A Joint-Training Two-Stage Method For Remote Sensing Image Captioning

Xiutiao Ye, Shuang Wang, *Member, IEEE*, Yu Gu, *Member, IEEE*, Jihui Wang, Ruixuan Wang, Biao Hou, *Member, IEEE*, Fausto Giunchiglia, and Licheng Jiao, *Fellow, IEEE*

***Abstract*— Compared with remote sensing image (RSI) captioning methods based on the traditional encoder–decoder model, two-stage RSI captioning methods include an auxiliary remote sensing task to provide prior information, which enables them to generate more accurate descriptions. In previous two-stage RSI captioning methods, however, the image captioning and the auxiliary remote sensing tasks are handled separately, which is time-consuming and ignores mutual interference between tasks. To solve this problem, we propose a novel joint-training two-stage (JTTS) RSI captioning method. We use multilabel classification to provide prior information, and we design a differentiable sampling operator to replace the traditional nondifferentiable sampling operation to index the multilabel classification result. In contrast to previous two-stage RSI captioning methods, our method can implement joint training, and the joint loss allows the error of the generated description to flow into the optimization of the multilabel classification via backpropagation. Specifically, we approximate the Heaviside step function with the steep logistic function to implement a differentiable sampling operator for the multilabel classification. We propose a dynamic contrast loss function for multilabel classification tasks to ensure that a certain margin is maintained between the probabilities of the positive label and the negative label during sampling. We design an attribute-guided decoder to filter the multilabel prior information obtained by the sampling operator to generate more accurate image captions. The results of extensive experiments show that the JTTS method achieves state-of-the-art performance on the RSI captioning dataset (RSICD), the University of California, Merced (UCM)-captions, and the Sydney-captions datasets.**

***Index Terms*— Image captioning, image understanding, joint training, multilabel attributes, remote sensing image (RSI).**

I. Introduction

REMOTE sensing image (RSI) captioning, which provides a method to convert complex geographic information from RSIs to text information, makes it easier for humans to utilize them. Therefore, RSI captioning has attracted increasing attention from researchers [1], [2], [3], [4].

The mainstream RSI captioning process follows an encoder–decoder sequence [5], which has the advantages of simple structure, flexible sentence length, and natural description. This process takes a convolutional neural network (CNN) as an encoder to map the image into feature vectors and takes a sequential model, such as a recurrent neural network (RNN) [6] or a long short-term memory (LSTM) network [7], as a decoder to transform the feature vectors into the description of the input image.

The methods that follow the encoder–decoder sequence can be further divided into single-stage methods and two-stage methods. The methods that improve the network structure and attention mechanisms in the encoder–decoder sequence are called one-stage methods [8], [9], [10]. The methods that add a stage to provide auxiliary information [11], [12], [13] are called two-stage methods. The additional stage usually deals with the task of image classification, image retrieval, or object detection. Due to the introduction of auxiliary information, two-stage methods can often generate more accurate descriptions than single-stage methods [14].

As the two stages are trained separately in the previous two-stage methods, the mutual interference between them has been troubling. On the one hand, weak task performance in the first stage will impact negatively on RSI captioning that follows [11] (not to mention it is often time-consuming and laborious to train the task). On the other hand, the errors in RSI captioning cannot be backpropagated to the first-stage task, which prevents the two tasks from working well together.

To solve this persistent problem, this article proposes a novel joint-training two-stage (JTTS) method, in which the tasks of the two stages are jointly trained and coordinated to generate accurate descriptions. Specifically, we design a joint-training structure to utilize the multilabel classification task in the first stage. Then, we propose a margin-based sampling operator (MSO) to solve the gradient explosion caused by steep logistic function during backpropagation. Additionally, we propose a dynamic contrast loss function into multilabel classification as a regularization term of binary cross-entropy (BCE) loss to take into account the distance between positive and negative labels, thereby ensuring the accuracy of the sampling results of the first-stage task. To make better use of the prior information provided by the multilabel classification, we improve the two-layer LSTM structure [15] and propose

Manuscript received 22 March 2022; revised 8 September 2022; accepted 26 October 2022. Date of publication 23 November 2022; date of current version 5 December 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62271377 and Grant 62171347, in part by the Key Research and Development Program of Shaanxi under Program 2021ZDLGY0106 and Program 2022ZDLGY0112, in part by the National Key Research and Development Program of China under Grant 2021ZD0110404, and in part by the Key Scientific Technological Innovation Research Project by Ministry of Education. *(Corresponding author: Shuang Wang.)*

Xiutiao Ye, Shuang Wang, Yu Gu, Jihui Wang, Ruixuan Wang, Biao Hou, and Licheng Jiao are with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China, Xidian University, Xi'an 710071, China (e-mail: shwang@mail.xidian.edu.cn).

Fausto Giunchiglia is with the Department of Information Engineering and Computer Science, University of Trento, 38123 Trento, Italy.

Digital Object Identifier 10.1109/TGRS.2022.3224244

an attribute-guided decoder to introduce three semantic gates to guide the generation of hidden states and captions.

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

1) We propose a JTTS method with an MSO that enables the two-stage RSI captioning to achieve joint training. Unlike the traditional two-stage RSI captioning method of separate training, we train the two-stage tasks together for the first time, producing more accurate image captions while eliminating the tedious pretraining process.
2) We propose dynamic contrast loss as a regular term of BCE loss to consider the relative distance between positive and negative labels in a multilabel classification task, which improves the accuracy of the MSO and enables the JTTS method to generate more accurate captions.
3) We propose an attribute-guided decoder with three semantic gate modules to guide the generation of hidden states as well as words in the two-layer LSTM decoder. Unlike existing methods, we use the semantic gate modules to filter the multilabel semantic prior information during language generation, thus enabling the decoder to better utilize the semantic information in the RSIs.
4) We conduct extensive experiments on the RSICD, the University of California, Merced (UCM)-captions, and the Sydney-captions datasets to validate the superior performance of our proposed method.

The remainder of this article is organized as follows. Section II introduces related work on RSI captioning. Section III describes the details of the proposed JTTS method. Section IV introduces the experiments and analysis on the three datasets. Section V summarizes this article.

II. Related Work

According to their implementation, mainstream RSI captioning methods can be divided into two categories: single-stage methods and two-stage methods.

A. Single-Stage Methods

The single-stage method is commonly designed with end-to-end structures that do not include auxiliary prior information. For example, Qu et al. [8] proposed a deep multimodal neural network, in which a CNN is used as the encoder to extract the image features, and the RNN is used as the decoder to transfer the extracted features into comprehensive descriptions. In addition, Qu et al. [8] built two public RSI captioning datasets, UCM-captions and Sydney-captions, both of which became benchmark datasets for subsequent research. Taking into account the characteristics of RSIs, Lu et al. [9] proposed the largest benchmark dataset for RSI captions, called the RSI captioning dataset (RSICD). They used various image captioning methods to provide benchmarks for the RSICD. Li et al. [16] explored the overfitting problem caused by CE loss in RSI captioning and proposed a novel truncation CE (TCE) loss to reserve probability margins for nontarget words, which helps to generate more flexible and concise descriptions for RSIs. Li et al. [17] proposed a multilevel attention model. Their proposal contains three attention structures, representing the attention to different areas of the image, to different words, and to semantics. Li et al. [18] proposed a recurrent attention mechanism to encode the input image into a context-aware feature representation. Zhang et al. [19] proposed a global visual feature-guided attention (GVFGA) mechanism and a linguistic state-guided attention (LSGA) mechanism to filter the filter out redundant information in the image features and the irrelevant information in the fused visual-textual feature, respectively. Hoxha and Melgani [20] introduced a novel decoder based on support vector machines to replace the RNN decoder in the conventional encoder–decoder framework. Compared with previous methods, it needs fewer annotated samples for training and requires less training and testing time.

In general, although the existing single-stage methods are simple and easy to train, they do not consider the prior information in RSIs. Therefore, they are often inferior to the two-stage methods in terms of the accuracy of generated captions.

B. Two-Stage Methods

In the first stage of the two-stage RSI captioning method, an auxiliary remote sensing task that differs from image captioning is set up to obtain prior information. In the second stage, a description is generated based on the combination of prior information and image features extracted by the CNN. The task in the first stage may vary; for example, it may involve image classification, image retrieval, or object detection. Consequently, an auxiliary task model is needed, such as an image classification model or an object detection model, for different tasks.

Several studies have been conducted using the two-stage method. To understand RSIs, a geospatial relation captioning method was proposed by Chen et al. [12]. A label-attention-mechanism method was proposed by Zhang et al. [14]. They established an image classification task to predict the scene category of the image and used the prediction results to guide the calculation of visual attention during the description generation process. Wang et al. [11] designed a novel retrieval topic recurrent memory network that uses topic information extracted from an RSI as guidance to generate a description. A novel summarization-driven RSI captioning approach was proposed by Sumbul et al. [21]. They first pretrained a pointer-generator network [22] for summarization and then combined the standard captions with the summarized captions to generate a comprehensive description of the image. Wang et al. [13] proposed a word-sentence framework, including a word extractor and a sentence generator, to improve the explainability of the RSI captioning method. Similar to our method, their framework uses a CNN-based multilabel classifier to provide prior information to the sentence generator. In their method, however, the word extractor and sentence generator are trained separately, and the mutual influence between tasks cannot be considered. Zhao et al. [23] proposed an RSI captioning method based on a structured attention mechanism that achieves weakly supervised image segmentation while dealing with the image captioning problem. In their work, class-agnostic image segmentation [24] can

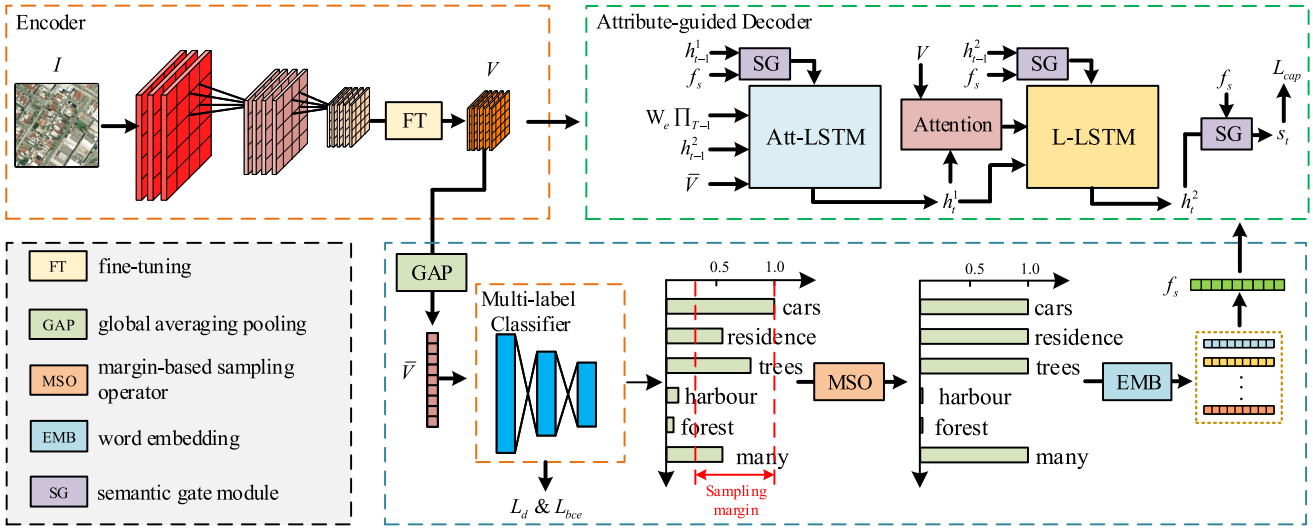

Fig. 1. Overall framework of our proposed JTTS method. For the input image, the pretrained encoder extracts the image features, and the multilabel classifier predicts the multilabel attributes. Then, MSO samples the multilabel attributes and embeds them as attribute features. Finally, the attribute and image features are input into the attribute guidance decoder to obtain the corresponding image captions.

be seen as an auxiliary task that provides prior information for the image captioning model. Zhao [25] conducted a systematic review of RSI captioning work. In this article, he reviewed the existing RSI captioning methods from various perspectives and offered insightful suggestions for potential future research directions.

