remote sensing

MDPI

Article

A Patch-Level Region-Aware Module with a Multi-Label Framework for Remote Sensing Image Captioning

Yunpeng Li [1,2], Xiangrong Zhang [3], Tianyang Zhang [3,*], Guanchun Wang [3], Xinlin Wang [3] and Shuo Li [3]

[1] Jiangsu Province Engineering Research Center of Photonic Devices and System Integration for Communication Sensing Convergence, Wuxi University, Wuxi 214105, China; ypli@cwxu.edu.cn
[2] Jiangsu Province Engineering Research Center of Integrated Circuit Reliability Technology and Testing System, Wuxi University, Wuxi 214105, China
[3] Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, School of Artificial Intelligence, Xidian University, Xi'an 710071, China; xrzhang@mail.xidian.edu.cn (X.Z.); wangguanchun@xidian.edu.cn (G.W.); wangxinlin@xidian.edu.cn (X.W.); lishuo@xidian.edu.cn (S.L.)
* Correspondence: zhangtianyang@xidian.edu.cn

Abstract: Recent Transformer-based works can generate high-quality captions for remote sensing images (RSIs). However, these methods generally feed global or grid visual features to a Transformer-based captioning model for associating cross-modal information, which limits performance. In this work, we investigate unexplored ideas for remote sensing image captioning task, using a novel patch-level region-aware module with a multi-label framework. Due to an overhead perspective and a significantly larger scale in RSIs, a patch-level region-aware module is designed to filter the redundant information in the RSI scene, which benefits the Transformer-based decoder by attaining improved image perception. Technically, the trainable multi-label classifier capitalizes on semantic features as supplementary to the region-aware features. Moreover, modeling the inner relations of inputs is essential for understanding the RSI. Thus, we introduce region-oriented attention, which associates region features and semantic labels, omits the irrelevant regions to highlight relevant regions, and learns related semantic information. Extensive qualitative and quantitative experimental results show the superiority of our approach on the RSICD, UCM-Captions, and Sydney-Captions. The code for our method will be publicly available.

Keywords: remote sensing image captioning; salient regions; multi-label classification; multi-head attention

check for updates

Citation: Li, Y.; Zhang, X.; Zhang, T.; Wang, G.; Wang, X.; Li, S. A Patch-Level Region-Aware Module with a Multi-Label Framework for Remote Sensing Image Captioning. *Remote Sens.* **2024**, *16*, 3987. https://doi.org/10.3390/rs16213987

Academic Editor: Jon Atli Benediktsson

Received: 6 September 2024
Revised: 18 October 2024
Accepted: 20 October 2024
Published: 27 October 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Generating a sentence about a remote sensing image (RSI), referred to as remote sensing image captioning (RSIC), requires a comprehensive cross-modality understanding and visual-semantic reasoning between RSIs and human-annotated captions. Thus, the RSIC task is connected with vision and natural language communities, which plays an important role in understanding RSIs and has a wide range of potential applications. Therefore, it has aroused extensive interesting. With the explosive progress of deep neural networks [1], a large amount of encoder–decoder networks have achieved notable successes in RSIC.

For an encoder–decoder structure in RSIC, Convolutional Neural Networks (CNNs) generally play the role of encoder and Long Short-Term Memory (LSTM) [2] is built as a decoder. Zhang et al. [3] employed a powerful CNN as an encoder and LSTM performed sequence modeling. The emergence of the attention mechanism promotes the development of attention-based RSIC methods. Lu et al. [4] firstly integrated the CNN-LSTM paradigm with an attention mechanism, realizing dynamic weighted visual regions via the queries of LSTM. Due to the outstanding effects of the attention mechanism, the diverse attention mechanisms have been widely adopted in RSIC models. A novel attribute attention mechanism [5] aimed to extract semantic features as guidance information, and

the attribute-driven regional features have efficiently improved the performance of existing algorithms. Semantic features are also focused by multi-level attention [6], which was proposed with three attention layers, namely, visual attention, semantic attention, and cross-modal attention.

1.1. Challenges in RSIC Tasks

In contrast to natural images, RSIs are taken from a high altitude and capture a significantly larger scale scene [7]. Traditional RSI tasks, such as scene classification [8,9], scene segmentation [10–12], object detection [13,14], and change detection [15,16], are primarily used for pixel-level and object-level visual interpretation. However, RSIC is devoted to a comprehensive understanding of multi-scale, multi-directional semantic content [17,18] in RSI. There are still some difficulties faced in RSIC tasks originating from (1) the complex objects and background information in an RSI, (2) the semantic prior knowledge, and (3) the overlooked inner relationships and cross-modal interaction.

Regarding the first difficulty, a CNN [19] is an effective feature mapping block that learns diversity information in RSIs. It may not perform well on such images. This is because the CNN is migrated from the natural image classification task on the ImageNet dataset. To remedy this defect, Shen et al. [20] proposed a Two-stage Multi-task Learning Model (VRTMM), which consists of a CNN encoder and a Transformer-based decoder (see Figure 1a). Note that their CNN encoder is finetuned on an RSI classification dataset. For expressing the multi-scale information [21,22], Li et al. [23] leveraged an efficient spatial pyramid with different dilated convolutions to model the subtle multi-scale information. Zhang et al. [24] expected higher vision quality; the improved global visual feature-guided attention can remove out the redundant information in RSI. Zhao et al. [25] considered the superiority of local region features compared with pixel-level features, combining obtained segmentation proposals with CNN features for object-level structured features. Due to numerous objects in the image, the structured features are still inefficient for association with the language model. Thus, how to further derive the core regional features in RSIs is still under-explored.

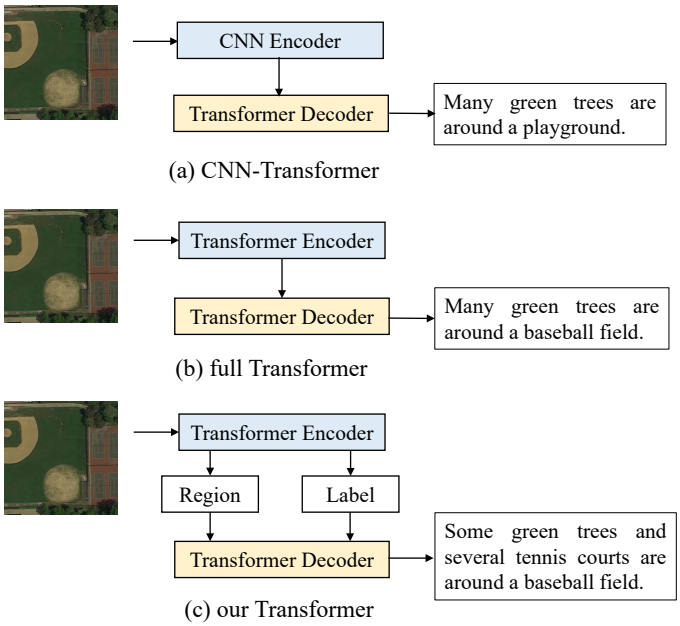


Figure 1. Comparison of different RSIC models. (**a**) The CNN-Transformer pipeline. (**b**) The full Transformer pipeline. (**c**) Our proposed Transformer model.

Regarding the second difficulty, although some attribute-assisted methods achieve remarkable progress, they heavily depend on a pre-trained scene or a multi-label classi-

fication task for the semantic embedding vectors, while the image visual features may be interrelated with labels from the unknown classes. For example, Zhang et al. [26] directly predicted a specific scene label by a single-label classification branch, and the label attention mechanism (LAM) reduced the difficulty of learning semantic information in the RSI. Inspired by the LAM model, some researchers were not satisfied with one label, containing less semantic information. Wang et al. [27] introduced multi-label classification into two-stage RSIC and further entangled a multi-label semantic feature with pixel-level or patch-level features. Considering that the detected labels are not the only element in the vocabulary, Huang et al. [28] proposed a word classification network to generate semantic words in ground-truth sentences. The sequence-level words serve as a more reasonable signal for the caption inference stage. However, these works transfer visual features to high-level semantic information with an individual network, and the phenomenon of semantic ambiguity appeared with unseen RSI content. In this study, the "attribute" and "label" represent an object's class.

Regarding the third difficulty, existing CNN-LSTM methods often directly integrate visual features (or fused features) with simple attention mechanisms, and the LSTM-based decoder cannot well capture the long-range temporal correlations for all words in a sentence. More recently, Transformer-based architecture, excellent in various tasks, has been introduced into RSIC tasks. It attempts to explore more meaningful prior knowledge, e.g., inner relation about visual features and generated words, and cross-modal interaction. For example, a novel MEmory-guided Transformer was adopted by Gaurav et al. [29]; an improved memory-guided block was contained in a multi-layer decoder, which could correlate any distinct visual and linguistic information. Similarly, Zia et al. [30] employed a CNN-Transformer framework for RSIC, namely CNN-T. Furthermore, Gou et al. [31] applied a full Transformer network; patch-level features had content information in RSIs, and a mask-guided decoder improved the language generation steps. A showcase is shown in Figure 1b. Although these models could capture the long-range correlations with the inherent advantages of the Transformer structure, they only consider the correlation of query and key modalities, and ignore the interaction with other modalities.

