remote sensing

MDPI

Article

Cross-Modal Retrieval and Semantic Refinement for Remote Sensing Image Captioning

Zhengxin Li [1,2,3], Wenzhe Zhao [1,2,*], Xuanyi Du [1,3], Guangyao Zhou [1,2] and Songlin Zhang [1,2]

[1] The Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China
[2] Key Laboratory of Spatial Information Processing and Application System Technology, Chinese Academy of Sciences, Beijing 100190, China
[3] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 101408, China
* Correspondence: zwz@mail.ie.ac.cn

Abstract: Two-stage remote sensing image captioning (RSIC) methods have achieved promising results by incorporating additional pre-trained remote sensing tasks to extract supplementary information and improve caption quality. However, these methods face limitations in semantic comprehension, as pre-trained detectors/classifiers are constrained by predefined labels, leading to an oversight of the intricate and diverse details present in remote sensing images (RSIs). Additionally, the handling of auxiliary remote sensing tasks separately can introduce challenges in ensuring seamless integration and alignment with the captioning process. To address these problems, we propose a novel cross-modal retrieval and semantic refinement (CRSR) RSIC method. Specifically, we employ a cross-modal retrieval model to retrieve relevant sentences of each image. The words in these retrieved sentences are then considered as primary semantic information, providing valuable supplementary information for the captioning process. To further enhance the quality of the captions, we introduce a semantic refinement module that refines the primary semantic information, which helps to filter out misleading information and emphasize visually salient semantic information. A Transformer Mapper network is introduced to expand the representation of image features beyond the retrieved supplementary information with learnable queries. Both the refined semantic tokens and visual features are integrated and fed into a cross-modal decoder for caption generation. Through extensive experiments, we demonstrate the superiority of our CRSR method over existing state-of-the-art approaches on the RSICD, the UCM-Captions, and the Sydney-Captions datasets

Keywords: semantic retrieving; attention mechanism; image captioning; remote sensing

Citation: Li, Z.; Zhao, W.; Du, X.; Zhou, G.; Zhang, S. Cross-Modal Retrieval and Semantic Refinement for Remote Sensing Image Captioning. *Remote Sens.* **2024**, *16*, 196. https://doi.org/10.3390/rs16010196

Academic Editor: Shuying Li

Received: 6 November 2023
Revised: 20 December 2023
Accepted: 28 December 2023
Published: 3 January 2024

1. Introduction

Remote sensing image captioning (RSIC) is a cutting-edge technology that bridges the gap between complex geographic information captured in RSIs and human comprehension by converting it into text-based descriptions. With the proliferation of remote sensing data from satellites, drones, and other sources, the demand for effective and interpretable methods to extract meaningful insights from these images has grown exponentially. RSIC offers a transformative solution by providing human-readable and contextual descriptions of the content within remote sensing images (RSIs), making them more accessible and actionable for various applications [1]. As a result, RSIC has emerged as a prominent and burgeoning research area, drawing significant attention and interest from the scientific community [2–6].

RSIC has made remarkable strides with the adoption of the encoder–decoder sequence as the mainstream approach. This process involves leveraging convolutional neural networks (CNNs) to extract essential image features, which are subsequently transformed

into natural language descriptions using sequential models like recurrent neural networks (RNNs) [7] or long short-term memory (LSTM) networks [8].

To improve the performance of RSIC models, attention mechanisms have been integrated into the encoder–decoder framework [9–12]. By capturing intricate relationships and dependencies between visual and textual modalities, attention mechanisms have significantly contributed to enhancing the overall performance and interpretability of RSIC models. To further enhance the capabilities of RSIC, researchers have introduced auxiliary information, giving rise to the development of two-stage methods [13–16]. These approaches incorporate an additional stage that provides supplementary information to the RSIC model, generally focusing on enhancing vision–language alignment by capturing fine-grained semantic information. As a result, these two-stage methodologies frequently outperform their attention-based one-stage counterparts, yielding improved performance [17,18].

Despite the benefits of using pre-trained detectors or classifiers to obtain supplementary information, there are limitations that can hinder their overall effectiveness. Primarily, the prowess of semantic comprehension in pre-trained detectors/classifiers is confined within the boundaries of pre-defined semantic/class labels, which may not fully encompass the intricacies of RSIs. Moreover, these pre-trained models are not optimized during the sentence decoding process, making it challenging to emphasize visually salient semantic information effectively in the generated captions. Consequently, effectively emphasizing visually significant semantic information while discerning and filtering out misleading semantic information becomes a complex endeavor, leading to less contextually relevant and coherent captions [19].

Therefore, to alleviate the limitations mentioned above and to enhance the scalability and generalization of image encoders for RSIC, this article proposes a novel cross-modal retrieval and semantic refinement (CRSR) method. Departing from the utilization of pre-trained detectors/classifiers, we leverage the power of the CLIP model [20], which has demonstrated remarkable performance in various vision–language multimodal tasks [21–23], employing it as a cross-modal retrieval tool to extract semantically relevant sentences from a pool of captioning sentences as supplementary information. Considering the consistent format of description sentences in RSIC datasets, we initially fine-tune the pre-trained CLIP model on the current RSIC datasets with a mask strategy, which directs the focus towards semantically relevant information retrieval in RSIs during the fine-tuning process. To address the complexity of RSIs, which often contain complicated and varied semantic information leading to semantic-irrelevant and conflicting retrievals, a semantic refinement module is introduced. We utilize a masked cross-attention mechanism, incorporating the image features of the RSIs to effectively filter out semantically irrelevant words. The introduced mask mechanism is specially designed to handle multiple conflicting semantic information in RSIs, which can lead to contradictory caption generation. To further utilize the image features obtained from the CLIP model, a Transformer Mapper network is introduced. This module employs learnable queries to predict essential words beyond retrieved supplementary information, expanding the representation of image features through a simple projection layer and enabling the model to capture semantically meaningful visual information more effectively. Thus, the model can better comprehend the intricate details and relationships within the RSIs. Ultimately, the refined semantic information and enriched image features are fed into a cross-modal decoder, which generates accurate and coherent captions enriched with contextual information and descriptive details.

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

1. We propose a new CRSR method incorporating the CLIP-based retrieval model and a semantic refinement module that effectively addresses the limitations of existing two-stage approaches. We firstly obtain the semantic information through a fine-tuned retrieval model. Then, the semantic refinement module is introduced for filtering out misleading words through a masked cross-attention mechanism.

2. We introduce a Transformer Mapper network, which is designed to provide a comprehensive representation of the image features that extends beyond the retrieved information using attention mechanisms and learnable queries for semantic prediction. The projected image feature with learnable queries employs self-attention to capture intricate relationships and dependencies within the image features, which provides a particular focus on the overlooked semantic region, enabling it to effectively analyze and understand more semantic details present in RSIs.
3. The extensive experiments conducted on three diverse datasets, RSICD, UCM-Captions, and Sydney-Captions, provide compelling evidence to validate the superior performance of our proposed CRSR method. We demonstrate that our approach achieves higher captioning accuracy compared to other state-of-the-art methods on the three benchmark datasets.

The subsequent organization of this paper is as follows. Section 2 provides an overview of related work on RSI captioning. Section 3 presents the details of the proposed CRSR method. Section 4 introduces the experiments and analysis conducted on three datasets. Section 5 provides a summary of this article.

2. Related Works

In this section, we review the related works in the field of RSIC and explore the advancements and research efforts that have been made. The following RSIC methods are divided into two categories: one-stage methods and two-stage methods.

2.1. One-Stage Methods

The mainstream one-stage methods in RSIC have predominantly adopted the encoder–decoder sequence architecture. Among the pioneers in this field, Qu et al. [24] proposed a deep multimodal neural network model, which is widely recognized as the classical encoder–decoder structure for RSIC. The model utilizes a CNN as the encoder to extract essential image features and an RNN as the decoder to transform these features into natural language descriptions. Similarly, Lu et al. [2] also explored and demonstrated the effectiveness of the encoder–decoder structure in RSIC. They further contributed to the field by conducting a series of experiments on three benchmark datasets released by Lu et al. [2] and Qu et al. [24]. Li et al. [25] proposed a multi-level attention model that closely imitates human attention mechanisms, comprising three attention structures for different areas of the image, words, and vision–semantic interactions. Huang et al. [26] proposed a denoising-based multi-scale feature fusion (DMSFF) mechanism to improve caption quality, which aggregates multiscale features using denoising operations during visual feature extraction. Li et al. [27] proposed a novel recurrent attention mechanism that integrates competitive visual features, allowing for the utilization of both static visual features and multiscale features. Li et al. [28] proposed a new truncation cross entropy (TCE) loss to reserve probability margins for non-target words, thereby reducing the risk of overfitting and improving the generalization capability of the model. Zhang et al. [29] introduced the global visual feature-guided attention (GVFGA) mechanism, leading to more descriptive and contextually relevant image captions. The introduced linguistic state (LS) and LS-guided attention (LSGA) mechanisms further refine the fusion of visual and textual features by filtering out the irrelevant information from the fused visual–textual feature. Hoxha et al. [30] proposed a new decoder structure by introducing a decoder based on support vector machines (SVMs) instead of RNNs. The SVM-based decoder proved to be effective, especially in scenarios with limited annotated samples, and offers the advantage of reduced training and testing time.