In contrast to the existing two-stage RSI captioning methods, the JTTS method proposed in this article does not require separate pretraining for the task in the first stage. We use an MSO to perform joint training of the multilabel classification and the description generation, thus achieving greater synergy between them. In addition, we design three semantic gate modules to filter the prior information of the classification results in the first stage to guide the description generation process better.

III. Methodology

This section introduces in detail the proposed JTTS method. As shown in Fig. 1, features of the input image are extracted by the encoder, and the multilabel attributes are predicted by a multilabel classifier. Then, multilabel attributes are sampled by MSO, and the attributes are embedded as attribute features. Finally, the attribute features and image features are entered into the attribute-guided decoder to obtain the corresponding image captions.

A. Image Feature Representation

For our JTTS method, image feature representation is critical to the quality of the generated description. In view of the advantages of the amount of data in the ImageNet dataset, we use a CNN pretrained on the ImageNet dataset as the image encoder. For a given RSI I, the process of feature extraction can be expressed as follows:

$$x = \mathrm{CNN}(I) \tag{1}$$

where $x \in R^{C\times H\times W}$ is the image patch feature extracted by the CNN.

To adapt to subsequent tasks, we fine-tune the extracted image features through the fully connected layer and transform the channel dimension from C to D, where D is the dimension of hidden states in the language model

$$V = W_f x + b_f \tag{2}$$

where $V \in R^{D\times H\times W}$ is the fine-tuned feature. $W_f \in R^{D\times C}$ and $b_f \in R^D$ are learned parameters.

To obtain the multilabel probability distribution of the image, we first design a multilayer perceptron (MLP) to predict the multilabel attributes contained in the image

$$\overline{V} = \mathrm{GAP}(V) \tag{3}$$

$$y_a = \mathrm{MLP}(\overline{V}) \tag{4}$$

where GAP(·) is the global average pooling operator and MLP(·) is an MLP with two cascaded FC-ReLU-dropout units and one FC-sigmoid unit. $\overline{V} \in R^D$ is the global average pooling of V, $y_a \in R^{|\Gamma|}$ represents the probability distribution of multilabel classification, and $|\Gamma|$ represents the size of the dictionary.

B. Margin-Based Sampling Operator

Previously, the tasks in the two-stage RSI captioning method had to be separately trained, because the sampling operation of the auxiliary RSI tasks, such as image classification and image retrieval, is not differentiable. To achieve end-to-end joint training of the two-stage RSI captioning method, we design an MSO to sample the results of the auxiliary multilabel classification task.

The logistic function belongs to an important category of smooth sigmoid functions. It was first used to model population and cell growth. Although there have been many studies on the logistic function, it has not been used in differentiable

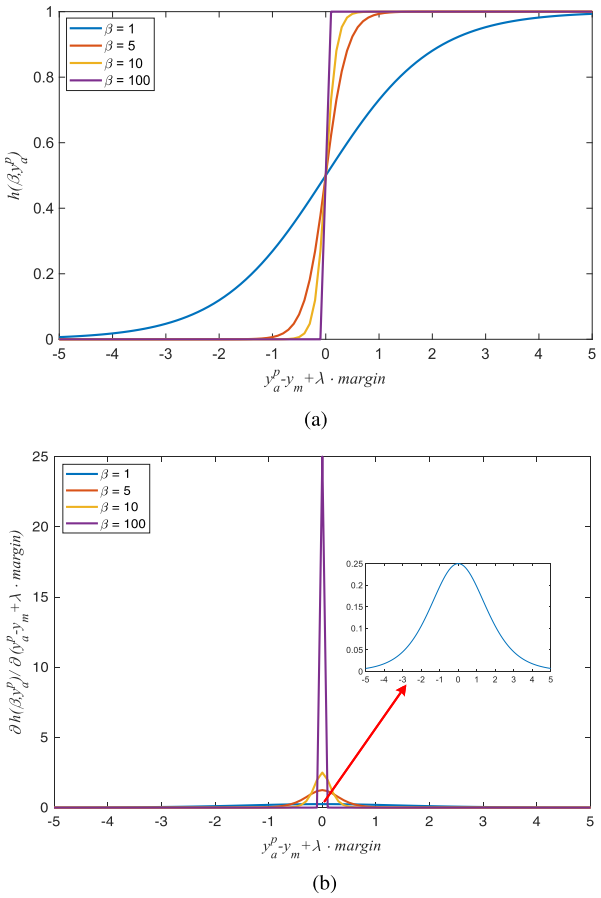

Fig. 2. Illustration of the logistic sampling operator and its derivative when β is 1, 5, 10 and 100. (a) Curves of logistic sampling operator. (b) Curves of derivative of logistic sampling operator.

sampling for multilabel classification tasks. For the probability distribution y_a output of the multilabel classifier, the logistic sampling operator is defined as follows:

$$h(\beta, y_a) = \frac{1}{1 + \exp(-\beta(y_a - y_m + \lambda \cdot \text{margin}))} \tag{5}$$

where β is the reaction rate, which determines the steepness of the sampling function; y_m represents the maximum value of y_a (the probability of the most probable attribute of the input image); λ represents the sampling coefficient; and margin represents the interval between the mean positive and hardest negative in the dynamic contract loss.

The sampling curves of $\beta = \{1, 5, 10, 100\}$ are shown in Fig. 2(a). As β increases, the logistic sampling operator becomes steeper and approximates the Heaviside step function.

However, this operator cannot be directly used to sample the results of multilabel classification, because the differential of y_a by the sampling operator is positively correlated with β, and the specific expression is as follows:

$$\frac{\partial h(\beta, y_a^p)}{\partial y_a^q} = \frac{\beta \exp(-\beta(y_a^p - y_m + \lambda \cdot \text{margin}))}{(1 + \exp(-\beta(y_a^p - y_m + \lambda \cdot \text{margin})))^2} \cdot \left(\frac{\partial y_a^p}{\partial y_a^q} - \frac{\partial y_m}{\partial y_a^q}\right). \tag{6}$$

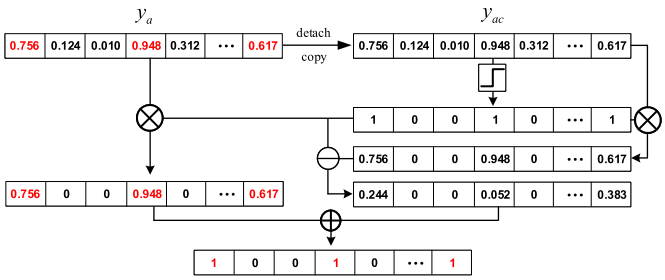

Fig. 3. Illustration of MSO.

Since $y_m = \max(y_a)$ cannot directly obtain the derivative, we can use the following differentiable expression to approximate the maximum value function:

$$y_m = \max(y_a) = \lim_{k \to \infty} \frac{1}{k} \log \Sigma_{\gamma=1}^{|\Gamma|} \exp(k y_a^\gamma) \tag{7}$$

$$\frac{\partial y_m}{\partial y_a} = \lim_{k \to \infty} \text{softmax}(k y_a) = \text{onehot}(\text{argmax}(y_a)) \tag{8}$$

where onehot$(\cdot)$ represents a one-hot operator, and argmax$(\cdot)$ refers to the index corresponding to the maximum value in the vector.

As a result, we have turned the originally nondifferentiable operation into a differentiable operation. The specific derivations of (8) are given in the Appendix. By combining (8) with (6), we can get the partial derivative of $h(\beta, y_a)$ to y_a. For the nonmaximum position, that is, when $q \neq \text{argmax}(y_a)$ the expression is as follows:

$$\frac{\partial h(\beta, y_a^p)}{\partial y_a^q} = \begin{cases} \dfrac{\beta f(y_a^p)}{(1 + f(y_a^p))^2}, & p = q \\ 0, & p \neq q \end{cases} \tag{9}$$

where $f(y_a)$ equals to $\exp(-\beta(y_a - y_m + \lambda \cdot \text{margin}))$.

For the maximum position, i.e., $q = \text{argmax}(y_a)$, its partial derivative is as follows:

$$\frac{\partial h(\beta, y_a^p)}{\partial y_a^q} = \begin{cases} 0, & p = q \\ \dfrac{-\beta f(y_a^p)}{(1 + f(y_a^p))^2}, & p \neq q. \end{cases} \tag{10}$$

We can find that the differential value of the nonmaximum position is consistent with the value when y_m is a constant. In this case, we can draw a curve in the simplest form of (9), as shown in Fig. 2(b). It can be found from (9) and Fig. 2(b) that the maximum slope κ of the sampling operator is positively correlated with β and $\kappa = \beta/4$. Therefore, when the logistic function is approximated as a Heaviside step function, that is, when β is a large value, the gradient explosion problem may arise when training the network.

To solve the problem described earlier, we convert the logistic sampling operator, as shown in Fig. 3, and obtain the MSO. Inspired by [26] and [27], we define y_{ac} as the detached copy of y_a and divide the sampling process into a two-branch structure. Specifically, we perform logistic sampling on y_{ac} and calculate the binary mask m_a and the shifting mask b_a. Then, we sample y_a through m_a and b_a. Since the sampling process of y_{ac} is not in the computation graph of the network, there is

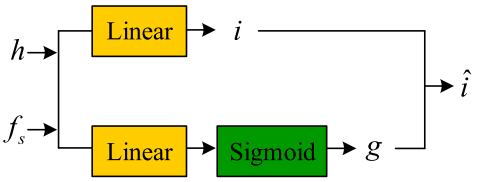

Fig. 4. Illustration of semantic gate module.

no need to calculate the gradient in the backpropagation, thus avoiding the gradient explosion problem

$$\begin{aligned} m_a &= h(\beta, y_{\text{ac}}) \\ &= \frac{1}{1+\exp(-\beta(y_{\text{ac}} - y_{\text{mc}} + \lambda \cdot \text{margin}))} \quad (11) \\ b_a &= m_a - y_{\text{ac}} \circ m_a \quad (12) \\ y_s &= m_a \circ y_a + b_a \quad (13) \end{aligned}$$

where $\circ$ represents the elementwise multiplication operation.

In this case, the derivative of y_s with respect to y_a is as follows:

$$\frac{\partial y_s^p}{\partial y_a^q} = \frac{\partial \left(m_a^p \circ y_a^p\right)}{\partial y_a^q} + \frac{\partial b_a^p}{\partial y_a^q} = \begin{cases} m_a^p, & p = q \\ 0, & p \neq q. \end{cases} \quad (14)$$

Since y_{ac} is the detached copy of y_a, it is equivalent to cutting off the backpropagation of the gradient in y_a and y_{ac}, so we can regard m_a and b_a as constant vectors when calculating the derivative. In this way, we have solved the gradient explosion problem that may occur during training.

The multilabel semantic features f_s can then be obtained by matrix multiplication

$$f_s = \frac{W_e y_s}{\Sigma_{\gamma=1}^{|\Gamma|} y_s^{\gamma}} \quad (15)$$

where $W_e \in R^{E \times |\Gamma|}$ is the embedding matrix and E represents the word embedding size.

