remote sensing

MDPI

Article

A Mask-Guided Transformer Network with Topic Token for Remote Sensing Image Captioning

Zihao Ren [1], Shuiping Gou [1], Zhang Guo [2,*], Shasha Mao [1] and Ruimin Li [2]

[1] Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, School of Artificial Intelligence, Xidian University, Xi'an 710071, China; zhren_1@stu.xidian.edu.cn (Z.R.); shpgou@mail.xidian.edu.cn (S.G.); ssmao@xidian.edu.cn (S.M.)

[2] Academy of Advanced Interdisciplinary Research, Xidian University, Xi'an 710071, China; rmli@xidian.edu.cn

* Correspondence: guozhang@xidian.edu.cn

Abstract: Remote sensing image captioning aims to describe the content of images using natural language. In contrast with natural images, the scale, distribution, and number of objects generally vary in remote sensing images, making it hard to capture global semantic information and the relationships between objects at different scales. In this paper, in order to improve the accuracy and diversity of captioning, a mask-guided Transformer network with a topic token is proposed. Multi-head attention is introduced to extract features and capture the relationships between objects. On this basis, a topic token is added into the encoder, which represents the scene topic and serves as a prior in the decoder to help us focus better on global semantic information. Moreover, a new Mask-Cross-Entropy strategy is designed in order to improve the diversity of the generated captions, which randomly replaces some input words with a special word (named [Mask]) in the training stage, with the aim of enhancing the model's learning ability and forcing exploration of uncommon word relations. Experiments on three data sets show that the proposed method can generate captions with high accuracy and diversity, and the experimental results illustrate that the proposed method can outperform state-of-the-art models. Furthermore, the CIDEr score on the RSICD data set increased from 275.49 to 298.39.

Keywords: image captioning; remote sensing; transformer network; global semantic information; topic token

check for updates

Citation: Ren, Z.; Gou, S.; Guo, Z; Mao, S.; Li, R. A Mask-Guided Transformer Network with Topic Token for Remote Sensing Image Captioning. *Remote Sens.* **2022**, *14*, 2939. https://doi.org/10.3390/rs14122939

Academic Editors: Miltiadis D. Lytras and Andreea Claudia Serban

Received: 5 April 2022
Accepted: 16 June 2022
Published: 20 June 2022

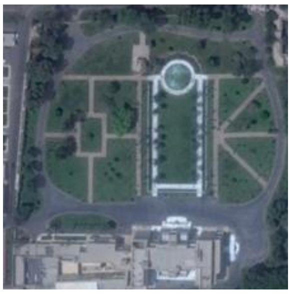
Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

With the popularity of remote sensing technology, applications based on remote sensing images have attracted widespread attention, including target detection [1], classification [2,3], and segmentation tasks [4,5]. As demand continues to expand, the sole use of such applications is inadequate to express the content and semantic information in the image [6]. Remote sensing image captioning aims to describe abundant image content using natural language [6]; compared with image target detection, classification, and segmentation tasks, it is more complex and challenging, as it requires representing the semantic relationships among objects and generating an exact and descriptive sentence [6,7]. Remote sensing image captioning is vital for numerous application fields, such as geographic information retrieval [8], disaster monitoring [9], and image understanding [6].

Remote sensing image captioning is an important branch of image captioning, and the most popular methods of image captioning mainly fall into three general categories: retrieval-based [10,11], detection-based [12,13], and encoder–decoder-based methods [14]. Methods based on retrieval require a given corresponding image-caption data set, such that captions can be directly generated by matching with a corresponding image. These methods demand large amounts of data and the generated captions are often quite limited. Detection-based methods mainly detect fixed objects in an image and then associate them through pre-set sentence patterns. The use of a fixed statement pattern makes the method unable to express the relationships among objects. Encoder–decoder methods are the mainstream at present. The encoder is used to extract image features and map them to the corresponding word embedding space, following which the decoder generates sentences according to the encoded image features. The most common encoder–decoder framework combines CNN (convolutional neural networks) and LSTM (long short-term memory) networks.

In remote sensing images, the scale, distribution, and number of objects often vary [6]. To further illustrate the difficulties in remote sensing image captioning, we chose an example in the RSICD data set, as shown in Figure 1. Variously sized objects with scattered distribution can be observed. If the bounding boxes of Faster-R-CNN [15] (Faster region convolutional neural networks) are used to represent the object features, they will overlap each other to a large extent and will contain a great deal of background noise. In this situation, most researchers can only use the output of a pre-trained CNN (e.g., ResNet [16], VGG [17], and so on) to extract image features, which may result in a loss of region features. In addition, it can be found in the descriptions that high-level semantic words, such as "square", are necessary to describe the image content. If only utilizing the image region or object features, we can only obtain low-level semantic words, such as "pond", "trees", and "buildings". It can be shown that high-level semantic words mainly rely on the global relationships among objects, but not local image features. Such global relations can be interpreted as global semantic information. In most CNN-based methods, the mean pooling operation is applied to capture the global semantic information, but this may lead to neglecting the relationships between objects and scenes, causing key information loss and influencing the high-level semantic words; therefore, CNN using mean pooling are inadequate when it comes to capturing the global semantic information required for remote sensing image captioning.



(1) the square with a circular pond and lawn is surrounded by some trees .

(2) surrounded by some trees there are a round pond and lawns on the square .

(3) the square with a circle pond and lawn is surrounded by some trees .

(4) this square is surrounded by rows of trees and several buildings .

(5) a square covered by green meadows is near several buildings and some green trees .

Figure 1. An example in the RSICD data set. An image is shown on the left and its corresponding descriptions are on the right.

With the continuous performance advancement in image captioning, improving the diversity of captioning has become another research tendency. In general, the training strategy for image captioning is divided into two stages: Cross-Entropy (CE) training in the beginning and self-critical (SC) training [18] for fine-tuning. The CE training stage calculates the Cross-Entropy loss function at the word level, while the SC training stage regards the sentence generation process as a sequential decision problem and fine-tunes the model through reinforcement learning [19]. It has been shown that SC training is effective for improving accuracy [18] and suppressing train–test condition bias (or *exposure bias*) [20] in the CE training stage; however, the models obtained via SC training tend to generate templated sentences, showing reduced sentence diversity. For remote sensing image captioning, the description is relatively simpler than in natural image description. As such, the lack of diversity in the generated sentences is more obvious, and so, improving the diversity of remote sensing description is vital. However, this aspect has been rarely explored in existing studies.

Hence, in order to handle the problems mentioned above, a new method—named mask-guided Transformer network with topic token—of remote sensing image captioning is proposed in this paper. It aims to generate captions with high accuracy and diversity, which has been rarely explored in the past. In the proposed method, we propose a topic encoder, which is composed of a topic token and a Transformer(we use term "Transformer" as a substitution of "Transformer network") encoder. The topic token is used to obtain the global semantic information and the multi-head attention mechanism in the Transformer is used to extract region and object features from the images, which is more suitable for remote sensing images with multi-scale objects and complex scenes. For the decoder, we propose a semantic-guided decoder, in which the global semantic information captured by the topic token is applied to help guide the generation of high-level semantic words.

In addition, a new training strategy to improve the diversity of generated sentences is proposed, named Mask–Cross-Entropy (Mask-CE). With the Mask-CE strategy, some words in the sequence are randomly masked in each training iteration, and the next word is predicted based on fewer ground-truth words, which effectively suppresses the *exposure bias* and enhances the robustness of the method. Furthermore, the random mask operation can boost the diversity of the training sequence context, which is useful for preventing over-fitting and enhancing the learning ability. Experiments on three remote sensing data sets show that this training strategy can outperform the evaluation metric score of CE with SC while maintaining diversity at the same time.

The key contributions of this paper are as follows:

- We propose a full Transformer network to improve the accuracy of generated captioning, which can better represent the remote sensing features of large-scale and crowded objects in contrast with the common CNN and LSTM framework;
- To better express the global semantic information in the image, a topic token is proposed, which can interact with the image features and can be directly used in the decoder to guide sentence generation;
- A new Mask–Cross-Entropy training strategy is proposed in order to reduce the model training time, and enables the generated descriptions to have high accuracy and diversity.

The remainder of this article is organized as follows: In Section 2, some previous representative related studies are introduced. Next, the overall structure and details of the mask-guided Transformer with a topic token method are described in Section 3. Then, the experiment and analysis on three data sets are detailed in Section 4. Finally, we summarize our conclusions and directions for future work in Section 5.

2. Related Work

In this section, we review some existing image captioning works. First, we introduce some representative works in natural image captioning, which is the main research aspect in the image captioning field. Then, some remote sensing image captioning works corresponding to the characteristics of remote sensing images are discussed. Finally, some training strategies to improve the description diversity are also mentioned.

2.1. Natural Image Captioning

Encoder–decoder is the main framework used in natural image captioning. NIC (neural image caption) [14] is a typical encoder–decoder network, which uses CNN to extract the global image features and feeds them to a long short-term memory (LSTM) network [21] in the first time step in order to avoid over fitting. However, the image features are not re-used in the later time step, leading to insufficient utilization of image features. To address this drawback, the Bottom-Up and Top-Down attention mechanism network [22] utilizes the mean values of Faster-RCNN [15] object features to obtain region features. In order to fully exploit the image features, the decoder is composed of two LSTM: at each time step, the first LSTM uses the previous hidden state and the global mean value of all region features to update the hidden value, while the second LSTM uses the weighted average value of region features based on the hidden state to generate the next word. In the Bottom-Up and Top-Down attention mechanism network, the input encoded image features in the decoder vary with time, and can be used in each time step.

To enhance the interactions between sentences and images, an attention on attention (AOA) module has been proposed in AoANet [23]. After an attention mechanism module, the AOA module first generates an "information vector" and "attention gate" using the attention value and the query context vector, respectively, then utilizes them to perform another attention operation to obtain the final output, which can reduce useless information in the model. The AOA network applies two attention modules to get better results, which also shows the strong learning ability associated with attention mechanisms.

The Transformer model [24] has made great breakthroughs in natural language processing through the use of a multi-head attention mechanism. Considering the ability of Transformer in processing sequences, researchers have attempted to use Transformer as an image captioning decoder. A successful example of replacing the LSTM decoder with Transformer is the Meshed Memory Transformer model [25], which improves both the image encoding and the language generation steps by utilizing the multi-level relationships in Transformer. Additionally, the multi-head attention can also be used to explore cross-modal relations. Object Relation Transformer [26] applies multi-head attention to exploring the relation between the object features and corresponding geometric features; the generated description can represent the relative positions of objects more accurately.

