remote sensing

MDPI

Review

A Review of Deep Learning-Based Remote Sensing Image Caption: Methods, Models, Comparisons and Future Directions

Ke Zhang [1,2,*], Peijie Li [1] and Jianqiang Wang [1]

[1] Department of Electronic and Communication Engineering, North China Electric Power University, Baoding 071003, China; 220222215024@ncepu.edu.cn (P.L.); 220222215019@ncepu.edu.cn (J.W.)
[2] Hebei Key Laboratory of Power Internet of Things Technology, North China Electric Power University, Baoding 071003, China
* Correspondence: zhangkeit@ncepu.edu.cn

Abstract: Remote sensing images contain a wealth of Earth-observation information. Efficient extraction and application of hidden knowledge from these images will greatly promote the development of resource and environment monitoring, urban planning and other related fields. Remote sensing image caption (RSIC) involves obtaining textual descriptions from remote sensing images through accurately capturing and describing the semantic-level relationships between objects and attributes in the images. However, there is currently no comprehensive review summarizing the progress in RSIC based on deep learning. After defining the scope of the papers to be discussed and summarizing them all, the paper begins by providing a comprehensive review of the recent advancements in RSIC, covering six key aspects: encoder–decoder framework, attention mechanism, reinforcement learning, learning with auxiliary task, large visual language models and few-shot learning. Subsequently a brief explanation on the datasets and evaluation metrics for RSIC is given. Furthermore, we compare and analyze the results of the latest models and the pros and cons of different deep learning methods. Lastly, future directions of RSIC are suggested. The primary objective of this review is to offer researchers a more profound understanding of RSIC.

Keywords: remote sensing; image caption; encoder–decoder framework; attention mechanism; reinforcement learning; auxiliary task; large visual language model; few-shot learning

check for updates

Citation: Zhang, K.; Li, P.; Wang, J. A Review of Deep Learning-Based Remote Sensing Image Caption: Methods, Models, Comparisons and Future Directions. *Remote Sens.* **2024**, *16*, 4113. https://doi.org/10.3390/rs16214113

Received: 6 September 2024
Revised: 23 October 2024
Accepted: 31 October 2024
Published: 4 November 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

The world today is confronting daunting challenges posed by rapid population growth, depletion of non-renewable resources [1] and deteriorating ecological environments. To tackle these challenges, there is an urgent need for humans to continuously and repeatedly observe the earth from space to gain a more comprehensive understanding of its geography and changes. Remote sensing technology offers an unprecedentedly powerful tool for this purpose, enabling us to more precisely detect and harness resources, more effectively manage the environment and more scientifically plan industrial production layouts. Consequently, remote sensing technology has become an indispensable scientific technology in today's society.

The advancement of computer science, sensors, aerospace and spatial positioning technologies has led to a rapid improvement in modern remote sensing technology. The ability to offer high-resolution images at the level of 0.3 to 0.5 m has become possible by the maturation of high-resolution remote sensing satellite technologies [2], such as WorldView3 [3], WorldView4 [4] and Pleiades Neo [5]. These satellites can provide high-quality data assistance for a range of applications by recording incredibly fine features and details of the Earth's surface. These high-quality, time-sensitive, high-resolution and high-definition remote sensing images have become the main data source for human beings to obtain geospatial information. The information contained in remote sensing images is

extremely rich, and it has become a major topic to figure out how to make sense of the vast amount of information for humans' production and existence.

Remote sensing image-processing techniques have produced fruitful research results in various application areas, such as scene classification [6,7], change detection [8,9], object detection [10,11], semantic segmentation [12,13] and geographic image retrieval [14,15]. Although these tasks are good at recognizing or classifying objects in a scene, they cannot capture the relationships between objects, nor can they perform high-level semantic reasoning based on such scenes to obtain information in natural language that is easier to understand and utilize. With the advancement of modern artificial intelligence technology, the scope of remote sensing applications has been further expanded. As a result, obtaining natural language captions from remote sensing images and interpreting them at a higher semantic level have emerged to be a novel subject.

Remote Sensing Image Caption (RSIC) is extended on the basis of natural image caption, where we input a satellite remote sensing image to obtain a sentence about describing the geographic location information of the objects within the image. This task is widely used in the military field and civilian field. In the military field, existing algorithms can be used to realize military reconnaissance and intelligence collection [16]. Significant military targets, such as ports, airports, ships and other significant military targets, can be promptly identified from the remote sensing image first, and the wide-format battlefield image can be accurately and comprehensively understood by using RSIC technology, so as to realize real-time decoding of the battlefield geography and dynamic intelligence generation and to generate faster battlefield text information. In civilian applications, RSIC can precisely provide easy-to-understand information on disaster assessment [17,18], vegetation cover [19], crop growth [20] and urban changes [21] to facilitate agricultural research and urban planning. For example, after the occurrence of natural disasters (such as earthquake, flood, fire, etc.), remote sensing images can quickly obtain images of the disaster area, assess the damage degree and impact scope of the disaster, and provide decision support for rescue operations. Or high-resolution remote sensing images can provide detailed urban geographic information, including roads, buildings, green spaces, etc., which is helpful for urban planning departments to carry out scientific and reasonable urban planning.

Compared to nature image caption, RSIC encounters different challenges, which are listed as follows.

- Different scales. Remote sensing images are typically acquired from a distance, whereas natural images are typically taken closer to the item. As a result, objects in remote sensing images are not only greater in size but also harder to distinguish.
- Different object distribution. Objects in natural images are often located in the center of the image, while the distribution of objects in remote sensing images is extremely complex and likely to be in the corners of the picture.
- Different number of objects. Natural images often have only a few key figures or animals, while remote sensing images usually involve a huge number of buildings.
- Different viewpoints of the shot, making it challenging to discern between foreground and background. The foreground and background of a natural image are instantly recognizable by human eyes, and we can naturally find the main subject and construct sentences. In contrast, remote sensing images are recorded in God's perspective [22], where the main objects in the image are similar in size to the background or are directly hidden from it, making them difficult to find for humans and even harder for machines.

In recent years, an increasing number of scholars have been working to provide technical solutions, large-scale datasets and ideas with potential applications for RSIC. Gradually, pertinent information, techniques and assessment measures have come together to establish a stand-alone system distinct from other domains. Numerous review papers of other remote sensing image-processing tasks have been published. For example, researchers in [23–25] conducted a review on remote sensing image classification, while studies

in [26–28] focused on a review of semantic segmentation for remote sensing images, and some scholars reviewed remote sensing image-change detection [29–31]. But there are few review articles on RSIC that can provide researchers with a systematic analysis of the current state of research, method comparisons and trends in this area. Although Zhao [32] conducted a preliminary investigation into RSIC, which gave researchers a brief understanding of the field, there were not many papers in this field at that time, and early investigations seem outdated from today's perspective. Bashmal et al. [33] conducted a review on the application of remote sensing images in natural language processing (NLP), but they did not specialize in RSIC tasks and did not summarize the development under different deep learning methods.

This paper summarizes the frequently used methods, models, datasets and evaluation metrics in RSIC. We reviewed RSIC by using different deep learning methods as an entry point. In Section 2, we provide a brief explanation of the sources of the articles discussed in this paper and the criteria used for filtering them, in Section 3, we comprehensively introduce the latest research progress achieved in RSIC under different methods and in Section 4 we give a brief explanation of the most commonly used datasets and evaluation metrics of RSIC. The performance of different models in terms of metrics is compared on different datasets in Section 5; we also analyze the pros and cons of seven methods in this section. The future directions in RSIC are summarized and presented in Section 6. Lastly, we conclude our work in Section 7.

2. Methods

This paper contains 97 relevant articles in total. This number accounts only for the articles related to the tasks discussed in this paper and excludes those solely cited for the purpose of explaining problems. To obtain high-quality and highly pertinent literature, the papers are comprised of two main components. The majority of these papers were sourced from Google Scholar, with keywords including but not limited to "remote sensing image caption", "remote sensing image captioning", "remote sensing image description" and "remote sensing image change caption or captioning". Additionally, a smaller number of articles were identified through references in papers from peers that did not contain the specified keywords but addressed related tasks; these papers can be included in discussions after reading by authors, such as those concerning remote sensing large visual language models. To ensure the quality of the selected literature, we conducted a preliminary filtering process.