C. Attribute-Guided Description Generation

We propose an attribute-guided decoder with three semantic gate modules to filter the prior information of multilabel semantic attributes and guide the generation of RSI captions. Specifically, we use a two-layer LSTM architecture composed of an attention LSTM and a language LSTM as a backbone. The attention LSTM with hidden state $h_t^1 \in R^D$ is used to compute visual attention, and the language LSTM with hidden state $h_t^2 \in R^D$ is used to generate the word at time t. In contrast to previous work, our model generates visual attention and the corresponding description after the objects and attributes in the image have been obtained. To achieve this, we design three semantic gate modules so that the multilabel prior information can be used to guide the generation of h_{t-1}^1 in the attention LSTM, h_2^{t-1} in the language LSTM, and the output word s_t at time t.

As shown in Fig. 4, each semantic gate module consists of an information vector i and an attention gate g, which are generated by two separate linear transformations. The information vector i of h_{t-1}^1, h_{t-1}^2, and s_t can be calculated as follows:

$$\begin{aligned} i_{t-1}^1 &= W_i^1 h_{t-1}^1 + W_i^{s_1} f_s + b_i^1 \quad (16) \\ i_{t-1}^2 &= W_i^2 h_{t-1}^2 + W_i^{s_2} f_s + b_i^2 \quad (17) \\ i_t^o &= W_i^o h_t^2 + W_i^{s_o} f_s + b_i^o \quad (18) \end{aligned}$$

where f_s represents the multilabel semantic features. $W_i^{\{1,2,o,s_1,s_2,s_o\}} \in R^{D \times D}$ and $b_i^{\{1,2,o\}} \in R^D$ are learned parameters.

Then, the attention gate g of h_{t-1}^1, h_{t-1}^2, and s_t is calculated as follows:

$$\begin{aligned} g_{t-1}^1 &= \sigma\left(W_g^1 h_{t-1}^1 + W_g^{s_1} f_s + b_g^1\right) \quad (19) \\ g_{t-1}^2 &= \sigma\left(W_g^2 h_{t-1}^2 + W_g^{s_2} f_s + b_g^2\right) \quad (20) \\ g_t^o &= \sigma\left(W_g^o h_t^2 + W_g^{s_o} f_s + b_g^o\right) \quad (21) \end{aligned}$$

where $W_g^{\{1,2,o,s_1,s_2,o\}} \in R^{D \times D}$ and $b_g^{\{1,2,o\}} \in R^D$ are learned parameters. σ denotes the sigmoid activation function.

The semantic gates apply the attention gates to the information vectors by using elementwise multiplication to obtain the attended information

$$\begin{aligned} \hat{i_{t-1}^1} &= i_{t-1}^1 \circ g_{t-1}^1 \quad (22) \\ \hat{i_{t-1}^2} &= i_{t-1}^2 \circ g_{t-1}^2 \quad (23) \\ \hat{i_t^o} &= i_t^o \circ g_t^o \quad (24) \end{aligned}$$

where $\circ$ represents the Hadamard product operation.

The workflow of the attribute-guided decoder is shown in the top right in Fig. 1. First, the first-layer LSTM, i.e., the Att-LSTM, is used to extract the contextual features. The input of the Att-LSTM consists of the hidden state h_{t-1}^2 at the time step $t-1$ of the second-layer LSTM, the output of the first semantic gate module $\hat{i_{t-1}^1}$, the global visual features $\overline{V}$, and the word embedding feature of the input word Π_t

$$h_t^1 = \text{LSTM}_1\left(h_{t-1}^1, \left[h_{t-1}^2, \hat{i_{t-1}^1}; \overline{V}; W_e \Pi_t\right]\right). \quad (25)$$

Then, the global visual feature $\overline{V}$ and the hidden state of the Att-LSTM h_t^1 are jointly used to calculate the visual attention weights

$$\begin{aligned} a_t^j &= W_a \tanh\left(W_{\text{va}} \overline{V} + W_{\text{ha}} h_t^1\right) \quad (26) \\ \alpha_t &= \text{softmax}(a_t) \quad (27) \\ \hat{V}_{\text{img}} &= \sum_{j=1}^{H \times W} \alpha_t^j V_{\text{img}}^j \quad (28) \end{aligned}$$

where W_{va}, $W_{\text{ha}} \in R^{D \times D}$ are learned parameters. $\alpha_t \in R^{H \times W}$ represents the visual attention weights and $\hat{V}_{\text{img}} \in R^D$ represents the weighted visual features of the input image.

The input of the second-layer LSTM, i.e., L-LSTM, consists of the hidden state of Att-LSTM h_t^1 at time step t, the output of the second semantic gate module $\hat{i_{t-1}^2}$, and the weighted visual features $\hat{V}_{\text{img}}$

$$h_t^2 = \text{LSTM}_2\left(h_{t-1}^2, \left[h_t^1; \hat{i_{t-1}^2}; \hat{V}_{\text{img}}\right]\right). \quad (29)$$

We denote the generated image captions as $S = [s_1, s_2, \ldots, s_t, \ldots, s_T]$. Combining (24) and (29), the word

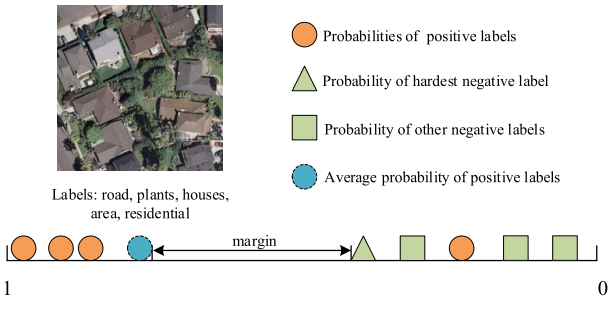

Fig. 5. Illustration of the dynamic contrast loss function for multilabel classification tasks.

s_t can be generated by the output of the third semantic gate module $\hat{i}_t^o$ at time t

$$s_t = \text{argmax}\left(\text{softmax}\left(W_o \hat{i}_t^o + b_o\right)\right) \tag{30}$$

where $W_o \in R^{|\Gamma| \times D}$ and $b_o \in R^{|\Gamma|}$ are learned parameters.

D. Training and Objectives

Most previous multilabel classification tasks used only BCE loss as the objective function. Although BCE loss can make the output of the multilabel classifier close to the distribution of the ground truth, it does not take into account the problem of label imbalance in multilabel classification and cannot control the probability distribution of positive and negative samples. Therefore, only using BCE loss as the objective function of multilabel classification cannot ensure the accuracy of MSO sampling results. We propose a dynamic contrast loss function for multilabel classification that controls the margin between positive and negative labels and reduces the imbalance of positive and negative labels during training. The schematic of the dynamic contrast loss function for multilabel classification is shown in Fig. 5. Our method calculates the average probability of positive labels and constrains its distance from the hardest negative label. In this way, the problem of label imbalance is reduced, thereby improving the sampling accuracy of MSO. Specifically, the expression of dynamic contrast loss function for multilabel classification is as follows:

$$L_d = \max\left(\frac{\sum_{\gamma=1}^{\Gamma} y^\gamma y_a^\gamma}{\sum_{\gamma=1}^{\Gamma} y^\gamma} - \max((1-y) \circ y_a) + \text{margin}, 0\right) \tag{31}$$

where y represents the ground truth of multilabel classification.

The above-mentioned dynamic contrast loss function constrains that the distance from the average of the positive labels to the hardest label is greater than the margin. It is worth mentioning that when λ in (11) is 1.0, the MSO samples the labels in the probability range $[y_{\text{mc}} - \text{margin}, y_{\text{mc}}]$. Since $y_{\text{mc}} >= (\sum_{\gamma=1}^{\Gamma} y^\gamma y_a^\gamma)/(\sum_{\gamma=1}^{\Gamma} y^\gamma)$ is always true, the accuracy of the sampling results is ensured. Moreover, when $\lambda > 1$, the accuracy of the sampling results will decrease, but the diversity will increase. Similarly, when $\lambda < 1$, the accuracy of the sampling results will increase, but the diversity may decline.

We retain the following BCE loss to accelerate the training of multilabel classification tasks:

$$L_{\text{bce}} = \frac{1}{\Gamma} \sum_{\gamma=1}^{\Gamma} \left(y^k \log\left(y_a^\gamma\right) + (1-y^\gamma) \log\left(1-y_a^\gamma\right)\right). \tag{32}$$

For RSI captioning, we train the proposed model by optimizing the CE loss

$$L_{\text{cap}} = -\sum_{t=1}^{T} \log\left(p_\theta\left(s_t^* | s_{1:t-1}^*\right)\right) \tag{33}$$

where $s_{1:t-1}^*$ denotes the target ground truth sequence.

Then, the total loss function is as follows:

$$L_{\text{total}} = L_{\text{cap}} + L_{\text{bce}} + L_d. \tag{34}$$

IV. Experiments

We conducted extensive numerical experiments to verify the effectiveness of our JTTS method. This section first introduces the datasets and the evaluation metrics used. Then, the experimental settings are specified. We also conducted a series of ablation experiments to verify the effect of each submodule in our method. We compared the experimental results of our proposed method with those of state-of-the-art methods. Finally, we discussed the selection of hyperparameters in our experiments.

A. Datasets

To verify the effectiveness of our proposed method, we conduct extensive experiments using the RSICD, the UCM-captions, and the Sydney-captions datasets.

1) *UCM-Captions:* This dataset, based on the UC Merced land-use dataset [28], was proposed in [8]. The dataset contains 2100 high-resolution aerial images (21 scenes with 100 images each). All the images are 256 × 256 pixels, with a pixel resolution of 0.3048 m. Each image is annotated with five descriptions, giving 10 500 sentences.
2) *Sydney-Captions:* This dataset, based on the Sydney dataset [29], was also proposed in [8]. The images were acquired from the Sydney area of Google Earth. It contains a total of 613 high-resolution RSIs with seven categories. All the images have been cropped to 500 × 500 pixels (with a pixel resolution of 0.5 m). Five different descriptions are included for each image, giving a total of 3065 sentences.
3) *RSICD:* This dataset was provided in [9]. It is by far the largest public RSI captioning dataset. The images were collected from Google Earth, Baidu Map, MapABC, and Tianditu. It contains 10 921 images covering 30 scene categories. The images are 224 × 224 pixels. A total of 24 333 different sentences are provided. Lu et al. [9] extended the number of descriptions to 54 605 by randomly duplicating the existing sentences, ensuring each image has five descriptions.

From these three datasets, we used 80% of the data for training, 10% for evaluation, and 10% for testing. Specifically,

we used the same settings as the papers that proposed these datasets [8], [9]. These divisions are consistent with the comparison methods, ensuring a fair comparison.

B. Evaluation Metrics

To evaluate the performance of our proposed method and the quality of the generated sentences, we used nine metrics, including BLEU-n ($n = 1, 2, 3, 4$), recall-oriented understudy for gisting evaluation (ROUGE_L), METEOR, consensus-based image description evaluation (CIDEr), semantic propositional image caption evaluation (SPICE), and Sm.