2.2. Remote Sensing Image Captioning

Multi-modal encoder–decoder [6] was the first remote sensing image captioning model, which is based on the NIC model [14]. To solve the data shortage associated with remote sensing image captioning tasks, Qu et al. [6] constructed the UCM (UC Merced) and Sydney data sets. To further enrich the data sets, Lu et al. [27] have annotated the RSICD data set. Compared with UCM and Sydney, the RSICD data set contains more images with abundant descriptions.

Due to the characteristics of remote sensing images, Faster-R-CNN [15] cannot perform well for remote sensing object feature extraction. In order to better extract the features of variously scaled objects in the image, structured attention [28] has been proposed to obtain the image segmentation proposal features. It applies a selective search to find the segmentation proposals and multiplies them with CNN features to obtain the structured features. The structured features can reflect the local region features well; however, the selective search does not work well when there are numerous objects in the image.

The size of objects in remote sensing image varies, which can lead to omissions during feature extraction. To address this problem, the denoising-based multi-scale feature fusion network [29] first filters noise in the image with two fully connected layers, following which the encoder output is obtained by fusing the outputs of CNN features at three scales. The Multi-Level Attention Model [30] uses three attention blocks to represent attention to different areas of the image, attention to different words, and attention to vision and semantics. The above two models take a multi-scale method as the entry point to extract different object features, but ignore the high semantic information.

Considering the large-scale variation and richness of objects in remote sensing images, researchers have attempted to extract global semantic information to facilitate word generation. The mean pooling operation is a common way to capture such information [6,27]. VRTMM [31] (Variational autoencoder and reinforcement-learning-based two-stage multitask learning model) uses the output of VGG16 [17] (visual geometry group 16) with a soft-max layer to represent the semantic features. RASG [32] (recurrent attention and semantic gate) utilizes a recurrent attention mechanism and semantic gate to generate better image features corresponding to the current word.

Multi-head attention has proven to be effective in extracting the complex relationships between different objects [24], which fits well with the characteristics of remote sensing images. Inspired by this, in our study, we explore the use of the Transformer encoder to extract the image features and a new way to capture global semantic information.

Furthermore, as the objects in remote sensing images are complex and abundant, common descriptions with high accuracy cannot represent the content well. Therefore, improving the diversity in remote sensing image captioning is necessary, but the above studies have not explored this aspect deeply.

2.3. Training Strategy and Improve Diversity

The use of Cross-Entropy (CE) and Self-Critical (SC) training is an effective training strategy to enhance the accuracy performance. In the initial training stage, the Cross-Entropy loss is used to pre-train the decoder at the word level. One limitation of CE is that, at each sampling time step, the previous words are the ground truth, but not the previously generated words. However, in the evaluation and test stage, the current predicted word only relies on the pre-predicted words, which may be different from the ground truth; this discrepancy is called *exposure bias* [20].

To solve this problem, a self-critical training strategy has been proposed [18]. In the SC strategy, the previous words are the pre-predicted words, which is the same as in the actual inference stage. The SC uses evaluation metrics such as BLEU [33] and CIDEr [34] as the reward. In this training strategy, the model tends to generate sentences with higher rewards, so the model often has a higher accuracy, but this will also result in a loss of sentence diversity.

To improve the diversity of captioning, Dai et al. [35] and Wang et al. [36] used a GAN (Generative adversarial network) [37] and a VAE (Variational Auto Encoder) [38], respectively; however, the two methods will result in accuracy losses to a certain extent. SLL-SLE [39] (sequence-level learning via sequence-level exploration)adds diversity evaluations in the optimization target to promote diversity. Wang et al. [40] have proposed the Self-CIDEr metric to evaluate the diversity in captioning and use it as the reward score in reinforcement learning. To combine the accuracy and diversity scores, a partial off-policy learning scheme [41] has been proposed. These latter three methods mainly conduct fine-tuning in the reinforcement learning stage, while our Mask–Cross-Entropy (Mask-CE) training strategy mainly explores optimizing the first Cross-Entropy loss function.

3. Method

In this section, we introduce the mask-guided Transformer with a topic token for remote sensing image captioning in detail. Its overall framework is shown in Figure 2, where N_e and N_d are the number of encoder and decoder layers. The proposed method consists of three main parts: a topic encoder to extract image features with global semantic information; a semantic-guided decoder for word generation; and the Mask–Cross-Entropy training strategy, which improves the diversity. More details are given in the following sections.

3.1. Topic Encoder

Faster-R-CNN has been widely used as a feature encoder in natural image captioning, but the extracted features from Transformer have been rarely explored. Further, the features obtained from Faster-R-CNN do not fit remote sensing images well, and we believe that the multi-head attention in Transformer may be a better fit with the characteristics of remote sensing images, as it can explore the relationships among different image regions. Based on this, we explore the use of Transformer to build the feature encoder for our remote sensing image captioning model.

Consider an input image of size $h \times w \times 3$. As the Transformer model takes the sequence as input, it is necessary to convert each image into a sequence composed of $N = \frac{h}{s} \times \frac{w}{s}$ patches, where s is the size of the single patch. Note that s was set as 16 in our experiments. Then, the patches are input to the embedding layer in order to map the features to the same space as the word embedding.

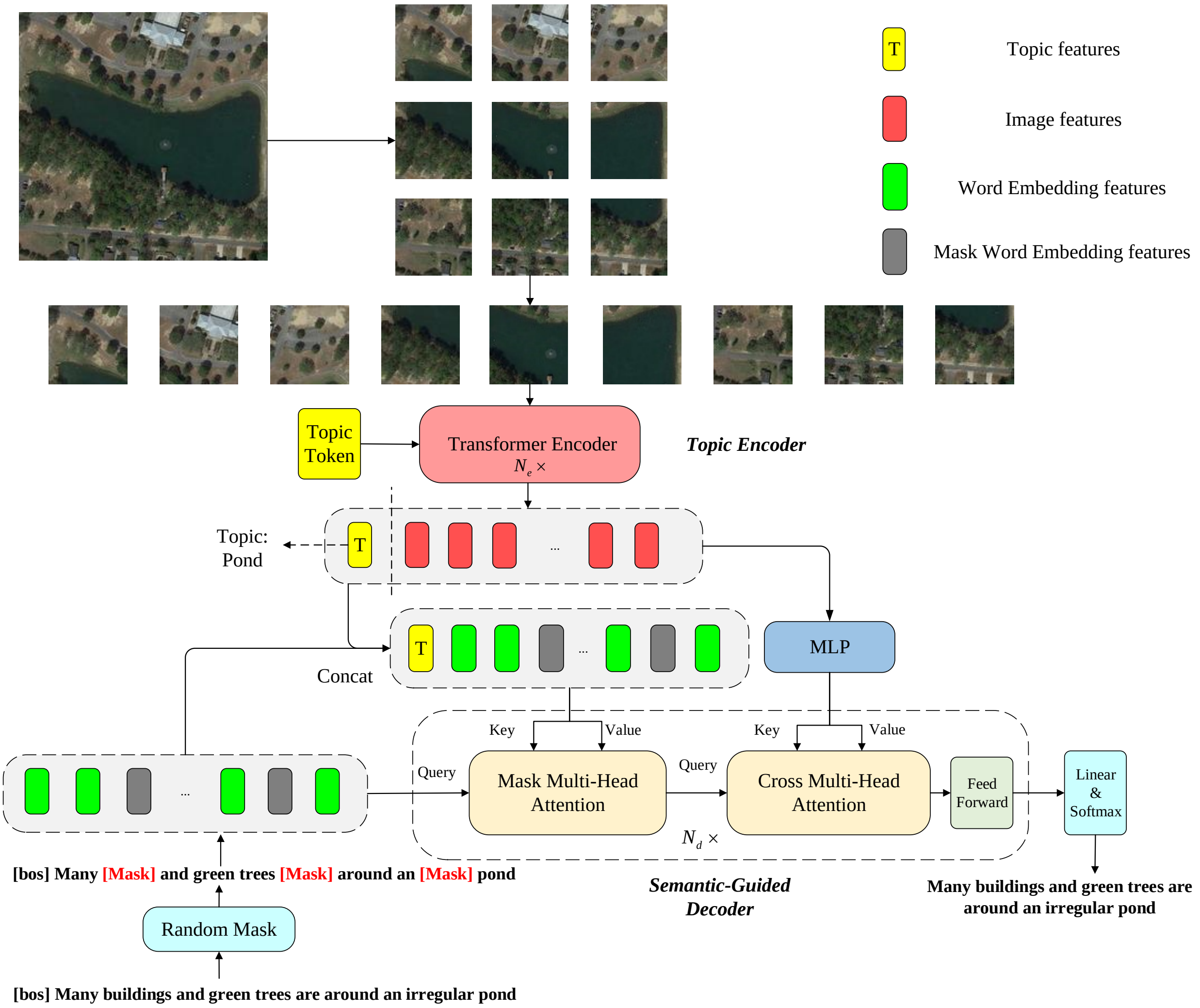

T
Topic features
Image features
Word Embedding features
Mask Word Embedding features
Topic Token
Transformer Encoder
$N_e \times$
Topic Encoder
Topic: Pond
T
...
Concat
MLP
Key
Value
Query
Mask Multi-Head Attention
Cross Multi-Head Attention
Feed Forward
Linear & Softmax
$N_d \times$
Semantic-Guided Decoder
[bos] Many [Mask] and green trees [Mask] around an [Mask] pond
Random Mask
[bos] Many buildings and green trees are around an irregular pond
Many buildings and green trees are around an irregular pond

Figure 2. The structure of proposed mask-guided Transformer with topic token. For the sake of clarity, the AddNorm and embedding operations are not shown.

For remote sensing image captioning, the captured global semantic information is crucial, as it is used to represent the complex relationships among objects and the whole scene. In Figure 3, we show some images which illustrate the importance of global semantic information. In the figure, some different high-level semantic words (also called topics), such as the "industrial area", the "commercial area", and the "residential area", have corresponding images which are all composed of many buildings with similar local characteristics, but different in the arrangement of buildings and the surrounding scene. Therefore, it is observed that the topics expressing global semantic information can be obtained from the large-scale scene, but not the local patch features.