Generally, the filtered papers exhibit the following characteristics:

1. The results presented in the paper are unreasonable or significantly deviate from findings reported in peers' papers.
2. Full-text access to certain articles could not be obtained from their respective publishers.
3. Only image caption applications in remote sensing are included, while traditional image caption tasks are not.
4. Papers that employ auxiliary tasks for generating remote sensing image captions may be included. However, those utilizing remote sensing image captions to assist other tasks will be excluded.
5. Regarding papers on remote sensing image retrieval, understanding, interpretation, analysis and other related topics, we cannot judge by a simple paper title, we will check the evaluation metrics used in experimental results. Papers which use evaluation metrics that do not belong to this area may fall outside this paper's scope.

To provide researchers with a perception of trends in the field, we tallied the journals from which our articles originated and the percentage of years during which those journals were published. These are shown in Figures 1 and 2, respectively.

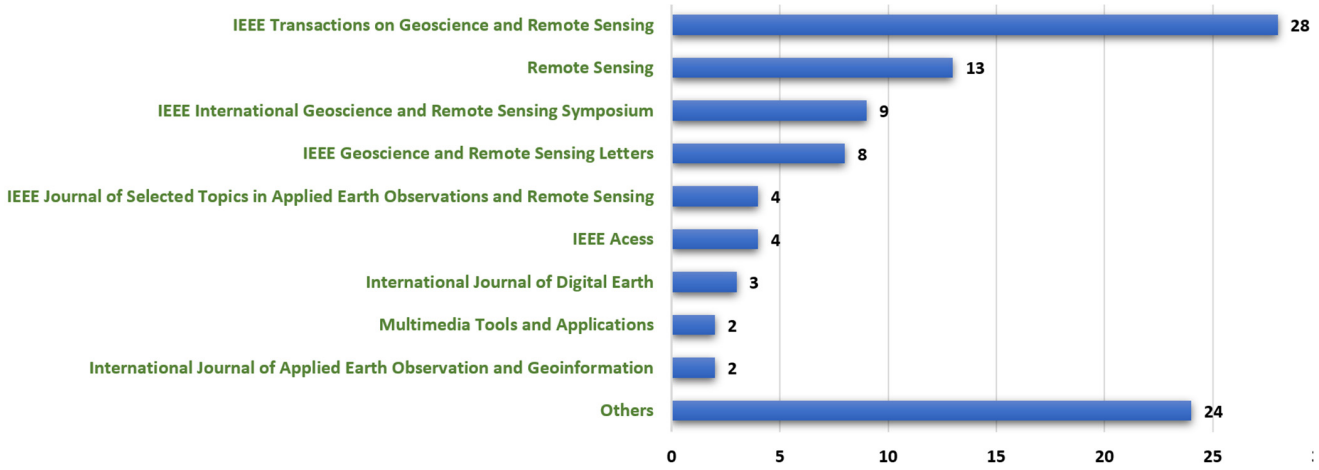


Figure 1. Statistics of journal sources.

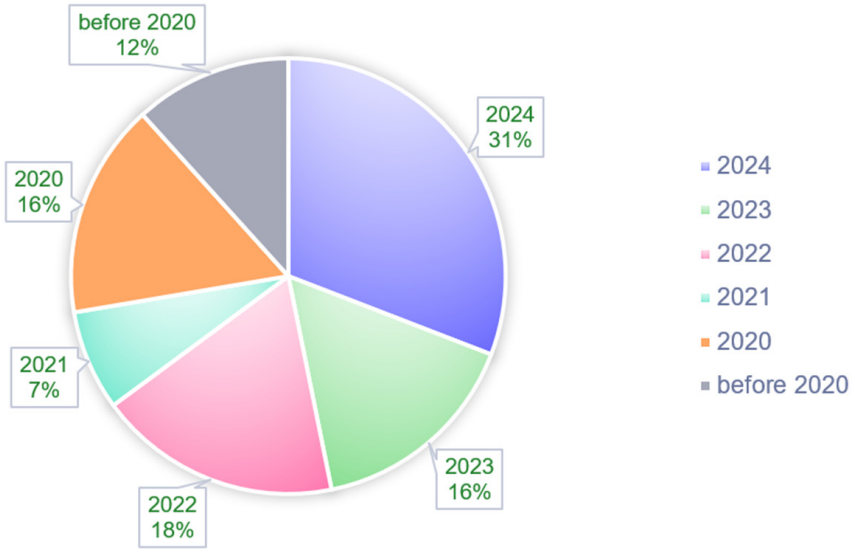


Figure 2. The proportion of the year of papers.

As is evident from Figure 1, the majority of papers in this fields are published in authoritative journals in remote sensing, while a certain number are published in authoritative journals of other fields.

As illustrated in Figure 2, there has been a significant increase in the number of papers within the field of RSIC in recent years. In 2024 in particular, a large number of relevant papers have been published, indicating the flourishing development of this field.

3. Review of RSIC Based on Deep Learning

In this section, we provide a detailed review of the achievements of RSIC in recent years. We introduce the current progress of RSIC from the aspects of encoder–decoder framework, attention mechanism, reinforcement learning, auxiliary task learning, large visual language models (LVLMs) and few-shot learning. Each of these will be detailed in the following eight sections. The structure of this section is depicted in Figure 3.

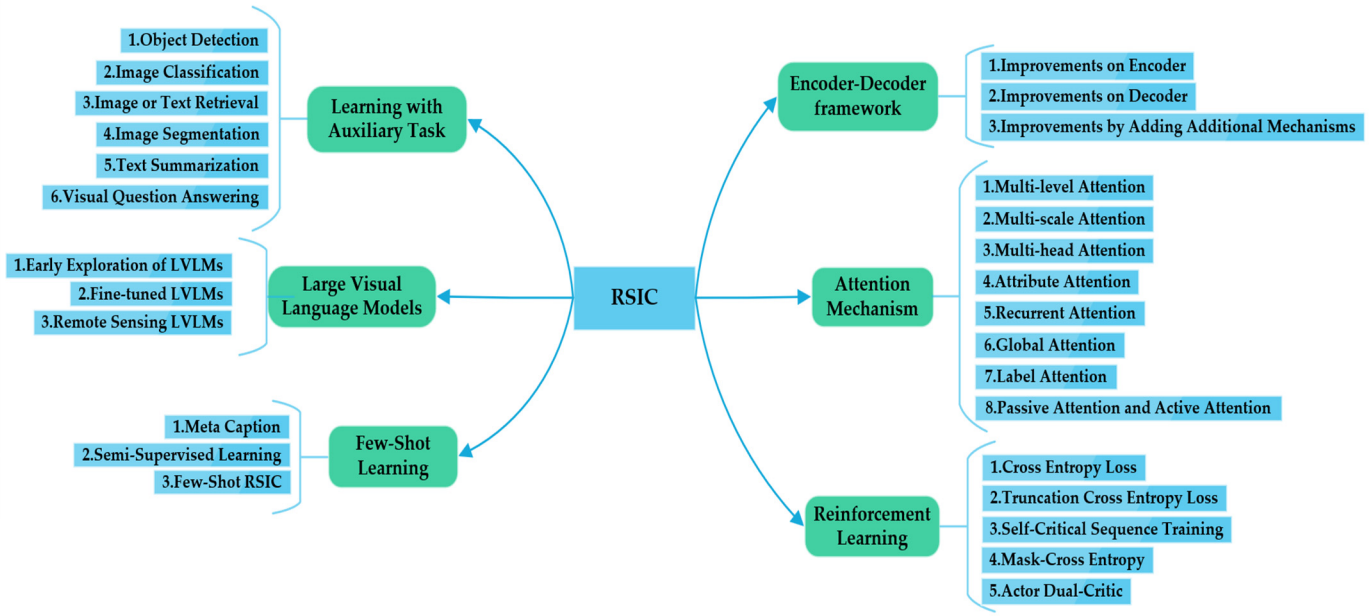


Figure 3. The structure of Section 3.