Without considering the auxiliary information in RSIs, the existing encoder–decoder-based methods mentioned above often underperform compared to the two-stage methods in terms of the accuracy of the generated captions.

2.2. Two-Stage Methods

To further utilize the information from image features, Chen et al. [15] included the expression of geospatial relations of geo-objects in the images. Zhang et al. [17] proposed a novel label-attention mechanism that incorporates label information from RSIs. A new model with an attribute attention mechanism was proposed for generating semantic descriptions by Zhang et al. [13]. Wang et al. [14] considered a collection of topic words representing common information across sentences using a retrieval topic recurrent memory network that leverages a topic repository to guide sentence generation. Wang et al. [31] introduced a new explainable word–sentence framework including a word extractor responsible for identifying valuable words through a word classification task, while the sentence generator organizes these words into a well-formed sentence. A summarization-driven approach was proposed by Sumbul et al. [32] to address the information deficiency issue in image–language mapping. Kandala et al. [33] introduced an auxiliary decoder that is trained for multilabel scene classification to improve the encoder's training process. Zhao et al. [18] presented a fine-grained, structured attention-based method that utilizes the structural characteristics of semantic contents. This method achieves both image captioning and weakly supervised segmentation within a unified framework. Ye et al. [16] proposed a novel joint-training two-stage (JTTS) method that combines image captioning and an auxiliary multilabel classification task into a joint training process, allowing mutual interference between tasks to be considered. The proposed method utilizes a dynamic contrast loss function and an attribute-guided decoder to filter the multilabel prior information and generate more accurate image captions. Yang et al. [34] introduced a meta captioning framework, which leverages meta learning to extract meta features from two support tasks (natural image classification and remote sensing image classification) and transfers them into the target task of RSIC. Zhang et al. [35] represented a multi-source interactive stair attention mechanism that separately models the semantic information of preceding sentences and visual regions of interest. The stair attention divides the attentive weights into three levels, allowing for better focus on different regions in the search scope. Additionally, CIDEr-based reward reinforcement learning [36] is used to enhance the quality of the generated sentences. Du et al. [37] proposed a Deformable Transformer with scaled attention for multi-scale features extracted from the foreground and background separately. Li et al. [38] introduced a semantic concept extractor with visual–semantic co-attention for cross-modal interaction. Moreover, attentive vectors and semantic-level relational features are utilized within a consensus exploitation (CE) block.

3. Methodology

In this section, we provide a detailed description of the proposed CRSR method. As depicted in Figure 1, our approach utilizes the Transformer Mapper network for encoding visual information extracted from the image, predicting essential words through learnable queries. Simultaneously, the semantic refinement module is employed to retrieve and refine semantic tokens from the current input. The encoded visual features, along with the predicted query tokens and refined semantic tokens, are combined and fed into a Transformer-based cross-modal decoder for sentence generation.

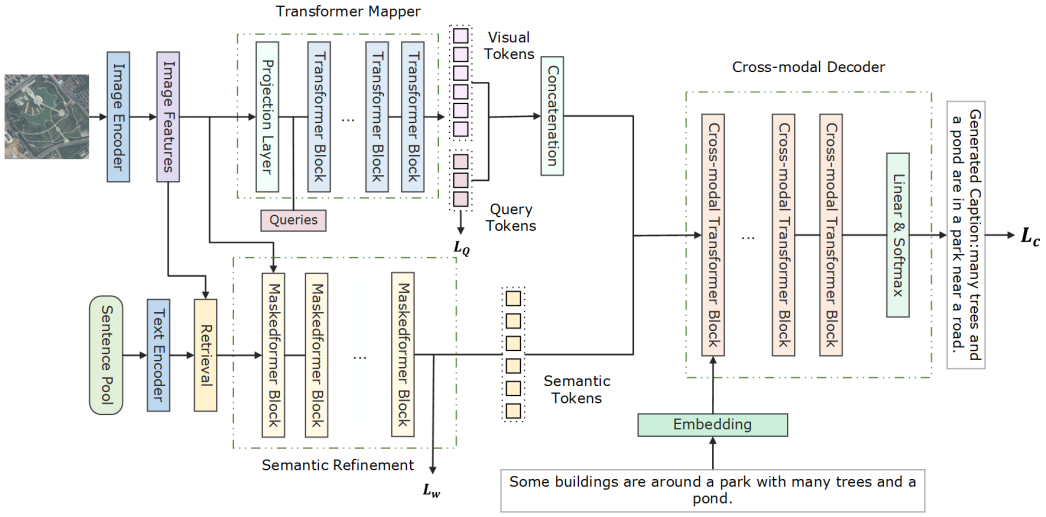

Figure 1. The overall structure of our CRSR model. The image features and text features are extracted by the CLIP image encoder and text encoder from the input image and sentence pool, respectively. Image features are transformed into a sequence of visual tokens through Transformer Mapper network with learnable queries. Based on the cross-modal retrieval, the retrieved relevant sentences are separated as a series of words and fed into the semantic refinement module to filter out irrelevant words. Finally, the obtained visual tokens with query tokens and semantic tokens are both fed into a cross-modal decoder to generate the corresponding image captions.

3.1. Semantic Refinement

Existing methods in image captioning often rely on pre-trained object detectors or classifiers to capture semantic information, which is directly fed into the sentence decoder. However, these approaches are limited by the pre-defined labels and lack of interaction between the components. To alleviate these limitations, we utilize the pre-trained CLIP model, which has been well trained on diverse and large-scale data, as the cross-modal retrieval tool to retrieve and filter the words that are associated with the current input.

We first fine-tune the CLIP model through training on the images and captions in the RSIC training dataset as image–text pairs. However, simply calculating the similarity between the image features of the input image and the text features of the whole paired caption to fine-tune the CLIP model may lead to semantic confusion problems during the retrieval process. For instance, when processing an image with the caption "there is a small tennis court with a wide road beside", another sentence "there is a small white storage tank on the lawn with a road beside" can also receive a high similarity score, even though it carries different semantic content, causing misleading semantic information retrieval. To address this issue, we introduce a semantic mask module during the fine-tuning process, through which we mask semantic-irrelevant words in the caption to limit the impact of these words during the training process. After calculating the similarity between the image features of current input and the masked text features from the entire sentence pool, which includes all training captions in current dataset, we retrieve the top-K semantically relevant sentences from the sentence pool for each input image. We then decompose these sentences into a set of words while further filtering out the semantic-irrelevant words.

Despite our precautions, there may still be instances of misleading semantic information present, and these can certainly influence the accuracy of caption generation. Therefore, we employ a cross-modal refinement network to further filter the decomposed words. Specifically, for the input image, I, $w_I = [w_I^1, \ldots, w_I^{N_w}]$ represents the decomposed words, which are then mapped into a new semantic embedding space, resulting in the semantic features denoted as $V_s = [w_s^1, \ldots, w_s^{N_w}]$, where N_w denotes the length of the decomposed words. Next, the generated semantic features are passed into the cross-modal refinement module, which consists of a stacked configuration of N_c Masked Transformer

blocks. Specifically, the generated semantic features contain comprehensive information about the input image, which may not be present in every caption among the five descriptions for the image. Moreover, considering that conflicting semantic information exists in different caption sentences, such as the words "plane" and "planes" in two captions of the same image, like "many planes are parked between several buildings and many runways" and "several runways are scattered in the airport while every plane is on their position", we implement a Masked Transformer structure. We randomly mask retrieved words with contradictory meanings, with semantically irrelevant words masked randomly for current input, contributing to improved semantic comprehension. Within each Transformer block, every input semantic feature is contextually encoded via self-attention. Subsequently, the semantic features are further enhanced by leveraging the interaction between the semantic features and the image features v_I through cross-attention; the ith Transformer block is implemented as follows:

$$y_s^i = V_s^i + MultiHead(norm(V_s^i), V_s^i, V_s^i) \tag{1}$$

$$V_s^{i+1} = F(y_s^i + MultiHead(norm(y_s^i), v_I, v_I)) \tag{2}$$

where V_s^{i+1} denotes the output of the ith layer of the stacked N_c Transformer blocks and $F(\cdot)$ presents the feed forward layer. Accordingly, the output semantic tokens of the final Transformer block are denoted as $V_s^{N_c} = [w_s^{(N_c)i}|_{i=0}^{N_w}]$, which are then leveraged for predicting the filtered semantic words.

During the training process, we address the optimization of the semantic refinement module by treating the filter task as object prediction problems. In particular, the obtained semantic tokens $V_w^{N_c} = [w_s^{(N_c)1}, \ldots, w_s^{(N_c)N_w}]$ are passed through a projection layer to filter out the misleading tokens by predicting the token distribution in the semantic vocabulary of $N_v + 1$ words with a special token representing misleading semantic information. The resulting semantic predictions are denoted as $P_s = [P^1, \ldots, P^{N_w}]$, representing the predictions of each retrieved semantic token.