1.2. Possible Solution Based on a Full Transformer

To deal with the above-mentioned limitations and further explore a full Transformer RSIC framework, we propose a novel patch-level region-aware module with a multi-label network. In specific, the proposed work is a full Transformer-based encoder–decoder architecture for RSIC tasks, which comprises many stacked identical layers in contrast to the CNN-LSTM structure. Figure 1c represents a broad overview of our approach. In the encoder, the raw RSI is firstly converted into a sequence of patches. Then the patches are processed through the Transformer-based encoder to generate feature representation along with class tokens. The Transformer-based features derive the interaction with all of the patches, which can afford some simple and useful prior information. Due to the large variation of objects in RSI, redundant information will be incurred in encoded features. The designed patch-level region-aware module selects effective regions with position information for each patch-level features. It is worth mentioning that salient areas also retain the information of other abandoned patches. To solve the second issue, the class tokens are passed through a multi-label classifier to infer core objects with labels (e.g., noun forms), in which every label is certain. Moreover, it is more conducive to provide prior knowledge with high semantic relevance for our one-stage RSIC method. In addition, the Transformer-based decoder associates patch-level region-aware features and predicted label embeddings with region-oriented attention. For replacing the tucker-fused features, the salient features and label embedding are fed into the region-oriented attention module separately. The experiments, validating using popular RSICD, UCM-Captions, and Sydney-Captions datasets, shows competitive performance compared with related state-of-the-art (SOTA) methods.

1.3. Novel Contributions

Our novel contributions on the literature for RSIC are four-fold:

- We propose a novel patch-level region-aware module with a multi-label framework for RSIC by unifying both the detected object's class and patch-level salient features. Furthermore, the expanded Transformer-based decoder with a region-oriented attention block enhances the cross-modal association learning and affords supplementary hints for query-to-key correlation.
- To extract visual features for multi-scale RSIs, RSI patches are encoded by a Transformer-based encoder for patch-level features and class tokens. Meanwhile, a patch-level region-aware module is designed to seek the core object features guided by the relationships between patch-level features and class tokens, which decides to replace the image's global or redundant representation.
- Rejecting pre-trained algorithm migration for multi-label classification, we directly apply a multi-label classifier on the class token features from the Transformer-based encoder. It can alleviate the negative effects of the potentially noisy labels during the whole model training phase compared with pre-trained algorithms.
- The integration of regional features and semantic features into a novel region-oriented attention block aims to capture more cross-modal interactional information, which is crucial for accurately characterizing the complex content of RSIs and thus for achieving accurate sentence predictions.

2. Related Work

RSIC is an immature sequence-generation task, which has attracted increasing research attention, and researchers have also proposed many methods with promising performance. Recent advances in RSIC focus on the problems of feature extraction, multimodal information interaction, and reasoning ability. According to whether the LSTM structure is utilized in the language model, current methods for RSIC can be generally divided into two types: (1) LSTM-based methods and (2) Transformer-based paradigms. Representative methods are introduced in Sections 2.1 and 2.2, respectively.

2.1. LSTM-Based RSIC Methods

Nowadays, the CNN-LSTM pipeline is still a popular framework in RSIC tasks. The RSI is represented by the fixed-size grid features extracted from a CNN model, such as VGG16 or ResNet101 [19]. However, not all features of the image are related to the query of the LSTM-based decoder, while some of them should be filtered out before generating the embedding vector. Therefore, the attention mechanism is introduced to tackle this problem. The early attempt made by Lu et al. [4] was to firstly detect the most relevant image regions via attention mechanism and then utilized an LSTM to form a sentence. Furthermore, Li et al. [6] constructed visual-level attention on spatial features, semantic-level attention on the generated sentence fragment, and cross-modal attention on two obtained attentive vectors. Another novel approach [32] attempted the role of sound information and validated the performance of the audio features combined with visual features at each decoding stage. This method aims at fixing the class-dependent label in the audio data. For directly extracting highly abstract semantic features from the RSI, Zhang et al. [5] extracted the classification vectors in the CNN as attribute vectors and then used attribute attention to process the CNN-based spatial features, yielding a set of attribute-refined features. Subsequently, the LAM [26] employed a pre-trained scene classifier for scene class prior probabilities, which were sent to the attention mechanisms. When the RSI is annotated with a single label, there is only one regional type associated with the scene label assignment. Thus, Wang et al. [27] proposed MLSFF with a teacher model for multi-label classification, which casts on learning major specific labels for objects and constructs importance-aware fusion features. Due to the large scale for RSIs, a multi-scale feature extractor was created in [33] to alleviate the scale-diversity problem, and recurrent attention was employed with a semantic gate [23] for balancing different contributions between multi-modal features

while predicting different words. Yang et al. [34] concentrated on hierarchical features and proposed effective cross-modal feature alignment, which outperformed the inefficient utilization of visual texture and semantic features. Additionally, regional features were used by Zhao et al. [25] via a segmentation proposal generation module, in which the structured attention calculates weights on provided regional information. Zhang et al. [24] made contributions on filtering out redundant information with global feature guidance, and the global information generally learned the essence of content in RSIs. The linguistic guidance in LSTM-based decoder modeled the visual–textual attention process. To overcome RSIC data insufficiency, Yang et al. [35] introduced meta learning into the RSIC task, and the improved performance depended on a superior feature extractor with multi-stage training. It can alleviate the issue of insufficient datasets (e.g., Sydney-Captions), so that the feature extractor benefits from meta learning and reducing the number of required resources. Theoretically, the attention mechanism allows the model to automatically select the areas that are most relevant to the output words.

2.2. Transformer-Based RSIC Methods

Inspired by the achievements of the Transformer structure in computer vision (CV) and natural language processing (NLP), several Transformer-based networks have been utilized in the RSIC task. The VRTMM [20] was a CNN-Transformer architecture, which encodes image content with the pre-trained CNN, producing convincing results on RSICD by adopting a Transformer-based decoder. Following the technical route of the Transformer [36] in neural machine translation, a word-driven Transformer (WDT) [28] firstly learned a series of semantic words through a word generator, and the Transformer-based encoder then concurrently read input words and outputted semantic features, which serves as an encoded information source to the Transformer-based decoder. Thereafter, CapFormer [37] was designed with a pure Transformer architecture, including a vision Transformer and a Transformer-based decoder. However, those tasks are focused on obtaining superior discriminative visual features. A global–local captioning model (GLCM) [38] introduced both global and local features into the RSIC model, and their attention-based decoding network was a Transformer block including self-attention and co-attention. A pixel-based analysis does not model the spatial relationships of objects in the input image and their arrangements and do not highlight the semantic meaning. In the Mask-Guided Transformer (MGT) network [31], global semantic information from the topic token highly abstracted semantic features, which shows a great potential in Transformer-based decoder. And, the plane to hierarchy (P-to-H) [39] used selective search to detect visual and semantic maps and two features were established connection via deformable Transformer, modeling multi-scale features and performing intraclass interactive learning. With the popularity of the CNN-based feature extractor, Zhuang et al. [40] proposed the Transformer-based encoder–decoder combined with grid features to improve RSIC performance. Gajbhiye et al. [29] proposed a memory-guided Transformer as a linguistic decoder to decode multi-attentive features [41], and the input of the Transformer decoder was spatial and channel attentive features from CNN features. Considering the multi-scale features by Zia et al. [30] for the RSI, topic-sensitive word embedding was also auxiliary information to capture the polysemous nature of words in training captions.

Therefore, we took advantage of the full Transformer framework and designed new modules to solve focused problems in RSIC tasks. For instance, there is a large gap among different types of objects in RSIs. The overall structure of the proposed network is shown in Figure 2. In our model, the patch-level region-aware module is proposed to aggregate more valuable features about objects in different positions for expected regional features. In addition, the class token from the Transformer-based encoder is fed into the co-training multi-label classifier for achieving accurate land-cover classes, which can infer seen object classes and provide the semantic expression of RSIs. Furthermore, the Transformer-based decoder uses a region-oriented attention module to aggregate cross-modal attention cues.

Through the integration between different modalities, the network can be guided to learn the high-level cross-modal relations to reasoning the sentence in a more plausible way.

