the CNN+LSTM framework leads to a significant improvement in the model results, with models using label information to guide the attention mechanism performing better. This suggests that using only visual features to guide sentence description is often insufficient, and models incorporating attribute and label information can perform better. However, the visual text alignment model proposed in this study, which includes a MS module and a TR module, demonstrates the best results, highlighting its superiority over other models.

The performances of the proposed model, MC-Net, smaller datasets, UCM-Captions (2100 images), and Sydney were compared on four distinct datasets, as depicted in Tables 5–8. The experimental results demonstrate that MC-Net performed better on the captions (613 images), compared to the two larger datasets, RSICD with 10,921 images and the other NWPU-Captions with 31,500 images. This indicates that the performance of image description generation tasks is correlated with the size of the dataset and the scale of the constructed vocabularies, especially given the diverse information and annotations provided by different professionals in the NWPU-Caption dataset. The RSICD dataset also posed a challenge owing to its similar visual categories, such as sparse, medium, and dense residential areas. Nevertheless, our proposed model outperformed the comparison models for all four datasets, demonstrating its generalizability.

3.4.3. The visualization of captioning results

We also present the captioning results of the description sentences using the decoding model based on visual text alignment achieved through gating and adaptive mechanisms. Figure 5 shows examples of the captioning results, wherein the category information is marked in green, relevant proxemics are marked in blue, and incorrect or missing words in the description process are