3.1. Encoder–Decoder Framework

Most of the achievements in RSIC originate from the encoder–decoder framework [34]. In image caption tasks, encoders are often used to extract image features, while decoders employ the features that are extracted to produce sentences. In the process of generating an image caption in early models, the Convolutional Neural Network (CNN) [35] first embeds the input image into a fixed-length feature vector that effectively captures its high-level characteristics, rendering it suitable for various visual tasks. This extracted feature vector is subsequently utilized as input to the Recurrent Neural Network (RNN) [36], which generates a textual caption pertinent to the content of the image. During model training, the cross-entropy [37] loss function is employed to assess the difference between the words predicted by the model and ground truth. Through the backpropagation algorithm, model parameters are optimized based on this loss value, thereby progressively minimizing errors and enhancing both the accuracy and the quality of generated captions. Traditional encoder–decoder models lack effective explanations and are often criticized as "black box models". However, if they are seen as a word extractor and a sentence generator [38], it is much easier to understand. The word extractor is responsible for extracting valuable descriptive words for remote sensing images, while the sentence generator is responsible for organizing these words into a coherent sentence. Many scholars have devoted themselves to improving the encoder–decoder framework and achieved good results.

3.1.1. Improvements on Encoder

Similar to natural image caption, early RSIC models [39] also employed a CNN network as an encoder and an RNN network as a decoder. During this period, three foundational datasets for the RSIC were also publicly released. Despite their simple model structure, these early RSIC models were capable of generating satisfactory captions for remote sensing images. Initially, researchers focused on enhancing the encoder by utilizing various CNN architectures to extract image features [40] because higher quality feature extraction can lead to improved performance. There are some examples including ResNet-50 [41], ResNet-101 [42], VGG-16 [43], VGG-19 [44] and DenseNet121 [45].

To a certain extent, the depth of the network correlates positively with the quality of the extracted features. The advancements presented in studies [41–45] are primarily attributed to the augmentation of network depth. The structure of VGG is more concise and consistent, consisting of multiple small convolutional kernels stacked together, and the small kernel convolution facilitates the extraction of detailed features in the remote sensing

images. Conversely, ResNet has efficiently addressed the gradient-vanishing issue in deep neural network training by introducing residual connection [46], enabling the construction of exceptionally deep network structures. Table 1 shows the approximate GFLOPs and parameters of traditional networks, which can be used to measure the operational complexity of networks.

Table 1. The approximate GFLOPs and params of different CNNs.

CNN	GFLOPs	Params
ResNet-50 [41]	4.1	25.6 M
ResNet-101 [42]	7.6	44.5 M
VGG-16 [43]	15.5	138 M
VGG-19 [44]	19.6	143 M
DenseNet121 [45]	8	8 M

As shown in the above table, the deepening of the network allows the model's computational ability to be strengthened, but at the same time the overfitting problem occurs beyond a certain limit, which affects the feature-extraction effect, especially for remote sensing images, where the datasets are usually small, and the overfitting problem is more likely to occur. Compared with ResNet, DenseNet mitigates overfitting by reducing the parameter count, thereby potentially achieving superior results on smaller datasets. A2DenseNet [47] incorporates dual attention mechanisms and enables one to extract weak remote sensing image features and multiscale information. It connects features at different depths with attention structures to learn the aligned representations between intricate image details and words more effectively.

Simply improving the encoder by iterating the depth of the network or optimizing the network model is not feasible, as the operation of the encoder will always be constrained by factors such as datasets and computational power. Improving the structure at encoder to assist in enhancing feature extraction has opened up a new avenue for improvement. Incorporating denoising-based multi-scale feature fusion (DMSFF) [48] into the encoder model is beneficial for extracting comprehensive visual features from remote sensing images. This is done by selecting multi-layer feature maps in the CNN to be fused after spatial-wise denoising (SD) and channel-wise denoising (CD), respectively, as the final feature representation. The conventional approaches for remote sensing image feature extraction typically rely on pretrained CNNs from the ImageNet dataset [49] as encoders, which may produce suboptimal results due to their limited adaptability to complex remote sensing environments. To address this issue, fine-tuning the encoder using a variational autoencoder (VAE) is proposed [50]. This process involves not only image classification and reconstruction but also incorporates large-scale remote sensing classification datasets during training to enhance the model's adaptation to small RSIC datasets. VAE solves overfitting problems and harnesses both low-level and high-level features of remote sensing images, while low-level features often contain detailed information, high-level features encompass semantic information [22]. The integration of these aspects significantly enhances RSIC performance. The procedure for extracting remote sensing image features using the pre-trained CLIP model is similar [51]. As CLIP is trained on a diverse range of images rather than specifically on remote sensing imagery, direct utilization of CLIP for remote sensing image feature extraction may cause suboptimal outcomes. After fine-tuning with the RSICD dataset [40], the image and text encoders of CLIP can be separately leveraged to extract remote sensing image features and textual features.

3.1.2. Improvements on Decoder

As advancements in the encoder continue, there have been significant developments in the decoder as well. The primary function of the decoder is to translate the features extracted by the encoder into natural language, thereby achieving the conversion from features to text. Initially, RNN, Long Short-Term Memory (LSTM) [52] and Gated Recurrent

Unit (GRU) [53] were widely used as decoders in RSIC. For example, Chouaf et al. [54] used CNN to extract features from bi-temporal images, and then these features were merged into a unique vector by a fusion block. The decoder used RNN to generate captions of changes in different remote sensing objects. While RNN is adept at processing sequential data, it often encounters challenges with vanishing or exploding gradients when handling excessively long sequences. LSTM effectively addresses gradient-explosion issues through its specialized gating mechanism, memory cell design and combinable gradient-clipping techniques. As a variant of RNN with lower computational complexity than LSTM, GRU also possesses capabilities for mitigating gradient vanishing or exploding issues. However, it still faces challenges when dealing with extremely lengthy sequences. Transformer [55] has demonstrated strong capabilities in NLP, particularly in sequence-to-sequence tasks [56] such as machine translation [57] and text summarization [58]; it has exhibited notable advantages in image caption [59,60]. Owing to its strong generalization ability, it has been applied to various tasks and domains in deep learning including RSIC.

The datasets commonly used in RSIC are relatively small, and the traditional encoder–decoder framework typically demands a substantial number of annotated samples with high computational requirements. Support vector machine (SVM) [61] is a well-known classifier in the field of remote sensing image processing. Unlike deep learning methods, SVM operates on the principle of margin maximization [62], which makes it less likely to overfit. In addition, the model complexity and number of parameters of CNN are higher than those of traditional machine learning methods like SVM and random forest [63], and they require fewer hyperparameters to obtain accurate results, thus reducing the complexity and calculation cost of the model. So utilizing an SVM as the decoder rather than RNN or LSTM can mitigate overfitting issues while accelerating both training and inference times. This approach is characterized by exceptionally brief training durations and rapid reasoning capabilities. This makes SVM well-suited for tasks on small datasets. Hoxha et al. [64] proposed two baseline methods for remote sensing image change caption (RSICC) based on the encoder–decoder framework. The purpose of RSICC is to describe the changes that occur between two different objects in the same geographical area. Unlike change-detection techniques, RSICC focuses more on understanding semantic-level changes in images rather than pixel-level changes. Figure 4 is the structure of a simple RSICC model. Generally, the dual-time features are extracted using CNNs, aggregated through specific methods, and then decoded by a decoder to generate change captions.

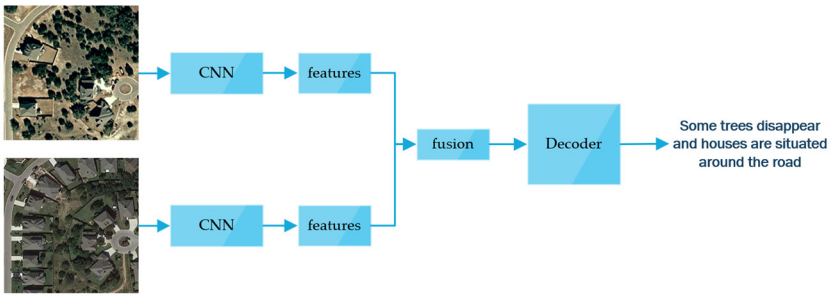


Figure 4. The structure of a simple RSICC model.