To extract the global semantic information, we propose a topic token module ($[T]$). It is similar to [Class] token in Bert [42] (Bidirectional encoder representations from Transformers) and Vit (Vision Transformer) [43], the difference being that $[T]$ is not only used for classification, but also guides word generation in the decoding stage. Compared with the mean pooling operation on image features [6,27] or the classification result of the softmax layer [31], $[T]$ is a trainable parameter and the self-attention calculates its attention value with each image path feature. Thus, its output V_T can better express global semantic information.

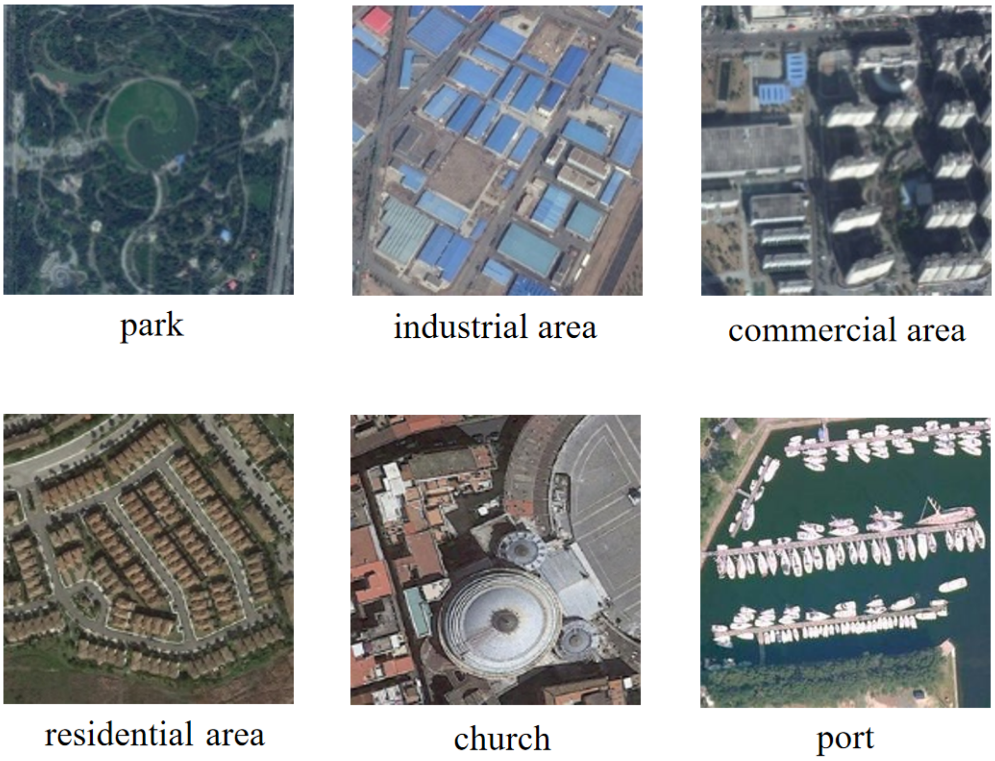


Figure 3. Some remote sensing images and corresponding topics.

After adding the topic token in image patch features, it is necessary to add position encoding to represent the sequence order. Thus, the input of the encoder can be written as:

$$V_1 = [T, p_1, p_2, \ldots, p_n, \ldots, p_N] + p_{pos}, \tag{1}$$

where T is the topic token, p_n is the nth image patch, N is the number of image patches, p_N is the last image patch, and p_{pos} is the position coding.

The structure of the topic encoder is shown in Figure 4. The main part of the topic encoder is the multi-head attention layer, where the operation in this layer is calculated as:

$$\begin{aligned} \mathrm{MHA}(Q, K, V) &= \mathrm{Concat}(\mathrm{head}_1, \mathrm{head}_2, \ldots, \mathrm{head}_H)W \\ \mathrm{head_i} &= \mathrm{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right), \\ \mathrm{Attention}(Q, K, V) &= \mathrm{soft\,max}\left(\frac{QK^T}{\sqrt{d}}\right)V, \end{aligned} \tag{2}$$

where MHA denotes the multi-head attention; H is the number of attention heads; $\mathrm{head_i}$ is the ith attention head; W is the weight vector in the multi-head attention; Q, K, V are the query, key, and value vectors, respectively; W_i^Q, W_i^K, W_i^V are trainable parameters; and d is the scale factor.

In the final encoder, the output is separated into two parts, which can be expressed as:

$$\begin{aligned} V_T &= V_{N_e}[1], \\ V_I &= V_{N_e}[2, \ldots, N+1], \end{aligned} \tag{3}$$

where V_T is the first element in V_{N_e} and represents the topic feature, while V_I is the rest of V_{N_e} and represents the image feature, N is the number of image patches, and V_{N_e} is the output of topic encoder.

As topics are based on the large-scale scene and the whole caption, it is difficult to directly train the topic token with image captioning data sets. Hence, in order to train the whole network faster and better, a remote sensing topic classification task is used to pre-train the encoder and the topic token. The first part involves setting topic labels, where annotators provide a class name set for the whole data set for further usage. In this sense, RSICD [27] provides 30 category names, UCM [6] provides 21 names, and Sydney [6] provides 7 names (e.g., airport, parking lot, dense residential area). In the RSICD data set, there is a category name corresponding to each image, which we directly set as the topic label. For the UCM and Sydney data sets, we use the following strategy to find the topic label corresponding to the image: if a class name appears in more than three captions of an image, it is set as the topic label of the image. An example of this strategy is shown in Figure 5. In this way, a remote sensing topic classification data set was built. In the pre-training classification task, to ensure that the topic token can extract the global semantic information, only the output feature of the topic token is used as input to the classification layer. Then, the pre-trained weights are used to initialize the encoder and topic token.

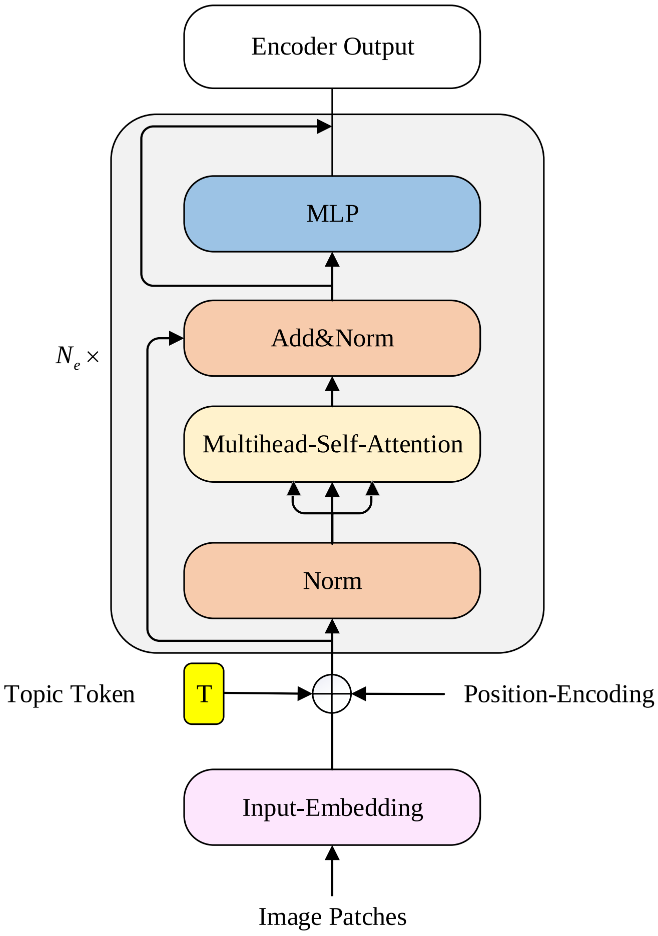
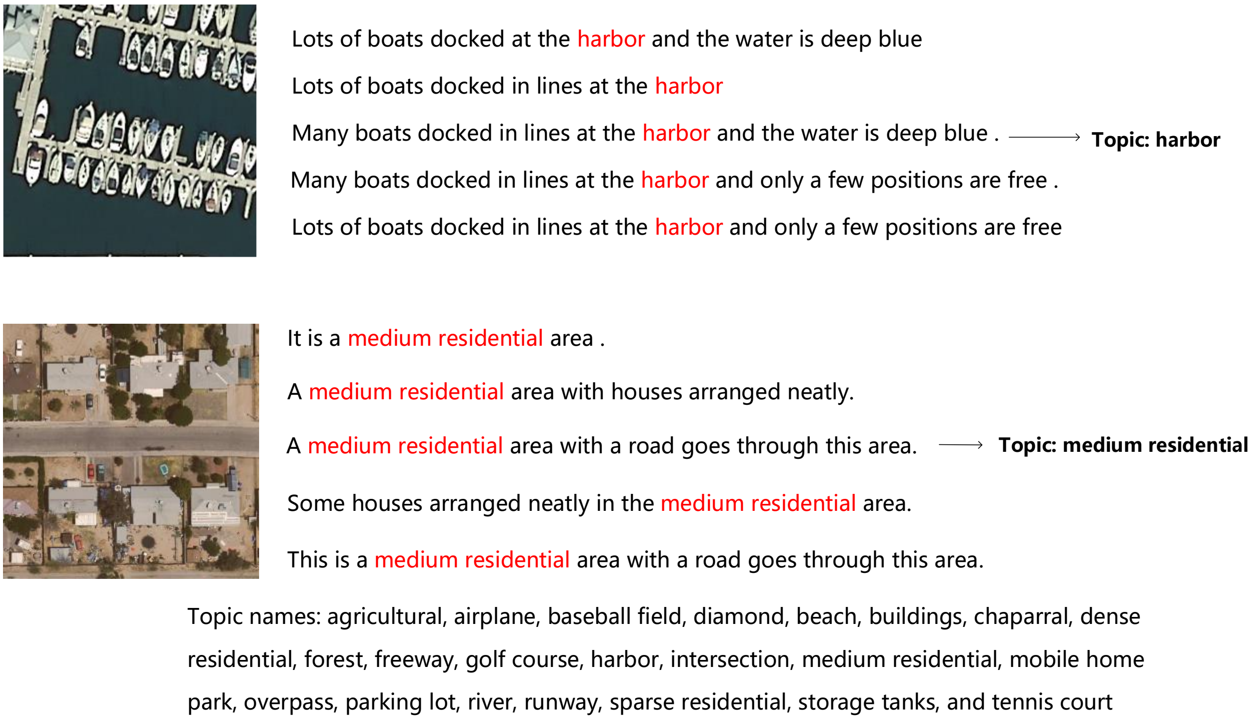


Figure 4. The structure of topic encoder.