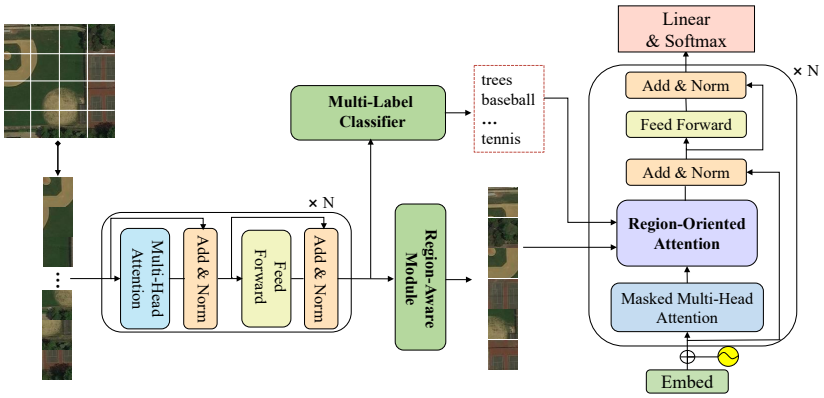


Figure 2. Overall flowchart of our proposed method.

3. Proposed Method

We firstly describe the pure Transformer framework for RSIC in Section 3.1, and then elaborate on our proposed methodology in Sections 3.2–3.4.

3.1. Pure Transformer Framework

Given an image I with its corresponding ground truth caption $y_{1:T}$, which is represented as a series of words $\{y_1, y_2, \ldots, y_{t-1}\}$, using the pre-trained Vit [42] to encode the RSI and a Transformer as the decoder, the output from Vit can be written as:

$$V_T = Vit(I) \tag{1}$$

it should be noted that V_T contains two different features: the patch-level visual feature and the class-level semantic feature. Let $V_T = [C, v_1, v_2, \ldots, v_m]$, where C is the class token, v_i is the features for patch I_i , and m denotes the total number of patches.

As for the decoder, the positional embedding firstly is added for word embedding. Then, it will be six stacked identical layers in the Transformer's decoder. Each layer is composed of masked multi-head attention (m-MHA), multi-head cross attention (MCA), and a feed-forward network (FFN) sequentially. There is an AddNorm between every sublayer. The m-MHA and MCA rely on a multi-head attention (MHA) module by calibrating input features as query (Q), key (K), and value (V), which is defined as follows:

$$Head_i = \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d}}\right) V_i \tag{2}$$

$$MHA(Q, K, V) = Concat(Head_1, \ldots, Head_H) W_s \tag{3}$$

where Q_i, K_i, and V_i represent query, key, and value vectors in $head_i$, and W_s is a learnable matrix for projecting the concatenation outputs of all heads. At the t time step, the predicted word's probability distribution y_t can be obtained using linear projection and the softmax function. The equation is as follows:

$$y_t = \text{softmax}\left(W_o O^t\right) \tag{4}$$

where W_o is a trainable linear layer. Finally, the pure Transformer model maximizes the cross entropy loss (XE) of each time step t, which is obtained using the chain rule:

$$L_{XE} = \frac{1}{T} \sum_{t=1}^{T} -\log\left(p_t^\theta(y_t | y_{1:t-1}, I)\right) \tag{5}$$

where θ denotes the parameters of the model.

3.2. Patch-Level Region-Aware Extractor

How to derive regional features in RSIs is still under-explored. To capture regional features, the current method [25] primarily employs segmentation proposal features with an attention mechanism. However, the structured attention does not take the visual relationships into account, and segmentation proposals are not always robust with ambiguous semantic optimization. We also hope to distill an effective target area for further semantic matching and inferring. Considering that the series of image patches can effectively represent the object information contained in the RSI.

Generally, the pixel-level visual features would be coarse and not with position information. How to use location to accurately locate object area is a interesting problem. In our work, a simple method is to interact the class token with all patch-level features for most object-level irrelevant information, preserving the information of the different scale areas as much as possible. Specifically, the patch-level region-aware module consists of the single Transformer sub-layer and region positioning, as shown in Figure 3. The input is the product V_T and embedded into three different sets of vectors (Q_v , K_v, and V_v), which are fed through the MHA layer. The similarity scores between queries Q_{vi} and keys K_{vi} in $head_i$ are calculated as follows:

$$s_i = \text{softmax}\left(\frac{Q_{vi}K_{vi}^T}{\sqrt{d}}\right) \tag{6}$$

$$S = Concat(s_1, \ldots, s_H) \tag{7}$$

where S denotes the similarity of queries Q_v and keys K_v , and d is the scaling factor. The output of the Transformer block in patch-level region-aware module is defined as:

$$V_s = AddNorm(FFN(AddNorm(SV_T))) \tag{8}$$

where $V_s = [C_s, v_{s1}, v_{s2}, \ldots, v_{sm}]$. Let $V_I = [v_{s1}, v_{s2}, \ldots, v_{sm}]$. $AddNorm(\cdot)$ denotes the composition of the residual connection and normalization. After that, S and V_I are input to the region positioning operation. Firstly, the s_i is composed of $m+1$ feature vectors, which could be expressed as:

$$s_i = \left[s_{i(0)}, s_{i(1)}, \ldots, s_{i(m)}\right] \tag{9}$$

where the number of vectors $m+1$ is based on the number of elements in V_T. However, the first element in $s_{i(0)}$ denotes similarity scores of $\langle class, class \rangle$. The similarity scores s_i' of $\langle class, region \rangle$ is expressed as:

$$s_i' = \left[s_{i(1)}, \ldots, s_{i(m)}\right] \tag{10}$$

$$S' = \frac{1}{H}\sum_{i=1}^{H} s'_i \tag{11}$$

where S' represents the averaged similarity scores of $\langle class, region \rangle$ for all heads. Then, we sort the weight values of S' for Top-K weights along with regional indexes and apply K indexes with V_I to obtain the salient regional map V_R.

3.3. Co-Training Multi-Label Classifier

When extracting specific semantic representation, a pre-trained multi-label classification network is exploited to capture the word-level representations. However, multi-label noise can distort the learning process of the visual–semantic relationship, which can be associated with wrong or non-existent labels. To solve this issue, an extra multi-label classifier layer projects the class token to semantic labels and utilizes the ground-truth labels to train the extra multi-label classifier in an end-to-end system.

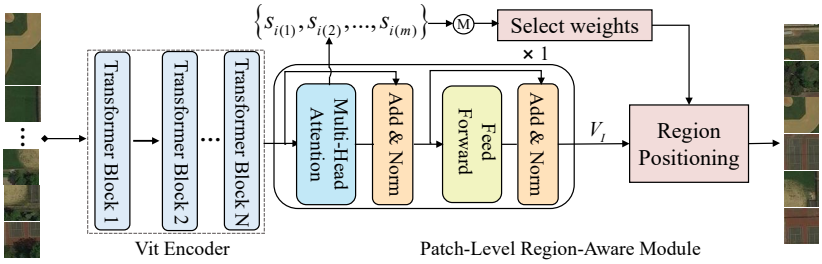


Figure 3. The architecture for the patch-level salient region module, which is leveraged to select patch-level regions to form effective object-level features. It consists of the single Transformer-based sub-layer and region positioning. "M" represents mean operation.

As we know, semantic labels are the keywords in the generated sentence. Therefore, taking the ground-truth labels as guidance, the multi-label classifier can be capable of generating correct semantic classes presented in RSIs. However, tagging RSIs with multi-labels is time-consuming, complex, and costly in operational scenarios. In details, we filter the visual words ("river", "airplane", etc.) as ground-truth labels, which are extracted from ground-truth captions. Then we select the top-ranked words to form the label set. The selected labels and other words with frequencies jointly construct the whole vocabulary. Inspired by [5], the class features C are passed through the multi-label classifier for class prediction. The output of multi-label classifier is associated with the label probability distribution $\mathbf{p}$ over ground-truth labels, which is defined as:

$$\mathbf{p} = \sigma\Big((W_a A_l)^T \otimes W_c C\Big) \tag{12}$$

where W_a and W_c are trainable parameters, A_l denotes the predicted label feature, $\otimes$ denotes the matrix multiplication, and $\sigma(\cdot)$ is the sigmoid function. In general, a given RSI can be associated with multiple semantic labels. Thus, several labels with top-D probability in $\mathbf{p}$ are considered as semantic classes. Thus, the top-D words $L = \{l_1, l_2, \ldots, l_D\}$ can be directly used by the Transformer-based decoder. The focal loss (FL) [43] is adopted for optimizing multi-label classifier throughout our network training, which is defined as follows:

$$L_{FL} = \frac{1}{D}\sum_{d=1}^{D}\left[\delta(1-p_d)^{\gamma}\log(p_d) + (1-\delta)p_d^{\gamma}\log(1-p_d)\right] \tag{13}$$

where p_d is the probability for the d-th label, and δ and γ are the focusing parameters for the positive labels.