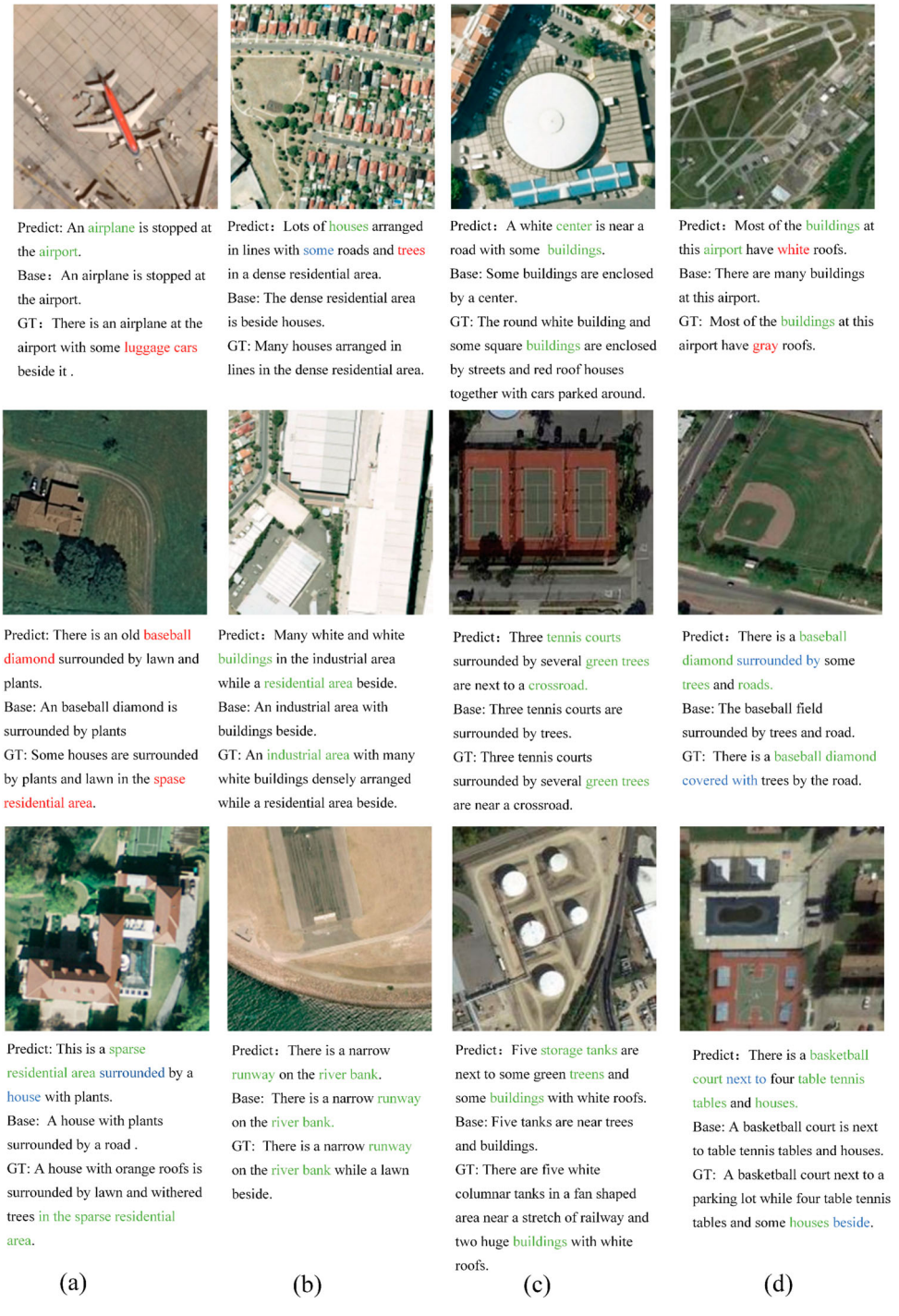

Figure 5. Captions results by MC-Net on four datasets. The first to the fourth columns are selected examples from UCM-Captions, Sydney-Captions, RSICD, and NWPU-Captions, respectively.

marked in red. The proposed MC-Net model performs well on the high-resolution remote sensing image captioning dataset, generating syntactically correct and similar statements to ground-truth sentences.

The model can accurately describe the attribute information of the main objects in the images, such as aircraft, buildings, and tennis courts, and their relationships, while generating a novel vocabulary. Specifically, the narrow runway in Figure 5(b) and white roofs in the subfigure (c) accurately represent the target attribute information. The visualization result in Figure 5(c) can accurately describe the five storage tanks, which indicates that the multi-scale feature extraction module of the model can effectively extract the multi-scale information of the image. In addition, 'near a road', 'next to four table tennis tables' and other relational terms accurately describe the scene information in the image, which indicates that the multi-scale feature fusion module of the model can effectively obtain the global contextual information of the image. The MS effectively extracts multi-scale information, whereas the TR module effectively obtains global contextual information. However, we also notice some challenges, such as the difficulty in detecting small targets like 'cars' in the first row of Figure 5(a). These experiments show that more discriminative image features and effective use of contextual features are crucial for image captioning and scene understanding. Achieving alignment of visual features and textual information is also important for generating grammatically accurate and natural statements with diverse semantics.

3.4.4. Parameters analysis

We have made a parameter sensitivity analysis as shown in Table 9. It shows captioning performance with different H values. H means the number of multi-Head in the GL module. The accuracy of the generated image description sentences varies with the number of heads, and a better experimental result can be achieved when the number of heads is 4, in which the CIDEr reaches 3.355, which is higher than that of LM's evaluation method, indicating that better performance can be achieved using the GL-based multi-scale feature aggregation algorithm.

3.4.5. Speed performance

The time costs of our model is shown in Table 10. To evaluate the efficiency of our method, we, respectively, calculated inference speed (images per second), Memory Access Costs(MACs) and the number of parameters on the UCM-Captions dataset. Comparing the results of the experiments, it can be seen that our method has more MACs and more parameters than the baseline model due to the addition of a multi-scale feature and global modeling module. Considering both time cost and performance factors, our method trades a significant performance gain for a relatively small time cost.

3.4.6. Different scale of contextual information analysis

In this paper, we have analyzed the features of the input multi-scale feature extraction module in the image encoder. In order to analyze the effect of image features of different scales, ablation

Table 9. Captioning performance Comparision with different multi-heads(H) values on the UCM-Captions.

H	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE_L	CIDEr	SPICE
1	0.842	0.774	0.718	0.658	0.425	0.779	3.214	0.496
2	0.843	0.782	0.724	0.672	0.431	0.782	3.336	0.511
4	**0.845**	**0.784**	**0.732**	**0.679**	**0.449**	**0.786**	**0.355**	**0.520**
8	0.844	0.781	0.728	0.675	0.442	0.783	3.351	0.518

Table 10. Comparison of our methods in terms of inference speed (images per second), MACs and parameters. All results are reported based on the UCM-Captions.

Method	Testing Time/Epoch(s)	MACs(G)	Params(M)
Baseline	72.00	20.29	20.91
Baseline+MSF	76.00	20.36	21.54
Bsaeline+MSF+GM(Ours)	80.00	20.40	22.47

Table 11. The comparison results with different scale of contextual information on the UCM-Captions.