After pre-training the topic encoder, considering the modality gap between the image and the sentence, it is necessary to fine-tune the encoder for the image captioning task. However, we find that the accuracy is decreased if the encoder and decoder are fine tuned together. To solve this issue, we add an extra MLP (Multilayer Perceptron) layer [44] after the encoder, freeze the encoder, and only train the extra MLP while fine-tuning. The extra MLP is a kind of embedding layer, which maps the image features to a deeper space so it can explore potential semantic associations between images and sentences.

To better extract image features, we utilize the multi-head attention in Transformer, which fits the characteristic of large-scale and numerous objects in remote sensing images. We also propose a topic token to extract the global semantic information, which is a trainable and independent vector and can interact with each image path feature.



Lots of boats docked at the harbor and the water is deep blue

Lots of boats docked in lines at the harbor

Many boats docked in lines at the harbor and the water is deep blue . ⟶ **Topic: harbor**

Many boats docked in lines at the harbor and only a few positions are free .

Lots of boats docked in lines at the harbor and only a few positions are free

It is a medium residential area .

A medium residential area with houses arranged neatly.

A medium residential area with a road goes through this area. ⟶ **Topic: medium residential**

Some houses arranged neatly in the medium residential area.

This is a medium residential area with a road goes through this area.

Topic names: agricultural, airplane, baseball field, diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks, and tennis court

Figure 5. Two examples of extracting topics from data sets. The total topic names are also shown.

3.2. Semantic-Guided Decoder

The semantic-guided decoder is mainly composed of two multi-head attention blocks. Section 3.1 showed that using only local patch features cannot represent the global semantic information in remote sensing images. To generate better captions, we designed the semantic-guided attention and cross-attention modules. The two parts of the output in the encoder (in Equation (3)) are used in the two attention modules, respectively. The semantic-guided decoder generates the next word token, based on the two parts above and the pre-predicted words.

In the semantic-guided attention module, as the new words should only rely on the previously predicted words, the mask attention is applied; that is, the attention weights of the words not generated should be quite small, and the new words are not influenced by them.

The difference of the semantic-guided attention module, compared to the conventional mask self-attention module, is that we add the topic semantic features V_T in the key and value vectors. While using the Vit features as the backbone feature extractors to carry out downstream tasks, the two parts in Equation (3) are not separated, as we mainly use the image features. However, in remote sensing image captioning, the global semantic information is also necessary. Based on this, considering that the number of topic features is only 1 and there are N local patch features, the important topic features will be hard to notice if not separated from patch features. Additionally, the topic features are high-level semantic word-embedding vectors, containing the information in word space, so we incorporate the topic features into the mask multi-head attention module.

Specifically, V_T is concatenated with the word embedding vector to form the key and value vector. When each word embedding vector performs the mask attention operation, the extra V_T will not be influenced and will be noticed at each time step, which is helpful in generating words specific to the global semantic information. After adding V_T, the new semantic-guided attention block can be expressed as

$$\begin{aligned} &X = \mathrm{AddNorm}(\mathrm{MaskMHA}(Q, K, V)),\\ &Q = X,\\ &K = V = [V_T, X], \end{aligned} \tag{4}$$

where X represents the output of the previous decoder (corresponding to the sentence word-embedding in the first decoder layer) and $[,]$ denotes the concatenate operation. Compare with the conventional self-attention block (whose Q, K, V have the same value and are all equal to X), extra global semantic information is added in Equation (4).

V_T is taken as prior knowledge and is a high-level semantic word-embedding vector based on the large scene. On this basis, the sequence can be used to obtain the global semantic information early and correctly, rather than mining in the later large-scale image patch features. For example, in two similar scenes, such as the 'dense residential area' and 'commercial area' shown in Figure 3, where there are many buildings in the two images, the topic token can help to generate some high-semantic-level words, such as 'dense residential' and 'commercial.' In the model without the topic token, the model needs to excavate the inner information, and the final result may not be accurate.

The second multi-head attention layer is a cross-modal attention between words and the image features. It aims to generate the corresponding description according to a given image feature. The cross-modal attention is expressed as

$$X = \mathrm{AddNorm}(\mathrm{MHA}(X, V_I, V_I)), \tag{5}$$

where X represents the output of the semantic-guided attention block and V_I is the image feature. The output is used as the input to the next decoder. In the final decoder, the caption is obtained using a fully connected layer and a soft-max layer.

To guide the decoder generates high-level semantic words, we modify the conventional self-attention by using the topic feature. In the proposed semantic-guided decoder, the global semantic information from topic token can be utilized and guide the word generation.

3.3. Mask–Cross-Entropy Training Strategy

For remote sensing image captioning, dense and multiple objects are hard to express using a common sentence pattern. Furthermore, as the description is simpler than that for natural images, when the model generates sentences with limited diversity, the lack of diversity among sentences will be more obvious. Therefore, improving the diversity of captioning is urgent, but has been rarely explored in the past. We address this problem in this section.

Masking is an effective learning strategy which has already been used in MAE [45] (Masked autoencoders) and BEIT [46] (bidirectional encoder representation from image Transformers) to enhance the model learning ability and to explore the relationships among objects. Inspired by this, to suppress the *exposure bias* [20] in CE, we propose a Mask–Cross-Entropy (Mask-CE) training strategy. The main idea of Mask-CE is randomly masking some words in the ground truth. Specifically, in each iteration, we randomly replace some words in the sequence with [Mask] before the word-embedding layer, where the [Mask] is another word token, similar to [bos] (beginning of a sequence). The loss function of Mask-CE is written as:

$$Loss = -\sum_{t=1}^{L} \log\left(p_\theta\left(y_t \middle| {y^{\mathrm{mask}}}_{1:t-1}\right)\right), \tag{6}$$

where the y denotes the ground-truth words, θ is the model parameter, and y^{mask} is the mask sentence, T is the length of ground-truth sentence. For each predicted time step, the probability p_θ of the next word y_t is obtained according to the given previous mask-sentence words.

The difference of Mask-CE from the Mask strategies used in MAE and BEIT is that the [Mask] token only affects the words behind it but not itself while predicting words, and the model needs to predict all words, not only words with the [Mask] token. In comparison, in MAE and BEIT, the model mainly reconstructs the masked patches, and the normal patches will not be affected.

Moreover, different from the *exposure bias* problem in CE, in the training stage of Mask-CE, some ground-truth words are unknown; thus, the model needs to predict the word based on some noisy words ([Mask]) and fewer ground-truth words. In the actual prediction, the model can also only rely on its own prediction, which may not be the ground truth. Therefore, the training environment of Mask-CE is closer to the actual prediction, such that the *exposure bias* can be suppressed.

The key point of the [Mask] operation is to explore the word relations, and it can be observed that the predicted word often relies on some keywords. If keywords are replaced with [Mask], Mask-CE will enforce the model to predict the word from other words, whose relation may not yet be explored, but which is useful. As the position of [Mask] is random, with an increasing number of iterations, the [Mask] token will cover every word in the sentence, which means that the relationships between words are better explored. In short, Mask-CE can overcome the *exposure bias* in CE and combines some of the core context of SC, and so it can ensure that the model generates captions with high accuracy and diversity.

4. Experiments and Analysis

4.1. Data Set and Setting

4.1.1. Data Set

Three open public remote sensing image captioning data sets were used to test the proposed method. The split of the training, validation, and testing sets was 80%, 10%, and 10%, respectively. To ensure fairness in the experimental comparison, we split the data sets following publicly available splits [6,27]. The details of the three data sets are provided in the following.

(1) The Sydney data set [6] contains 613 images with a size of 500 × 500 pixels. The images are from Australia, obtained from Google Earth, and comprise seven categories.

(2) The UCM data set [6] is based on the UC Merced Land-Use data [47], which contains 21 categories and 2100 images with size of 256 × 256 pixels.

(3) The RSICD data set [27] contains 10,921 images in 30 categories, where the size of each image is 224 × 224 pixels. Compared with other data sets, RSICD has a larger number of images and richer descriptions.
Each image in these three data sets has five corresponding human-annotated descriptions.

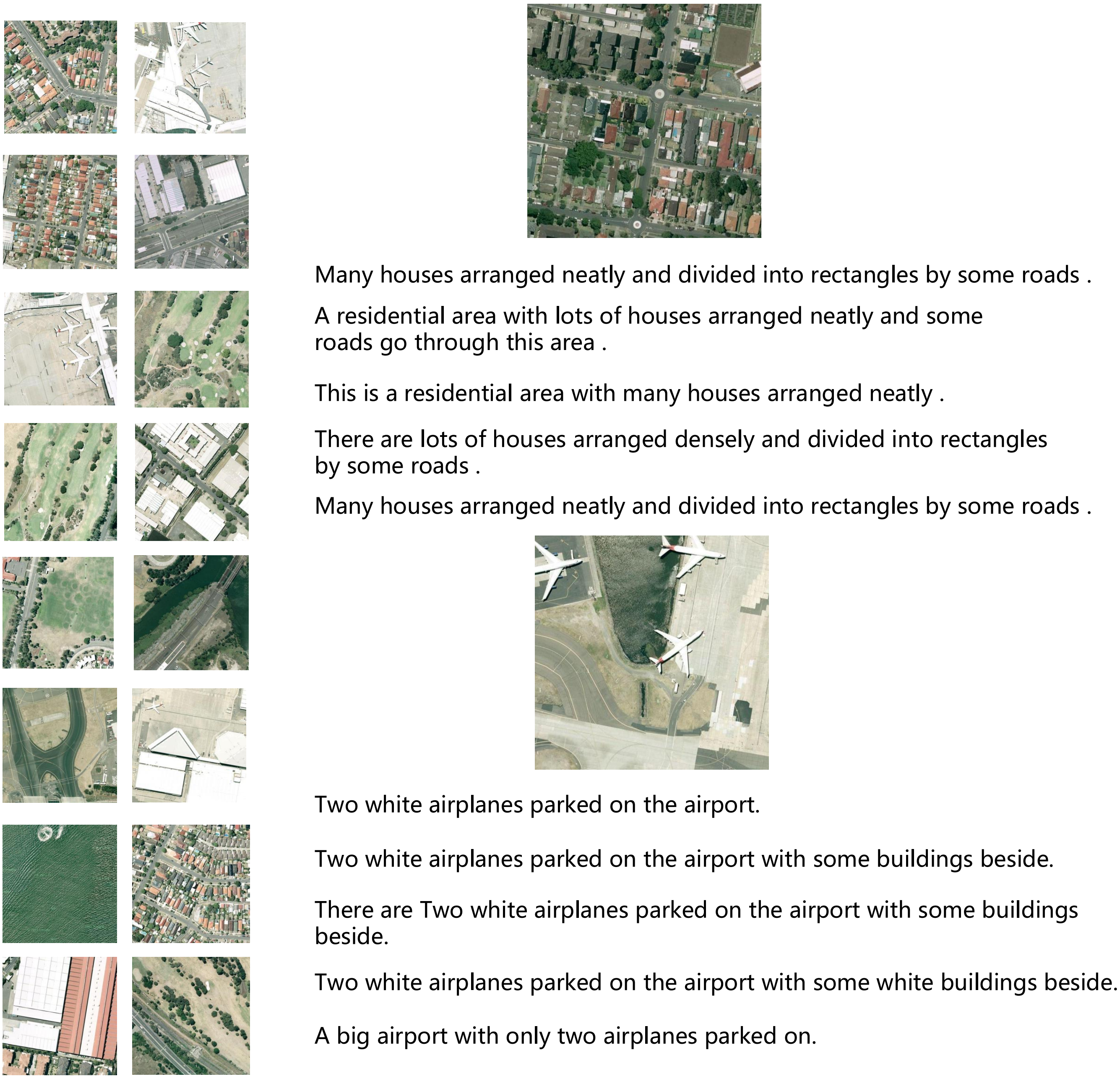
In order to show the differences among the data sets more intuitively, some images and corresponding descriptions of the RSICD and Sydney data sets are illustrated in Figures 6 and 7, respectively. It can be observed that the categories and descriptions are more complex in the RSICD data set than in the Sydney data set, while the resolution in Sydney is higher.