In the two baselines of Hoxha et al.'s work, the difference was in the decoder part. The first method used a classic RNN decoder, while the second method used a decoder based on SVM. Moreover, employing a decoder comprised of multiple SVM networks (SVMs) [65] can mitigate overfitting issues even when sample size is limited, yielding not only improved results but also requiring less training and inference time. In their experiment, the image features are mapped in an embedding space whose dimension is 128. And n is the number of SVM cascades in the decoder, then they conducted experiments with parameter n from 0 to 6, and the results obtained showed significant improvement when n was less than 4, and when n was more than 4, the improvement was no longer

significant, and in some cases, it began to decline. The structure of a decoder based on SVMs is shown in Figure 5.

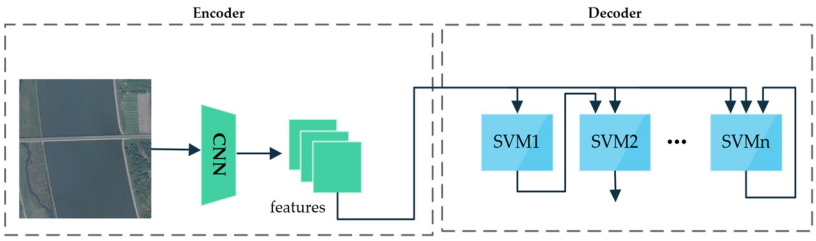


Figure 5. An RSIC model with multiple SVM networks as a decoder.

As sentences lengthen, LSTM may lose some information, possibly due to limited information capacity. This issue can be solved by integrating an external memory. The persistent memory mechanism (PMM) [66] is capable of decoding and mapping the output sequence information to an external memory for updating and retrieval. It maintains and outputs memory information over extended periods without being affected by the issue of vanishing gradients. A single-layer LSTM may lose essential contextual information for the RSIC. In contrast, the bidirectional LSTM (BiLSTM) [47] facilitates both forward and backward propagation, allowing the collection of information from previous data along the backward propagation path. This capability enables the hidden layer to capture past and future contextual cues, thereby enhancing model performance compared to a single-layer LSTM while simultaneously increasing model complexity.

Transformer can be used as a decoder to handle and catch long-distance dependencies between pixels, making it more accurate and comprehensive in describing images. Wang et al. [67] proposed CapFormer, which was a pure transformer model. They used a vision transformer [68] to obtain image representations and a transformer decoder to convert image features into sentences sequentially. Li et al. [42] developed a transformer model suitable for small datasets, and they added an extra dropout layer after the output of each sublayer in the decoder, making it suitable for scenarios with limited RSIC datasets. The authors in [41] leveraged a pre-trained BERT [69] model in the decoder to produce word embeddings enriched with contextual information. In order to effectively utilize multi-scale information and generate accurate and detailed sentences, researchers in [70] proposed the multi-layer aggregated transformer (MLAT). The experimental results showed that MLAT significantly solved multi-scale problems and improved the model's performance in sentence accuracy and diversity.

3.1.3. Improvements by Adding Additional Mechanisms

In addition to enhancing the models at both the encoder and decoder, introducing specific mechanisms into the encoder–decoder framework without altering the model structure can also facilitate RSIC. The multi-scale cropping training mechanism [71] is designed to enable the encoder to capture more detailed multi-scale features, thus mitigating overfitting. It involves dividing an image into three pictures of varying scale sizes and then resizing them to obtain multiple scale features. These are subsequently fed into an encoder–decoder framework for RSIC. Ramos et al. [72] have modified the traditional output mechanism of discrete probability distribution in RSIC. The discrete results pose challenges in optimizing training effects through backpropagation, so they transformed the output of the model into continuous word embeddings and optimized the model during training by comparing semantic similarity with reference captions in training data. Even with a basic model and minimal data, improved performance was achieved. Chen et al. [73] proposed a new mechanism on a transformer model with a caption-type controller (CTC) to adapt to various caption styles simultaneously, where necessary caption types can be incorporated into learnable sentence features using optional CTC blocks. To meet the

demands of various users and adjust to increasingly complicated remote sensing situations, we can construct different kinds of captions using various CTCs.

In order to correct potential errors that may occur during caption generation, Hoxha et al. [74] proposed two postprocessing strategies based on encoder–decoder structure. The first one is to detect and correct potential errors once the caption is fully generated from the decoder. In the second strategy, they detect potential errors at every moment in the process of generation, effectively preventing the accumulation of errors.

3.2. Attention Mechanism

The remarkable success of attention mechanisms in NLP has stimulated researchers' interest in image caption [75]. Xu et al. [76] pioneered the application of attention to image caption. The attention-based image caption algorithm focuses on different regions of the input image and assigns weights accordingly, enabling the model to gain a deeper understanding of the content and structure of the image. For instance, in the CNN + RNN architecture, we initially extract image features using a CNN network. During the decoding phase, as each word in the description is generated, the attention mechanism assesses the similarity between the current hidden state of the RNN and each feature region. Specifically, this model employs current hidden states alongside representations of feature areas to compute attention scores that indicate the significance of each feature area to the model. By applying the softmax function, these scores are normalized into attention weights, enabling the model to dynamically prioritize which features warrant greater focus. Attention mechanisms have gained significant advancements in natural image caption, including soft and hard attention [76], bottom-up and top-down attention [77], attention on attention [78], adaptive attention [79] and so on. Most of these developments are applicable to RSIC and serve as baselines [80]. However, remote sensing images differ significantly from natural images due to challenges related to distinguishing small targets amidst a large number of objects; traditional attention mechanisms often fail to produce satisfactory results. It is vital to enhance attention mechanisms that can adapt to the complicated remote sensing environments. As RSIC continues its development trajectory, an increasing number of advanced attention mechanisms are integrated into this field. Different users may need different caption styles.

3.2.1. Multi-Level Attention

The current focus of attention mechanisms is primarily on images, but in some cases, images may not be highly sensitive to attention. For instance, some remote sensing images are captured in a simplistic manner, featuring only plain grass or fields. Enhancing their efficacy solely through traditional attention methods proves challenging. Li et al. [81] also employed attention mechanisms for words, vision and semantics to constitute a multi-level attention mechanism. By considering the role of attention mechanisms across multiple dimensions, this approach aligns more closely with human cognition. Li et al. [82] introduced visual semantic co-attention (VSCA), which integrates visual and semantic attention vectors with semantic relationship features within a consensus exploitation (CE) module to acquire cross-modal consistent perceptual knowledge.

A better context vector is essential for exploring deeper semantic information of images. However, current remote sensing images not only lack strong capturing capabilities for multi-scale information, but also fail to effectively utilize contextual features. The multi-level and context-aware attention network (MLCA-Net) proposed by Cheng et al. [83] introduced multi-scale attention and context attention in the encoder. Unlike traditional attention mechanisms, MLCA-Net not only extracts multi-scale features but also leverages the context attention network to extract contextual features, focusing on complex scene context mining and semantic alignment between visual scenes and attribute information. The structure of MLCA-Net is depicted in Figure 6. Another model, MC-Net [84], is built upon MLCA-Net by incorporating a feature-aggregation module composed of multi-layer

transformers. This module integrates image features from different scales to capture global information, thereby enhancing the understanding of contextual information.

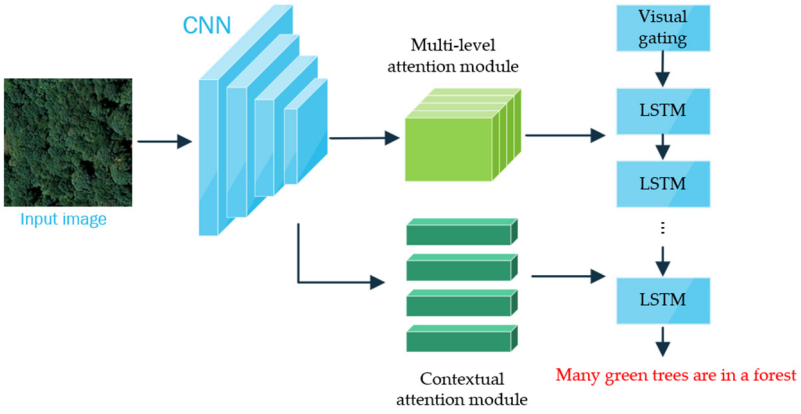


Figure 6. The framework of MLCA-Net.