3.4. Region-Oriented Attention

After generating the patch-level salient features and the meaningful labels, it may be careless to fuse the visual and label features directly, as in previous works [27]. Such fused features have a serious impact on subsequent processing and analysis in the decoder due to the degraded transcription performance. To tackle this issue, a region-oriented attention layer is designed, which balances the different contributions on region-aware features and semantic embeddings while predicting different words.

Our caption generator consists of a stack of N identical Transformer-based layers. The initial layer receives the sequence of embedded target caption with positional encoding. Each layer has a masked MHA over the sentence words and the region-oriented attention module between label embeddings and salient visual features in this order, followed by the FFN layer. The masked MHA layer at the l-th layer receives an input sequence embedding Y^t at time step t , yielding M_l^t. Moreover, the region-oriented attention module performs on the cross-modal embeddings in parallel. As shown in Figure 4, it consists of two parts: label attention and regional ones, where the former aims to find the associated label by

using inter-semantic information, and the latter captures the task-relevant parts based on obtaining salient regional features. Herein, their details are as follows.

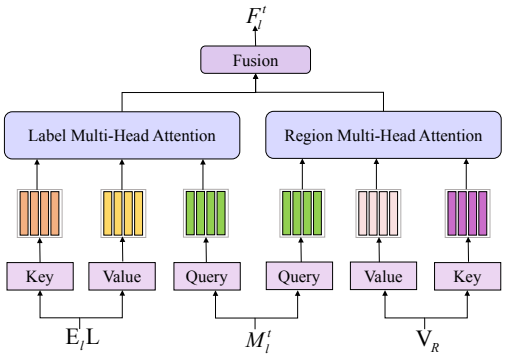


Figure 4. An illustration of the region-oriented attention module.

Label multi-head attention (LMHA) is capable of supplying semantic guidance to carry out reasonable cross-modality inference, therefore correcting misclassification and inconsistently parsing results. According to the fact that some objects may have a stronger connection, while others may not. Therefore, a different treatment for each label embedding is necessary for exploring the object's relationships. In this block, LMHA is computed with M_l^t as $Q_{l,L}^t$ and the label embeddings $E_l L$ as $K_{l,L}^t$ and $V_{l,L}^t$, yielding attended features L_l^t, which is defined as:

$$L_l^t = LMHA\left(Q_{l,L}^t, K_{l,L}^t, V_{l,L}^t\right) \tag{14}$$

where E_l is a matrix for embedding labels into semantic features.

Region multi-head attention (RMHA) allows the model to consider which patch-level region to omit or focus on by inter-visual relation information. Not all salient regions are equally important for multi-level information transmission in the Transformer-based decoder. Based on this finding, the RMHA module is utilized to find the task-relevant regional features by using inter-regional information. In this block, RMHA is computed with M_l^t as $Q_{l,R}^t$ and with the regional features V_R as $K_{l,R}^t$ and $V_{l,R}^t$, yielding attended features V_l^t, which is defined as:

$$V_l^t = RMHA\left(Q_{l,R}^t, K_{l,R}^t, V_{l,R}^t\right) \tag{15}$$

Next, an adaptive fusing operation "add" L_l^t and V_l^t, which is defined as:

$$\mu = sigmoid\left(W_\mu\left(\frac{L_l^t + V_l^t}{2}\right)\right) \tag{16}$$

$$F_l^t = \mu L_l^t + (1-\mu)V_l^t \tag{17}$$

where μ is a tuning parameter controlling the fusion ratio on L_l^t and V_l^t, and W_μ are learnable parameters. The F_l^t denotes the fused cross-modal features. Finally, F_l^t is fed to the FFN layer, obtaining the resultant vectors O_l^t as follows:

$$O_l^t = AddNorm\left(FFN\left(AddNorm\left(F_l^t\right)\right)\right) \tag{18}$$

For improved training and optimizing the designed model, total loss is defined as follows:

$$L = L_{XE} + \lambda L_{FL} \tag{19}$$

where the hyper-parameters λ are set with a value of 0.2.

4. Experiments

4.1. Datasets

The three RSIC datasets were randomly split into 80% for training, 10% for validating, and 10% for testing. **RSICD** [4] contains 10,921 images with a size of 224 × 224 pixels. In addition, each image is marked with one of 30 categories. **UCM-Captions** [44] contains 21 categories with 2100 images. The size of each image is 256 × 256 pixels. **Sydney-Captions** [44] is the smallest RSIC dataset and comprises 613 images with a size of 500 × 500 pixels, segregating into 7 categories. Noted that each image in these datasets includes five image description sentences artificially annotated. For UCM-Captions and Sydney-Captions, we chose to resize them to 224 × 224 in our experiments.

4.2. Model Implementations

In the experiments, we used a full Transformer framework as a backbone for the proposed network. Among different versions of Transformer, each of which includes different numbers of layers, 6 layers were stacked in our model as the Transformer-based encoder and decoder. For the MAH module, eight different attention heads are set up. Moreover, we used a pre-trained Vit with RSI classification on RSICD, which can enhance the patch-level feature generalization performance. The word embedding size was 768 for the masked MAH module. We used the Adam [45] optimizer with an initial learning rate of 3×10^{-5}. The learning rate decays with 0.8 every 5 epochs. The batch sizes were set to 32. We report the results of training obtained after 25 epochs. Beam search was adopted in our model and was set to 3. The model training and further experiments are conducted on the single Nvidia GeForce GTX 1080Ti GPU.

Additionally, the RSICD vocabulary contains 1806 words, in which the total number of class labels is 250. Limited to the scale of the dataset, we exploited 110 words and 90 words as class labels occurring in the UCM vocabulary and the Sydney vocabulary, respectively. The UCM vocabulary is constructed with 342 words. The Sydney vocabulary contains 238 words.

4.3. Evaluation Metrics

In order to make a fair comparison with other methods, we made verifications on eight automatic evaluation metrics: BLEU-n [46], METEOR [47], ROUGE-L [48], SPICE [49], and CIDEr [50]. For all metrics, a higher score indicates improved performance. Specifically, BLEU-n is an n-gram precision score widely adopted for corpus level comparisons. The n-gram is used to assess the correctness of the generated sentence to the ground truth sentence, where n is from 1 to 4. ROUGE-L is similar with the concept of BLEU-n, which calculates the recall of the Longest Common Sub-sequence (LCS) shared by candidate and reference sentences. METEOR is capable of generating an alignment on all references based on WordNet synonyms and stemmed tokens for judging the word correlation. However, SPICE focuses on the graphic-based semantic representation for predicted descriptions. Out of other metrics, CIDEr is the one especially designed for vision-to-language tasks, such as video captioning, NIC, and RSIC. To capture a human judgment of consensus, CIDEr considers the word frequency as the weight and measures the weighted cosine similarity of words between generated n-grams and annotated captions.

4.4. Evaluation Results and Analysis

(1) Quantitative Comparison

In Tables 1–3, we report the performance of our proposed method in comparison with other SOTA methods on three publicly available datasets. Generally, the CIDEr is an excellent evaluation metric, which can better represent the quality of generated sentences. Thus, the CIDEr scores are listed in the last column for mastering performance trends on different models, and the best scores for all metrics are marked in bold. It is easy to see that the shown competitive performances belong to our model. To be specific, these

compared methods are categorized into two kinds: LSTM-based methods and Transformer-based methods.

Table 1. Comparison of our method and other state-of-the-art methods on the RSICD dataset.

Methods	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE-L	SPICE	CIDEr
SAT	0.6707	0.5438	0.4550	0.3870	0.3203	0.5724	0.4539	2.4686
FC-ATT	0.6671	0.5511	0.4691	0.4059	0.3225	0.5781	0.4673	2.5763
LAM(SAT)	0.6753	0.5537	0.4686	0.4026	0.3254	0.5823	0.4636	2.5850
Sound-a-a	0.6196	0.4819	0.3902	0.3195	0.2733	0.5143	0.3598	1.6386
GA	0.6779	0.5600	0.4781	0.4165	0.3258	0.5929	-	2.6012
Struc-ATT	0.7016	0.5614	0.4648	0.3934	0.3291	0.5706	-	1.7031
VRTMM	0.7813	0.6721	0.5645	0.5123	0.3737	0.6713	-	2.715
WDT	0.7240	0.5861	0.4933	0.4250	0.3197	0.6260	-	2.0629
SCAMET	0.7681	0.6309	0.5352	0.4611	0.4572	**0.6979**	-	2.4681
CNN-T	0.7980	0.6470	0.5690	0.4890	0.2850	-	-	2.4040
GLCM	0.7767	0.6492	0.5642	0.4937	0.3627	0.6769	-	2.5491
HCNet	0.7863	0.6754	0.5863	0.5122	0.3837	0.6877	-	2.8916
P-to-H	0.7581	0.6416	0.5585	0.4923	0.3550	0.6523	0.4579	2.5814
MGT	0.7931	0.6874	0.5960	0.5131	**0.3878**	0.6900	-	2.9231
Ours	**0.8022**	**0.6924**	**0.6018**	**0.5257**	0.3815	0.6919	**0.4966**	**2.9471**

The "-" represents that the score are not reported in some compared methods.