4.1.2. Evaluation Metrics

Following the general evaluation metrics, we employed the following captioning metrics: BLEU [33], Rouge [48], Meteor [49], and CIDEr [34]. Meanwhile, the Self-CIDEr [40] score was used to compare the diversity performance in the ablation study.

BLEU: BLEU is a commonly used evaluation index in the field of machine translation. BLEU directly counts how many words in the ground truth have been successfully predicted. To evaluate sentence integrity, BLEU uses four sub-indices to evaluate the words in n-grams.

Rouge-L: The basic concept of Rouge-L is similar to that of BLEU. Rouge uses the recall score for evaluation, but not the accuracy score (as in BLEU).



Many houses arranged neatly and divided into rectangles by some roads .

A residential area with lots of houses arranged neatly and some roads go through this area .

This is a residential area with many houses arranged neatly .

There are lots of houses arranged densely and divided into rectangles by some roads .

Many houses arranged neatly and divided into rectangles by some roads .

Two white airplanes parked on the airport.

Two white airplanes parked on the airport with some buildings beside.

There are Two white airplanes parked on the airport with some buildings beside.

Two white airplanes parked on the airport with some white buildings beside.

A big airport with only two airplanes parked on.

Figure 6. Some images and two image-captioning pairs in Sydney data set.

Meteor: Different from the direct comparison in BLEU, Meteor takes into account word forms and synonyms, and uses the F1 score as the evaluation metric, which combines the recall and precision.

CIDEr: CIDEr sums the weighted cosine similarity of words in different n-grams, considering the word frequency as the weight. Generally speaking, the general words in a sentence often contain little semantic information, and so they have low weight. CIDEr can accurately express semantic information, as it captures the keywords in the sentence well.

Self-CIDEr: Self-CIDEr first calculates the CIDEr score based on the sampled captions, and then uses CIDEr as the kernel to calculate the singular vector decomposition (SVD) over autocorrelation matrices. This score mainly evaluates the diversity among sampled captions with respect to the CIDEr score.

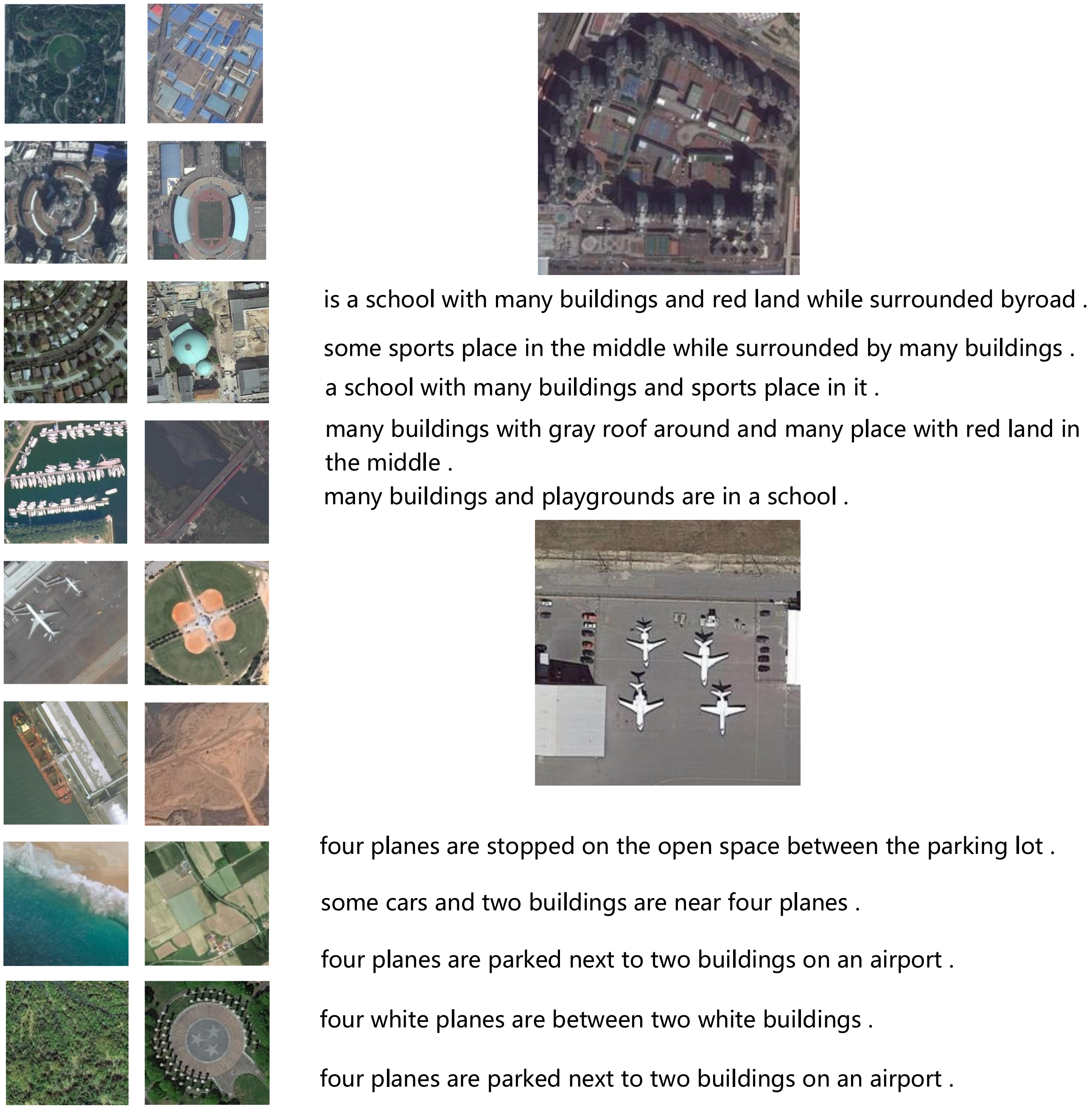

is a school with many buildings and red land while surrounded byroad .
some sports place in the middle while surrounded by many buildings .
a school with many buildings and sports place in it .
many buildings with gray roof around and many place with red land in the middle .
many buildings and playgrounds are in a school .
four planes are stopped on the open space between the parking lot .
some cars and two buildings are near four planes .
four planes are parked next to two buildings on an airport .
four white planes are between two white buildings .
four planes are parked next to two buildings on an airport .

Figure 7. Some images and two image-captioning pairs in RSICD data sets. It can be observed that the large scene has various objects and the captions are abundant.

Among the above metrics, BLEU and Rouge are mainly used to evaluate sentence integrity and consistency, but are less effective at measuring the semantic information. Meteor can better explore the semantics and similarities in sentences. CIDEr and Self-CIDEr are mainly used in image captioning, in order to evaluate whether the semantic words are well-captured. For any of the above metrics, a higher score indicates a better performance.

4.1.3. Training Details

In the topic encoder, the number of encoder layers N_e was 12. For UCM and RSICD data sets, to use the pre-trained weights on ImageNet [50] as the initial weights, the images were resized to 224 × 224 pixels while, for the Sydney data set, we chose to resize them to 384 × 384 pixels. In the decoder, the number of decoder layers was two, and the word embedding size was 768. For each attention module, we set up eight different attention heads. In the training stage, the initial learning rate was set as 3×10^{-5}, which decayed by 0.8 every three epochs. The number of training epochs was set as 20. In the inference stage, the beam search [51] method was applied, where the beam size was set to three.

4.1.4. Compared Models

Most state-of-the-art models in natural image captioning cannot be directly applied to remote sensing image captioning, as they are based on the existing Faster-RCNN bounding box features, which do not fit well in remote sensing images. Under this condition, we compared our model with several state-of-the-art models in remote sensing image captioning. In the following models—except for VRTMM—the main framework is based on CNN + LSTM.

(1) *Soft attention and Hard attention* [27]: The basic framework is VGG-16 + LSTM for both of these models. The soft attention gives weights to different image feature parts, according to the hidden state, while the hard attention samples different image parts and optimizes them by reinforcement learning.

(2) *Structured-attention* [28]: The main framework of Structured-attention is ResNet50 + LSTM. The structured attention module utilizes the structural characteristics in remote sensing images and obtains image region structured segmentation features through a selective search [52].

(3) *AttrAttention* [53]: An attribute attention mechanism is proposed to obtain the high-level attributes from VGG-16, while the encoded features are a combination of image features and the attribute features.

(4) *MLA* [30]: In MLA (Multi-level attention model), a multi-level attention mechanism is set to choose whether to apply the image or the sequence as the main information to generate the new word.

(5) *RASG* [32]: The decoder in RASG is based on an Up–Down attention model [22]. A recurrent attention and semantic gate are proposed in the decoder, in order to help integrate visual features and attention features to generate a better context vector.