1) *BLEU-n:* Bilingual evaluation understudy (BLEU) [30] is a metric first used to evaluate the performance of machine translation methods. It measures the co-occurrences of n-grams in generated sentences and ground truth sentences. Commonly used values of n are 1, 2, 3, and 4.
2) *METEOR:* Metric for evaluation of translation with explicit ordering (METEOR) is used to evaluate the accuracy of machine translation [31]. The metric is calculated by generating an alignment between the generated sentences and ground truth sentences. Unlike BLEU-1, METEOR takes into account the uni-gram precision and the uni-gram recall.
3) *ROUGE_L:* It is a metric for evaluating automatic summarization and machine translation [32]. ROUGE_L is the F-measure of the longest common subsequence between the generated sentences and the ground truth sentences.
4) *CIDEr:* It is specially designed for image captioning tasks [33]. It applies term frequency-inverse document frequency (TF-IDF) weights to n-grams in the generated and ground truth sentences.
5) *SPICE:* It is a principled metric for evaluating image captioning that takes semantic content into account [34]. This method converts the generated sentences and ground truth sentences into a graphic-based semantic representation to evaluate the quality of the generated descriptions.
6) Sm is the arithmetic mean of BLEU-4, METEOR, ROUGE_L, and CIDEr. It was proposed in the 2017 AI Challenger[1] to evaluate the quality of the generated sentences

$$\text{Sm} = \frac{1}{4}(\text{BLEU-4} + \text{METEOR} + \text{ROUGE\_}L + \text{CIDEr}). \quad (35)$$

C. Experimental Settings

We performed the experiments on NVIDIA Quadro RTX 5000 with PyTorch version 1.6.0. We used ResNet-101 [35] pretrained on the ImageNet dataset as the encoder. For the multilabel classification task, we extracted high-frequency nouns and adjectives from the target ground truth sequences of the datasets as the multilabel classification labels. More specifically, due to the different sizes of the UCM-Captions dataset, the Sydney-captions dataset, and the RSICD, we selected words that appear more than 100, 50, and 200 times, respectively, as the ground truth labels. The reaction rate β was set to 100. The best values on the validation set for all datasets are margin $= 0.2$ and $\lambda = 0.5$. In the attribute-guided description generation step, we used GloVe [36] to embed the multilabel classification results into feature vectors. The hidden state dimensions of the LSTMs were set to 1000. We used RAdam [37] to optimize the entire model with the learning rate 5e-4. The batch size was set to 10. All models were trained for a total of 30 epochs.

[1] https://challenger.ai/competition/caption

D. Exploring the Efficient CNN Structure

To explore the efficient CNN structure for our method, in this section, we discuss several most commonly used CNN encoders in RSI captioning task, i.e., AlexNet [38], VGG-16 [39], VGG-19 [39], GoogleNet [40], and ResNet-101 [35].

The results of the comparison experiments under different CNN encoder structures are shown in Table I. On the UCM-Captions dataset, different CNNs significantly impact the experimental results. Among them, our method performs best when ResNet-101 is used as the encoder, and the SPICE and Sm scores are 0.5231 and 1.4437, respectively. It is worth noting that although the performance of our method is lower than that of ResNet-101 when using other CNNs, it still achieves competitive performance compared with the previous methods. On the Sydney-captions dataset, there is not much difference in performance when using different CNNs. It may be because the number of images in the Sydney-captions dataset is relatively small, and all CNNs can obtain good image feature representations on this dataset. On the RSICD, our method achieved the best results when ResNet-101 was used as the encoder, with SPICE and Sm scores of 0.4877 and 1.0922, respectively.

Considering the performance of our method under different CNNs on the three datasets, we find that ResNet-101 is the most suitable encoder for our method. It is possible that ResNet-101 can extract more discriminative feature representations than other CNNs, which is very important to our method because the features extracted by CNN in our method are entered into both the decoder and the multilabel classifier.

E. Comparison With State-of-the-Art Methods

To demonstrate the effectiveness of our proposed method, exhaustive comparative experiments were conducted with the following 17 state-of-the-art methods: vector of locally aggregated descriptors (VLAD)-LSTM [9], scale-invariant feature transform (SIFT)-LSTM [9], collective semantic metric learning framework (CSMLF) [41], FC-ATT + LSTM [10], SM-ATT + LSTM [10], soft attention [9], hard attention [9], Sound-a-a [42], SAT (LAM) [14], ADAPTIVE (LAM) [14], the TCE loss-based method [16], the word-sentence framework [13], Recurrent-ATT [18], GVFGA + LSGA [19], structured attention [23], SVM-D BOW [20], and SVM-D CONC [20]. VLAD-LSTM and SIFT-LSTM use handcrafted features to represent the image, and the remaining methods use CNNs as encoders to extract the image features. These

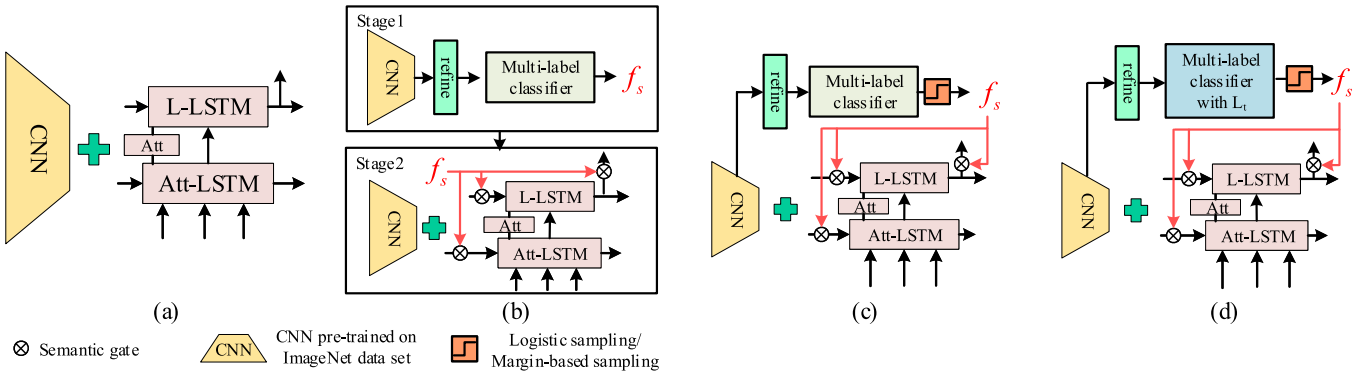

Fig. 6. Illustration of ablation study. (a) "b." (b) "b. + sg/s." (c) "b. + sg + ls" and "b. + sg + MSO." (d) "b. + sg + MSO + L_d." "b. + sg + MSO + L_d." The "b" denotes the structure of the baseline model. "b. + sg/s" denotes the baseline model combined with the semantic gate modules, which the models in two tasks are separately trained. "b. + sg + ls" denotes the joint-training image captioning model with logistic sampling. "b. + sg + MSO" denotes the joint-training image captioning model with MSO. "b. + sg + MSO + L_d" denotes the model introduces the dynamic contrast loss function as the objective function for a multilabel classification task based on "b. + sg + ls" model. "b. + sg + MSO + L_d" denotes the model introduces the dynamic contrast loss function as the objective function for a multilabel classification task based on "b. + sg + MSO" model.

TABLE I
COMPARISON RESULTS OF DIFFERENT CNN ENCODER STRUCTURES

Dataset	CNN Encoder	B-1	B-2	B-3	B-4	M	R	C	S	Sm
UCM-Captions	AlexNet	0.8446	0.7869	0.7377	0.6942	0.4526	0.7894	3.5213	0.5097	1.3644
	VGG-16	0.8474	0.7941	0.7440	0.6930	0.4650	0.8069	3.5544	0.5190	1.3798
	VGG-19	0.8463	0.7929	0.7500	0.7085	0.4647	0.8038	3.5767	0.4942	1.3884
	GoogleNet	0.8551	0.8057	0.7625	0.7225	0.4651	0.8051	3.5686	0.5203	1.3903
	ResNet-101	**0.8696**	**0.8224**	**0.7788**	**0.7376**	**0.4906**	**0.8364**	**3.7102**	**0.5231**	**1.4437**
Sydney-Captions	AlexNet	0.8333	0.7600	0.6933	0.6319	0.4271	0.7539	2.7691	**0.4716**	1.1455
	VGG-16	0.8415	0.7725	0.7052	0.6406	0.4313	0.7636	**2.8073**	0.4505	1.1607
	VGG-19	0.8391	0.7686	0.7034	0.6438	0.4335	0.7605	2.7831	0.4478	1.1552
	GoogleNet	0.8283	0.7491	0.6732	0.6032	0.4233	0.7508	2.7730	0.4488	1.1376
	ResNet-101	**0.8492**	**0.7797**	**0.7137**	**0.6496**	**0.4457**	**0.7660**	2.8010	0.4679	**1.1656**
RSICD	AlexNet	0.7759	0.6622	0.5743	0.5039	0.3711	0.6667	2.6881	0.4746	1.0575
	VGG-16	0.7601	0.6413	0.5486	0.4727	0.3598	0.6589	2.5736	0.4705	1.0163
	VGG-19	0.7620	0.6458	0.5569	0.4858	0.3556	0.6554	2.5999	0.4596	1.0242
	GoogleNet	0.7662	0.6589	0.5749	0.5055	0.3696	0.6658	2.6934	0.4773	1.0586
	ResNet-101	**0.7893**	**0.6795**	**0.5893**	**0.5135**	**0.3773**	**0.6823**	**2.7958**	**0.4877**	**1.0922**

CNNs are pretrained on the ImageNet dataset, as is the case in our method. The details of these methods are as follows.

VLAD-LSTM and SIFT-LSTM were both proposed in [9]. They use handcrafted features (VLAD features [43] and SIFT features [44]) to represent the input image and use an LSTM as the decoder to generate image captions. Specifically, the RSI is first resized to 224 × 224 and then segmented into 16 patches, each with a size of 56 × 56. Then, an SIFT feature is obtained for each patch by principal component analysis of the original SIFT features. Finally, 16 SIFT features are arranged into a vector to represent the image. The VLAD feature is obtained by aggregating the SIFT features.

CSMLF was proposed in [41], in which metric learning was introduced to learn the latent semantic embeddings of the input image and corresponding captions.

FC-ATT + LSTM and SM-ATT + LSTM were both proposed in [10]. Both methods extracted high-level features to represent attributes using a CNN, and they used the features to guide the calculation of attention. FC-ATT + LSTM extracts the high-level features from the output of the last fully connected layer, and SM-ATT + LSTM extracts the high-level features from the output of the softmax layer.

Hard attention and soft attention were both proposed in [9]. The authors proposed a new benchmark dataset, the RSICD, for the RSI captioning task, and they used the encoder–decoder models based on the "soft" and "hard" attention mechanisms proposed in [45], to evaluate the performances of the two methods on the RSICD.

Sound-a-a was proposed in [42]. The method used sound information as the active attention for generating more accurate RSI captions.

SAT (LAM) and ADAPTIVE (LAM) were both proposed in [14]. It proposed a label-attention-mechanism method, which used the word embedding vector of a predicted label to guide the calculation of an attention mask, to exploit redundant image features of the RSI.

The TCE loss-based method was proposed in [16]. By reserving a probability margin for the nontarget words, a novel TCE loss was proposed to alleviate the overfitting problem caused by CE loss in RSI captioning.