Table 2. Comparison of our method and other state-of-the-art methods on the UCM-Captions dataset.

Methods	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE-L	SPICE	CIDEr
SAT	0.7995	0.7365	0.6792	0.6244	0.4171	0.7441	0.4951	3.1044
FC-ATT	0.8102	0.7330	0.6727	0.6188	0.4280	0.7667	0.4867	3.3700
LAM(SAT)	0.8195	0.7764	0.7485	0.7161	0.4837	0.7908	0.5024	3.6171
Sound-a-a	0.7093	0.6228	0.5393	0.4602	0.3121	0.5974	0.3837	1.7477
GA	0.8319	0.7657	0.7103	0.6596	0.4436	0.7845	0.4853	3.327
Struc-ATT	0.8538	0.8035	0.7572	0.7149	0.4632	0.8141	-	3.3489
VRTMM	0.8394	0.7785	0.7283	0.6828	0.4527	0.8026	-	3.4948
WDT	0.7931	0.7237	0.6671	0.6202	0.4395	0.7132	-	2.7871
SCAMET	0.8460	0.7772	0.7262	0.6812	**0.5257**	0.8166	-	3.3773
CNN-T	0.8390	0.7690	0.7150	0.6750	0.4460	-	-	3.2310
GLCM	0.8182	0.7540	0.6986	0.6468	0.4619	0.7524	-	3.0279
HCNet	0.7686	0.7109	0.6573	0.6102	0.3980	0.7172	-	2.4714
P-to-H	0.8230	0.7700	0.7228	0.6792	0.4439	0.7839	0.4852	3.4629
MGT	**0.8839**	**0.8359**	**0.7909**	**0.7482**	0.4872	**0.8369**	-	3.6566
Ours	0.8557	0.8013	0.7567	0.7163	0.4754	0.8153	**0.5134**	**3.6965**

The "-" represents that the score are not reported in some compared methods.

For LSTM-based methods, such as SAT [4], FC-ATT [5], LAM(SAT) [26], Sound-a-a [32], GA [24], Struc-ATT [25], and HCNet [34], the modified CNN-LSTM architecture is used as the backbone. The core difference among LSTM-based methods lies in their explored attention mechanisms. As shown in Tables 1–3, FC-ATT outperforms the SAT and sound-a-a on all metrics. This demonstrates that using image-guided attributes can build connections between vision and language to better exploit textual information. Therefore, the LAM uses a pre-trained scene classifier to tag input images with a specific word. The results in Tables 1–3 show that the LAM obtains 0.87%, 24.71%, and 11.04% on the CIDEr metric, respectively. However, the GA adopts a GVFGA module, which filters out the redundant feature in the encoder stage for providing more effective visual features. By analyzing the results in Tables 1–3, one can see that the GA surpasses the FC-ATT and LAM in most evaluation metrics. The reason is that the salient visual features is sufficient for the networks. Moreover, Struc-ATT builds structure features in the encoder for seeking the pixel-level segmented object regions. Struc-ATT obtains comparative or improved results on Sydney-Captions compared with the results of RSICD. For HCNet, all obtained metric scores are the highest (around 3.52 on CIDEr). However, for the RSICD and UCM-Captions,

the image content is more complex than Sydney-Captions. It is difficult to match visual features with semantic vectors, and the improvement is quite limited. Such experimental results completely prove the effectiveness and superiority of embedding multiple labels and focusing on object-level regions.

Table 3. Comparison of our method and other state-of-the-art methods on the Sydney-Captions dataset.

Methods	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE-L	SPICE	CIDEr
SAT	0.7391	0.6402	0.5623	0.5248	0.3493	0.6721	0.3945	2.2015
FC-ATT	0.7383	0.6440	0.5701	0.5085	0.3638	0.6689	0.3951	2.2415
LAM(SAT)	0.7405	0.6550	0.5904	0.5304	0.3689	0.6814	0.4038	2.3519
Sound-a-a	0.7484	0.6837	0.6310	0.5896	0.3623	0.6579	0.3907	2.7281
GA	0.7681	0.6846	0.6145	0.5504	0.3866	0.7030	0.4532	2.4522
Struc-ATT	0.7795	0.7019	0.6392	0.5861	0.3954	0.7299	-	2.3791
VRTMM	0.7443	0.6723	0.6172	0.5699	0.3748	0.6698	-	2.5285
WDT	0.7891	0.7094	0.6317	0.5625	0.4181	0.6922	-	2.0411
SCAMET	0.8072	0.7136	0.6431	0.5846	0.4614	0.7258	-	2.3570
CNN-T	0.8220	0.7410	0.6620	**0.5940**	0.3970	-	-	2.7050
GLCM	0.8041	0.7305	0.6745	0.6259	0.4421	0.6965	-	2.4337
HCNet	**0.8826**	**0.8335**	**0.7885**	**0.7449**	**0.4865**	**0.8391**	-	**3.5183**
P-to-H	0.8373	0.7771	0.7198	0.6659	0.4548	0.7860	**0.4839**	3.0369
MGT	0.8155	0.7315	0.6517	0.5796	0.4195	0.7442	-	2.6160
Ours	0.7816	0.6980	0.6268	0.5628	0.4044	0.7231	0.4637	2.5920

The "-" represents that the score are not reported in some compared methods.

In our model, an extra multi-label classifier is added to extract specific semantic features. The multi-label classifier is trainable with the whole model. Compared with the pre-trained multi-label classification network, the trainable multi-label classifier can capture the semantic information more appropriately and accurately. To extract object-level features and cater to characteristics for RSI, the patch-level region-aware module following the transformer-based encoder is applied in our model, in which the position about core patch-level features locates salient objects. Moreover, our model refers to a region-oriented attention layer for the matching between salient regions and semantic label embeddings with MHA for the cross-modal linguistic-aware representation. It is not hard to find that the adopted Transformer structure in our method outperforms LSTM-based methods, and our model can obtain the best results on all metrics.

Note that VRTMM [20], SCAMET [29], CNN-T [30], GLCM [38], and P-to-H [39] are based on the CNN-Transformer architecture, while WDT [28] and MGT [31] are a full Transformer architecture. As can be seen from Tables 1–3, VRTMM is superior to other relevant methods, including GLCM, SCAMET (multi-attention encoder with a Transformer-based decoder), CNN-T (multi-scale features with a Transformer-based decoder), and P-to-H (multi-scale features from the foreground and background), except for Sydney-Captions, whose results were slightly below the published results of the CNN-T but superior to those of the SCAMET. This is because the encoder of the VRTMM is pretrained with a multi-task model for encoding RSIs, which can better deal with complex scenes in RSIs compared with conventional algorithms. Due to the detected noisy words by the word classifiers, the WDT with a Transformer-based generator has the worst results of all compared methods. Different from the WDT, the Transformer-based encoder utilized in MGT learns the relationship between image patches and improves the richness of captions by better exploring multi-modality relationships. This model has achieved a great performance improvement compared with the Transformer-based methods on the three datasets. Furthermore, our model individually outperforms the MGT by about 1.6% and 6.26% in terms of BLEU4 and CIDEr on RSICD, respectively. Our model trained with UCM-Captions and Sydney-Captions brings little performance improvements. As we know, few words are included in the UCM vocabulary and Sydney vocabulary; semantic richness is lacking. However, Sydney-Captions has the smallest image–caption pairs with different sentence lengths. Thus, a small improvement on Sydney-Captions is shown compared

with the other two datasets. Above all, the experimental results confirm the effectiveness of our model, which focuses on not only the specific labels but also the objects that are most relevant to the caption. This phenomenon also suggests that using a full Transformer architecture can better capture the abundant relationships among objects compared with CNN-LSTM methods.

(2) Qualitative Comparison

Some examples of multi-scale scenes from UCM-Captions are shown in Figure 5. The larger objects in Figure 5a,c can be correctly predicted by the baseline and our model. However, the "cars" in the sentence from the baseline in Figure 5b is wrong. In Figure 5d, there is a small house, which could be learned by our model. Therefore, it can be seen that our model is not sensitive to multi-scale targets.