Traditional attention mechanisms typically handle the focus of word vectors to the visual area through a weighted average approach, which may not be conducive to precise focus on key areas. Zhang et al. [85] proposed the multi-source interactive staircase attention mechanism (MSISAM), a staircase strategy to categorize attention weights into three levels: the core region, surrounding peripheral region and other outer regions. Different regions within the search range were allocated varying levels of attention. This method effectively emphasizes the most crucial regions in an image while correspondingly diminishing the weights of distant regions, thereby establishing a more tight and accurate mapping relationship between images and text. Instance-aware [86] technology is applied to separate semantic information in scenes from ground objects and their relationships. Through instance-aware technology, specific object features encompass not only their own details but also integrate adjacent patch information and global features to form a multi-instance level feature set. This strategy significantly enhances accuracy and flexibility in representing features at both spatial and semantic levels. The cross-hierarchy attention mechanism [87] can flexibly adapt to different levels of feature inputs and prompt decoders to dynamically focus on various semantic levels and specific instances during decoding. This dynamic adjustment not only improves RSIC accuracy but also strengthens the model's semantic understanding and scale adaptability when handling extreme remote sensing scenes.

3.2.2. Multi-Scale Attention

The size, shape and structure of objects in remote sensing images may vary. To better capture these changes and enhance our comprehension and analysis of image content, multi-scale features are frequently used in RSIC. In addition, multi-scale features aid in addressing issues stemming from scale variations, target distance and shading occlusion. Normally, different features can be observed at different scales; finer-grained sampling can unveil more intricate details, while coarser-grained sampling can reveal overarching trends. Combining multi-scale features enables the extraction of both global contextual information and local detailed insights, which is widely applied in RSIC [43,70,71,88].

Natural images are usually taken at a short distance, while remote sensing images are usually obtained over a wide scale, and the divergent acquisition methods make it difficult for models pretrained on natural images to capture multi-scale remote sensing targets. For example, the lack of height information caused by the shooting distance makes some remote sensing targets like grass harder to discern. Wang et al. [89] proposed a multi-scale multi-interaction RSIC model. After the CNN backbone is domain-adapted [90], features from different stages are collected as multi-scale feature representations, which are beneficial for obtaining better multi-scale features. The multi-interaction network is designed for the connection between different levels of features, where the information

flow can interact more fully. Features will become more distinguishable by calculating the similarity score between them, fusing features with high similarity and increasing the distance between features of different categories.

The scales of objects in remote sensing images vary greatly and different targets are located at different spatial positions, Yuan et al. [91] proposed a multi-level attention module to concentrate on various spatial positions and scales, which can adaptively extract specific spatial and scale features. The differences in spatial positions of objects across various viewpoint can also result in multi-scale problems, and the combination of multi-scale features formed in different viewpoints is beneficial for obtaining image information more comprehensively. Meng et al. [92] utilized a multi-view attention network (MAN) to realize the integration of features from different perspectives. By using MAN, more semantically rich features can be acquired.

3.2.3. Multi-Head Attention

The multi-head attention mechanism is an extended version of the self-attention mechanism. It divides the input data into multiple heads, conducts a self-attention calculation on each head, and eventually concatenates the results of the multiple heads to obtain the final output. Utilizing the multi-head attention mechanism enables the model to be more efficient when handling long sequence data. Figure 7 is the architecture of multi-head attention in a common transformer model, where Q, K and V stand for Query, Key and Value, respectively, which are the key concepts to implement the multi-head self-attention mechanism. Ne represents the number of layers in the encoder and Nd represents the number of layers in the decoder. Transformer is generally a 6 × 6 structure [55], but the number of layers can be changed according to the actual situation [42].

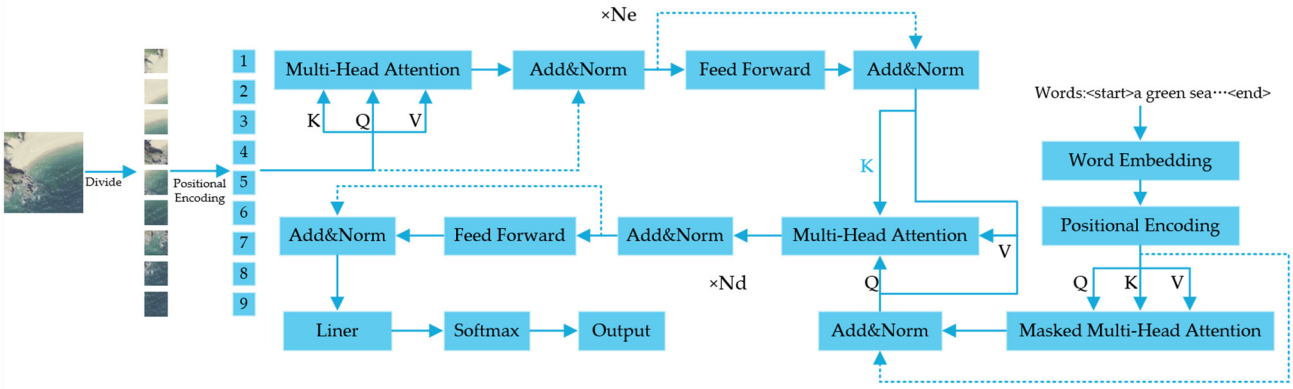


Figure 7. The framework of transformer with multi-head attention.

The attention mechanism in transformer has always been a focal point for experts and scholars to enhance the model, and a large amount of advanced research is based on multi-head attention. Zia et al. [93] proposed a transformer model with adaptive multi-head attention mechanism, whose decoder can dynamically weigh the contributions of multi-scale image features to the next word prediction to generate fine-grained captions. Inspired by the memory mechanism in [94], Gajbhiye et al. [95] constructed a transformer language decoder based on memory-guided attention. Memory elements were inserted into the multi-head attention mechanism of the decoder to extract intrinsic semantic relationships from spatial and channel attention feature maps. This helped the decoder to accurately describe the visual content in remote sensing. Li et al. [96] made aerial remote sensing images stereoscopic by separating the foreground and background. They designed a deformable transformer with deformable scale dot-product attention, enabling the acquisition of multi-scale features from both foreground and background through robust interactive learning capabilities. In order to solve the problems of multi-scale, multi-class and directional fuzziness of remote sensing images Wu et al. [97] make full use of transformer's semantic understanding ability and propose a cross-modal reasoning dual transformer model (TrTr-

CMR). They used a swin transformer [98] with a shifted window partitioning scheme as an encoder, which utilizes alternative window-based multi-head attention and shifted window-based multi-head attention to effectively extract visual features. This method is beneficial to the extraction of multi-scale visual features and the discovery of the intrinsic relationship of objects

Remote sensing images contain numerous similar objects, making it challenging to identify key elements and generate appropriate captions. Meng et al. [99] focused on prior knowledge-guided transformer (PKG-Transformer). They used transformer as a decoder to integrate the prior knowledge-guided attention (PKA) mechanism. This attention mechanism established the relationship between the scene areas and selected the objects more relevant to the scene. Despite achieving good results, the PKG-Transformer fails to make full use of the multi-scale features, so the generated captions are not comprehensive enough. They improved their work by proposing a multi-scale grouping transformer (MG-Transformer) [100]. Firstly, the multi-scale feature module (MFA) composed of CNN and CLIP was used to extract regional features and image embedding, and then combined with attention through semantic correlation (SC) module. Finally, a global grouping attention (GGA) mechanism was introduced to enable each attention head to focus on different local information. Compared with PKG-Transformer, it achieved better effects.

The current RSIC datasets solely encompass images and their corresponding captions, lacking essential region annotations, which severely constrains the advancement of RSIC. In order to fulfill the need for region-related datasets, Zhao et al. [101] presented region annotations for UCM captions and Sydney captions. This offered a unique technique for feature extraction and utilization in RSIC. They also proposed a region attention transformer (RAT) model to address the issue of information loss caused by pooling of regions of interest (RoI). Guo et al. [102] proposed a two-stage feature-enhancement method, which employed the multi-head attention mechanism to establish associations between different regions of the image, they also utilized a local attention module to capture the structural information not addressed by the multi-head attention mechanism. Both regional features and grid features have their own strengths and weaknesses. In order to more efficiently exploit regional information, they also suggested a cooperative connection transformer (CCT) [103] featuring regional feature mapping and fusion attention modules. They added cooperative connection attention to improve the decoder's awareness of various features and reduce the noise and redundancy that result from feature extraction and fusion.