Method				Metrics							
1 × 1	3 × 3	5 × 5	7 × 7	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE_L	CIDEr	SPICE
✓				0.806	0.724	0.656	0.592	0.403	0.744	3.009	0.462
✓	✓			0.821	0.746	0.672	0.623	0.421	0.756	3.128	0.485
✓	✓	✓		0.834	0.762	0.684	0.656	0.432	0.768	3.236	0.497
✓	✓	✓	✓	0.845	0.784	0.732	0.679	0.449	0.786	3.355	0.520

experiments consider the combination of different scales (1 × 1, 3 × 3, 5 × 5 and 7 × 7). Table 11 shows the experimental results, it can be seen that divided into the only 1 × 1 scale feature extraction of the experimental results of image captioning lowest, as the scale of fusion increases, gradually obtaining the information of larger sensory field, the better the performance of the generated image description sentences, while the experimental results of fusing four scales at the same time is the best, the difference between the two is about 10%, indicating that different scales image features affect the performance of image captioning.

4. Conclusion

In this study, we have proposed a novel network called MC-Net for remote-sensing image captioning. MC-Net addresses the multiscale and complex background problems in practical applications. The proposed MC-Net consists of a multi-scale feature extraction module and a feature fusion module. Additionally, an adaptive visual-text alignment mechanism is used to generate accurate and fluent descriptive sentences. Ablation experiments confirm the availability of each module. Comparative experiments demonstrate the generalizability and robustness of MC-Net. Despite the superior performance achieved by our proposed model, we believe that the MS and FA modules are not the only way to address the multiscale and background complexity of remote sensing images, and we encourage more approaches to explore. Future research will explore more fine-grained image captioning approaches to improve the captioning results of complex large scenes and extend our research to unsupervised learning to automatically generate fine-grained semantic description utterances for an abundance of unlabeled complex images for better use in real scenes.

Acknowledgments

We thank the anonymous reviewers and the editors for constructive comments that helped improve the manuscript.

Data availability statement

The data that support the findings of this study are available from the corresponding author, upon reasonable request.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work is supported by the National Natural Science Foundation of China under Grant (42090012, 42171434), in part by Guangxi Science and Technology Program under Grant GuiKe 2021AB30019, 03 special research and 5G projects of Jiangxi Province in China under Grant 20212ABC03A0, Zhuhai Industry University Research Cooperation Project of China under Grant ZH22017001210098PWC, Sichuan Science and Technology Program under Grant 2022YFN0031, and Zhizhuo Research Fund on Spatial-Temporal Artificial Intelligence under Grant ZZJJ202202.

ORCID

Haiyan Huang 0000-0002-9931-9884
Zhenfeng Shao 0000-0003-4587-6826
Xiao Huang 0000-0002-4323-382X

References

Anderson, Peter, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. "Spice: Semantic Propositional Image Caption Evaluation." In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part V 14*, 382–398. Springer.

Anderson, Peter, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. "Bottom-up and Top-down Attention for Image Captioning and Visual Question Answering." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6077–6086.

Aneja, Jyoti, Aditya Deshpande, and Alexander G. Schwing. 2018. "Convolutional Image Captioning." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5561–5570.

Banerjee, Satanjeev, and Alon Lavie. 2005. "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments." In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65–72.

Cheng, Qimin, Haiyan Huang, Yuan Xu, Yuzhuo Zhou, Huanying Li, and Zhongyuan Wang. 2022. "NWPU-Captions Dataset and MLCA-Net for Remote Sensing Image Captioning." IEEE Transactions on Geoscience and Remote Sensing 60:1–19.

Cheng, Qimin, Haiyan Huang, Lan Ye, Peng Fu, Deqiao Gan, and Yuzhuo Zhou. 2021. "A Semantic-Preserving Deep Hashing Model for Multi-Label Remote Sensing Image Retrieval." Remote Sensing 13 (24): 4965. https://doi.org/10.3390/rs13244965.

Du, Shouji, Shihong Du, Bo Liu, and Xiuyuan Zhang. 2021. "Incorporating DeepLabv3+ and Object-based Image Analysis for Semantic Segmentation of Very High Resolution Remote Sensing Images." International Journal of Digital Earth 14 (3): 357–378. https://doi.org/10.1080/17538947.2020.1831087.

Farhadi, Ali, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. "Every Picture Tells a Story: Generating Sentences from Images." In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*, 15–29. Springer.

Feng, Jiangfan, Panyu Chen, Zhujun Gu, Maimai Zeng, and Wei Zheng. 2023. "MDSNet: A Multiscale Decoupled Supervision Network for Semantic Segmentation of Remote Sensing Images." International Journal of Digital Earth 16 (1): 2844–2861. https://doi.org/10.1080/17538947.2023.2241435.