For the ith retrieved semantic token, the ground truth label is denoted as $y^i \in \mathbb{R}^{N_v+1}$, and its objective is measured with cross-entropy loss as follows:

$$L_w = -\frac{1}{N_w} \sum_{i=0}^{N_w} \sum_{v=0}^{N_v+1} y_v^i log(P_v^i) \tag{3}$$

3.2. Transformer Mapper

In caption tasks, the quality of the extracted image features and their representation play a crucial role in caption generation. To leverage the benefits of image feature representations in the well-pre-trained CLIP model, we employ the visual encoder of the CLIP model to extract the image visual feature, denoted as v_I, from the provided image I. Considering that the extracted feature v_I contains dense image information with a dimension of 512, we utilize a Transformer Mapper network to map the CLIP feature to a set of k visual vectors and yield the enriched visual tokens.

While the semantic refinement module significantly improves the quality of semantic tokens, it is important to note that some vital information may still be overlooked at the beginning of the semantic retrieval process. Taking inspiration from [39], we introduce a series of learnable queries, denoted as $Q = [q^1, \ldots, q^l]$. These queries are combined with the extracted image features and sent into the Transformer Mapper module for predicting omitted semantic information from the image features and guiding the projected visual features to encompass more than just partial semantic tokens.

Specifically, our Transformer Mapper module consists of a projection layer followed by a stacked configuration of N_r Transformer blocks with multi-head attention. Initially,

we obtain the k visual vectors $V_I = [v_I^1, \ldots, v_I^k]$ through the application of the projection layer, denoted as P:

$$v_I^1, \ldots, v_I^k = P(v_I) \tag{4}$$

Then, the learnable queries and the k visual vectors are combined as $W_I = [v_I^1, \ldots, v_I^k, q^1, \ldots, q^l]$. And we proceed to implement the stacked N_r Transformer blocks; the ith Transformer block operates as

$$x_I^i = W_I^i + MultiHead(norm(W_I^i), V_I^i, V_I^i) \tag{5}$$

$$W_I^{i+1} = x_I^i + MLP(norm(x_I^i)) \tag{6}$$

where $norm$ denotes the layer normalization, $MultiHead(\cdot)$ represents the multi-head attention mechanism utilized in each Transformer block, and $MLP(\cdot)$ is an MLP with two cascaded FC-ReLU-dropout units.

$\tilde{W}_I = [\tilde{v}_I^1, \ldots, \tilde{v}_I^k, \tilde{q}_I^1, \ldots, \tilde{q}_I^l]$ denotes the final outputs of the Transformer Mapper module, in which $\tilde{Q}_I = [\tilde{q}_I^1, \ldots, \tilde{q}_I^l]$ predicts the omitted words. For the ground truth set of omitted words y, assuming that prediction queries of the number l is greater than the ground truth word count N_o. To establish a bipartite matching between these two sets, we employ the Hungarian algorithm for the permutation of prediction query set with the lowest cost, which is calculated as follows:

$$\tilde{\alpha} = argmin \sum_{i}^{N_o} y_i log(\tilde{q}_{i_j}) \tag{7}$$

$$L_Q = -\sum_{i=0}^{N_o} y_i log(\tilde{q}_{\tilde{i}_j}) \tag{8}$$

Through minimizing the cost $\tilde{\alpha}$, the prediction queries and ground truth words are matched pairs. Here, $\tilde{i}_j$ denotes the current matched query of the ground truth word y_i after the overall matching process. For all matched pairs, its objective is measured with cross-entropy loss L_Q, and the ground truth of the unmatched queries is also set as $N_v + 1$, followed by the settings in semantic refinement module.

By expanding the extracted visual features and harnessing learnable queries for omitted semantic information beyond the retrieval process via the attention mechanism, the Transformer Mapper network effectively handles the intricacies of visual information, enabling the acquisition of more contextually relevant and well-structured visual representations.

3.3. Cross-Modal Decoder

After obtaining the enriched visual tokens $\tilde{V}_I$ with query tokens $\tilde{Q}_I$ from the visual encoder and the refined semantic tokens $\tilde{V}_s$ from the semantic refinement module, we now focus on integrating these representations into the Transformer-based cross-modal decoder for sentence generation. To be specific, the target caption of the current input image is denoted as $S = [s_0, \ldots, s_{L-1}]$, where L is the word number of the sentence. The sentence S is then passed through the word embedding layer, which converts input words into dense vector representations denoted as $\tilde{S}^0 = [\tilde{s}_0^0, \ldots, \tilde{s}_{L-1}^0]$, which is represented as the initial "Input Embed" in Figure 2. In each Transformer block, a multi-head attention layer is utilized to capture the attention between the dense vector representations and current input embedding with the concatenated tokens $\tilde{W}_I = [\tilde{V}_I, \tilde{Q}_I]$ and semantic tokens $\tilde{V}_s$, respectively. After the attention layer, both features are combined

and fused through a fusion layer. At t time step, the ith cross-modal Transformer block is implemented as follows:

$$z_t^i = multihead(\tilde{s}_t^i, \tilde{S}_{0:t}^i, \tilde{S}_{0:t}^i) \tag{9}$$

$$att_{(I)t}^i = multihead(z_t^i, \tilde{W}_I, \tilde{W}_I) \tag{10}$$

$$att_{(s)t}^i = multihead(z_t^i, \tilde{V}_s, \tilde{V}_s) \tag{11}$$

$$att_{(f)t}^i = Fusion(concat(att_{(I)t}^i, att_{(s)t}^i)) \tag{12}$$

where $att_{(f)t}^i$ represents the fusion of cross-modal visual features and semantic features, and Fusion refers to the projection layer. Then, the caption vectors are combined with the fusion visual semantic features $att_{(f)t}^i$ and passed through the feed forward layer with the sigmoid function to obtain the weight γ. Finally, the output of the ith Transformer block $\tilde{s}_t^{i+1}$ is calculated as follows:

$$\gamma = Sigmoid(Fusion(concat(z_t^i, att_{(f)t}^i))) \tag{13}$$

$$\tilde{s}_t^{i+1} = \tilde{s}_t^i + norm(\gamma z_t^i + (1-\gamma) att_{(f)t}^i) \tag{14}$$

Accordingly, the generated caption token of the final Transformer block at time t is denoted as $\tilde{s}_t^{N_d}$ and used to predict the c_{t+1} word of the decoded sentence $C = [c_1, \ldots, c_{L-1}]$.

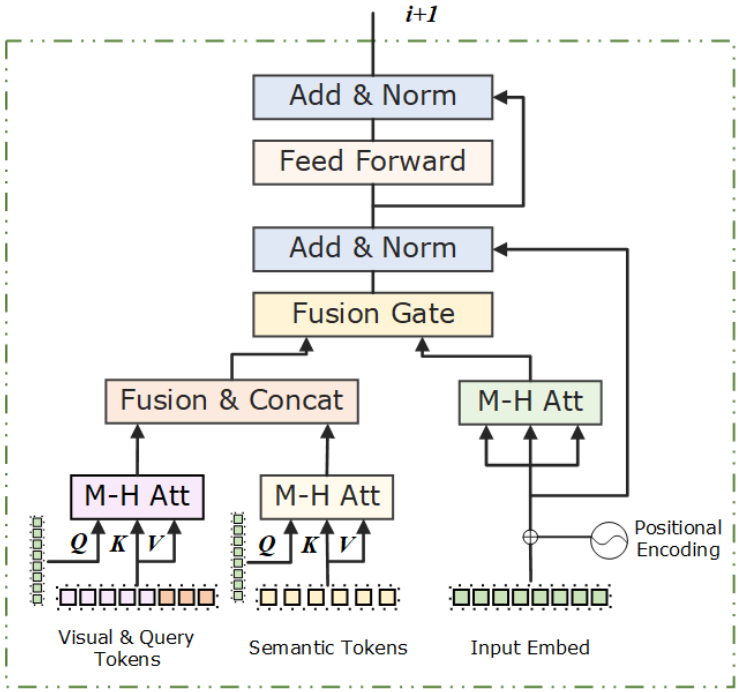

Figure 2. The structure of the ith cross-modal transformer block of the decoding module. The "M-H Att" denotes the multi-head attention layer. The "Input Embed" denotes the input of the ith block.

For caption generation, we train the proposed model by optimizing cross-entropy loss:

$$L_c = -\sum_{t=0}^{L-1} log(p_\theta(s_t|s_{0:t-1})) \tag{15}$$

where θ represents the training parameters, $s_t \in S$ is the target caption, as mentioned earlier, and $p_\theta(s_t|s_{0:t-1})$ is the predicted probability of generating token s_t given the preceding tokens $s_{0:t-1}$.

The overall objective of our CRSR model is given by the combination of the proxy objective in the semantic refinement module L_w , the query loss L_Q in Transformer Mapper, and sentence generation loss L_c:

$$L = L_w + L_Q + L_c \tag{16}$$

4. Experiments and Analysis

In this section, we present the experimental evaluation of our method and discuss the results. We start by introducing the publicly available datasets that were used in our experiments, as well as the evaluation metrics commonly employed in the field. Then, we provide details of the implementation setup, including the model configuration and training procedure. Additionally, we perform a series of ablation experiments to evaluate the individual contributions of each module in our method and to compare the performance with the state-of-the-art approaches.

4.1. Datasets

In our experiments, we utilized three publicly available RSIC datasets: RSICD, UCM-Captions, and Sydney-Captions. These datasets were used to evaluate the performance of our proposed method.