In order to comprehensively understand the presence and absence of different regions at different scales in dual-time images and provide a comprehensive description of these changes, Cai et al. [104] proposed an interactive change-aware transformer network (ICT-Net). In the encoder, the cross multi-head attention (cross-MHA) mechanism is used to learn the most distinctive representations and identify changes in interest between pairs of features. In the decoder, a cross gated-attention (CGA) module is proposed to model the relationship between semantic features and multi-scale change features, which improves the performance of change caption. Similarly, in order to extract the difference information, Zhou et al. [105] proposed a shallow feature embedding (SFE) module and a cross-attention guided difference (CAGD) module, where the SFE module uses channel attention to the shallow features embedded in each time stage, while the CAGD module uses multi-head cross-attention and self-attention mechanisms to encode the difference between two single-time features, improving the representation and performance of the single-stream visual feature extractor for remote sensing image-change caption tasks. In addition, Liu et al. [106] also used multi-head cross attention in their proposed progressive scale-aware network (PSNet) [106] to enhance multi-scale features of dual-time images.

3.2.4. Attribute Attention

Remote sensing images contain not only abundant feature information, but also a large amount of attribute information. Although the resolution of remote sensing images at present needs to be improved, there is still a significant semantic gap between low-

level features and high-level semantics in remote sensing images. Incorporating attribute information into an attention mechanism can help bridge this gap. Attribute information is the detailed information of each part of remote sensing images such as classification, color, state, quantity and so on. Compared to traditional attention mechanisms that only use feature information, combining attribute attention and original attention [107] and giving them different weights will help increase the model's utilization of image information. In addition, there is also much attribute information in the text. Peng et al. [108] constructed an attribute generator (AG) to integrate and analyze the information in the text, record the extracted object information as attribute information and assist in the RSIC task.

3.2.5. Recurrent Attention

An ineffective attention mapping can reduce the performance of RSIC, and the existing decoder is prone to accumulating errors during the process of generating sentences. Therefore, it is necessary to learn more semantic reference information at each time step to form an informative word state, which is crucial for narrowing the gap between images and text. Li et al. [109] proposed a recurrent attention mechanism that could capture high-level attention maps and categorized the attention results into visual and non-visual information. Using the attention results with rich contextual relevance from the previous time step, the significant visual features are adaptively retained, while the non-visual features which usually refer to the visual sentinel [79] are weakened. The LSTM used for sentence generation can enhance its performance with the recurrent attention mechanism to produce more accurate sentences.

3.2.6. Global Attention

The emergence of attention mechanisms has brought significant results. However, contemporary attention-based RSIC models only establish relationships between local features, but ignore global visual features and do not remove redundant feature components, which may lead to generating sentences that are weakly related to the scene. To address this issue, Zhang et al. [110] developed a global visual feature guided attention (GVFGA) mechanism. GVFGA introduced global visual features and fused them with local visual features, and then filtered out redundant feature components in the fused image features through an attention gate that utilizes global visual features, resulting in providing more prominent image features. Similarly, in order to fully utilize the advantages of global and local features, Wang et al. [111] proposed an attention-based global–local captioning model (GLCM) to obtain a global–local visual feature representation of RSIC. For visual features extracted from remote sensing images, global features provide both complete and comprehensive visual correlations of all words in a sentence, while local features can individually emphasize the differentiation of these words. Therefore, not only are global features important for RSIC, but local features also make sense for distinguishing words. The acquired global–local features not only contain complete and comprehensive relationships between features and objects, but also include more independent and distinctive visual information for each word when generating captions.

3.2.7. Label Attention

Traditional attention mechanisms only use visual information about remote sensing images, without considering using label information to guide the calculation of attention masks. A new label-attention mechanism (LAM) [112] was proposed by Zhang et al. to utilize the label information of high-resolution remote sensing images to describe a given image. The label information provides guidance for the attention masks, enabling them to focus on the interesting regions of the input image based on the image's category. Image features guided by labels expose more object information and relationships related to the label. Therefore, it is easy to generate a complete caption of the input image.

Remote sensing images frequently feature numerous objects of varying scales and lower resolution, often tightly packed together. After several down-sampling processes,

small to medium-sized targets in the feature map may resemble single points, compromising feature extraction and potentially causing the targets to be overlooked or inaccurately described. To address this, Cheng et al. [113] introduced the Small Target Attention (STA) module. By utilizing embeddings of small target semantic labels, STA enhances the network's sensitivity to these targets, guiding the creation of more precise attention maps.

3.2.8. Passive Attention and Active Attention

Attention mechanisms can be categorized into passive attention and active attention, with the active attention method providing a more targeted approach to describe images.

Traditional RSIC models based on attention mechanisms are driven by marked sentences, which is called passive attention. Sound can represent the attention of different observers, which is active and proactive. Lu et al. employed a sound active attention mechanism framework [114] to generate more specific RSIC. The sound was modeled by Mel frequency cepstral coefficients (MFCCs) [115], and the image was encoded by a convolutional neural network (CNN). To handle the continuous nature of the sound, a sound module and an attention module based on GRU were designed. Finally, the sound-guided image features processed by the attention module were imported into the output module to generate descriptive sentences. Experimental results demonstrated that this method can produce sentences capturing human focus.

3.3. Reinforcement Learning

Introducing the reinforcement learning method into RSIC can enhance the effectiveness of the model without requiring additional datasets. This is particularly beneficial in addressing the challenge of limited availability of datasets. In the terminology of reinforcement learning [116], the RSIC model is considered as an agent, and the parameters of this agent determine the policy. At each time step, the agent executes the policy to select an action, which will be a prediction for the next word in RSIC. Once the end of the sequence is reached, the agent receives a reward. In this scenario, the reward is the evaluation metric. The purpose of training is to optimize the agent's parameters to maximize the expected reward [117]. In this process, words and image features are considered as environment, while state refers to the state of the model. The pros and cons of different reinforcement methods are shown in Table 2.

3.3.1. Cross Entropy Loss

Cross entropy (CE) loss [37] is widely used in the encoder–decoder framework to represent the difference information between two probability distributions. In the decoder, all words in the candidate vocabulary list are assigned a probability after a softmax layer. The goal of CE loss is to optimize the probability of the target word to 1 and the probability of the non-target word to 0. The CE loss can be formulated as:

$$L_{CE} = -\sum_{i=1}^{N} y^{(i)} * \log_{y_o}(i) \tag{1}$$

where L_{CE} denotes the CE loss, and $y^{(i)}$ and $y_o^{(i)}$ denote the label of ground truth and its factual output probability, respectively. N denotes the total number of classes.

3.3.2. Truncation Cross Entropy Loss

In the traditional training step, the probability of the target word at each time step is forced to be optimized to 1 by CE loss. However, due to the variability and ambiguity of RSIC, the target word can be replaced by other words such as its synonyms. This optimization strategy can lead to overfitting of the network. A new truncated cross entropy (TCE) loss was proposed by Li et al. [118] to effectively mitigate the overfitting problem.

The TCE loss is denoted as L_{TCE} and consists of two components: a common PCE loss and a truncation loss. PCE loss is essentially the CE loss when used in RSIC [119]. L_{TCE} is formulated as:

$$L_{TCE} = \begin{cases} -{y_t}^{S_t} * log_{y_{to}} s_t, if\ {y_t}^{S_t} < 1 - \gamma \\ -log(1-\gamma), otherwise \end{cases} \quad (2)$$

where γ is the value of the truncation threshold that reserves a margin for non-target words during training. When γ is equal to 0, the TCE loss is actually the PCE/CE loss. ${y_t}^{s_t}$ and ${y_{to}}^{s_t}$ are the output probability and target probability of s_t, respectively.

The TCE loss is developed based on the PCE loss. However, unlike the PCE loss, an upper bound of $1 - \gamma$ is set in TCE loss to determine which losses will be selected in the optimization process. Assuming that the output probability of the target word exceeds $1 - \gamma$, the loss function will switch from the PCE loss to the truncation loss. As a result, ${y_t}^{s_t}$ will be optimized to $1 - \gamma$ instead of 1, which means that the probability margin of γ will be specifically reserved for the remaining non-target words in the vocabulary other than the target word. In this way, the overfitting phenomenon caused by CE loss can be effectively mitigated.