TABLE II
COMPARISON RESULTS WITH 17 STATE-OF-THE-ART METHODS ON THE UCM-CAPTIONS DATASET

Method	B-1	B-2	B-3	B-4	M	R	C	S	Sm
VLAD-LSTM [9]	0.7016	0.6085	0.5496	0.5030	0.3464	0.6520	2.3131	-	0.9536
SIFT-LSTM [9]	0.5517	0.4166	0.3489	0.3040	0.2432	0.5235	1.3603	-	0.6078
PCSMLF [41]	0.4361	0.2728	0.1855	0.1210	0.1320	0.3927	0.2227	-	0.2171
FC-ATT+LSTM [10]	0.8135	0.7502	0.6849	0.6352	0.4173	0.7504	2.9958	-	1.1997
SM-ATT+LSTM [10]	0.8154	0.7575	0.6936	0.6458	0.4240	0.7632	3.1864	-	1.2549
Soft Attention [9]	0.7454	0.6545	0.5855	0.5250	0.3886	0.7237	2.6124	-	1.0624
Hard Attention [9]	0.8157	0.7312	0.6702	0.6182	0.4263	0.7698	2.9947	-	1.2023
Sound-a-a [42]	0.7484	0.6837	0.6310	0.5896	0.3623	0.6579	2.7281	0.3907	1.0845
SAT (LAM)* [14]	0.8195	0.7764	0.7485	0.7161	0.4837	0.7908	3.6171	0.5024	1.4019
ADAPTIVE (LAM)* [14]	0.817	0.751	0.699	0.654	0.448	0.787	3.280	0.503	1.2923
TCE loss-based method [16]	0.8210	0.7622	0.7140	0.6700	0.4775	0.7567	2.8547	-	1.1897
Word-sentence framework [13]	0.7931	0.7237	0.6671	0.6202	0.4395	0.7132	2.7871	-	1.1400
Recurrent-ATT [18]	0.8518	0.7925	0.7432	0.6976	0.4571	0.8072	3.3887	0.4891	1.3377
GVFGA+LSGA [19]	0.8319	0.7657	0.7103	0.6596	0.4436	0.7845	3.3270	0.4853	1.3037
SVM-D BOW [20]	0.7635	0.6664	0.5869	0.5195	0.3654	0.6877	2.7142	-	1.0717
SVM-D CONC [20]	0.7653	0.6947	0.6417	0.5942	0.3702	0.6877	2.9228	-	1.1437
Structured attention [23]	0.8538	0.8035	0.7572	0.7149	0.4632	0.8141	3.3489	-	1.3353
Ours	**0.8696**	**0.8224**	**0.7788**	**0.7376**	**0.4906**	**0.8364**	**3.7102**	**0.5231**	**1.4437**

"*" means the pre-training process of the method is different from those of other baselines.

The word-sentence framework was proposed in [13]. By dividing the task into two stages, word extraction and sentence generation, it provides a more intuitive way for the RSI captioning task.

Recurrent-ATT was proposed in [18]. It utilized a recurrent attention mechanism to encode the input image into context-aware feature representation and employed different dilated convolutions to capture multiscale features.

GVFGA + LSGA was proposed in [19]. By exploiting GVFGA and LSGA mechanisms, it filtered out redundant information in image features and irrelevant information in the fused visual-textual features.

SVM-D BOW and SVM-D CONC were both proposed in [20]. They replaced the RNN decoder in the traditional encoder–decoder framework with a support vector machine for decoding image features into image captions. SVM-D BOW and SVD-D CONC denote the methods for encoding sentences based on the bag-of-words model and word concatenation, respectively.

Structured attention was proposed in [23]. It utilized selective search [24] to segment the input image into a set of class-agnostic segmentation proposals. Then, it used a structured attention module to guide the model to focus on structured features during training.

1) Results on the UCM-Captions Dataset: Table II shows the comparison results of the methods described earlier and our method on the UCM-Captions dataset. Our method achieves the highest performance for all nine metrics. For the CIDEr score, which is used to evaluate the image captioning task, our method achieved 3.7102, improving 2.5% over the previous state-of-the-art methods. For the comprehensive score Sm, our method achieved 1.4437, which is 3.0% higher than the previous state-of-the-art methods.

2) Results on the Sydney-Captions Dataset: The comparative results on the Sydney-captions dataset are shown in Table III. Similar to the results obtained from the UCM-Captions, our method achieves the highest scores for all nine metrics. Although our method improves by only 0.6% in the METEOR score compared with the "TCE loss-based method," our method's comprehensive performance has obvious advantages over the previous state-of-the-art methods. Compared with the "recurrent-ATT" method, which has the previous highest Sm score, our method achieves a 7.6% improvement on Sm. Moreover, our method achieves a 6.5% improvement on the CIDEr score compared with the "recurrent-ATT" method.

3) Results on the RSICD: Table IV shows the results of comparative experiments on the RSICD. Our method achieves the highest performance on the RSICD, scoring first in all nine metrics. For the CIDEr metric, which specifically reflects the performance of the image captioning tasks, our method achieves 2.7958, a 1.5% improvement over the second-highest score ("recurrent-ATT" method). In addition, for the Sm score, which comprehensively evaluates the performance of the RSI captioning task, our method has improved by 1.8% over the "recurrent-ATT" method.

F. Ablation Study

To verify the effectiveness of each submodule in our proposed model, we conducted extensive ablation experiments on the RSICD, the UCM-captions, and the Sydney-captions datasets. The baseline model uses a CNN pretrained on the ImageNet dataset as the encoder and the two-layer LSTM model [15] as the decoder. For convenience, we denote the baseline model as "b." The structure of the baseline model is shown in Fig. 6(a). We have conducted the following five ablation experiments.

1) *"b. + sg/s:"* The baseline model is combined with the semantic gate modules when the multilabel classifier and the image captioning model are separately trained.
2) *"b. + sg + ls:"* The multilabel classifier and the captioning model are jointly trained, and the logistic sampling operator in (5) is used to sample the classification results.

TABLE III
COMPARISON RESULTS WITH 17 STATE-OF-THE-ART METHODS ON THE SYDNEY-CAPTIONS DATASET

Method	B-1	B-2	B-3	B-4	M	R	C	S	Sm
VLAD-LSTM [9]	0.4913	0.3472	0.2760	0.2314	0.1930	0.4201	0.9164	-	0.4402
SIFT-LSTM [9]	0.5793	0.4774	0.4183	0.3740	0.2707	0.5366	0.9873	-	0.5422
CSMLF(ft) [41]	0.5998	0.4583	0.3869	0.3433	0.2475	0.5018	0.9378	-	0.5076
FC-ATT+LSTM [10]	0.8076	0.7160	0.6276	0.5544	0.4099	0.7114	2.2033	-	0.9698
SM-ATT+LSTM [10]	0.8143	0.7351	0.6586	0.5806	0.4111	0.7195	2.3021	-	1.0003
Soft Attention [9]	0.7322	0.6674	0.6223	0.5820	0.3942	0.7127	2.4993	-	1.0471
Hard Attention [9]	0.7591	0.6610	0.5889	0.5258	0.3898	0.7189	2.1819	-	0.9541
Sound-a-a [42]	0.7093	0.6228	0.5393	0.4602	0.3121	0.5974	1.7477	0.3837	0.7794
SAT (LAM)* [14]	0.7405	0.6550	0.5904	0.5304	0.3689	0.6814	2.3519	0.4038	0.9832
ADAPTIVE (LAM)* [14]	0.7323	0.6316	0.5629	0.5074	0.3613	0.6775	2.3455	0.4243	0.9729
TCE loss-based method [16]	0.7937	0.7304	0.6717	0.6193	0.4430	0.7130	2.4042	-	1.0449
Word-sentence framework [13]	0.7891	0.7094	0.6317	0.5625	0.4181	0.6922	2.0411	-	0.9285
Recurrent-ATT [18]	0.8000	0.7217	0.6531	0.5909	0.3908	0.7218	2.6311	0.4301	1.0837
GVFGA+LSGA [19]	0.7681	0.6846	0.6145	0.5504	0.3866	0.7030	2.4522	0.4532	1.0231
SVM-D BOW [20]	0.7787	0.6835	0.6023	0.5305	0.3797	0.6992	2.2722	-	0.9704
SVM-D CONC [20]	0.7547	0.6711	0.5970	0.5308	0.3643	0.6746	2.2222	-	0.9480
Structured attention [23]	0.7795	0.7019	0.6392	0.5861	0.3954	0.7299	2.3791	-	1.0226
Ours	**0.8492**	**0.7797**	**0.7137**	**0.6496**	**0.4457**	**0.7660**	**2.8010**	**0.4679**	**1.1656**

"*" means the pre-training process of the method is different from those of other baselines.

TABLE IV
COMPARISON RESULTS WITH 17 STATE-OF-THE-ART METHODS ON THE RSICD

Method	B-1	B-2	B-3	B-4	M	R	C	S	Sm
VLAD-LSTM [9]	0.5004	0.3195	0.2319	0.1778	0.2046	0.4334	1.1801	-	0.4990
SIFT-LSTM [9]	0.4859	0.3033	0.2186	0.1678	0.1966	0.4174	1.0528	-	0.4587
CCSMLF [41]	0.5759	0.3959	0.2832	0.2217	0.2128	0.4455	0.5297	-	0.3524
FC-ATT+LSTM [10]	0.7459	0.6250	0.5338	0.4574	0.3395	0.6333	2.3664	-	0.9492
SM-ATT+LSTM [10]	0.7571	0.6336	0.5385	0.4612	0.3513	0.6458	2.3563	-	0.9537
Soft Attention [9]	0.6753	0.5308	0.4333	0.3617	0.3255	0.6109	1.9643	-	0.8156
Hard Attention [9]	0.6669	0.5182	0.4164	0.3407	0.3201	0.6084	1.7925	-	0.7654
Sound-a-a [42]	0.6196	0.4819	0.3902	0.3195	0.2733	0.5143	1.6386	0.3598	0.6864
SAT (LAM)* [14]	0.6753	0.5537	0.4686	0.4026	0.3254	0.5823	2.5850	0.4636	0.9738
ADAPTIVE (LAM)* [14]	0.6664	0.5486	0.4676	0.4070	0.3230	0.5843	2.6055	0.4673	0.9800
TCE loss-based method [16]	0.7608	0.6358	0.5471	0.4791	0.3425	0.6687	2.4665	-	0.9892
Word-sentence framework [13]	0.7240	0.5861	0.4933	0.4250	0.3197	0.6260	2.0629	-	0.8584
Recurrent-ATT [18]	0.7729	0.6651	0.5782	0.5062	0.3626	0.6691	2.7549	0.4719	1.0732
GVFGA+LSGA [19]	0.6779	0.5600	0.4781	0.4165	0.3285	0.5929	2.6012	0.4683	0.9848
SVM-D BOW [20]	0.6112	0.4277	0.3153	0.2411	0.2303	0.4588	0.6825	-	0.4032
SVM-D CONC [20]	0.5999	0.4347	0.3355	0.2689	0.2299	0.4557	0.6854	-	0.4100
Structured attention [23]	0.7016	0.5614	0.4648	0.3934	0.3291	0.5706	1.7031	-	0.7491
Ours	**0.7893**	**0.6795**	**0.5893**	**0.5135**	**0.3773**	**0.6823**	**2.7958**	**0.4877**	**1.0922**

"*" means the pre-training process of the method is different from those of other baselines.

3) *"b. + sg + MSO:"* The multilabel classifier and the image captioning model are jointly trained, and the sampling operator adopts the MSO in (11)–(13).
4) *"b. + sg + MSO + L_d:"* On the basis of "b. + sg + ls," the dynamic contrast loss function is introduced into the multilabel classification as part of the objective function.
5) *"b. + sg + MSO + L_d (ours):"* The method proposed in this article, that is, on the basis of "b. + sg + MSO," the dynamic contrast loss function is introduced into the objective function of the multilabel classification.

It is worth mentioning that, since β is set to 100, the gradient explosion problem may occur in the backpropagation in "b. + sg + ls" and "b. + sg + MSO + L_d," which may easily lead to the collapse of the model training. To solve this problem, we introduce the strategy of gradient clipping, which restricts the gradient in the range of $[-1, 1]$ during backpropagation. After adding the gradient constraint, the "b. + sg + ls" and "b. + sg + MSO + L_d" models can be trained stably.

We analyzed the effect of each submodule in combination with the experimental results in Table V.