4.1.1. RSICD

The RSICD dataset, introduced in [2], is the largest publicly available dataset for remote sensing image captioning. It consists of 10,921 images collected from various mapping platforms, such as Google Earth, Baidu Map, MapABC, and Tianditu. The images in the dataset are 224 $\times$ 224 pixels and cover 30 different scene categories. Initially, the dataset provided 24,333 unique sentences as captions for the images. To increase the amount of information in the dataset, the captions were expanded by randomly duplicating the existing sentences, resulting in a total of 54,605 captions, with each image having five descriptions.

4.1.2. UCM-Captions

The UCM-Captions dataset [24] is derived from the UC Merced land-use dataset [40]. It comprises 2100 high-resolution aerial images, with 100 images per scene and a total of 21 scenes. The images have a resolution of 256 $\times$ 256 pixels. Each image in the dataset is accompanied by five descriptions, resulting in a total of 10,500 sentences for the dataset.

4.1.3. Sydney-Captions

The Sydney-Captions dataset is based on the Sydney dataset [41] and was introduced in the same study [24]. The dataset consists of 613 high-resolution RSIs acquired from the Sydney area of Google Earth. The images are cropped to a size of 500 $\times$ 500 pixels, with a pixel resolution of 0.5 m. For each image in the dataset, five different descriptions are provided, resulting in a total of 3065 sentences.

4.1.4. Datasets Alignment

To ensure a fair comparison and align with the settings used in the papers that introduced the datasets [2,24], we divided the data into three sets: training, evaluation, and testing. Following a standard practice of allocating 80% of the data for training, 10% for evaluation, and the remaining 10% for testing. For this study, we utilized an updated version of the three benchmark datasets provided by [25]. The key details and modifications made to the datasets are shown in Table 1. With the error descriptions and words removed, the vocabulary size was also changed.

Table 1. Datasets Comparison.

Datasets	Categories	Mean Caption Length	Vocab Size (before)	Vocab Size (after)	Image Numbers
Sydney-Captions	21	11.5	315	298	2100
UCM-Captions	7	13.2	231	179	613
RSICD	30	11.4	2695	1252	10,000

4.2. Evaluation Metrics

We utilized a comprehensive set of nine evaluation metrics to evaluate the performance of our proposed method and assess the quality of the generated sentences. These metrics include BLEU-n (n = 1, 2, 3, 4), Recall-Oriented Understudy for Gisting Evaluation (ROUGE_L), Metric for Translation Evaluation with Explicit Ordering (METEOR), Consensus-Based Image Description Evaluation (CIDEr), Semantic Propositional Image Caption Evaluation (SPICE), and Sm.

BLEU-n: The Bilingual Evaluation Understudy [42] is a widely used metric originally developed for evaluating machine translation systems. It measures the co-occurrence of n-grams (contiguous sequences of n words) between the generated sentences and the reference (ground truth) sentences. The value of n can be chosen as 1, 2, 3, or 4, representing unigrams, bigrams, trigrams, and four-grams, respectively.

ROUGE_L: The ROUGE_L metric [43] is a widely used evaluation metric in the fields of automatic summarization and machine translation. It measures the F-measure of the longest common subsequence (LCS) between the generated sentences and the reference (ground truth) sentences. The LCS represents the longest sequence of words that appears in both the generated and reference sentences. By computing the F-measure based on the LCS, it provides an indication of how well the generated sentences capture the key information and content of the reference sentences.

METEOR: This method is a widely used metric for evaluating the quality of machine translation output [44]. It measures the similarity between a generated sentence and a reference sentence by considering word-to-word matches and aligning the words in both sentences. The final METEOR score is calculated as the harmonic mean of precision and recall, providing a balanced evaluation of the generated sentence quality.

CIDEr: This metric is specifically designed for evaluating image captioning tasks [45]. It takes into account the term frequency-inverse document frequency (TF-IDF) weights of n-grams in both the generated and ground truth sentences. By applying TF-IDF weights, CIDEr captures the importance of specific n-grams in the context of the entire corpus of captions. It considers not only the presence of relevant n-grams but also their rarity across the dataset. This allows CIDEr to provide a more comprehensive evaluation of the generated captions, taking into account both the accuracy of the generated descriptions and their distinctiveness compared to other captions.

SPICE: SPICE metric [46] constructs tuples from both the candidate (generated) captions and reference captions and then calculates the F-score based on the matching tuples. Unlike traditional n-gram-based metrics, SPICE focuses on capturing the semantic meaning of captions rather than relying on specific word sequences. It represents objects, attributes, and relationships in a graph-based representation, which makes it less sensitive to the specific choice of n-grams.

Sm: Sm is a metric proposed in the 2017 AI Challenger competition to evaluate the quality of generated sentences. It is the arithmetic mean of four popular evaluation metrics: BLEU-4, METEOR, ROUGE_L, and CIDEr. We also take the SPICE metric into consideration:

$$S_m = \frac{1}{5}(\text{BLEU}-4 + \text{METEOR} + \text{ROUGE\_L} + \text{CIDEr} + \text{SPICE}) \tag{17}$$

4.3. Experimental Details

In our CRSR model, the visual encoder, semantic refinement module, and sentence decoder are constructed with $N_r = 6$, $N_c = 3$, and $N_d = 6$ Transformer blocks with a

hidden state size of 512. The projection length of the visual mapper network is set to $k = 10$ for three datasets. In the semantic retrieval and query prediction task, we extract high-frequency nouns and adjectives from the ground truth sequences of the datasets to serve as labels for filtering and prediction. To ensure meaningful and representative labels, we consider the variations in dataset sizes and the occurrence of words within the captions. Specifically, we select words that appear more than 15, 50, and 50 times in the Sydney, UCM, and RSICD datasets, respectively. The frequency of word appearances is calculated based on the condition that a word must appear at least three times within each image's five captions. And the query length l for predicting omitted words is set as the average length of the overall overlooked words in the current dataset. For the Sydney, UCM, and RSICD datasets, l is set to 4, 2, and 7.

During training, we employ beam search decoding with a beam size of three. The modulating factor α for the semantic refinement loss is set to $\alpha = 0.1$. The entire architecture is optimized for 20 epochs with a batch size of 16. The Adam optimizer [47] is employed with a learning rate of 4×10^{-5} (warmup: 20,000 iterations). The experiments are conducted on a Tesla V100 GPU using PyTorch version 1.13.1.

4.4. Experiments on Image Encoder

In this section, we investigate the efficiency of different image features extracted from the pre-trained image extractors and explore the impact of varying projection lengths in the Transformer Mapper network. We conducted these experiments on all three datasets, RSICD, UCM-Captions, and Sydney-Captions, to evaluate the impact of different feature extractors and Transformer Mapper projection length on the overall caption generation performance.

4.4.1. Different Image Feature Extractors

We performed a comprehensive comparison of different image feature extractors in the CLIP model. Specifically, we evaluated the performance of the following feature extractors: RN50, RN50x4, RN101, ViT-B/16, and ViT-B/32 [20]. The results of the comparison experiments under different image feature extractors are shown in Table 2. The best results are highlighted in bold.

Table 2. Comparison results of different image feature extractors.

Dataset	Model	$B-1$	$B-2$	$B-3$	$B-4$	M	R	C	S	S_m
Sydney-Captions	RN50	0.7827	0.7241	0.6775	0.6369	0.4141	0.7351	2.7790	0.4614	1.0053
	RN50x4	0.7826	0.7226	0.6701	0.6265	0.3984	0.7260	2.7286	0.4680	0.9895
	RN101	0.7935	0.7284	0.6792	0.6358	0.3951	0.7230	2.7214	0.4789	0.9908
	ViT-B/16	0.7756	0.7137	0.6612	0.6146	0.3958	0.7303	2.8420	0.4670	1.0100
	ViT-B/32	**0.7994**	**0.7440**	**0.6987**	**0.6602**	**0.4150**	**0.7488**	**2.8900**	**0.4845**	**1.0397**
UCM-Captions	RN50	0.8959	0.8425	0.7941	0.7491	0.4794	0.8425	3.8020	0.5253	1.2797
	RN50x4	0.9034	0.8553	**0.8137**	**0.7751**	**0.4994**	0.8472	3.7401	**0.5448**	1.2813
	RN101	0.9026	0.8559	0.8097	0.7638	0.4853	0.8544	**3.8206**	0.5179	1.2884
	ViT-B/16	0.8928	0.8451	0.7981	0.7520	0.4919	**0.8587**	3.7753	0.5277	1.2811
	ViT-B/32	**0.9060**	**0.8561**	0.8122	0.7681	0.4956	0.8586	3.8069	0.5201	**1.2899**
RSICD	RN50	0.8063	0.7029	0.6178	0.5458	0.3911	0.7027	2.9708	0.5085	1.0238
	RN50x4	0.8056	0.7025	0.6173	0.5458	0.3909	0.7011	2.9630	0.5133	1.0228
	RN101	0.7998	0.6981	0.6122	0.5392	0.3921	0.6993	2.9863	0.5130	1.0260
	ViT-B/16	0.8136	0.7077	0.6207	0.5484	0.3956	0.7052	3.0062	0.5190	1.0349
	ViT-B/32	**0.8192**	**0.7171**	**0.6307**	**0.5574**	**0.4015**	**0.7134**	**3.0687**	**0.5276**	**1.0537**

For the RSICD and Sydney-Captions datasets, different image feature extractors significantly affect the experiment results. The best CIDEr score obtained for the RSICD dataset is 3.0687, while the best SPICE score is 0.5276 when ViT-B/32 is used as the feature extractor. Similarly, for the Sydney-Captions dataset, our model achieves the best BLEU1-4, METEOR, ROUGE-L, and SPICE scores based on ViT-B/32. What is remarkable is that even when employing other feature extractors, our model maintains state-of-the-art performance, showcasing its robustness and effectiveness across different image feature representations.