3.3.3. Self-Critical Sequence Training

RSIC models trained using CE loss may lead to a mismatch between the optimized target and the evaluation metric. Rennie et al. [120] proposed self-critical sequence training (SCST), which used the evaluation metric generated by the model as a benchmark to optimize the model's parameters. The goal is to minimize the difference between the captions generated by models and high-quality reference captions, and use the differences between them to calculate the gradient and optimize. SCST can effectively improve accuracy and suppress exposure bias [121] in CE training. In addition, SCST is widely applicable to a large number of RSIC models and it is the most frequently used reinforcement learning method in image caption.

The key formula for SCST gradient optimization is:

$$\frac{\partial \mathrm{L}(\theta)}{\partial \mathrm{S}_t} = (r(\omega^s) - b)\left(p_\theta(\omega_t|h_t) - 1_{\omega_t^s}\right) \quad (3)$$

During this procedure, $r(\omega^s)$ represents the reward obtained by the model during testing. Consequently, samples with rewards higher than b will be elevated or increase its probability, while those with lower rewards will be suppressed. The selection of baseline b significantly influences the optimization direction of the model [122].

3.3.4. Mask-Cross Entropy

Although SCST contributes to enhancing the quality of generated sentences and suppressing exposure bias during CE training, it cannot help to generate more diverse sentences. Ren et al. [123] introduced a Mask-CE training strategy aimed at enhancing the diversity of RSIC. The main idea of Mask-CE is to randomly mask some words in the ground truth. Before each iteration of the word-embedding layer, randomly replace some words with [Mask] in the sequence, where [Mask] is an alternative word token similar to [bos] (the token for the beginning of the sequence). This approach can effectively suppress exposure bias and enhances robustness [124], and random masking operations can effectively augment the contextual diversity within the training sequence, which helps with averting overfitting and bolstering the learning capacity of the model.

The loss function for Mask-CE is formulated as:

$$Loss = -\sum_{t=1}^{L} log\left(p_\theta\left(y_t \middle| y_{1:t-1}^{mask}\right)\right) \quad (4)$$

In this equation, y represents the word of ground truth, θ is the model parameter, y^{mask} denotes the masked word and t is the length of the reference sentence. For each

prediction step, the probability p_θ of the next word y_t is calculated through the previous masked words.

3.3.5. Actor Dual-Critic

SCST is trained by maximizing the similarity between ground truth and generated sentences. Although it is simple and effective, it cannot directly maximize the metrics we need like CIDEr. Actor–critic (AC) [125] is a classic reinforcement learning method that can directly optimize the non-differentiable quality indicators of interest. The process is to train the actor model separately first, then freeze the parameters of the actor model, and use the generated samples as the input of the critic model, ensuring that the generated samples from both the critic and actor are consistently equal. Finally, train the critic and actor separately at each iteration, although in practice it is not possible to strictly train the critic for every step taken by the actor. However, it is worth learning the awareness of intentionally synchronizing the training pace of the two networks. The principle of AC is similar to generative adversarial networks (GAN) [126] used in an image-generation [127] task, where a generator (G) and a discriminator (D) are the basic components. After the generator creates an image, the discriminator trains to distinguish between real and fake images, while the generator learns to produce convincing fake images by using feedback from the discriminator. AC and GAN share many similarities [128].

The datasets of RSIC exhibit high inter-class similarity between images and captions, so models often generate semantically unrelated captions. Chavhan et al. [129] proposed an actor dual-critic training strategy that used an actor to predict sentences for an image and a critic to provide rewards. The innovation lay in employing an additional encoder–decoder RNN as the critic, which generated a caption translated back into features that should be present in remote sensing image and rewarded the actor by assessing the similarity of this reverse translation. The proposed model generated sentences on test data that are very similar to ground truth, and successfully generated better captions in many significant cases.

Table 2. Pros and cons of different reinforcement learning models.

Models	Advantages	Disadvantages
CE [37]	Most frequently used; effectively improves performance	Overfitting; exposure bias
TCE [118]	Solving the overfitting problem of CE; generate more flexible and concise captions	After the probability of truncating words decreases, other words may not be able to learn accurately, either
SCST [120]	Using metrics to motivate the model to generate sentences with higher rewards	Sentence diversity is neglected
Mask-CE [123]	Solved the overfitting problem of CE; combined with the advantages of SCST; improved the accuracy and diversity of caption generation	When keywords are masked, it relies more on the prediction of the model
ADC [129]	The generated sentences have stronger semantic relevance	Easy to lead to local optimal solutions and difficult convergence

3.4. Learning with Auxiliary Task

Sometimes we develop auxiliary tasks to improve the primary task because we want to profit from other tasks while concentrating on the primary one. Usually closely tied to the primary task, the auxiliary task is able to catch up on details that the primary task ignores. In addition to being widely employed in other domains including biomedicine [130],

computer vision [131], speech recognition [132] and data mining [133], this technique has also made significant advancements in remote sensing image processing. The study on learning with auxiliary tasks in RSIC is summarized in this part. The pros and cons of different models are presented in Table 3.

3.4.1. Object Detection

Object detection can usually obtain region-level features of objects in an image [101], which is very beneficial to the accurate positioning of described objects in RSIC and the description of spatial position relationships between objects.

Shi et al. [16] pioneered an investigation into whether a machine can automatically generate humanlike captions for remote sensing images. They utilized a full convolutional network model (FCN) [134], with the input consisting of three parts corresponding to three sub-tasks: critical object detection, environmental element analysis and landscape analysis. Through these subtasks, they achieved multi-level image understanding and generated captions for remote sensing images. The results indicate that despite some challenges, the future development of RSIC is promising.

The authors in [82] proposed two multi-scale methods to address the challenges of multi-scale problems. In addition to the first method called multi-scale attention (MSA), they also introduced a multi-task attention (MFA) method. The MFA method integrates object-level features and scene-level features, and incorporates an object-detection task as an auxiliary task to enhance the contextual features. Pixel-level change captioning requires precise identification and accurate localization of pixel-level changes. Liu et al. [135] utilized a change-detection model trained on existing change-detection datasets to generate change masks as pseudo-labels. They then proposed an RSICC network that leverages these pseudo-labels as an auxiliary change-detection branch to enhance the model's capability to accurately generate change captions. Similarly, to accurately localize the visual change region in multi-temporal images, the existing high-quality remote sensing change-detection model P2V-CD [136] was employed for training on the LEVIR-CD dataset as a visual change detector. This approach proposed by Li et al. [137] can extract the visual change information at the object level while preserving the spatial position details of the changed objects, which is more conducive to the accuracy of the change caption model.

The existing RSIC methods ignored the accurate representation of numerical data. Ref. [138] first employed a counting module to detect objects in remote sensing images and ascertain their quantities. After captions which may not include object count information are generated, the object counts are integrated into the captions through a digital editing module, leading to more precise and informative captions.

3.4.2. Image Classification

Usually, describing an object revolves around a noun. This implies that any error in the choice of noun will significantly influence the entire sentence. Considering that the labels in multi-label classification primarily represent real objects found within those images, multi-label classification was applied to remote sensing images in [139] to enhance the accuracy of real object representation as well as the generation of more detailed image captions. Similarly, in order to improve the likelihood of recognizing words related to the image content and inform the model to predict based on the types of landscapes shown in the image, Kumar et al. [140] used multi-label classification to figure out the probability of various target categories for a given image. They then combined this probability or weight with extracted features. In their proposed UAVIC dataset, they obtained better results.

RSIC and its auxiliary task in [139,140] were processed separately, making it relatively time-consuming to enhance the main task by executing the auxiliary task first. Kandala et al. [141] proposed an RSIC model based on transformer and using multi-label classification as a supplementary task. This was the first research to jointly address multi-label classification and RSIC, and it improved the main task by standardizing the features learned from transformer. Two decoders are connected with the transformer encoder; one is for

RSIC and the other is for label classification. Under the circumstances, no prior knowledge about image labels is required. Since the two decoders are trained at the same time, the time cost of auxiliary task calculation is negligible. However, the shortage is that even their model trains two tasks at the same time; there is less interaction between them; the real-time interaction between them is neglected. The difference appeared in Ye et al.'s innovative joint training two-stage (JTTS) approach [142] for RSIC, which eliminated the tedious pretraining procedure and produced higher-quality captions. They also incorporated the multi-label classification task as an auxiliary task to supply previous information that improved the model's capacity to produce precise captions. In this model, both tasks are trained jointly, resulting in greater synergy between them. The multi-label classification tasks predicted label attributes and embedded them as attribute features by using a marginal sampling operator (MSO). Finally, the attribute features and image features are fed into an attribute-guided decoder to obtain the corresponding RSIC.