1) Effect of Attribute-Guided Decoder: Comparison of the experimental results of "b." and "b. + sg/s" in Table V shows that adding semantic gate modules in the decoding stage to introduce the multilabel semantic prior information improves the performance on three datasets. Specifically, in the UCM-Captions dataset, the SPICE and Sm scores of "b. + sg/s" increase by 3.9% and 3.4%, respectively, compared with those of "b.;" in the Sydney-Captions dataset, the SPICE and Sm scores of "b. + sg/s" are increased by 1.8% and 3.1%, respectively, compared with those of the "b.;" in the RSICD, the SPICE and Sm scores of "b. + sg/s" are increased by 3.6% and 1.1%, respectively, compared with those of "b." These results show that the attribute-guided decoder can effectively use multilabel semantic information and is suitable for different datasets.

2) Effect of Joint Training: The experimental results of "b. + sg + ls" and "b. + sg/s" in Table V show that, compared

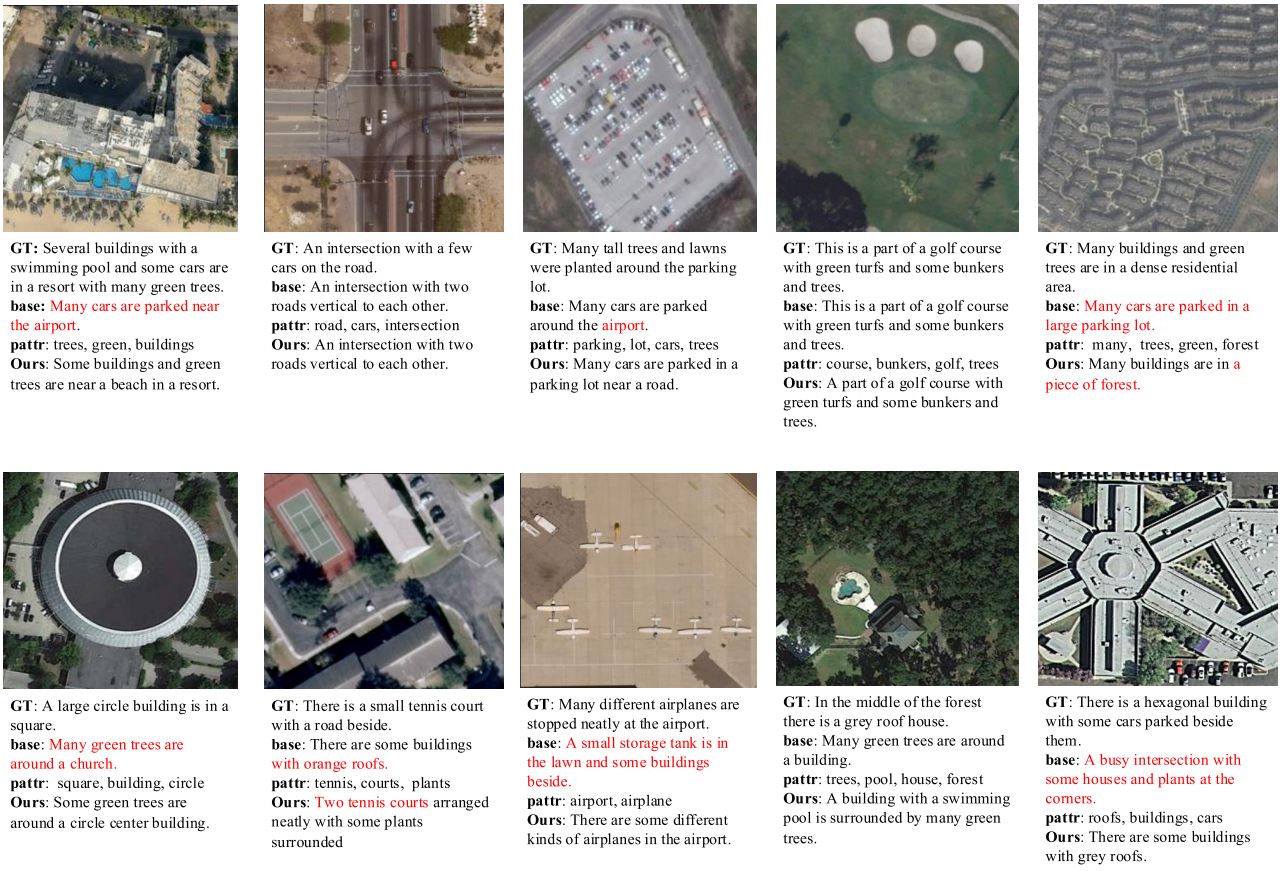

Fig. 7. Captioning examples with the predicted multilabel attributes and the sentences generated by the baseline model and our proposal. The "GT" denotes the ground truth sentence. The "base" denotes the sentence generated by the baseline model. The "pattr" denotes the multilabel attributes predicted by the multilabel classifier. The "Ours" denotes the sentence generated by our method. The words in red are words that are not matched with the corresponding image.

with the previous method of training two-stage tasks, the joint-training of two-stage tasks by using logistic sampling improves the performance of the model on the RSICD and the UCM-Captions datasets. However, in the Sydney-captions dataset, "b. + sg/s" behaves similar to "b. + sg + ls." This may be because the gradient distribution of the logistic sampling method is very steep when β is 100, as shown in Fig. 2(b). Although the model can be made trainable through the strategy of gradient clipping, the derivative of the logistic function with a β of 100 is zero in some situations. In the zero gradient interval, the joint training of multilabel classification and image caption tasks will degenerate into an effect similar to that of separate training.

3) Effect of Margin-Based Sampling: The results of "b. + sg + MSO" and "b. + sg + ls" in Table V show that after improving the sampling method, the performance of the model on the three datasets improved. This is because the MSO not only solves the problem of gradient explosion but also alleviates the problem of the gradient of the steep logistic function sometimes being zero in joint training.

4) Effect of the Dynamic Contrast Loss Function: We established two groups of comparative experiments to verify the effectiveness of introducing the dynamic contrast loss function to multilabel classification. The results of "b. + sg + MSO + L_d" and "b. + sg + ls" demonstrate that the performance of the model on the three datasets improved after introducing the dynamic contrast loss function. Similarly, the performance of "b. + sg + MSO + L_d" compared with "b. + sg + MSO" also improved on the three datasets. In particular, after introducing the dynamic contrast loss function, the most obvious improvement is on the Sydney-captions dataset. The Sm score of "b. + sg + MSO + L_d" is 5.1% higher than that of "b. + sg + ls" and the Sm score of "b. + sg + MSO + L_d" is 3.7% higher than that of "b. + sg + MSO."

G. Parameter Sensitivity Analysis

There are two significant hyperparameters in the JTTS method, margin and λ. Margin corresponds to the distance between the average positive labels and the most difficult negative label in the dynamic contract loss, and λ corresponds to the sampling coefficient in the MSO. In this section, we performed a sensitivity analysis of the JTTS method in terms of these two hyperparameters, which not only helps to evaluate the extent to which various hyperparameter selections affect the performance of our method but also provides guidance for the hyperparameter initialization when transferring our method to a new RSI dataset in the future.

TABLE V
ABLATION STUDY RESULTS ON THE THREE DATASETS

Dataset	Model	B-1	B-2	B-3	B-4	M	R	C	S	Sm
UCM-Captions	b.	0.8241	0.7678	0.7204	0.6791	0.4466	0.7731	3.3548	0.4706	1.3134
	b.+sg/s	0.8500	0.7966	0.7458	0.6965	0.4438	0.7911	3.5036	0.4890	1.3588
	b.+sg+ls	0.8522	0.7979	0.7486	0.7035	0.4712	0.8091	3.5784	0.5146	1.3905
	b.+sg+MSO	0.8678	0.8109	0.7611	0.7159	0.4669	0.8162	3.6668	0.5229	1.4165
	b.+sg+ls+L_d	0.8690	0.8152	0.7658	0.7198	0.4761	0.8229	3.6490	0.5196	1.4169
	ours	**0.8696**	**0.8224**	**0.7788**	**0.7376**	**0.4906**	**0.8364**	**3.7102**	**0.5231**	**1.4437**
Sydney-Captions	b.	0.7932	0.7155	0.6389	0.5698	0.4047	0.7208	2.5313	0.4445	1.0567
	b.+sg/s	0.8173	0.7491	0.6814	0.6201	0.4237	0.7366	2.5756	0.4466	1.0890
	b.+sg+ls	0.8174	0.7371	0.6585	0.5900	0.4176	0.7406	2.6228	0.4476	1.0928
	b.+sg+MSO	0.8293	0.7506	0.6764	0.6106	0.4325	0.7559	2.6974	0.4533	1.1241
	b.+sg+ls+L_d	0.8333	0.7588	0.6930	0.6319	0.4388	0.7543	2.7436	0.4588	1.1421
	ours	**0.8492**	**0.7797**	**0.7137**	**0.6496**	**0.4457**	**0.7660**	**2.8010**	**0.4679**	**1.1656**
RSICD	b.	0.7580	0.6479	0.5614	0.4912	0.3541	0.6549	2.5790	0.4558	1.0198
	b.+sg/s	0.7643	0.6481	0.5670	0.4942	0.3612	0.6627	2.6058	0.4725	1.0310
	b.+sg+ls	0.7803	0.6669	0.5753	0.4991	0.3704	0.6714	2.6842	0.4729	1.0563
	b.+sg+MSO	0.7833	0.6702	0.5773	0.4997	0.3714	0.6733	2.7193	0.4812	1.0659
	b.+sg+ls+L_d	0.7753	0.6653	0.5755	0.5010	0.3695	0.6740	2.7280	0.4738	1.0681
	ours	**0.7893**	**0.6795**	**0.5893**	**0.5135**	**0.3773**	**0.6823**	**2.7958**	**0.4877**	**1.0922**

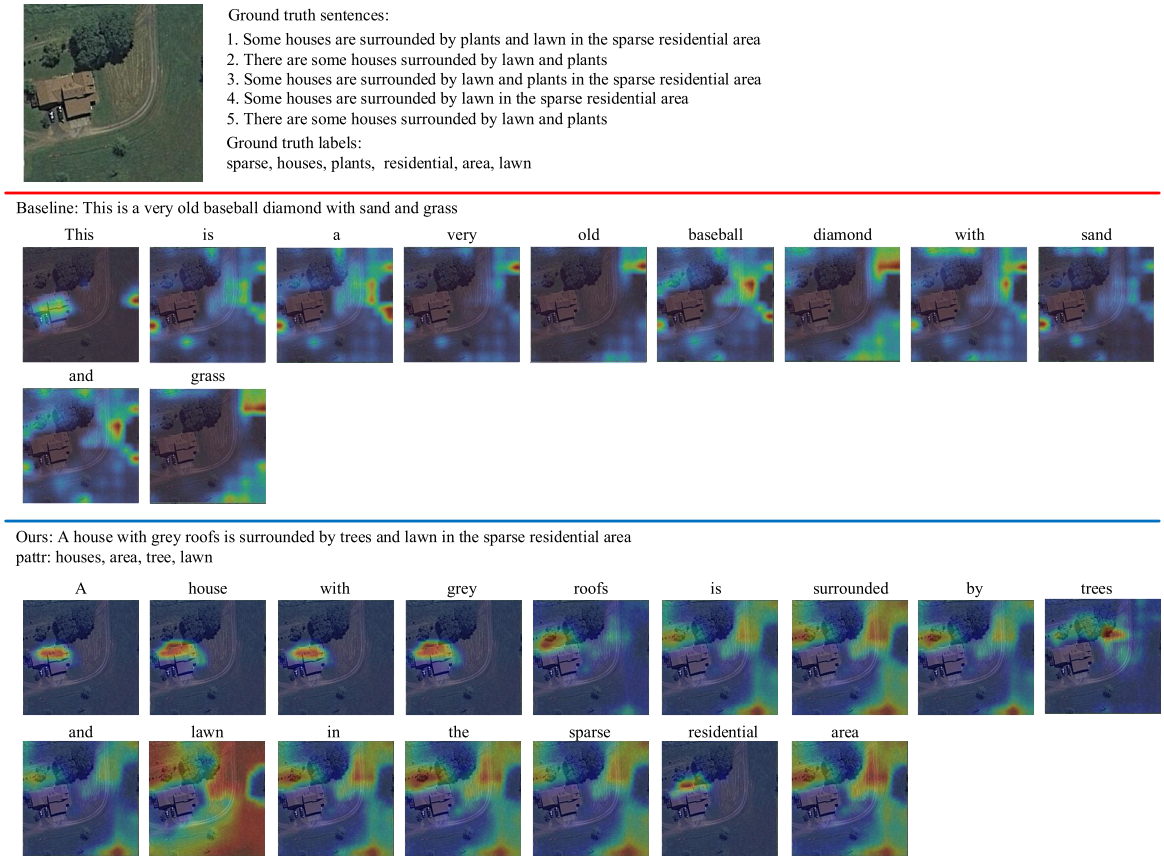

Fig. 8. Visualized attention results of the baseline model and our proposal. In the figures, the closer it is to red, the higher the degree of attention, and the closer it is to blue, the lower the degree of attention.