(6) *VRTMM* [31]: In VRTMM, image features are captured by a variational auto-encoder model. Furthermore, VRTMM replaces the LSTM with Transformer as the decoder, thus achieving better performance.

4.2. Experimental Results

The evaluation scores of different models on the three data sets are shown in Tables 1–3. All experimental results followed the same open public data splits [6,27], and the best scores are marked in bold. The BLEU-*n* are four BLEU indices with *n*-grams. The different *n*-grams provide an assessment of whether *n* words are in a particular order. The CIDEr score is the most commonly used metric, which can evaluate whether the global semantic words are generated. It is easy to see that our model presented superior performance over the compared models in almost all of the metrics.

Table 1. Comparison scores (%) on UCM data set.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGEL	CIDEr
Soft-attention [27]	74.54	65.45	58.55	52.50	38.86	72.37	261.24
Hard-attention [27]	81.57	73.12	67.02	61.82	42.63	76.98	299.47
Struct-attention [28]	85.38	80.35	75.72	71.49	46.32	81.41	334.89
AttrAttention [53]	81.54	75.75	69.36	64.58	42.40	76.32	318.64
MLA [30]	84.06	78.03	73.33	69.16	**53.30**	81.96	311.93
RASG [32]	85.18	79.25	74.32	69.76	45.71	80.72	338.87
VRTMM [31]	83.94	77.85	72.83	68.28	45.27	80.26	349.48
Ours	**89.36**	**84.82**	**80.57**	**76.50**	50.81	**85.86**	**389.92**

Table 2. Comparison scores (%) on Sydney data set.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGEL	CIDEr
Soft-attention [27]	73.22	66.74	62.23	58.20	39.42	71.27	249.93
Hard-attention [27]	75.91	66.10	58.89	52.58	38.98	71.89	218.19
Struct-attention [28]	77.95	70.19	63.92	58.61	39.54	72.99	237.91
AttrAttention [53]	81.43	73.51	65.86	58.06	41.11	71.95	230.21
MLA [30]	81.52	74.44	67.55	**61.39**	**45.60**	70.62	199.24
RASG [32]	80.00	72.17	65.31	59.09	39.08	72.18	263.11
VRTMM [31]	74.43	67.23	61.72	56.99	37.48	66.98	252.85
Ours	**83.38**	**75.72**	**67.72**	59.80	43.46	**76.60**	**269.82**

Table 3. Comparison scores (%) on RSICD data set.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGEL	CIDEr
Soft-attention [27]	67.53	53.08	43.33	36.17	32.55	61.09	196.43
Hard-attention [27]	66.69	51.82	41.64	34.07	32.01	60.84	179.25
Struct-attention [28]	70.16	56.14	46.48	39.34	32.91	57.06	170.31
AttrAttention [53]	75.71	63.36	53.85	46.12	35.13	64.58	235.63
MLA [30]	77.25	62.90	53.28	46.08	34.71	69.10	236.37
RASG [32]	77.29	66.51	57.82	50.62	36.26	66.91	275.49
VRTMM [31]	78.13	67.21	56.45	51.23	37.37	67.13	271.50
Ours	**80.42**	**69.96**	**61.36**	**54.14**	**39.37**	**70.58**	**298.39**

Diversity is also a key point to evaluate the description ability of models. To further evaluate our method performance, in terms of accuracy and diversity, we show some sentences generated using our method in Figure 8, which will be discussed in detail in Section 4.4.

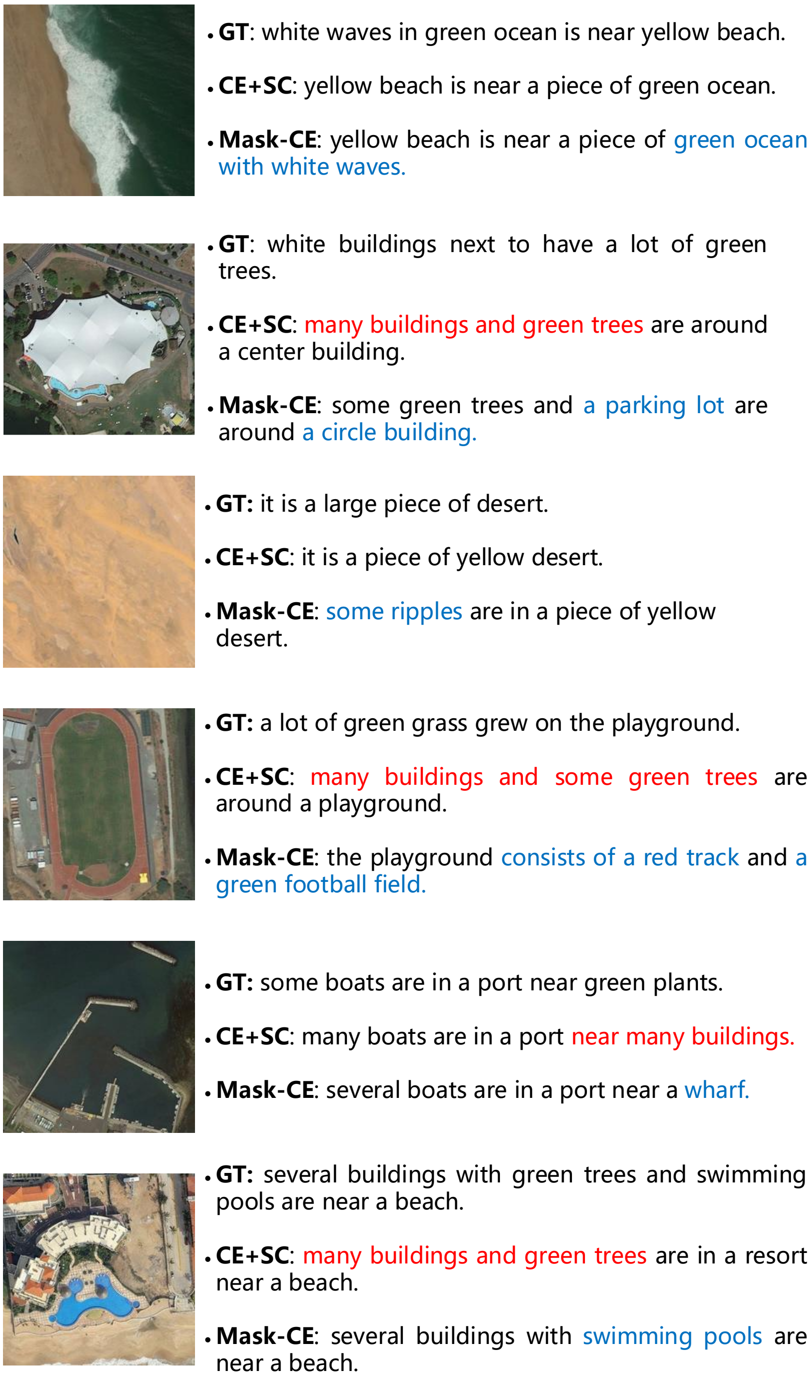
4.3. Discussion and Analysis

The core difference among the state-of-the-art models lies in their attention modules. Compared with the image features in Soft- and Hard-attention mechanisms, Structured-attention provides segmentation features to improve accuracy and decrease the training difficulty of the attention module. However, for the RSICD and Sydney data sets, whose images are difficult to segment through a selective search, the improvement is quite limited and, in some cases, it showed worse performance. The result of Struct-attention also demonstrated that obtaining the region object features is useful, but conventional algorithms cannot deal with the complex objects in remote sensing images. To extract image features considering remote sensing characteristics, the multi-head attention in Transformer was applied in our model encoder, which could capture the abundant relationships among objects in order to perform better in the three data sets under consideration.

The VRTMM mainly modifies the framework in the sentence decoding stage, replacing the LSTM with Transformer as the decoder. Compared with other models using an LSTM as the decoder, the Transformer decoder with multi-head attention can better explore the word relationships, leading to improvement in some scores. However, VRTMM also applied the Cross-Entropy and Self-Critical training strategy, which resulted in higher scores at the expense of losing diversity in the generated descriptions. In comparison, the Mask–Cross-Entropy training strategy was applied to train our model, and a balanced improvement in both aspects was observed. The diversity comparison is illustrated in Section 4.4; to evaluate the diversity, another evaluation metric—namely, Self-CIDEr—was also assessed.

To some extent, extra global semantic information was provided in AttrAttention, MLA, and RASG, which led to performance improvements to some extent. In our model, an extra topic token was added to extract this information. The topic token is trainable and can interact with both image and language features. Compared with the semantic information calculated by the mean pooling operation or soft-max layer, the topic token can capture the semantic information more accurately. In the evaluation metrics, the CIDEr score mainly evaluates whether the semantic words are captured. It was found that the

CIDEr score had the largest increase in our model, which also indicates that the use of the topic token is an effective method to express global semantic information.



- **GT**: white waves in green ocean is near yellow beach.
- **CE+SC**: yellow beach is near a piece of green ocean.
- **Mask-CE**: yellow beach is near a piece of green ocean with white waves.

- **GT**: white buildings next to have a lot of green trees.
- **CE+SC**: many buildings and green trees are around a center building.
- **Mask-CE**: some green trees and a parking lot are around a circle building.

- **GT:** it is a large piece of desert.
- **CE+SC**: it is a piece of yellow desert.
- **Mask-CE**: some ripples are in a piece of yellow desert.

- **GT:** a lot of green grass grew on the playground.
- **CE+SC**: many buildings and some green trees are around a playground.
- **Mask-CE**: the playground consists of a red track and a green football field.

- **GT:** some boats are in a port near green plants.
- **CE+SC**: many boats are in a port near many buildings.
- **Mask-CE**: several boats are in a port near a wharf.

- **GT:** several buildings with green trees and swimming pools are near a beach.
- **CE+SC**: many buildings and green trees are in a resort near a beach.
- **Mask-CE**: several buildings with swimming pools are near a beach.

Figure 8. Sentence comparison under different training strategies. The GT is one of the ground-truth sentences, CE with SC and Mask-CE are two different sentences generated in two corresponding training stages. Some common words are colored in red, and the diverse words are colored in blue.