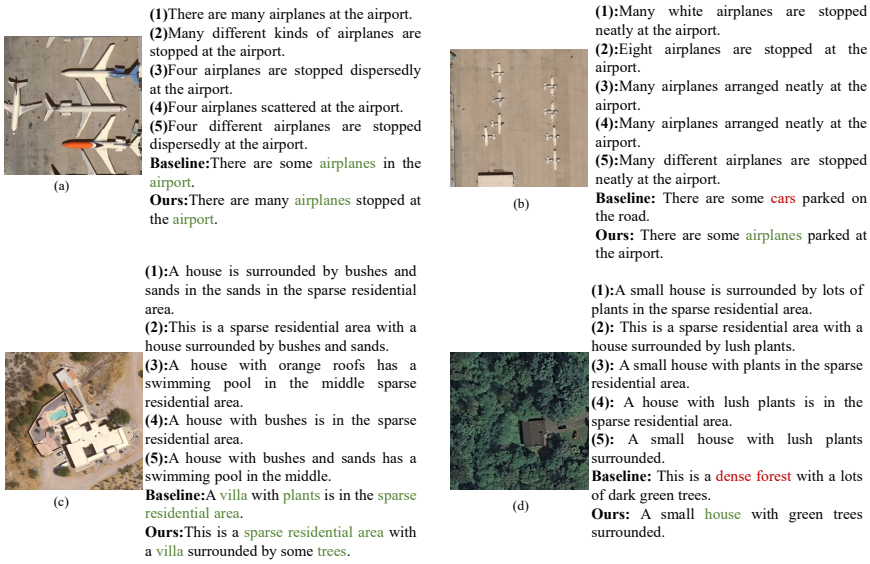


Figure 5. Some examples (**a**–**d**) of captions from baseline and our model on UCM-Captions.

In order to visually demonstrate the superiority of our multi-label classifier on three standard datasets, some label predictions are shown in Figure 6. It is worth noting that the predicted labels are arranged in descending order according to prediction probabilities. We predict three labels for each image to analyze the performance of our proposed method.

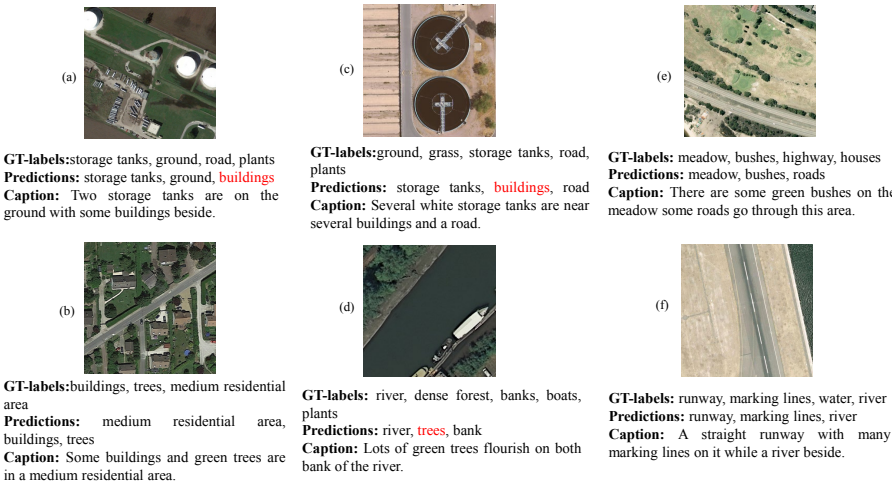


Figure 6. Examples for (**a**–**f**) with the ground truth (GT) labels, the predicted three labels and the generated caption with detected labels. The red words indicate mismatching or errors in label predictions.

The RSICD contains 21 land-cover categories. The increased variability and complexity within ground objects make the multi-label classifier more difficult to keep well-performed. Two examples are shown in Figure 6a,b. The dispersed minor object in Figure 6a has the wrong predictions, which shares a similar appearance to "buildings". However, the key object (i.e., "storage tanks") in the image is correctly depicted in the final description. In addition, the multiple labels generated from Figure 6b can provide a specific label, such as "medium residential area", "buildings", and "trees", which can guide the inference direction at the decoding stage.

In UCM-Captions, some noisy labels occur that misguide the decoder module, as shown in Figure 6c,d in the second column. However, the prediction for the salient object (i.e., "storage tanks") is robust. The other label prediction also has some problems that are distributed at the edges and corners in the given image. While noticing the labels in the caption (i.e., "buildings"), the message displaying the noisy image labels is associated with the high-frequency label words. The phenomenon in Figure 6d can also be considered a noisy label. However, when we understand the content of the image, we find that the generated caption is positive with respect to what the image wants to express.

The images selected from Sydney-Captions are shown in the third column. As we can see, the trainable multi-label classifier can accurately predict the labels contained in Figure 6e,f. Taking Figure 6f as an example, the main content in the image is "a runway with marking lines", and the caption also generates key information with available labels. The performance depends on our multi-label classifier, which can correctly identify associated multi-labels from the given RSI.

4.5. Ablation Experiments

To analyze the influence of each module, we compared the models with a transformer-based baseline (T1). Three different configurations of our model were added on the baseline to quantitatively analyze their effectiveness: the patch-level salient region module (T2), the multi-label classifier module (T3), and the region-oriented attention layer (T4).

(1) Quantitative Comparison

We show the the results by validating the Transformer-based baseline (T1), T1 with the patch-level salient region module (T2), T2 with the multi-label classifier (T3), and our full model (T4), respectively. Table 4 shows the results of the ablation study in terms of eight metric scores for the RSICD. From Table 4, one can see that that the highest metric scores are obtained when all modules of the model were included. It can be seen that simply stacking the patch-level region-aware module has a significant influence on most metrics (0.87% and 3.71% on BLEU4 and CIDEr metrics, respectively) compared with the T1 model. However, the multilayer Transformer-based encoder might gain visual information related to the global representation with exchanging patch-level context information. In contrast, the patch-level region-aware module in T2 can extract object-level features from mixed patch-level features. To pursue a plausible semantic layout, the reasonable category labels are transferred from the seen categories in an RSI. Such multiple words can appear with grammatically correctness in the caption. In particular, T3 is established to evaluate the usefulness of the trainable multi-label classifier module, where the multi-label classifier module is added based on the T2 model. By comparing T2, we see that the introduction of the T3 can considerably improve the CIDEr score from about 2.92 to 2.94. In order to verify the effectiveness of the region-oriented attention layer, we construct T4 models, whereas T1–T3 use one MHA module, respectively. The adopted MHA module in T3 fuses the representations of the two modalities by concatenation for learning. Unlike T3, the two modalities divided into parallel MHA branches are independent. Obviously, a region-oriented attention layer with parallel attention branches can learn superior complementary cross-modal information than one MHA, resulting in advanced model performance.

From Table 5, the training time and inference speed between T1 and T4 model are tested on UCM-Captions. The lower amount of training time shows an improved designed

RSIC model, while a greater inference speed is due to the increased parameters in T4 compared with the T1 model.

Table 4. Ablation study of our model on the RSICD dataset.

Methods	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE-L	SPICE	CIDEr
T1	0.7913	0.6786	0.5882	0.5126	0.3832	0.6872	0.4963	2.8826
T2	0.797	0.6869	0.5963	0.5213	0.3822	0.6866	0.4943	2.9197
T3	0.7931	0.6835	0.5952	0.5218	**0.3853**	0.6909	**0.4970**	2.9399
T4	**0.8022**	**0.6924**	**0.6018**	**0.5257**	0.3815	**0.6919**	0.4966	**2.9471**

Table 5. Comparison between our model and the baseline on training time and inference speed.

Methods	Training Time (min)	Inference Speed (Images/s)
T1	24.3 min	2.58
T4	20.5 min	2.72

(2) Qualitative Comparison

In Figure 7, the ablation models are able to generate captions with higher quality, in which the images are selected from RSICD. However, in some case, the baseline model fails to understand minor objects with correct or informative words. As shown in Figure 7a, T1 correctly expresses the description "Some planes are parked in an airport", while "a building" in the GT caption also should be identified and inferred. T2, T3, and T4 could learn the entities of "terminal", which indicates that the patch-level region-aware module plays a supporting role. A similar case can also be found in Figure 7e. T1 does not know the edge's information, and it is difficult to distinguish the features of the "road" from the ground. The results obtained by T2, T3, and T4 contain object-level ("cars") and scene-level ("parking lot") features, especially along the peripheral road. Meanwhile, in Figure 7c, T2 and T3 can avoid describing the wrong color, such as "green sea" generated by T1, while T4 could give more detailed descriptions than other ablation models. Further improvements will be achieved when the labels and salient regions are contained. As shown in Figure 7d, T2 and T3 could achieve more complex grammar structures in sentence generation than the T1 model. Furthermore, T4 with the region-oriented attention layer contains abundant cross-modal information, which was essential for the prediction of the small-scale object "cars" on the bridge. By comparing the descriptions shown in Figure 7b, it is apparent that T2, T3, and T4 could also describe an important amount of information in the scene (like "Two baseball fields"). In other cases, our ablation models achieved the same or slightly improved results than the T1 model, as shown in Figure 7f, which focuses on different relationships between objects in an image. Collectively, the qualitative results agree with the previous observations from the quantitative results.