The impact of margin on our method is shown in Table VI. For the experiments, we set $\lambda = 0.5$ as described in Section IV-C, and set the margin from 0.1 to 0.6. Overall, the performance of our method fluctuates with different margin choices but is still competitive with the previous state-of-the-art methods on the three datasets. On the Sydney-captions dataset, our method has the highest sensitivity to margin. In comparison with the performance at margin $= 0.2$, the Sm and SPICE scores decline at margin $= 0.6$ by 4.5% and 4.3%, respectively. It could be because the Sydney-captions dataset is the smallest of the three datasets and when the margin is large, it is easier to overfit the label distribution of the training set.

The impact of λ on our method is shown in Table VII. For the experiments, we set λ to 0.1, 0.3, 0.5, 1.0, and

TABLE VI
COMPARISON RESULTS WITH DIFFERENT MARGINS ON THE DIFFERENT DATASETS

Dataset	margin	B-4	M	R	C	S	Sm
UCM-Captions	0.1	0.7185	0.4755	0.8275	3.6372	0.5316	1.4147
	0.2	**0.7376**	**0.4906**	**0.8364**	**3.7102**	0.5231	**1.4437**
	0.3	0.7322	0.4799	0.8211	3.6988	0.5147	1.4330
	0.4	0.7327	0.4781	0.8274	3.6550	0.5101	1.4233
	0.5	0.7325	0.4846	0.8211	3.6758	0.5252	1.4285
	0.6	0.7286	0.4772	0.8082	3.6507	**0.5272**	1.4149
Sydney-Captions	0.1	0.6382	0.4355	0.7528	2.7467	0.4622	1.1433
	0.2	**0.6496**	**0.4457**	**0.7660**	**2.8010**	**0.4679**	**1.1656**
	0.3	0.6219	0.4291	0.7384	2.7196	0.4540	1.1273
	0.4	0.6303	0.4332	0.7449	2.7307	0.4610	1.1348
	0.5	0.6220	0.4352	0.7504	2.6856	0.4624	1.1233
	0.6	0.6150	0.4233	0.7462	2.6680	0.4477	1.1131
RSICD	0.1	0.4984	0.3679	0.6746	2.7044	0.4781	1.0613
	0.2	**0.5135**	**0.3773**	0.6823	**2.7958**	**0.4877**	**1.0922**
	0.3	0.5108	0.3725	0.6800	2.7518	0.4828	1.0788
	0.4	0.5113	0.3738	**0.6825**	2.7703	0.4825	1.0845
	0.5	0.5063	0.3755	0.6773	2.7455	0.4802	1.0762
	0.6	0.5018	0.3714	0.6749	2.7111	0.4817	1.0648

TABLE VII
COMPARISON RESULTS WITH DIFFERENT SIZE OF λ ON THE DIFFERENT DATASETS

Dataset	λ	B-4	M	R	C	S	Sm
UCM-Captions	0.1	0.7308	0.4766	0.8196	3.6657	0.5220	1.4232
	0.3	0.7231	0.4820	0.8215	3.6495	0.5157	1.4190
	0.5	**0.7376**	**0.4906**	**0.8364**	**3.7102**	0.5231	**1.4437**
	1.0	0.7324	0.4782	0.8274	3.6966	0.5127	1.4336
	2.0	0.7334	0.4771	0.8299	3.6772	**0.5275**	1.4294
Sydney-Captions	0.1	0.6335	0.4275	0.7552	2.6915	0.4462	1.1269
	0.3	0.6354	0.4354	0.7476	2.7016	0.4671	1.1300
	0.5	0.6496	**0.4457**	**0.7660**	2.8010	0.4679	**1.1656**
	1.0	**0.6518**	0.4327	0.7474	**2.8102**	**0.4729**	1.1605
	2.0	0.6288	0.4238	0.7539	2.6673	0.4405	1.1185
RSICD	0.1	0.5051	0.3696	0.6723	2.7242	0.4752	1.0678
	0.3	0.5060	0.3751	0.6765	2.7554	0.4833	1.0783
	0.5	**0.5135**	**0.3773**	**0.6823**	**2.7958**	**0.4877**	**1.0922**
	1.0	0.5116	0.3743	0.6807	2.7672	0.4830	1.0835
	2.0	0.4996	0.3734	0.6777	2.7041	0.4822	1.0637

2.0, respectively, and set the margin to 0.2, as described in Section IV-C. Similar to the analysis of margin, the performance of our method fluctuates with different λ choices but has an overall advantage over the previous state-of-the-art methods. Among the three datasets, our method has the highest sensitivity to λ on the Sydney-captions dataset. The performance of our method decreases significantly when λ is 0.1 and 2.0. As can be seen from (11), when λ is less than 1.0, sampling accuracy increases, while diversity drops and vice versa. From the experimental results, we can find that our method effectively balances accuracy and label diversity when lambda is set at 0.5 or 1.0.

H. Qualitative Analysis

To show the effectiveness of our proposed JTTS method, we selected some of the generated captions from all three datasets for qualitative analysis, as shown in Fig. 7. In the figure, "GT" represents the ground truth sentences of the image, "base" represents the description generated by the baseline model mentioned in Section IV-F, "pattr" represents the results of multilabel classification, and "Ours" represents the description generated by the JTTS method.

To facilitate comparison, we manually marked the inappropriate descriptions in red. The overall description results show that the descriptions generated by our method are more in line with the content of the image than those of the baseline model. In the first image in the first row, for example, our method accurately describes the scene as a resort near a beach, whereas the baseline model mistakes it for a parking lot next to an airport. Moreover, the results of multilabel classification can guide the generation of descriptions. The results in Fig. 7 show that most images with accurate multilabel predictions are accurately described. However, for the fifth image in the first row, the description generated by our method has some mistakes. As the multilabel classifier incorrectly predicts that the scene contains "forest," our method describes the scene as "many buildings are in a piece of forest." It shows that the results of multilabel classification significantly impact our method.

In addition, according to (25)–(27), we can find that the output of the semantic gate module can guide the calculation of visual attention. Therefore, we compared the results of the attention weights of our JTTS method and the baseline method when generating each word. The visualized results are shown in Fig. 8.

It can be seen from Fig. 8 that our method focused on the area of the house when generating the first word due to the introduction of semantic prior information. Then, it accurately focused on the area containing "trees" and "lawn" when describing the background. In the baseline method, the image is described as a baseball field. This error is reasonable for the algorithm because the fan-shaped contour in the upper left corner of the image is similar to a baseball diamond. However, the baseline method's visual attention is chaotic when generating captions. When it generates words corresponding to a baseball diamond, it focuses on the upper right corner of the image, which is irrelevant to the easily confused area. Our method can produce more accurate attention and image captions than the baseline through semantic gate modules and joint training.

I. Analysis of Training and Testing Time

In engineering deployment, algorithmic efficiency is significant. To evaluate the efficiency of our method, we, respectively, calculated the pretraining time, training time, testing time, floating point operations (FLOPs), and the number of parameters of SAT, baseline model, and our method on the UCM-Captions dataset, as shown in Table VIII.

From the results of the comparison experiments, we can find that since our method adds a multilabel classification branch and three semantic gate modules, the FLOPs and number of parameters of our method are more than the baseline model. However, our method is superior in terms of performance and total training time (training for 30 epochs) compared with the previous two-stage methods that train two tasks separately. In addition, the performance of our method is the best among the three methods. Thus, considering the time cost

TABLE VIII
COMPARISON RESULTS OF THE TRAINING AND TESTING TIME OF DIFFERENT METHODS ON THE UCM-CAPTIONS DATASET

Model	Pre-training Time (s)	Training Time / Epoch (s)	Training Time in Total (s)	Testing Time / Epoch (s)	FLOPs (G)	Params (M)	Sm
b.	-	48.6	1458	6.0	6.9	30.7	1.3134
b.+sg/s	460.3	55.3	2119	7.1	7.5	42.0	1.3588
Ours	-	60.6	1818	7.9	7.7	43.3	1.4437

TABLE IX
COMPARISON RESULTS OF DIFFERENT PRETRAINED DATASETS

Dataset	Pre-trained Dataset	B-1	B-2	B-3	B-4	M	R	C	S	Sm
UCM-Captions	ImageNet	0.8696	0.8224	0.7788	0.7376	0.4906	0.8364	3.7102	0.5231	1.4437
	NWPU-RESISC 45	**0.8908**	**0.8455**	**0.8042**	**0.7643**	**0.5052**	**0.8545**	**3.8538**	**0.5470**	**1.4945**
Sydney-Captions	ImageNet	0.8492	0.7797	0.7137	0.6496	0.4457	0.7660	2.8010	0.4679	1.1656
	NWPU-RESISC45	**0.8615**	**0.7980**	**0.7361**	**0.6774**	**0.4599**	**0.7833**	**2.9378**	**0.4960**	**1.2146**
RSICD	ImageNet	0.7893	0.6795	0.5893	0.5135	0.3773	0.6823	2.7958	0.4877	1.0922
	NWPU-RESISC45	**0.8010**	**0.6879**	**0.5947**	**0.5157**	**0.3852**	**0.6955**	**2.8224**	**0.4933**	**1.1047**

and performance factors, our method trades a relatively small time cost for a significant performance improvement.

V. CONCLUSION

In this article, we propose a JTTS method for RSI captioning to collaboratively train tasks in two-stage RSI captioning. We established multilabel classification to provide prior information for image description generation, and we designed a differentiable MSO to replace the traditional sampling process in multilabel classification. We propose a dynamic contrast loss function as a regularization term for the multilabel classification task to improve the accuracy of the sampling results. Our proposed method allows for joint training for both multilabel classification and language models. The results of extensive experiments show that our proposed method outperforms the current state-of-the-art methods on the RSICD, the UCM-captions, and the Sydney-captions datasets. However, one limitation of this work is that our method cannot generate fine-grained RSI captions based on practical application scenarios, such as road traffic management, disaster assessment, and so on, since the annotations of current RSI captioning datasets are coarse-grained. In the future, we will use fine-grained RSI classification datasets and corpus data to provide detailed prior information for the RSI captioning task to generate more detailed image captions, which enables the RSI captioning model to provide more helpful information for practical application scenarios. We also want to implement our method on MindSpore, a new deep-learning computational framework. These problems are left for future work.