In order to better explain the effect of multi-head attention, word–image attention maps in two of the multi-attention heads are visualized in Figure 9, where the red and green areas are the regions with higher attention weight. It can be seen that the attention maps are various in different attention heads, and some semantic word attentions are focused in different regions; thus, the generated words are based on the information of different regions information. With the increasing number of attention heads, although some information is useless in some attention heads, the multi-head attention strategy can learn more explicable and visible information from different aspects. Finally, the information will be integrated using a weight vector, in order to obtain the specific attention value which is effective for the task.

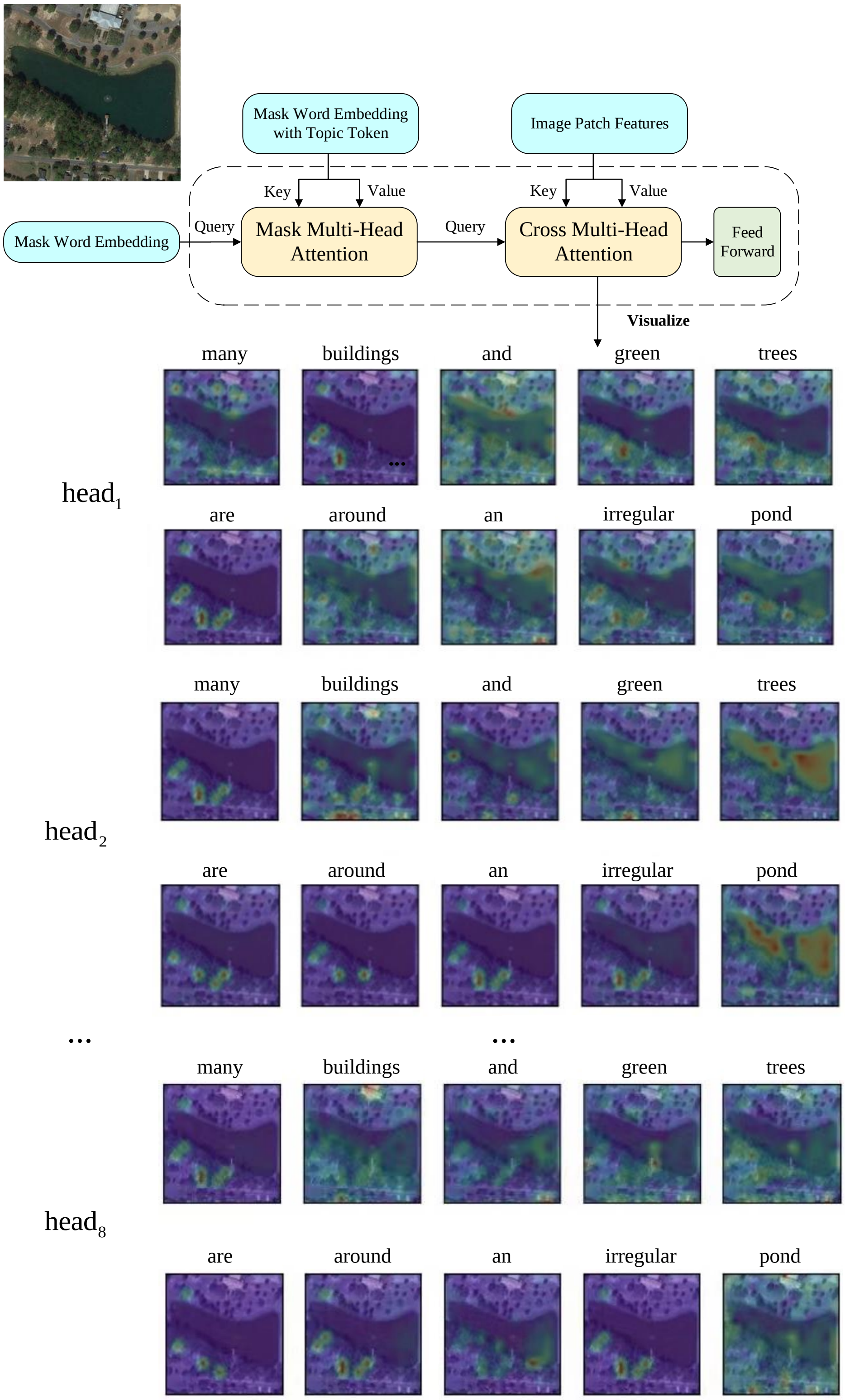


Figure 9. Visualization of cross-modal multi-head attention in decoder.

The drawback of our model is that the topic Transformer encoder relies on the size of the data volume. As the topic encoder is based on a 12-layer Transformer network, it requires a large volume of data for training. As shown in Table 2, although the size of the Sydney data set was only 613, compared with 10,921 in RSICD, the improvement was relatively limited.

4.4. Ablation Experiments

Ablation experiments were conducted to evaluate the effectiveness of three components in our model, including the topic token, the MLP layer between the encoder and decoder, and the Mask–Cross-Entropy training strategy. The ablation experiments were performed by the following ways.

- To find the effect of the topic token, we removed it from the topic encoder and kept other configurations unchanged.
- The MLP layer was removed in order to evaluate the importance of this fine-tuning layer, and the parameters of topic encoder were set to be frozen at the same time.
- We trained the model using an ordinary strategy and compared it with the performance obtained when trained with Mask–Cross-Entropy.

The ablation experiment results of topic token and MLP are shown in Table 4. It can be observed that removing either of the two modules degraded the performance of the model, and removing the topic token led to a more obvious reduction than removing MLP. This phenomenon indicates that the extra global semantic information expressed by the topic token was more important for visual content understanding. The MLP is mainly used for fine-tuning the encoded features in order to overcome the gap between classification and captioning tasks. The combination of global semantic information and fine-tuning encoded features can enhance the model, in order to generate words more comprehensively and precisely.

Table 4. Ablation experiment scores (%) on RSICD data set.

Topic	MLP	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGEL	CIDEr
×	×	79.38	67.58	58.50	51.82	38.13	68.68	287.97
×	✓	79.12	68.05	59.03	51.77	38.35	68.99	293.22
✓	×	79.47	69.28	60.33	52.86	38.53	69.31	295.73
✓	✓	**80.40**	**69.95**	**61.34**	**54.12**	**39.36**	**70.51**	**298.39**

The results with the different training strategies are shown in Table 5. To compare the diversity performance, the Self-CIDEr [40] score was assessed and, for fair comparison in sentence diversity, the beam search strategy was not used while calculating the Self-CIDEr score. It can be observed that the accuracy score of the model trained by Mask-CE was almost the same as the model trained using CE with SC, and both of them were better than the model only trained using CE. The diversity score with Mask-CE was near to that of CE and much higher than that of CE and SC.

Table 5. Comparison scores (%) of different strategies on RSICD data set. B, M, R, C, and Self-C represent BLEU, Meteor, Rouge-L, CIDEr, and Self-CIDEr, respectively.

Dataset	Strategy	B-1	B-2	B-3	B-4	M	R	C	Self-C
UCM	CE	88.39	83.59	79.09	74.82	48.72	83.69	365.66	**59.94**
	CE and SC	87.48	83.06	78.71	74.16	48.32	82.95	377.85	15.80
	Mask-CE	**89.36**	**85.17**	**80.57**	**76.50**	**51.38**	**85.85**	**389.92**	59.86
Sydney	CE	81.55	73.15	65.17	57.96	41.95	74.42	261.60	**63.94**
	CE and SC	82.19	75.23	69.02	**63.57**	43.61	75.43	**272.90**	8.33
	Mask-CE	**83.38**	**75.72**	**67.72**	59.80	43.46	**76.60**	269.82	60.61
RSICD	CE	79.31	68.74	59.60	51.31	38.78	69.00	292.31	**69.28**
	CE and SC	79.46	68.34	59.74	52.63	38.87	69.33	**301.76**	13.12
	Mask-CE	**80.42**	**69.96**	**61.36**	**54.14**	**39.37**	**70.58**	298.39	60.41

To further illustrate the effect of Mask-CE on sentence diversity, some sentence generation results based on different training strategies and the corresponding ground truth are exhibited in Figure 8. It can be seen that some phrases, such as “many buildings” or “green trees”, had a high probability of occurrence in CE with SC. In contrast, Mask-CE can effectively prevent this situation while promoting sentence diversity.

From the above experiments, it can be observed the Mask-CE method maintained diversity as much as possible without much loss of accuracy.

5. Conclusions

In this paper, a mask-guided Transformer with topic token is proposed for remote sensing image captioning. The model can generate sentences with high accuracy and diversity. At first, the topic encoder utilizes the topic token to interact with the scene and objects in the image to represent the global semantic information, and the semantic-guided decoder is introduced to generate words with the assistance of the topic token. Moreover, the Mask–Cross-Entropy training strategy replaces some input words with [Mask], forcing the model to better explore the word relationships. The experimental results of the proposed model on three data sets outperform those of state-of-the-art models, and it obtained the highest CIDEr score (298.39) in the RSICD data set.

Our method is designed for remote sensing images, mainly addressing the difficulty of extracting large-scene image features and global semantic information. The use of this framework can be also explored in other images with similar characteristics. Additionally, although the Mask–Cross-Entropy strategy was found to be effective in remote sensing image captioning tasks, it may also be fit for other image captioning tasks. Our method also has limitations, as the Transformer performance relies on the data volume: if the data size is limited, the performance will also decrease to some extent.

In the future, with the further development of remote sensing technology, the contents of image will be more complex, and so it will become more difficult to describe the images. Therefore, we plan to upgrade our model in the following aspects: (1) Considering the difference between a remote sensing image and a natural image, using the ordinary evaluation metrics may be insufficient, so we will explore more appropriate evaluation metrics in accordance with remote sensing image characteristics, such as evaluating words associated with relative object positions. (2) To describe objects more accurately, the fine-grained features should be extracted in the encoder; in this way, we intend to combine the Transformer with fine-grained learning. (3) Considering the shortcomings of our model, we will try to use some data augmentation methods and reduce some of the unnecessary model parameters to overcome the high data volume requirement.

Author Contributions: Conceptualization, Z.R. and S.G.; funding acquisition, R.L.; methodology, Z.R. and S.G.; software, Z.R.; supervision, S.G. and Z.G.; writing—original draft, Z.R.; writing—review and editing, S.G., Z.G. and S.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (Grant No. 62102296), the Natural Science Foundation of Shaanxi Province (Grant No. 2022JQ-661) and the Fundamental Research Funds for the Central Universities (Grant No. XJS222215, XJS222221).