In order to better explain the effect of the patch-level region-aware module, the higher attention weight on the original RSI is visualized in Figure 8. It can be seen that the selected weights are various in different scenes. However, the weights for significant objects are comprehensive enough. Note that the proposed model can focus on multi-scale objects (i.e., cars in Figure 8a, airplanes in Figure 8c, and houses in Figure 8e) while abandoning redundant closely related parts between objects when the salient analysis is performed (i.e., shown in Figure 8b,d,f). Therefore, the patch-level region-aware module can obtain the specific patch-level salient features, which is effective for RSIC tasks.

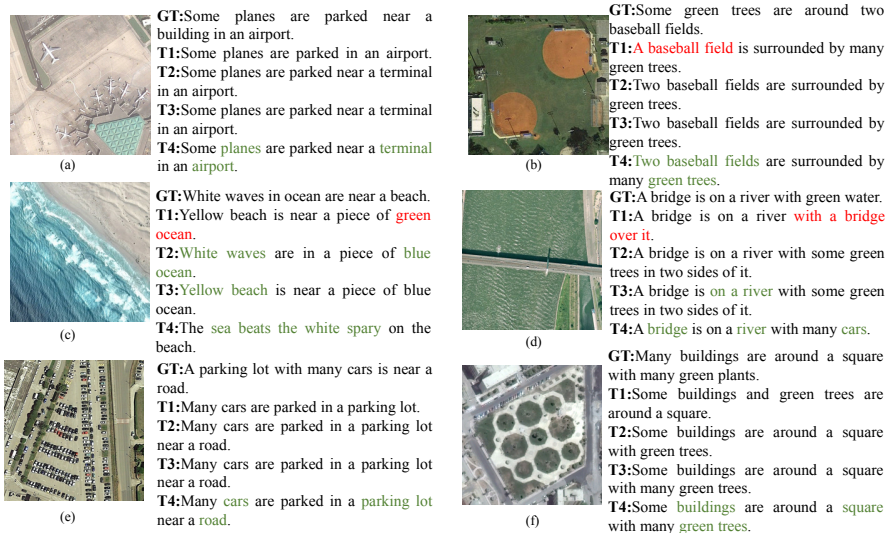


Figure 7. Visualization of the ablated models (T1, T2, T3, and T4) for (**a**–**f**). GT is one caption in the annotated sentences. The wrong words are indicated in red; green words are more accurate and rich semantics.

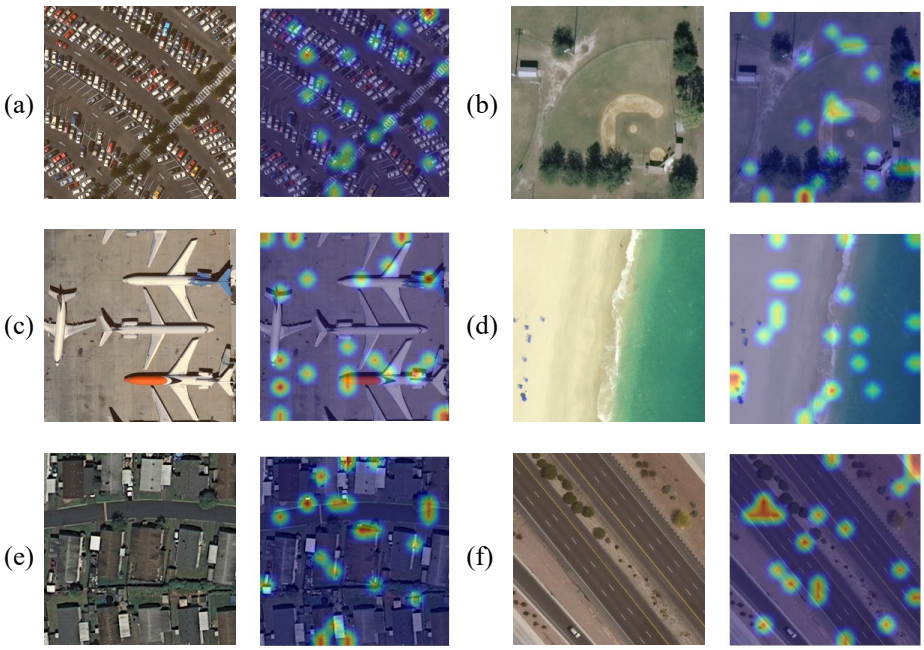


Figure 8. Patch-level region visualization. (**a**–**f**) denote some typical scenes of visible results. (**Left**) Original RSIs. (**Right**) Superimposed RSIs.

4.6. Parameter Analysis

The effect of different values for the parameter D can be found in Figure 9a. The figure shows that, when D is too few or too more, the value of this parameter affects the performance of the proposed network negatively. Choosing a value of $D = 2$ outperforms the case with one label. This is because too few labels would cause a lack of semantic diversity. Furthermore, choosing a value of $D = 3$ outperforms the case with a higher value of D. The CIDEr scores will decrease once the value of D increases. As an example, when the value of D is set to 5, the CIDEr score drops by 3.4%. For $D = 5$, the multi-label classifier will generate five plausible categories, which may contain unseen categories in other RSIs. The superfluous semantic information is very difficult to imagine and causes unsatisfactory performance. Thus, we opted for multiple labels and set $D = 3$ by default on three different benchmark RSIC datasets.

Figure 9b shows the experiment results of the number K for the selected salient regions. K is an important factor that delivers improved object-level representations to characterize different objects in RSIs. To explore the impact of parameter K, we varied K as follows: [10, 15, 20, 25, 30]. The CIDEr score increases following the increase in K (from 10 to 20), which means that more effective region-level region features are selected. $K = 20$ is a turning point. For $K > 20$, some background patches overwhelm most real category patches in the RSIs, which account for a large proportion. The region-oriented attention will be torn between mixed patch-level features and label features. In this case, inter-patch relationship across different objects is seldom considered. When K is too small, the performance also will decrease due to the lack of patch variety. Based on the experiment results, the optimal choice is $K = 20$. We also applied this choice on other RSIC datasets.

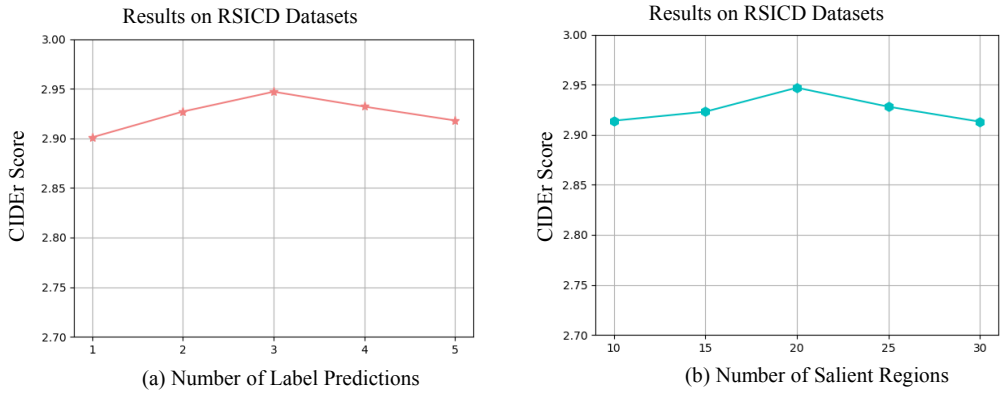


Figure 9. The visualizations of captioning performance (CIDEr) that are affected by parameters on RSICD.

5. Conclusions

This study demonstrates an RSIC technique, which explores the utility of patch-level salient regions and image-labels along with a Transformer-based decoder to generate meaningful captions of still RSIs. The proposed architecture generates salient regional features by matching intrinsic relationships between patch-level features and class tokens from the Transformer encoder. To further enrich the RSI representation with semantic features, we devise a multi-label classifier to map class tokens to explicit object classes. Different from the previous label-driven RSIC approaches, the multi-label classifier module is injected into the RSIC model in an end-to-end manner. Further, our network binds regional features and associated labels using region-oriented attention to explore semantically coherence. Specifically, region attention endows an ability to figure out which patch-level region to omit or focus on. Label attention improves the sentence generation with semantic relation information. The extensive experiments on the three popular benchmark caption datasets validate the effectiveness and generalization ability of our proposed framework. Especially, for the patch-level region-aware block cascaded with the baseline model, we obtained a reported CIDEr performance of about 2.95 on the RSICD test.