APPENDIX

A. Proof for (8) in This Article

Here, we provide the proof of (8) in this article. The proof of $(\partial y_m/\partial y_a) = \lim_{k\to\infty} \operatorname{softmax}(ky_a)$ in (8) is as follows:

$$\frac{\partial y_m}{\partial y_a^q} = \lim_{k\to\infty} \frac{1}{k} \frac{\partial \log \sum_{\gamma=1}^{\Gamma} \exp\left(ky_a^\gamma\right)}{\partial y_a^q} = \frac{\exp\left(ky_a^q\right)}{\sum_{\gamma=1}^{\Gamma} \exp\left(ky_a^\gamma\right)}. \tag{36}$$

Therefore, we can get the following derivation:

$$\frac{\partial y_m}{\partial y_a} = \left[\frac{\exp\left(y_a^1\right)}{\sum_{\gamma=1}^{\Gamma} \exp\left(ky_a^\gamma\right)}, \frac{\exp\left(y_a^2\right)}{\sum_{\gamma=1}^{\Gamma} \exp\left(ky_a^\gamma\right)}, \ldots, \frac{\exp\left(y_a^q\right)}{\sum_{\gamma=1}^{\Gamma} \exp\left(ky_a^\gamma\right)}\right] = \operatorname{softmax}(ky_a). \tag{37}$$

The proof of $\lim_{k\to\infty} \operatorname{softmax}(ky_a) = \operatorname{onehot}(\operatorname{argmax}(y_a))$ in (8) is as follows. Assuming that $y_a' = [y_a^1, y_a^2, \ldots, y_a^{|\Gamma|}] - \max(y_a)$, we can get the following expression:

$$\exp\left(ky_a'\right) = \left[\exp\left(ky_a^1 - k\max(y_a)\right), \exp\left(ky_a^2 - k\max(y_a)\right), \ldots, \exp\left(ky_a^{|\Gamma|} - k\max(y_a)\right)\right]. \tag{38}$$

When k tends to ∞, we can get the following expression:

$$\lim_{k\to\infty} \exp(ky_a^{'p}) = \begin{cases} 0, & p \neq \operatorname{argmax}(y_a) \\ 1, & p = \operatorname{argmax}(y_a). \end{cases} \tag{39}$$

Due to $\operatorname{softmax}(ky_a) = \operatorname{softmax}(ky_a')$, we have

$$\begin{aligned} \lim_{k\to\infty} \operatorname{softmax}(ky_a) &= \lim_{k\to\infty} \operatorname{softmax}\left(ky_a'\right) \\ &= \lim_{k\to\infty} \left[\frac{\exp\left(ky_a^1 - k\max(y_a)\right)}{\sum_{l=1}^{|\Gamma|} \exp\left(ky_a^l - k\max(y_a)\right)}, \ldots, \frac{\exp\left(ky_a^{|\Gamma|} - k\max(y_a)\right)}{\sum_{l=1}^{|\Gamma|} \exp\left(ky_a^l - k\max(y_a)\right)}\right] \\ &= \operatorname{onehot}(\operatorname{argmax}(y_a)). \end{aligned} \tag{40}$$

B. Discussion on Pretraining Dataset

Most previous works on RSI captioning used CNNs pretrained on the ImageNet dataset, but natural images in ImageNet are very different from RSIs. Therefore, we set up the

case of CNN pretrained on the NWPU-RESISC45 dataset, a remote sensing scene classification dataset, to compare with the case of CNN pretrained on the ImageNet dataset, and the results are shown in Table IX.

The comparison results in Table IX show that CNN pretrained on NWPU-RESISC45 can achieve better results than those pretrained on ImageNet. It proves that pretraining on an RSI dataset is more helpful for RSI captioning than pretraining on a natural image dataset.

REFERENCES

[1] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," *ACM Comput. Surveys*, vol. 51, no. 6, pp. 1–36, Nov. 2019.

[2] C. T. Recchiuto and A. Sgorbissa, "Post-disaster assessment with unmanned aerial vehicles: A survey on practical implementations and research approaches," *J. Field Robot.*, vol. 35, no. 4, pp. 459–490, Jun. 2018.

[3] Q. Liu, C. Ruan, S. Zhong, J. Li, Z. Yin, and X. Lian, "Risk assessment of storm surge disaster based on numerical models and remote sensing," *Int. J. Appl. Earth Observ. Geoinformation*, vol. 68, pp. 20–30, Jun. 2018.

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

[5] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2015, pp. 3156–3164.

[6] J. L. Elman, "Finding structure in time," *Cognit. Sci.*, vol. 14, no. 2, pp. 179–211, Mar. 1990.

[7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[8] B. Qu, X. Li, D. Tao, and X. Lu, "Deep semantic understanding of high resolution remote sensing image," in *Proc. Int. Conf. Comput., Inf. Telecommun. Syst. (CITS)*, Jul. 2016, pp. 1–5.

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

[10] X. Zhang, X. Wang, X. Tang, H. Zhou, and C. Li, "Description generation for remote sensing images using attribute attention mechanism," *Remote Sens.*, vol. 11, no. 6, p. 612, 2019.

[11] B. Wang, X. Zheng, B. Qu, and X. Lu, "Retrieval topic recurrent memory network for remote sensing image captioning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 256–270, 2020.

[12] J. Chen, Y. Han, L. Wan, X. Zhou, and M. Deng, "Geospatial relation captioning for high-spatial-resolution images by using an attention-based neural network," *Int. J. Remote Sens.*, vol. 40, no. 16, pp. 6482–6498, Aug. 2019.

[13] Q. Wang, W. Huang, X. Zhang, and X. Li, "Word—Sentence framework for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10532–10543, Dec. 2020.

[14] Z. Zhang, W. Diao, W. Zhang, M. Yan, X. Gao, and X. Sun, "LAM: Remote sensing image captioning with label-attention mechanism," *Remote Sens.*, vol. 11, no. 20, pp. 1–15, 2019.

[15] P. Anderson et al., "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1–10.

[16] X. Li, X. Zhang, W. Huang, and Q. Wang, "Truncation cross entropy loss for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5246–5257, Jun. 2020.

[17] Y. Li, S. Fang, L. Jiao, R. Liu, and R. Shang, "A multi-level attention model for remote sensing image captions," *Remote Sens.*, vol. 12, no. 6, p. 939, Mar. 2020.

[18] Y. Li et al., "Recurrent attention and semantic gate for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 13, 2021, Art. no. 5608816.

[19] Z. Zhang, W. Zhang, M. Yan, X. Gao, K. Fu, and X. Sun, "Global visual feature and linguistic state guided attention for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022.

[20] G. Hoxha and F. Melgani, "A novel SVM-based decoder for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.

[21] G. Sumbul, S. Nayak, and B. Demir, "SD-RSIC: Summarization-driven deep remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 6922–6934, Aug. 2020.

[22] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," in *Proc. Assoc. Comput. Linguist.*, 2017, pp. 1073–1083.

[23] R. Zhao, Z. Shi, and Z. Zou, "High-resolution remote sensing image captioning based on structured attention," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.

[24] J. Uijlings, K. van de Sande, and T. Gevers, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, pp. 154–171, Oct. 2013.

[25] B. Zhao, "A systematic survey of remote sensing image captioning," *IEEE Access*, vol. 9, pp. 154086–154111, 2021.

[26] F. Sammani and L. Melas-Kyriazi, "Show, edit and tell: A framework for editing image captions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4808–4816.

[27] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.

[28] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst. (GIS)*, 2010, pp. 270–279.

[29] F. Zhang, B. Du, and L. Zhang, "Saliency-guided unsupervised feature learning for scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 2175–2184, Apr. 2015.

[30] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2002, pp. 311–318.

[31] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. 2nd Workshop Stat. Mach. Transl. (StatMT)*, 2005, pp. 65–72.

[32] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Proc. Assoc. Comput. Linguistics*, 2004, pp. 74–81.

[33] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4566–4575.

[34] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: Semantic propositional image caption evaluation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 382–398.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[36] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *Proc. Conf. Empir. Methods Nat. Lang. Process.*, 2014, pp. 1532–1543.

[37] L. Liu et al., "On the variance of the adaptive learning rate and beyond," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–13.

[38] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.

[39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015, *arXiv:1409.1556*.

[40] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[41] B. Wang, X. Lu, X. Zheng, and X. Li, "Semantic descriptions of high-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1274–1278, Aug. 2019.

[42] X. Lu, B. Wang, and X. Zheng, "Sound active attention framework for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 1985–2000, Mar. 2020.

[43] H. Jegou, M. Douze, C. Schmid, and P. Perez, "Aggregating local descriptors into a compact image representation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3304–3311.

[44] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2003.

[45] K. Xu et al., "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Mach. Learn. Res.*, vol. 37, Jun. 2015, pp. 2048–2057.

Xiutiao Ye received the B.E. degree in intelligent science and technology and the M.Eng. degree in electronics and communication engineering from Xidian University, Xi'an, China, in 2015 and 2018, respectively, where he is currently pursuing the Ph.D. degree in electronic science and technology with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China.

His research interests include deep learning, remote sensing image captioning, and radar image interpretation.

Shuang Wang (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in circuits and systems from Xidian University, Xi'an, China, in 2000, 2003, and 2007, respectively.

She is currently a Professor with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China, Xidian University. Her research interests include sparse representation, image processing, synthetic aperture radar (SAR) automatic target recognition, remote sensing image captioning, and polarimetric SAR data analysis and interpretation.

Yu Gu (Member, IEEE) received the B.S. degree in computer science from Xidian University, Xi'an, China, in 2005, the M.S. degree in computer software and theory from Guangxi University, Nanning, China, in 2012, and the Ph.D. degree in artificial intelligence from Tilburg University, Tilburg, The Netherlands, in 2018.

He is currently a Lecturer with the Department of Artificial Intelligence, Xidian University. His research interests include cognitive computing, speech emotion recognition and audio signal processing, target recognition, remote sensing image captioning, and polarimetric SAR data analysis and interpretation.

Jihui Wang received the B.S. and M.S. degrees from Xidian University, Xi'an, China, in 2019 and 2022, respectively.

His research interests include deep learning, continual learning, and image captioning.

Ruixuan Wang received the B.S. degree from Xidian University, Xi'an, China, in 2020, where he is currently pursuing the M.S. degree with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China.

His research interests include deep learning, computer vision, and image captioning.

Biao Hou (Member, IEEE) received the B.S. and M.S. degrees in mathematics from Northwest University, Xi'an, China, in 1996 and 1999, respectively, and the Ph.D. degree in circuits and systems from Xidian University, Xi'an, in 2003.

Since 2003, he has been with the Key Laboratory of Intelligent Perception and Image Understanding of the Ministry of Education, Xidian University, where he is currently a Professor. His research interests include compressive sensing and synthetic aperture radar image interpretation.

Fausto Giunchiglia received the B.S., M.S., and Ph.D. degrees from the University of Genoa, Genoa, Italy, in 1981, 1983, and 1987, respectively.

He is currently a Full Professor of computer science with the University of Trento, Trento, Italy. His research interests include artificial intelligence, computational models of the mind, and implications on how the known is grounded in the unknown.

Licheng Jiao (Fellow, IEEE) was born in Shaanxi, China, in 1959. He received the B.S. degree from Shanghai Jiao Tong University, Shanghai, China, in 1982, and the M.S. and Ph.D. degrees from Xi'an Jiaotong University, Xi'an, China, in 1984 and 1990, respectively.

From 1984 to 1986, he was an Assistant Professor with the Civil Aviation Institute of China, Tianjin, China. From 1990 to 1991, he was a Post-Doctoral Fellow with the Key Laboratory for Radar Signal Processing, Xidian University, Xi'an, where he is currently the Director of the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China. His research interests include signal and image processing, nonlinear circuits and system theories, learning theory and algorithms, optimization problems, wavelet theory, and machine learning.