Data Availability Statement: The RSICD, UCM, and Sydney data sets can be obtained from (https://github.com/201528014227051/RSICD_optimal, accessed on 15 June 2022). The code is available at (https://github.com/Meditation0119/a-mask-guided-transformer-with-topic-token-for-remote-sensing-image-captioning.git, accessed on 15 June 2022).

Acknowledgments: The authors would like to express their gratitude to the editors and the anonymous reviewers for their insightful comments.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations and mathematical symbols are used in this manuscript:

CNN	Convolutional neural networks.
Faster-R-CNN	Faster region convolutional neural networks.
LSTM	Long- and short-term memory.
ResNet	Residual network.
VGG	Visual geometry group.
RSICD	Remote sensing image captioning data set.
CE	Cross-entropy.
SC	Self-critical.
Mask-CE	Mask–Cross-Entropy
NIC	Neural image caption.
AOA	Attention on attention.
UCM	UC Merced.
VRTMM	Variational autoencoder and reinforcement learning. based two-stage multitask learning model.
RASG	recurrent attention and semantic gate.
GAN	Generative adversarial network.
VAE	Variational auto encoder.
SLL-SLE	Sequence-level learning via sequence level exploration.
Vit	Vision Transformer.
Bert	Bidirectional rncoder representations from Transformers.
MLP	Multilayer perceptron.
MAE	Masked autoencoders.
BEIT	Bidirectional encoder representation from image Transformers.
BLEU	Biingual evaluation understudy.
Rouge-L	Recall-oriented understudy for gisting evaluation—Longest.
Meteor	Metric for Evaluation of translation with explicit ordering.
CIDEr	Consensus-based image description evaluation.
MLA	Multi-level attention model.
h	The height of the image.
w	The width of the image.
s	The height and width of the image patch.
N	The number of image patches.
V_1	Input of the topic encoder.
T	The proposed topic token.
p_{pos}	The position encoding.
p_n	The nth image patch.
N_e	The number of encoder layers.
N_d	The number of decoder layers.
MHA	The multi-head attention mechanism.
Q, K, V	The query, key, and value vector of attention mechanism.
W	Weight matrix.
d	The scale factor of attention mechanism.
V_T	The topic feature.
V_I	The image feature.
V_{N_e}	The output of topic encoder.
AddNorm	Residual connection and layer normalization.
X	The output of the previous decoder.
MaskMHA	The mask multi-head attention mechanism.
L	The length of ground-truth sentence.
t	Current time step.
y	The ground-truth sentence.
θ	The model parameter.
y^{mask}	The mask sentence.
p_θ	The probability of generating specific word.

References

1. Zhang, L.; Zhang, L.; Tao, D.; Huang, X.; Du, B. Hyperspectral remote sensing image subpixel target detection based on supervised metric learning. *IEEE Trans. Geosci. Remote Sens.* **2013**, *52*, 4955–4965. [CrossRef]
2. Xu, Y.; Du, B.; Zhang, L. Multi-source remote sensing data classification via fully convolutional networks and post-classification processing. In Proceedings of the IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 3852–3855.
3. Cheng, G.; Yang, C.; Yao, X.; Guo, L.; Han, J. When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 2811–2821. [CrossRef]
4. Baumgartner, J.; Gimenez, J.; Scavuzzo, M.; Pucheta, J. A new approach to segmentation of multispectral remote sensing images based on mrf. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 1720–1724. [CrossRef]
5. Wang, Q.; Gao, J.; Li, X. Weakly supervised adversarial domain adaptation for semantic segmentation in urban scenes. *IEEE Trans. Image Process.* **2019**, *28*, 4376–4386. [CrossRef]
6. Qu, B.; Li, X.; Tao, D.; Lu, X. Deep semantic understanding of high resolution remote sensing image. In Proceedings of the 2016 International Conference on Computer, Information and Telecommunication Systems (Cits), Istanbul, Turkey, 16–18 December 2016; pp. 1–5.
7. Shi, Z.; Zou, Z. Can a machine generate humanlike language descriptions for a remote sensing image? *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3623–3634. [CrossRef]
8. Lu, X.; Zheng, X.; Li, X. Latent semantic minimal hashing for image retrieval. *IEEE Trans. Image Process.* **2016**, *26*, 355–368. [CrossRef]
9. Liu, Y.; Wu, L. Geological disaster recognition on optical remote sensing images using deep learning. *Procedia Comput. Sci.* **2016**, *91*, 566–575. [CrossRef]
10. Ordonez, V.; Han, X.; Kuznetsova, P.; Kulkarni, G.; Mitchell, M.; Yamaguchi, K.; Stratos, K.; Goyal, A.; Dodge, J.; Mensch, A.; et al. Large scale retrieval and generation of image descriptions. *Int. J. Comput. Vis.* **2016**, *119*, 46–59. [CrossRef]
11. Kuznetsova, P.; Ordonez, V.; Berg, A.; Berg, T.; Choi, Y. Collective generation of natural image descriptions. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Jeju Island, Korea, 8–14 July 2012; Volume 1, pp. 359–368.
12. Kulkarni, G.; Premraj, V.; Ordonez, V.; Dhar, S.; Li, S.; Choi, Y.; Berg, A.C.; Berg, T.L. Babytalk: Understanding and generating simple image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 2891–2903. [CrossRef]
13. Gupta, A.; Mannem, P. From image annotation to image description. In Proceedings of the International Conference on Neural Information Processing, Doha, Qatar, 12–15 November 2012; Springer: Berlin/Heidelberg, Germany, 2012; pp. 196–204.
14. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3156–3164.
15. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of the Advance in Neural Information Processing Systems (NIPS), Montreal, QC, USA, 7–12 December 2015.
16. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
17. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
18. Rennie, S.J.; Marcheret, E.; Mroueh, Y.; Ross, J.; Goel, V. Self-critical sequence training for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7008–7024.
19. Li, Y. Deep reinforcement learning: An overview. *arXiv* **2017**, arXiv:1701.07274.
20. Ranzato, M.; Chopra, S.; Auli, M.; Zaremba, W. Sequence level training with recurrent neural networks. *arXiv* **2015**, arXiv:1511.06732.
21. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
22. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. 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, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6077–6086.
23. Huang, L.; Wang, W.; Chen, J.; Wei, X.Y. Attention on attention for image captioning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 4634–4643.
24. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; Volume 30.
25. Cornia, M.; Stefanini, M.; Baraldi, L.; Cucchiara, R. Meshed-memory transformer for image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10578–10587.
26. Herdade, S.; Kappeler, A.; Boakye, K.; Soares, J. Image captioning: Transforming objects into words. In Proceedings of the Conference on Neural Information Processing Systems (NIPS 2019), Vancouver, BC, Canada, 8–14 December 2019; Volume 32.
27. Lu, X.; Wang, B.; Zheng, X.; Li, X. Exploring models and data for remote sensing image caption generation. *IEEE Trans. Geosci. Remote Sens.* **2017**, *56*, 2183–2195. [CrossRef]
28. Zhao, R.; Shi, Z.; Zou, Z. High-resolution remote sensing image captioning based on structured attention. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–14. [CrossRef]

29. Huang, W.; Wang, Q.; Li, X. Denoising-based multiscale feature fusion for remote sensing image captioning. *IEEE Geosci. Remote. Sens. Lett.* **2020**, *18*, 436–440. [CrossRef]
30. Li, Y.; Fang, S.; Jiao, L.; Liu, R.; Shang, R. A multi-level attention model for remote sensing image captions. *Remote Sens.* **2020**, *12*, 939. [CrossRef]
31. Shen, X.; Liu, B.; Zhou, Y.; Zhao, J.; Liu, M. Remote sensing image captioning via Variational Autoencoder and Reinforcement Learning. *Knowl.-Based Syst.* **2020**, *203*, 105920. [CrossRef]
32. Li, Y.; Zhang, X.; Gu, J.; Li, C.; Wang, X.; Tang, X.; Jiao, L. Recurrent Attention and Semantic Gate for Remote Sensing Image Captioning. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–16. [CrossRef]
33. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318.
34. Vedantam, R.; Lawrence Zitnick, C.; Parikh, D. Cider: Consensus-based image description evaluation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4566–4575.
35. Dai, B.; Fidler, S.; Urtasun, R.; Lin, D. Towards diverse and natural image descriptions via a conditional gan. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2970–2979.
36. Wang, L.; Schwing, A.; Lazebnik, S. Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; Volume 30.
37. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Conference on Neural Information Processing Systems (NIPS 2014), Montreal, QC, Canada , 8–13 December 2014.
38. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv* **2013**, arXiv:1312.6114.
39. Chen, J.; Jin, Q. Better captioning with sequence-level exploration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10890–10899.
40. Wang, Q.; Chan, A.B. Describing like humans: On diversity in image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4195–4203.
41. Shi, J.; Li, Y.; Wang, S. Partial Off-Policy Learning: Balance Accuracy and Diversity for Human-Oriented Image Captioning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 10–17 October 2021; pp. 2187–2196.
42. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2019**, arXiv:1810.04805.
43. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
44. Ramchoun, H.; Ghanou, Y.; Ettaouil, M.; Janati Idrissi, M.A. Multilayer perceptron: Architecture optimization and training. *IJIMAI* **2016**, *4*, 26–30. [CrossRef]
45. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked autoencoders are scalable vision learners. *arXiv* **2021**, arXiv:2111.06377.
46. Bao, H.; Dong, L.; Wei, F. Beit: Bert pre-training of image transformers. *arXiv* **2021**, arXiv:2106.08254.
47. Yang, Y.; Newsam, S. 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, San Jose, CA, USA, 2–5 November 2010; pp. 270–279.
48. Lin, C.Y. Rouge: A package for automatic evaluation of summaries. In Proceedings of the Text Summarization Branches Out, Barcelona, Spain, 25–26 July 2004; pp. 74–81.
49. Banerjee, S.; Lavie, A. 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, Ann Arbor, MI, USA, 29 June 2005; pp. 65–72.
50. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 248–255.
51. Graves, A. Sequence transduction with recurrent neural networks. *arXiv* **2012**, arXiv:1211.3711.
52. Uijlings, J.R.; Van De Sande, K.E.; Gevers, T.; Smeulders, A.W. Selective search for object recognition. *Int. J. Comput. Vis.* **2013**, *104*, 154–171. [CrossRef]
53. Zhang, X.; Wang, X.; Tang, X.; Zhou, H.; Li, C. Description generation for remote sensing images using attribute attention mechanism. *Remote Sens.* **2019**, *11*, 612. [CrossRef]