This demonstrates the adaptability and generalization capabilities of our model, which can deliver competitive results regardless of the feature extractor used.

Thus, based on these comprehensive results, ViT-B/32 stands out as the preferred choice for image feature extraction in our model, as it consistently delivers outstanding performance across the overall three datasets, ensuring optimal caption generation results.

4.4.2. Transformer Mapper Projection Lengths

By varying the projection length in the Transformer Mapper network, we can effectively control the intricacies of the visual information that the model can capture. To investigate the impact of different projection lengths on caption generation performance, we conducted experiments with projection lengths of 5, 10, 15, and 20.

The results, as shown in Table 3, clearly demonstrate that the choice of projection length in the Transformer Mapper network has a significant impact on the model's caption generation performance. The best results of different length settings are highlighted in bold. Among the tested lengths, setting the projection length to 10 consistently yields the best results on all three datasets. This indicates that a projection length of 10 strikes an optimal balance between capturing relevant visual features and managing the complexity of the visual information. When the projection length is too short (e.g., 5), the model's visual representation lacks sufficient context and detail, leading to a degradation in caption quality. Conversely, when the projection length is too long (e.g., 15 or 20), the model may become overwhelmed with excessive visual information, leading to a scattered and less-focused representation, which also negatively impacts caption generation performance.

Table 3. Comparison results of different project lengths.

Dataset	Length	$B-1$	$B-2$	$B-3$	$B-4$	M	R	C	S	S_m
Sydney-Captions	5	0.7715	0.7130	0.6666	0.6300	0.4068	0.7372	2.8260	0.4596	1.0119
	10	**0.7994**	**0.7440**	**0.6987**	**0.6602**	**0.4150**	**0.7488**	**2.8900**	**0.4845**	**1.0397**
	15	0.7886	0.7220	0.6662	0.6187	0.3991	0.7232	2.7212	0.4557	0.9836
	20	0.7829	0.7148	0.6549	0.6068	0.4007	0.7261	2.7861	0.4756	0.9991
UCM-Captions	5	0.8851	0.8345	0.7911	0.7502	0.4871	0.8424	3.7313	0.5010	1.2624
	10	**0.9060**	0.8561	0.8122	0.7681	**0.4956**	**0.8586**	**3.8069**	0.5201	**1.2899**
	15	0.8924	0.8410	0.7976	0.7574	0.4899	0.8531	3.7736	**0.5272**	1.2802
	20	0.9034	**0.8580**	**0.8144**	**0.7689**	0.4860	0.8471	3.7945	0.4990	1.2791
RSICD	5	0.8050	0.6978	0.6072	0.5309	0.3957	0.7065	2.9726	0.5164	1.0245
	10	**0.8192**	**0.7171**	**0.6307**	**0.5574**	0.4015	**0.7134**	**3.0687**	**0.5276**	**1.0537**
	15	0.8110	0.7070	0.6205	0.5471	0.3973	0.7083	3.0261	0.5208	1.0399
	20	0.8054	0.7038	0.6203	0.5500	**0.4030**	0.7114	3.0364	0.5193	1.0440

By setting the projection length to 10, the model is allowed to effectively capture contextually meaningful visual representations and avoid overwhelming amounts of information, enabling it to generate more accurate and coherent captions. This emphasizes the importance of selecting an appropriate projection length to ensure optimal performance in the caption generation task and highlights the effectiveness of the Transformer Mapper network in handling visual information at a moderate granularity level.

4.5. Ablation Studies

In this section, we conduct an ablation study to investigate how each design in our CRSR model influences the overall performances on the RSICD, UCM-Captions, and Sydney-Captions datasets.

baseline model: In the baseline model (denoted as "bs"), we use a Transformer-based encoder–decoder structure, which utilizes only CLIP features as visual inputs and does not incorporate any supplemented semantic information. This serves as the foundation for our CRSR model, which incorporates additional components and modifications to enhance its performance.

bs+m: This denotes the baseline model with the Transformer Mapper network added to the visual encoder. The modification aims to enhance the visual encoding process, resulting in a more comprehensive and informative visual representation.

bs+mq: "mq" signifies the Transformer Mapper network with the inclusion of query prediction for additional semantic information. This guides the generated visual tokens to focus more on the critical regions of the image.

bs+sr: "sr" denotes the semantic refinement module, which introduces the retrieved words and the filtering of semantic tokens to our model.

bs+mq+sr: This denotes that both the Transformer Mapper network with learnable queries and the semantic refinement module are introduced into the baseline model. With the addition of both modules, our model can generate captions that have a more accurate and comprehensive structure.

We analyzed the effect of each submodule in combination with the experimental results. The results demonstrate the impact of each submodule on the overall performance of our model. The best results among three datasets are highlighted in bold.

As shown in Table 4, the comparison between the baseline model and "bs+m" model underscores the significance of incorporating the projection and attention mechanism of the Transformer Mapper network. This inclusion leads to increased BLEU1-4 and S_m scores across all three datasets, indicating that the model effectively captures the relationships and dependencies within the visual features. The BELU-4 scores increased by 0.67%, 2.07%, and 2.23% in the Sydney-Captions, UCM-Captions, and RSICD datasets, respectively. We observed more significant improvements in the larger dataset, which indicates that the Transformer Mapper network is particularly effective in capturing complex relationships and dependencies within the visual features when dealing with larger and more diverse datasets. The larger dataset provides a richer and more diverse set of visual information, benefiting from the self-attention mechanism of the Transformer Mapper network. As a result, the model can better understand the spatial and contextual information present in the images, leading to more accurate and informative captions. This highlights the scalability and generalization capabilities of the Transformer Mapper network, making it a valuable addition to the caption generation model for larger and more challenging datasets.

Table 4. Ablation study results on the three datasets.

Dataset	Model	$B-1$	$B-2$	$B-3$	$B-4$	M	R	C	S	S_m
Sydney-Captions	bs	0.7697	0.6914	0.6219	0.5584	0.3817	0.6933	2.4097	0.4283	0.8943
	bs+m	0.7754	0.6985	0.6297	0.5651	0.3894	0.7042	2.4429	0.4261	0.9055
	bs+mq	0.7947	0.7200	0.6546	0.5932	0.4019	0.7265	2.6237	0.4461	0.9583
	bs+sr	0.7873	0.7088	0.6425	0.5857	0.4035	0.7217	2.6867	0.4542	0.9704
	bs+mq+sr	**0.7994**	**0.7440**	**0.6987**	**0.6602**	**0.4150**	**0.7488**	**2.8900**	**0.4845**	**1.0397**
UCM-Captions	bs	0.8295	0.7660	0.7184	0.6747	0.4467	0.7820	3.5171	0.4898	1.1821
	bs+m	0.8434	0.7870	0.7414	0.7010	0.4639	0.8068	3.5358	0.5072	1.2029
	bs+mq	0.8623	0.8145	0.7703	0.7287	0.4694	0.8246	3.4941	0.5001	1.2034
	bs+sr	0.8918	0.8457	0.8015	0.7602	0.4909	0.8531	3.7335	0.5109	1.2697
	bs+mq+sr	**0.9060**	**0.8561**	**0.8122**	**0.7681**	**0.4956**	**0.8586**	**3.8069**	**0.5201**	**1.2899**
RSICD	bs	0.7823	0.6729	0.5837	0.5090	0.3899	0.7012	2.8694	0.5076	0.9954
	bs+m	0.8038	0.6943	0.6045	0.5313	0.3890	0.6964	2.8975	0.5063	1.0041
	bs+mq	0.8068	0.6978	0.6090	0.5352	0.3894	0.6985	2.9658	0.5095	1.0197
	bs+sr	0.7977	0.6906	0.6018	0.5288	0.3952	0.6956	2.9830	0.5189	1.0243
	bs+mq+sr	**0.8192**	**0.7171**	**0.6307**	**0.5574**	**0.4015**	**0.7134**	**3.0687**	**0.5276**	**1.0537**

Comparing the results of "bs+m" and "bs+mq", the additional inclusion of learnable queries significantly enhances the model's performance. By fusing the projected image features with learnable queries that predict critical semantic information, the attention mechanism further improves the extraction of semantically relevant features in the image. The overall metric scores exhibit great improvements on the Sydney-Captions and RSICD datasets, with S_m scores increasing by 5.28% and 1.56%, respectively, while there are relatively smaller improvements on UCM-Captions. This discrepancy can be attributed to the shorter query length of UCM-Captions compared to the other two datasets, which

could limit the potential for additional improvements. Given that the retrieval results on UCM-Captions already include enough semantic information, setting a longer query length for repeated semantic information is not necessary.