Fu, Kun, Yang Li, Wenkai Zhang, Hongfeng Yu, and Xian Sun. 2020. "Boosting Memory with a Persistent Memory Mechanism for Remote Sensing Image Captioning." Remote Sensing 12 (11): 1874. https://doi.org/10.3390/rs12111874.

Hoxha, Genc, and Farid Melgani. 2020. "Remote Sensing Image Captioning with SVM-based Decoding." In *IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium*, 6734–6737. IEEE.

Hoxha, Genc, and Farid Melgani. 2022. "A Novel SVM-Based Decoder for Remote Sensing Image Captioning." IEEE Transactions on Geoscience and Remote Sensing 60:1–14.

Hoxha, Genc, Farid Melgani, and Begüm Demir. 2020. "Toward Remote Sensing Image Retrieval Under a Deep Image Captioning Perspective." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 13:4462–4475. https://doi.org/10.1109/JSTARS.4609443.

Hoxha, Genc, Giacomo Scuccato, and Farid Melgani. 2023. "Improving Image Captioning Systems With Postprocessing Strategies." IEEE Transactions on Geoscience and Remote Sensing 61:1–13. https://doi.org/10.1109/TGRS.2023.3281334.

Huang, Wei, Qi Wang, and Xuelong Li. 2020. "Denoising-Based Multiscale Feature Fusion for Remote Sensing Image Captioning." IEEE Geoscience and Remote Sensing Letters 18 (3): 436–440. https://doi.org/10.1109/LGRS.8859.

Kulkarni, Girish, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. 2013. "Babytalk: Understanding and Generating Simple Image Descriptions." IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (12): 2891–2903. https://doi.org/10.1109/TPAMI.2012.162.

Li, Yunpeng, Xiangrong Zhang, Jing Gu, Chen Li, Xin Wang, Xu Tang, and Licheng Jiao. 2021. "Recurrent Attention and Semantic Gate for Remote Sensing Image Captioning." IEEE Transactions on Geoscience and Remote Sensing 60:1–16.

Li, Xuelong, Xueting Zhang, Wei Huang, and Qi Wang. 2020. "Truncation Cross Entropy Loss for Remote Sensing Image Captioning." IEEE Transactions on Geoscience and Remote Sensing 59 (6): 5246–5257. https://doi.org/10.1109/TGRS.2020.3010106.

Liu, Qingrong, Chengqing Ruan, Shan Zhong, Jian Li, Zhonghui Yin, and Xihu Lian. 2018. "Risk Assessment of Storm Surge Disaster Based on Numerical Models and Remote Sensing." International Journal of Applied Earth Observation and Geoinformation 68:20–30. https://doi.org/10.1016/j.jag.2018.01.016.

Lu, Xiaoqiang, Binqiang Wang, and Xiangtao Zheng. 2019. "Sound Active Attention Framework for Remote Sensing Image Captioning." IEEE Transactions on Geoscience and Remote Sensing 58 (3): 1985–2000. https://doi.org/10.1109/TGRS.36.

Lu, Xiaoqiang, Binqiang Wang, Xiangtao Zheng, and Xuelong Li. 2017. "Exploring Models and Data for Remote Sensing Image Caption Generation." IEEE Transactions on Geoscience and Remote Sensing 56 (4): 2183–2195. https://doi.org/10.1109/TGRS.2017.2776321.

Lu, Jiasen, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. "Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 375–383.

Ordonez, Vicente, Girish Kulkarni, and Tamara Berg. 2011. "Im2text: Describing Images Using 1 Million Captioned Photographs." In *Advances in Neural Information Processing Systems*, 24.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. "Bleu: A Method for Automatic Evaluation of Machine Translation." In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318.

Qu, Bo, Xuelong Li, Dacheng Tao, and Xiaoqiang Lu. 2016. "Deep Semantic Understanding of High Resolution Remote Sensing Image." In *2016 International Conference on Computer, Information and Telecommunication Systems (CITS)*, 1–5. IEEE.

Rouge, Lin C. Y. 2004. "A Package for Automatic Evaluation of Summaries." In *Proceedings of Workshop on Text Summarization of ACL, Spain*, Vol. 5.

Shi, Zhenwei, and Zhengxia Zou. 2017. "Can a Machine Generate Humanlike Language Descriptions for a Remote Sensing Image?." IEEE Transactions on Geoscience and Remote Sensing 55 (6): 3623–3634. https://doi.org/10.1109/TGRS.2017.2677464.

Sumbul, Gencer, Sonali Nayak, and Begüm Demir. 2020. "SD-RSIC: Summarization-Driven Deep Remote Sensing Image Captioning." IEEE Transactions on Geoscience and Remote Sensing 59 (8): 6922–6934. https://doi.org/10.1109/TGRS.2020.3031111.