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

Funding: This work was supported in part by the National Natural Science Foundation of China under Grants 62276197, 62006178, and 62171332, in part by the Shaanxi Province Innovation Capability Support Plan under Grant 2023-CX-TD-09, in part by the Postdoctoral Fellowship Program of CPSF under Grant GZC20241321 and GZC20232033, and in part by the Wuxi University Research Start-up Fund for Introduced Talents under Grant 2024r011.

Data Availability Statement: The RSICD, UCM-Captions, and Sydney-Captions datasets can be obtained from (https://pan.baidu.com/s/1bp71tE3#list/path=%2F, https://pan.baidu.com/s/1mjPToHq#list/path=%2F, https://pan.baidu.com/s/1hujEmcG#list/path=%2F, accessed on 1 September 2024).

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 conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

RSIC	Remote sensing image captioning.
CNN	Convolutional neural networks.
LSTM	Long short-term memory.
VGG	Visual geometry group.
ResNet	Residual network.
LAM	Label attention mechanism.
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.
GLCM	Global–local captioning model.
VRTMM	Variational autoencoder and reinforcement learning based two-stage multitask learning model.
MGT	Mask-guided Transformer.
MHA	Multi-head attention.
FFN	Feed-forward network.
I	The input remote sensing image.
V_T	The visual features from Vit encoder.
S	The similarity of queries Q_v and keys K_v.
s_i	The similarity score in $head_i$.
s_i'	The similarity score of $\langle class, region \rangle$ in $head_i$.
S'	The averaged similarity scores of $\langle class, region \rangle$ for all heads.
V_R	The salient regional map .
$\mathbf{p}$	The label probability distribution over ground-truth labels.
L_l^t	The attended features from label mlti-head attention.
V_l^t	The attended features from region mlti-head attention.
F_l^t	The fused cross-modal features.
O_l^t	The resultant vectors from F_l^t.
p_t^θ	The probability of generating specific word.
T	The max length in ground-truth sentence.
y_t	The generated word at t time.

References

1. Farooq, A.; Jia, X.; Hu, J.; Zhou, J. Transferable Convolutional Neural Network for Weed Mapping with Multisensor Imagery. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–16. [CrossRef]
2. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
3. Zhang, X.; Li, X.; An, J.; Gao, L.; Hou, B.; Li, C. Natural language description of remote sensing images based on deep learning. In Proceedings of the 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, USA, 23–28 July 2017; pp. 4798–4801.
4. 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]
5. 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]
6. 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]
7. 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]

8. Chen, W.; Ouyang, S.; Tong, W.; Li, X.; Zheng, X.; Wang, L. GCSANet: A global context spatial attention deep learning network for remote sensing scene classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 1150–1162. [CrossRef]
9. Wang, W.; Chen, Y.; Ghamisi, P. Transferring CNN With Adaptive Learning for Remote Sensing Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–18. [CrossRef]
10. He, X.; Zhou, Y.; Zhao, J.; Zhang, D.; Yao, R.; Xue, Y. Swin transformer embedding UNet for remote sensing image semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]
11. Wang, H.; Tao, C.; Qi, J.; Xiao, R.; Li, H. Avoiding Negative Transfer for Semantic Segmentation of Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]
12. Zhang, X.; Jiao, L.; Liu, F.; Bo, L.; Gong, M. Spectral Clustering Ensemble Applied to SAR Image Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2008**, *46*, 2126–2136. [CrossRef]
13. Ma, W.; Li, N.; Zhu, H.; Jiao, L.; Tang, X.; Guo, Y.; Hou, B. Feature split-merge-enhancement network for remote sensing object detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–17. [CrossRef]
14. Zhou, X.; Shen, K.; Weng, L.; Cong, R.; Zheng, B.; Zhang, J.; Yan, C. Edge-guided recurrent positioning network for salient object detection in optical remote sensing images. *IEEE Trans. Cybern.* **2022**, *53*, 539–552. [CrossRef] [PubMed]
15. Zhang, C.; Wang, L.; Cheng, S.; Li, Y. SwinSUNet: Pure transformer network for remote sensing image change detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–13. [CrossRef]
16. Lv, Z.; Wang, F.; Cui, G.; Benediktsson, J.A.; Lei, T.; Sun, W. Spatial–spectral attention network guided with change magnitude image for land cover change detection using remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–12. [CrossRef]
17. 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]
18. Yuan, Z.; Li, X.; Wang, Q. Exploring multi-level attention and semantic relationship for remote sensing image captioning. *IEEE Access* **2019**, *8*, 2608–2620. [CrossRef]
19. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
20. 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]
21. 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]
22. Ma, X.; Zhao, R.; Shi, Z. Multiscale methods for optical remote-sensing image captioning. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 2001–2005. [CrossRef]
23. 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]
24. Zhang, Z.; Zhang, W.; Yan, M.; Gao, X.; Fu, K.; Sun, X. Global visual feature and linguistic state guided attention for remote sensing image captioning. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–16. [CrossRef]
25. 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]
26. Zhang, Z.; Diao, W.; Zhang, W.; Yan, M.; Gao, X.; Sun, X. LAM: Remote sensing image captioning with label-attention mechanism. *Remote Sens.* **2019**, *11*, 2349. [CrossRef]
27. Wang, S.; Ye, X.; Gu, Y.; Wang, J.; Meng, Y.; Tian, J.; Hou, B.; Jiao, L. Multi-label semantic feature fusion for remote sensing image captioning. *ISPRS J. Photogramm. Remote Sens.* **2022**, *184*, 1–18. [CrossRef]
28. Wang, Q.; Huang, W.; Zhang, X.; Li, X. Word–sentence framework for remote sensing image captioning. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 10532–10543. [CrossRef]
29. Gajbhiye, G.O.; Nandedkar, A.V. Generating the captions for remote sensing images: A spatial-channel attention based memory-guided transformer approach. *Eng. Appl. Artif. Intell.* **2022**, *114*, 105076. [CrossRef]
30. Zia, U.; Riaz, M.M.; Ghafoor, A. Transforming remote sensing images to textual descriptions. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *108*, 102741. [CrossRef]
31. 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. [CrossRef]
32. Lu, X.; Wang, B.; Zheng, X. Sound active attention framework for remote sensing image captioning. *IEEE Trans. Geosci. Remote Sens.* **2019**, *58*, 1985–2000. [CrossRef]
33. Zhang, M.; Zheng, H.; Gong, M.; Wu, Y.; Li, H.; Jiang, X. Self-structured pyramid network with parallel spatial-channel attention for change detection in VHR remote sensed imagery. *Pattern Recognit.* **2023**, *138*, 109354. [CrossRef]
34. Yang, Z.; Li, Q.; Yuan, Y.; Wang, Q. HCNet: Hierarchical Feature Aggregation and Cross-Modal Feature Alignment for Remote Sensing Image Captioning. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–11. [CrossRef]
35. Yang, Q.; Ni, Z.; Ren, P. Meta captioning: A meta learning based remote sensing image captioning framework. *ISPRS J. Photogramm. Remote Sens.* **2022**, *186*, 190–200. [CrossRef]
36. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762.

37. Wang, J.; Chen, Z.; Ma, A.; Zhong, Y. Capformer: Pure Transformer for Remote Sensing Image Caption. In Proceedings of the IGARSS 2022—2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022; pp. 7996–7999.
38. Wang, Q.; Huang, W.; Zhang, X.; Li, X. GLCM: Global–local captioning model for remote sensing image captioning. *IEEE Trans. Cybern.* **2022**, *53*, 6910–6922. [CrossRef]
39. Du, R.; Cao, W.; Zhang, W.; Zhi, G.; Sun, X.; Li, S.; Li, J. From plane to hierarchy: Deformable transformer for remote sensing image captioning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 7704–7717. [CrossRef]
40. Zhuang, S.; Wang, P.; Wang, G.; Wang, D.; Chen, J.; Gao, F. Improving remote sensing image captioning by combining grid features and transformer. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]
41. Fu, K.; Li, Y.; Zhang, W.; Yu, H.; Sun, X. Boosting memory with a persistent memory mechanism for remote sensing image captioning. *Remote Sens.* **2020**, *12*, 1874. [CrossRef]
42. 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 16 $\times$ 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
43. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
44. 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), Kunming, China, 6–8 July 2016; pp. 1–5.
45. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
46. 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, 6–12 July 2002; pp. 311–318.
47. 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.
48. ROUGE, L.C. ROUGE: A package for automatic evaluation of summaries. In Proceedings of the Workshop on Text Summarization of ACL, Barcelona, Spain, 25 July 2004.
49. Anderson, P.; Fernando, B.; Johnson, M.; Gould, S. Spice: Semantic propositional image caption evaluation. In Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Part V 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 382–398.
50. Vedantam, R.; Zitnick, C.L.; 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.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.