With the added semantic refinement module, the experimental results of "bs" and "bs+sr" in Table 4 demonstrate notable improvements in caption generation. Incorporating the retrieval and filtering of semantic tokens results in improved metric scores across all three datasets. For instance, in the largest dataset, RSICD, BLEU-4, SPICE, and S_m scores increased by 1.98%, 1.13%, and 2.89%, respectively. In the UCM-Captions and Sydney-Captions datasets, there is even greater improvement with the semantic refinement module. Specifically, in UCM-Captions, SPICE and S_m scores are 1.49% and 7.03% higher, respectively. In Sydney-Captions, SPICE and S_m scores improved by 2.59% and 7.61%, compared to the baseline model. These results demonstrate the effectiveness of the semantic retrieval and refinement module in refining generated captions and enhancing overall performance.

Integrating both the Transformer Mapper network and the semantic refinement module into the model, the results of "bs+m+sr" in comparison to "bs+m" and "bs+sr" further affirm the cumulative benefits of both submodules. Notably, within the RSICD and Sydney-Captions datasets, there are significant improvements in BLEU1-4, CIDEr, and S_m metrics. Moreover, the CIDEr, SPICE, and S_m scores in the UCM-Captions dataset also exhibit compelling enhancements, showcasing the effectiveness of these combined submodules across diverse datasets.

4.6. Comparison with Other Methods

In this section, we conduct extensive comparative experiments with seventeen state-of-the-art methods to demonstrate the effectiveness of our proposed CRSR method. The overall experiments result of three datasets are shown in Tables 5–7. The best results among three datasets are highlighted in bold.

Table 5. Comparison results on the Sydney-Captions dataset.

Dataset	$B-1$	$B-2$	$B-3$	$B-4$	M	R	C	S	S_m
Soft-Att [2]	0.7322	0.6674	0.6223	0.5820	0.3942	0.7127	2.4993	-	-
Hard-Att [2]	0.7591	0.6610	0.5889	0.5258	0.3898	0.7189	2.1819	-	-
Up-Down [48]	0.8180	0.7484	0.6879	0.6305	0.3972	0.7270	2.6766	-	-
MLAM [25]	0.7900	0.7108	0.6517	0.6052	0.4741	0.7353	2.1811	0.4089	0.8809
Re-ATT [27]	0.8000	0.7217	0.6531	0.5909	0.3908	0.7218	2.6311	0.4301	0.9529
GVFGA + LSGA [29]	0.7681	0.6846	0.6145	0.5504	0.3866	0.7030	2.4522	0.4532	0.9091
SVM-D CONC [30]	0.7547	0.6711	0.5970	0.5308	0.3643	0.6746	2.2222	-	-
FC-ATT [13]	0.8076	0.7160	0.6276	0.5544	0.4099	0.7114	2.2033	0.3951	0.8355
SM-ATT [13]	0.8143	0.7351	0.6586	0.5806	0.4111	0.7195	2.3021	0.3976	0.8593
SAT (LAM-TL) [17]	0.7425	0.6570	0.5913	0.5369	0.3700	0.6819	2.3563	0.4048	0.8698
Adaptive (LAM-TL) [17]	0.7365	0.6440	0.5835	0.5348	0.3693	0.6827	2.3513	0.4351	0.8746
struc-att [18]	0.7795	0.7019	0.6392	0.5861	0.3954	0.7299	2.3791	-	-
JTTS [16]	**0.8492**	**0.7797**	0.7137	0.6496	0.4457	0.7660	2.8010	0.4679	1.0260
Meta-ML [34]	0.7958	0.7274	0.6638	0.6068	0.4247	0.7300	2.3987	-	-
SCST [35]	0.7643	0.6919	0.6283	0.5725	0.3946	0.7172	2.8122	-	-
DTFB [37]	0.8373	0.7771	**0.7198**	**0.6659**	**0.4548**	**0.7860**	**3.0369**	0.4839	**1.0855**
CASK [38]	0.7908	0.7200	0.6605	0.6088	0.4031	0.7354	2.6788	0.4637	0.9780
ours	0.7994	0.7440	0.6987	0.6602	0.4150	0.7488	2.8900	**0.4845**	1.0397

Regarding the Sydney-Captions dataset, our proposed CRSR model exhibits competitive performance when compared with state-of-the-art methods. Notably, it achieved the highest SPICE scores, although it lags slightly behind the DTFB method in overall performance. This achievement is significant given that the Sydney-Captions dataset is relatively small, consisting of only 613 RSIs, in comparison to other datasets. This limited data size might have contributed to the relatively lower scores obtained by our model. As shown in Table 5, despite this limitation, our model still gained a competitive score of 1.0397 for the comprehensive metric S_m compared with the overall methods.

Table 6. Comparison results on the UCM-Captions dataset.

Dataset	$B-1$	$B-2$	$B-3$	$B-4$	M	R	C	S	S_m
Soft-Att [2]	0.7454	0.6545	0.5855	0.5250	0.3886	0.7237	2.6124	-	-
Hard-Att [2]	0.8157	0.7312	0.6702	0.6182	0.4263	0.7698	2.9947	-	-
Up-Down [48]	0.8356	0.7748	0.7264	0.6833	0.4447	0.7967	3.3626	-	-
MLAM [25]	0.8864	0.8233	0.7735	0.7271	0.5222	0.8441	3.3074	0.5021	1.1806
Re-ATT [27]	0.8518	0.7925	0.7432	0.6976	0.4571	0.8072	3.3887	0.4891	1.1679
GVFGA + LSGA [29]	0.8319	0.7657	0.7103	0.6596	0.4436	0.7845	3.3270	0.4853	1.1400
SVM-D CONC [30]	0.7653	0.6947	0.6417	0.5942	0.3702	0.6877	2.9228	-	-
FC-ATT [13]	0.8135	0.7502	0.6849	0.6352	0.4173	0.7504	2.9958	0.4867	1.1339
SM-ATT [13]	0.8154	0.7575	0.6936	0.6458	0.4240	0.7632	3.1864	0.4875	1.1435
SAT (LAM-TL) [17]	0.8208	0.7856	0.7525	0.7229	0.4880	0.7933	3.7088	0.5126	1.2450
Adaptive (LAM-TL) [17]	0.857	0.812	0.775	0.743	**0.510**	0.826	3.758	**0.535**	1.2734
struc-att [18]	0.8538	0.8035	0.7572	0.7149	0.4632	0.8141	3.3489	-	-
JTTS [16]	0.8696	0.8224	0.7788	0.7376	0.4906	0.8364	3.7102	0.5231	1.2596
Meta-ML [34]	0.8714	0.8199	0.7769	0.7390	0.4956	0.8344	3.7823	-	-
SCST [35]	0.8727	0.8096	0.7551	0.7039	0.4652	0.8258	3.7129	-	-
DTFB [37]	0.8230	0.7700	0.7228	0.6792	0.4439	0.7839	3.4629	0.4825	1.1705
CASK [38]	0.8900	0.8416	0.7987	0.7575	0.4931	0.8578	**3.8314**	0.5227	**1.2925**
ours	**0.9060**	**0.8561**	**0.8122**	**0.7681**	0.4956	**0.8586**	3.8069	0.5201	1.2899

In the case of the UCM-Captions dataset, our CRSR model outperforms existing methods, achieving the highest scores in most of the metrics, with the exception of SPICE and CIDEr, where it remains competitive. A substantial improvement is observed in our model's BLEU1-4 scores compared to the previous state-of-the-art methods. Furthermore, the S_m metric score remains highly competitive when compared to the CASK method, demonstrating the capability to generate more descriptive and contextually relevant captions of our method.

Table 7. Comparison results on the RSICD dataset.

Dataset	$B-1$	$B-2$	$B-3$	$B-4$	M	R	C	S	S_m
Soft-Att [2]	0.6753	0.5308	0.4333	0.3617	0.3255	0.6109	1.9643	-	-
Hard-Att [2]	0.6669	0.5182	0.4164	0.3407	0.3201	0.6084	1.7925	-	-
Up-Down [48]	0.7679	0.6579	0.5699	0.4962	0.3534	0.6590	2.6022	-	-
MLAM [25]	0.8058	0.6778	0.5866	0.5163	0.4718	0.7237	2.7716	0.4786	0.9924
Re-ATT [27]	0.7729	0.6651	0.5782	0.5062	0.3626	0.6691	2.7549	0.4719	0.9529
GVFGA + LSGA [29]	0.6779	0.5600	0.4781	0.4165	0.3285	0.5929	2.6012	0.4683	0.8815
SVM-D CONC [30]	0.5999	0.4347	0.3355	0.2689	0.2299	0.4557	0.6854	-	-
FC-ATT [13]	0.6671	0.5511	0.4691	0.4059	0.3225	0.5781	2.5763	0.4673	0.8700
SM-ATT [13]	0.6699	0.5523	0.4703	0.4068	0.3255	0.5802	2.5738	0.4687	0.8710
SAT (LAM-TL) [17]	0.6790	0.5616	0.4782	0.4148	0.3298	0.5914	2.6672	0.4707	0.8946
Adaptive (LAM-TL) [17]	0.6756	0.5549	0.4714	0.4077	0.3261	0.5848	2.6285	0.4671	0.8828
struc-att [18]	0.7016	0.5614	0.4648	0.3934	0.3291	0.5706	1.7031	-	-
JTTS [16]	0.7893	0.6795	0.5893	0.5135	0.3773	0.6823	2.7958	0.4877	0.9713
Meta-ML [34]	0.6866	0.5679	0.4839	0.4196	0.3249	0.5882	2.5244	-	-
SCST [35]	0.7836	0.6679	0.5774	0.5042	0.3672	0.6730	2.8436	-	-
DTFB [37]	0.7581	0.6416	0.5585	0.4923	0.3550	0.6523	2.5814	0.4579	0.9078
CASK [38]	0.7965	0.6856	0.5964	0.5224	0.3745	0.6833	2.9343	0.4914	1.0012
ours	**0.8192**	**0.7171**	**0.6307**	**0.5574**	**0.4015**	**0.7134**	**3.0687**	**0.5276**	**1.0537**

Furthermore, on the RSICD dataset, which is the largest among the three datasets, our CRSR model continues to deliver remarkable performance. It achieves the highest scores in all evaluated metrics, outperforming the previous state-of-the-art methods. Notably, our model exhibits a substantial improvement of 13.44% in the CIDEr metric over the CASK method, which obtained the second-highest score. Moreover, there is a significant increase in all other evaluated metrics, reaffirming the CRSR model's superiority in generating high-quality captions for RSIs.