Ushiku, Yoshitaka, Masataka Yamaguchi, Yusuke Mukuta, and Tatsuya Harada. 2015. "Common Subspace for Model and Similarity: Phrase Learning for Caption Generation from Images." In *Proceedings of the IEEE International Conference on Computer Vision*, 2668–2676.

Vedantam, Ramakrishna, C. Lawrence Zitnick, and Devi Parikh. 2015. "Cider: Consensus-based Image Description Evaluation." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4566–4575.

Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. "Show and Tell: A Neural Image Caption Generator." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3156–3164.

Wang, Qi, Wei Huang, Xueting Zhang, and Xuelong Li. 2020. "Word–Sentence Framework for Remote Sensing Image Captioning." IEEE Transactions on Geoscience and Remote Sensing 59 (12): 10532–10543. https://doi.org/10.1109/TGRS.2020.3044054.

Wang, Binqiang, Xiaoqiang Lu, Xiangtao Zheng, and Xuelong Li. 2019. "Semantic Descriptions of High-Resolution Remote Sensing Images." IEEE Geoscience and Remote Sensing Letters 16 (8): 1274–1278. https://doi.org/10.1109/LGRS.8859.

Wang, Shuang, Xiutiao Ye, Yu Gu, Jihui Wang, Yun Meng, Jingxian Tian, Biao Hou, and Licheng Jiao. 2022. "Multi-Label Semantic Feature Fusion for Remote Sensing Image Captioning." ISPRS Journal of Photogrammetry and Remote Sensing 184:1–18. https://doi.org/10.1016/j.isprsjprs.2021.11.020.

Wang, Yong, Wenkai Zhang, Zhengyuan Zhang, Xin Gao, and Xian Sun. 2022. "Multiscale Multiinteraction Network for Remote Sensing Image Captioning." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 15:2154–2165. https://doi.org/10.1109/JSTARS.2022.3153636.

Wang, Binqiang, Xiangtao Zheng, Bo Qu, and Xiaoqiang Lu. 2020. "Retrieval Topic Recurrent Memory Network for Remote Sensing Image Captioning." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 13:256–270. https://doi.org/10.1109/JSTARS.4609443.

Xu, Kelvin, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention." In *International Conference on Machine Learning*, 2048–2057. PMLR.

Yang, Yi, and Shawn Newsam. 2010. "Bag-of-Visual-Words and Spatial Extensions for Land-use Classification." In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 270–279.

Yang, Qiaoqiao, Zihao Ni, and Peng Ren. 2022. "Meta Captioning: A Meta Learning Based Remote Sensing Image Captioning Framework." ISPRS Journal of Photogrammetry and Remote Sensing186:190–200. https://doi.org/10.1016/j.isprsjprs.2022.02.001.

Yuan, Zhenghang, Xuelong Li, and Qi Wang. 2019. "Exploring Multi-Level Attention and Semantic Relationship for Remote Sensing Image Captioning." IEEE Access 8:2608–2620. https://doi.org/10.1109/Access.6287639.

Zhang, Zhengyuan, Wenhui Diao, Wenkai Zhang, Menglong Yan, Xin Gao, and Xian Sun. 2019. "LAM: Remote Sensing Image Captioning with Label-Attention Mechanism." Remote Sensing 11 (20): 2349. https://doi.org/10.3390/rs11202349.

Zhang, Xueting, Qi Wang, Shangdong Chen, and Xuelong Li. 2019. "Multi-scale Cropping Mechanism for Remote Sensing Image Captioning." In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, 10039–10042. IEEE.

Zhang, Xiangrong, Xin Wang, Xu Tang, Huiyu Zhou, and Chen Li. 2019. "Description Generation for Remote Sensing Images Using Attribute Attention Mechanism." Remote Sensing 11 (6): 612. https://doi.org/10.3390/rs11060612.

Zhang, Lu, Jianming Zhang, Zhe Lin, Huchuan Lu, and You He. 2019. "Capsal: Leveraging Captioning to Boost Semantics for Salient Object Detection." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6024–6033.

Zhang, Zhengyuan, Wenkai Zhang, Menglong Yan, Xin Gao, Kun Fu, and Xian Sun. 2021. "Global Visual Feature and Linguistic State Guided Attention for Remote Sensing Image Captioning." IEEE Transactions on Geoscience and Remote Sensing 60:1–16. https://doi.org/10.1109/TGRS.2020.3040221.