4.7. Analysis of Training and Testing Time

In the context of practical applications, algorithmic efficiency holds paramount importance. For a comprehensive evaluation of the efficiency of our approach, we measured key parameters, including training time, testing time, and the total number of parameters with the S_m metric. The comparison was conducted between the baseline model and our proposed method using the RSICD dataset, and the results are presented in Table 8.

Table 8. Comparison of the training and testing time on the RSICD dataset.

Model	Training Time/Epoch(s)	Testing Time/Epoch(s)	Params(M)	S_m
bs	375	40	52.02	0.9954
bs+mq+sr(ours)	414	45	67.26	1.0537

Analyzing the results from the comparative experiments, it is evident that our method, incorporating a Transformer Mapper and semantic refinement module, leads to an increase in the number of parameters compared to the baseline model. Despite this increment in parameters, the performance gains in caption generation are substantial. Therefore, when weighing the time cost against performance factors, our method exhibits a favorable trade-off, incurring a relatively small increase in time cost for a significant improvement in performance.

4.8. Qualitative Analysis

4.8.1. Generated Captioning Results

As illustrated in Figure 3, we demonstrate several generated captions from all three datasets to conduct a qualitative analysis of our proposed CRSR method. In each row of the figure, we present the image with the ground truth (GT) sentences, the description generated by the baseline model (base) used in Ablation Studies, the description generated by our CRSR method (ours), the retrieved words (retr) based on CLIP, and the missing words (miss) that will be predicted by the learnable queries. To highlight the contributions of our CRSR method, the words that are uniquely generated by our model are marked in blue. Additionally, we identify and mark incorrect words, and words should be filtered from the retrieved words in red.

Comparing the captions generated by the baseline model with our CRSR method, we can observe that our approach effectively captures more precise and intricate information from the input image. As depicted in Figure 3, our model generates more comprehensive details, particularly concerning the overall scene categories, the quantity of objects, and the descriptive words that articulate relationships within the scene. For instance, in the first image of the first row, generated by our CRSR method, it accurately identifies “baseball field” and proceeds to generate descriptions like “many houses”, “surrounded”, and “sands” in the subsequent images of the same row. In addition, our model demonstrates its capacity to filter out misleading retrieved words and reconstruct missing words based on the prediction queries. To illustrate, in the first image of the second row and the second image of the third row, the retrieved words “closed” and “square” are effectively filtered out by our model. Moreover, in the subsequent images of the second row, our model generates missing words such as “yellow”, “near”, and “road”. The retrieved words, when appropriately filtered and integrated into the captioning process, enhance the overall coherence and relevance of the generated captions. In addition, the presence of missing words in our generated captions showcases the reconstruction ability of the Transformer Mapper module.

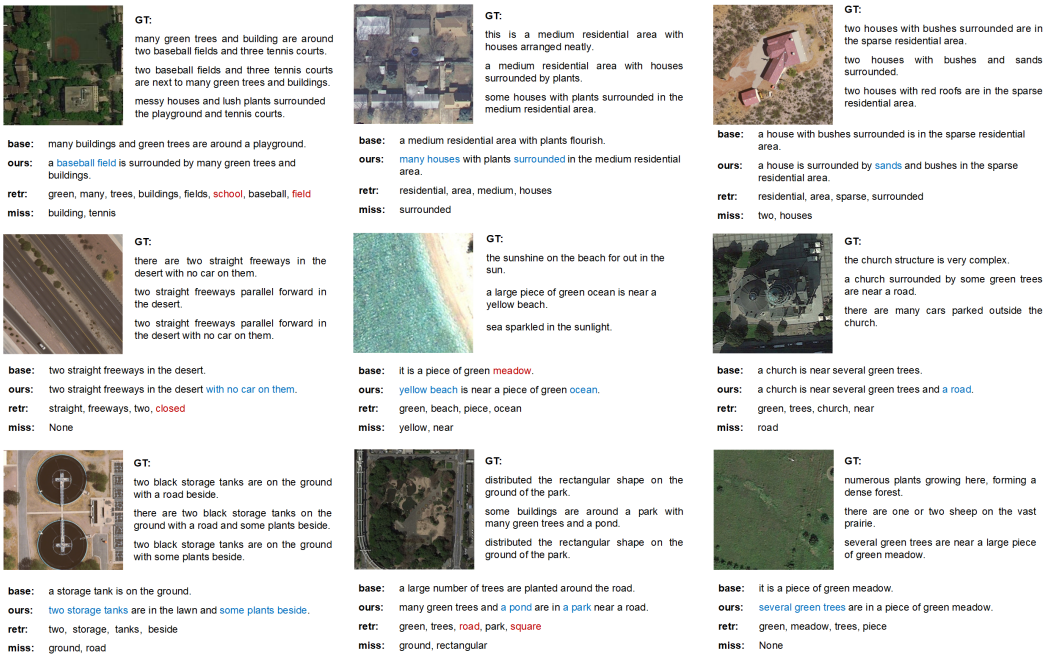

Figure 3. Captioning results of the baseline model and our proposed CRSR method. The "GT" denotes the ground truth captions. The "base" denotes the description generated by the baseline model. "ours" denotes the description generated by our CRSR method. "retr" denotes the retrieved words from the sentence pool, while "miss" denotes the overlooked words during retrieval. Blue words denote the scene uniquely captured by our model compared with baseline model. Red words denote the incorrectly generated words in description and retrieved words.

4.8.2. Visualization of Attention Weights

We provide visualizations of attention weights in Figure 4, offering insights into the attention mechanism of our CRSR model during caption generation. The y-axis in this figure represents the words retrieved through our retrieval method, while the x-axis corresponds to the generated captions. This visualization allows us to understand how the model prioritizes the retrieved semantic tokens, observing how the model filters out irrelevant words while utilizes existing words to generate more relevant captions.

It can be seen from Figure 4 that the model concentrates on the corresponding words when generating the caption sequence due to the introduction of semantic information. Notably, it accurately focuses on terms like "road" and "residential" when the same words are implemented. Furthermore, the presence of semantic tokens leads to heightened attention towards pertinent words during generation, as exemplified by "across", "dense", and "neatly". The incorporation of semantic information significantly influences attention and word prioritization during caption generation, which ensures that the generated captions are more aligned with the visual content of the input images, resulting in increased accuracy and relevance.

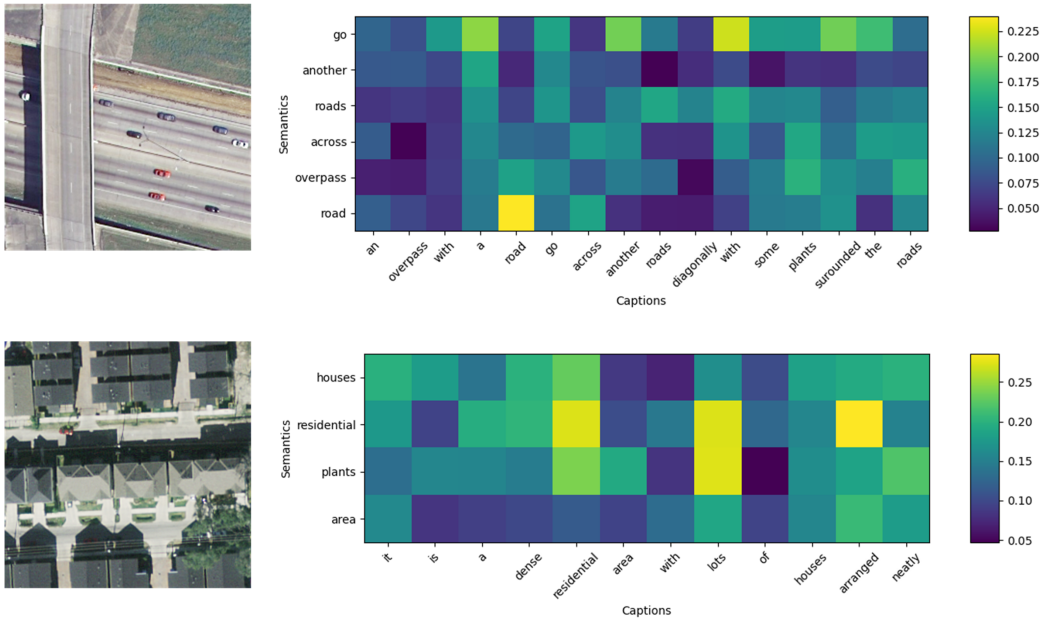

Figure 4. Visualized attention matrix between the semantic tokens and the captioning of the input images.

5. Conclusions

In this article, we introduce the CRSR image captioning method for RSIs, incorporating two distinct modules for enhanced caption generation. The semantic refinement module systematically organizes semantic information. We initially leveraged the powerful CLIP model, fine-tuned on RSIC datasets, for retrieving relevant words from the sentence pool, providing valuable supplementary information for RSIs. Subsequently, we further refined and filtered primary semantic information using a masked-attention strategy. Additionally, within the Transformer Mapper network, we introduce learnable queries to enhance the model's understanding of semantically relevant information and extend the representation of image features through a projection mechanism. The comprehensive experiment results from both quantitative and qualitative evaluations validate the superiority of our method over existing state-of-the-art approaches in RSIC.

However, there are still challenges and limitations in our method that need to be addressed in future research. Training the retrieval model separately from the captioning model introduces the risk of semantic misalignment between the retrieved words and the actual content of the RSIs. Moreover, the semantic refinement module still introduces some errors with contradictory semantic information during the refinement process, resulting in incorrect captions. To overcome these challenges, future research will focus on exploring improvements in the current retrieval method and semantic refinement module to better align the retrieval and captioning models.

Author Contributions: Z.L. designed the model, then implemented the model and wrote the paper. W.Z., X.D., G.Z., and S.Z. contributed to the supervision of the work, analysis of the method, and paper writing. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The RSICD, UCM-Captions, and Sydney-Captions datasets are available for download from https://github.com/201528014227051/RSICD_optimal (accessed on 11 December 2023).

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

References

1. 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]
2. Lu, X.; Wang, B.; Zheng, X.; Li, X. Exploring models and data for remote sensing image caption generation. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 2183–2195. [CrossRef]
3. Recchiuto, C.T.; Sgorbissa, A. Post-disaster assessment with unmanned aerial vehicles: A survey on practical implementations and research approaches. *J. Field Robot.* **2018**, *35*, 459–490. [CrossRef]
4. Tian, Y.; Sun, X.; Niu, R.; Yu, H.; Zhu, Z.; Wang, P.; Fu, K. Fully-weighted HGNN: Learning efficient non-local relations with hypergraph in aerial imagery. *ISPRS J. Photogram. Remote Sens.* **2022**, *191*, 263–276. [CrossRef]
5. Hossain, M.Z.; Sohel, F.; Shiratuddin, M.F.; Laga, H. A comprehensive survey of deep learning for image captioning. *ACM Comput. Surv.* **2019**, *51*, 1–36. [CrossRef]
6. Zhao, B. A systematic survey of remote sensing image captioning. *IEEE Access* **2021**, *9*, 154086–154111. [CrossRef]
7. Elman, J.L. Finding structure in time. *Cognit. Sci.* **1990**, *14*, 179–211. [CrossRef]
8. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
9. Sun, X.; Tian, Y.; Lu, W.; Wang, P.; Niu, R.; Yu, H.; Fu, K. From single- to multi-modal remote sensing imagery interpretation: A survey and taxonomy. *Sci. China Inf. Sci.* **2023**, *66*, 140301. [CrossRef]
10. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 2048–2057.
11. Cheng, G.; Li, Q.; Wang, G.; Xie, X.; Min, L.; Han, J. SFRNet: Fine-Grained Oriented Object Recognition via Separate Feature Refinement. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5610510. [CrossRef]
12. Niu, R.; Sun, X.; Tian, Y.; Diao, W.; Chen, K.; Fu, K. Hybrid Multiple Attention Network for Semantic Segmentation in Aerial Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5603018. [CrossRef]
13. 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]
14. Wang, B.; Zheng, X.; Qu, B.; Lu, X. Retrieval topic recurrent memory network for remote sensing image captioning. *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.* **2020**, *13*, 256–270. [CrossRef]
15. Chen, J.; Han, Y.; Wan, L.; Zhou, X.; Deng, M. Geospatial relation captioning for high-spatial-resolution images by using an attention-based neural network. *Int. J. Remote Sens.* **2019**, *40*, 6482–6498. [CrossRef]
16. Ye, X.; Wang, S.; Gu, Y.; Wang, J.; Wang, R.; Hou, B.; Giunchiglia, F.; Jiao, L. A Joint-Training Two-Stage Method For Remote Sensing Image Captioning. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4709616. [CrossRef]
17. 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]
18. Zhao, R.; Shi, Z.; Zou, Z. High-Resolution Remote Sensing Image Captioning Based on Structured Attention. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5603814. [CrossRef]
19. Sarto, S.; Cornia, M.; Baraldi, L.; Cucchiara, R. Retrieval-Augmented Transformer for Image Captioning. In Proceedings of the 19th International Conference on Content-Based Multimedia Indexing, Graz, Austria, 14–16 September 2022.
20. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the ICML, Online, 18–24 July 2021.
21. Shen, S.; Li, L.H.; Tan, H.; Bansal, M.; Rohrbach, A.; Chang, K.W.; Yao, Z.; Keutzer, K. How much can clip benefit vision-and-language tasks? *arXiv* **2021**, arXiv:2107.06383.
22. Jiasen, L.; Goswami, V.; Rohrbach, M.; Parikh, D.; Lee, S. 12-in-1: Multi-task vision and language representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 July 2020; pp. 10437–10446.
23. Mokady, R.; Hertz, A.; Bermano, A.H. ClipCap: CLIP prefix for image captioning. *arXiv* **2021**, arXiv:2111.09734.
24. 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. [CrossRef]
25. 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]
26. Huang, W.; Wang, Q.; Li, X. Denoising-based multiscale feature fusion for remote sensing image captioning. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 436–440. http://doi.org10.1109/LGRS.2020.2980933. [CrossRef]
27. 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]
28. Li, X.; Zhang, X.; Huang, W.; Wang, Q. Truncation cross entropy loss for remote sensing image captioning. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 5246–5257. [CrossRef]
29. 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.* **2022**, *60*, 5608816. [CrossRef]

30. Hoxha, G.; Melgani, F. A novel SVM-based decoder for remote sensing image captioning. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5404514. [CrossRef]
31. 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]
32. Sumbul, G.; Nayak, S.; Demir, B. SD-RSIC: Summarization-driven deep remote sensing image captioning. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 6922–6934. [CrossRef]
33. Kandala, H.; Saha, S.; Banerjee, B.; Zhu, X. Exploring Transformer and Multilabel Classification for Remote Sensing Image Captioning. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6514905. [CrossRef]
34. Yang, Q.; Ni, Z.; Ren, P. Meta captioning: A meta learning based remote sensing image captioning framework. *ISPRS J. Photogram. Remote Sens.* **2022**, *186*, 190–200. [CrossRef]
35. Zhang, X.; Li, Y.; Wang, X.; Liu, F.; Wu, Z.; Cheng, X.; Jiao, L. Multi-Source Interactive Stair Attention for Remote Sensing Image Captioning. *Remote Sens.* **2023**, *15*, 579. [CrossRef]
36. 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]
37. 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. Topics Appl. Earth Observ. Remote Sens.* **2023**, *16*, 7704–7717. [CrossRef]
38. Li, Y.; Zhang, X.; Cheng, X.; Tang, X.; Jiao, L. Learning consensus-aware semantic knowledge for remote sensing image captioning. *Pattern Recognit.* **2024**, *145*, 109893. [CrossRef]
39. Carion, N.; Massa, F.; Synnaeve, G. End-to-end object detection with transformers. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020; pp. 213–229.
40. Yang, Y.; Newsam, S. Bag-of-visual-words and spatial extensions for land-use classification. In Proceedings of the 18th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS), San Jose, CA, USA, 2–5 November 2010; pp. 270–279.
41. Zhang, F.; Du, B.; Zhang, L. Saliency-guided unsupervised feature learning for scene classification. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 2175–2184. [CrossRef]
42. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318.
43. Banerjee, S.; Lavie, A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the Second Workshop on Statistical Machine Translation (StatMT), Morristown, NJ, USA, 29–30 June 2005; pp. 65–72.
44. Lin, C.-Y. *Rouge: A Package for Automatic Evaluation of Summaries*; Association for Computational Linguistics: Barcelona, Spain, 2004; pp. 74–81.
45. Vedantam, R.; Zitnick, C.; Parikh, D. CIDEr: Consensus-based image description evaluation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 4566–4575.
46. Anderson, P.; Fernando, B.; Johnson, M.; Gould, S. SPICE: Semantic propositional image caption evaluation. *Proc. Eur. Conf. Comput. Vis.* **2016**, *9909*, 382–398.
47. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the ICLR, San Diego, CA, USA, 7–9 May 2015.
48. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6077–6086.