Another way to employ image classification as an auxiliary task is meta caption [143], a meta-learning approach for RSIC. This approach extracted meta features from two supporting tasks, natural image classification and remote sensing image classification, and conveyed the meta characteristics to RSIC. The two supporting tasks were trained with a large amount of labeled data to extract comprehensive meta-features that represent the visual characteristics of the images from a classification perspective. By utilizing these meta characteristics, the target task was able to effectively describe remote sensing images with a comparatively minimal amount of annotated data. A simple structure of meta caption that has used remote sensing image classification as an auxiliary task is shown in Figure 8.

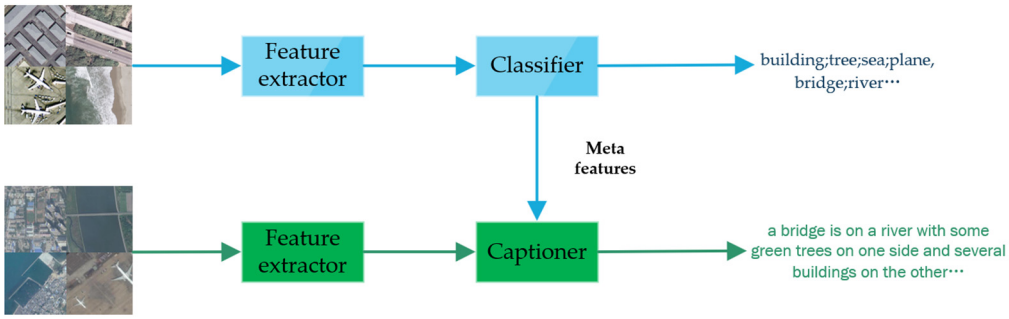


Figure 8. A simple expression of meta caption.

3.4.3. Image or Text Retrieval

In addition to using image-classification tasks to enhance accurate perception of target nouns, image- and text-retrieval tasks also work effectively in achieving this effect. Moreover, using retrieval models as auxiliary tasks is not limited by predefined labels. Hoxha et al. [144] combined RSIC and remote sensing image-retrieval tasks; they first used beam search [145] to obtain many sentences, and then retrieved images along with their reference captions that are comparable to the original image. To reduce global semantic errors caused by a word error, the best caption is selected based on the similarity between the reference caption and the generated caption. The research in [146] employed a similar approach, with the exception that they constructed a thesaurus through the retrieval of relevant topic words, and then obtained the topic words of the test images from the topic repository. Subsequently, the retrieved subject words were integrated into RNN to generate sentences. It is noteworthy that topic words can also be manually input, which demonstrates the controllability of caption generation and its potential in controlling. However, this method also reveals some drawbacks. The model extracts semantic topic words from the captions corresponding to remote sensing images as additional information about these images. Although topic words can help the model generate detailed titles, they were excessively fixed and had poor visual consistency.

Present RSIC techniques lack component interaction and are primarily restricted to predefined tags and semantic information of pretrained models. A novel cross-modal

retrieval and semantic refinement (CRSR) approach that combines a semantic refinement module with a clip-based retrieval model [51] was proposed by Li et al. They first fine-tuned the clip using the RSICD dataset to retrieve sentences related to remote sensing images, thereby providing auxiliary semantic information for sentence generation. In order to prevent semantic confusion in the retrieval process, a semantic refinement module is introduced to filter out misleading words through the masked cross-attention mechanism.

3.4.4. Image Segmentation

Image segmentation is crucial for capturing detailed pixel-level features in remote sensing images, which contributes to enhancing the accuracy. Zhao et al. [22] merged RSIC and semantic segmentation by creating a structured attention mechanism through semantic segmentation technology. The input image is initially segmented into a set of class-agnostic segmentation proposals, and then the structure of each segmentation proposal is encoded into the attention module. During training, structured attention can assist the model in precisely identifying highly structured semantic content. However, this method is limited to high-resolution remote sensing images because insufficient image resolution prevents extraction of segmentation proposals necessary for supporting the proposed structured attention mechanism. And the number of segmentation proposals used in the structured attention module remains constant; when the input image contains more semantic content than the predefined number of segmentation proposals, the structured attention module may not be able to solely concentrate on the most prominent areas. Cui et al. [147] contended that the application of RSIC in disaster prediction often caused error accumulation and inaccurate attention positioning, leading to risk misjudgment. Therefore, they proposed semantic gate and a bi-temporal long-short term memory network (SG-BiTLSTM), which leveraged U-Net to generate mask maps of landslides and other geographical objects, while introducing a novel two-temporal LSTM. This new approach incorporated three losses, language, prediction and attention, to train network parameters and alleviate cumulative errors when predicting.

3.4.5. Text Summarization

The training of RSIC may encompass superfluous information, which probably causes inadequate details for learning the mapping from the image domain to the language domain. The text-summarization task [58] can condense text into a more concise version that captures the essential information. Sumbul et al. [148] proposed a summarization-driven RSIC approach (SD-RSIC), which utilized automatic summarization techniques from NLP to tackle issues in image caption. They condensed standard captions into single representations for each remote sensing image to remove redundancy and then used a summarization model trained on a large text corpus to expand the vocabulary in RSIC, which lead to more precise captions for complex scenes. Adaptive weights were introduced between standard captions and summaries, and these were combined based on the semantic content of the images. However, it is crucial to acknowledge that summarization signifies another layer of detailed textual information, and the introduction of such complexity also involves time and space expenses.

3.4.6. Visual Question Answering

Visual question answering (VQA) [149] and RSIC both belong to the cross-modal task of visual language. The two tasks share similarities in many aspects, and sometimes more detailed information of images can be obtained in VQA. Murali et al. proposed a model that used VQA to help RSIC [150], which solved the issues that most remote sensing images have regarding great inter-class similarity and that may be difficult to distinguish. The VQA model aided in identifying key words within the caption, such as all objects present in the image and their respective quantities, thereby enhancing the informativeness of captions. The generation of RSIC can be enhanced through knowledge acquired from VQA.

Table 3. Pros and cons of various models using auxiliary tasks.

Authors	Auxiliary Task	Advantages	Disadvantages
Shi et al. [16]	Object detection	Investigate the feasibility of RSIC	Simple model-FCN
Li et al. [81]	Object detection	Enrich feature and semantic information by extracting object-level features	There is a deviation of dataset between the auxiliary task and original task, fewer categories
Liu et al. [135]	Change detection	The pseudo-labels strengthen the recognition and accurate location of the change of targets, and improve the accuracy of change caption	Require pretrained model; highly reliant on the precision of the change-detection model
Li et al. [137]	Change detection	Precisely identify the key areas of visual change	Require pretrained model; highly reliant on the precision of the change-detection model
Ni et al. [138]	Change detection	The numerical accuracy of the caption has been enhanced	Only valid for partial multi-quantity caption, mostly templated
Long et al. [139]	Multi-label classification	Generate more accurate captions about the object information	Require double training time; no real-time interaction
Zhao et al. [140]	Image classification	The probability obtained through classification enriches image features	Use pretrained models; generate less comprehensive captions
Kumar et al. [141]	Multi-label classification	Training two tasks simultaneously reduces computation time	Less interaction between the two tasks; ignoring the real-time impact
Kanadala et al. [142]	Multi-label classification	Reduces computation time; enables joint training for real-time interaction	Unable to generate fine-grained RSIC based on practical application scenarios
Ye et al. [143]	Image classification	Solved the problem of sparse RSIC datasets	Meta-learning may produce inaccurate captions in some situations
Yang et al. [144]	Image and text retrieval	Reduce global semantic errors caused by a single word error	The model structure is simple and adopts a CNN + RNN model
Wang et al. [146]	Text retrieval	Overcome long-term information dilution in RNN; reflect the controllability of captions	More fixed topic words lead to more fixed captions; sometimes result in meaningless captions
Li et al. [51]	Text retrieval	Not restricted by predefined labels; employ more refined and accurate semantic information	The risk of semantic misalignment between the retrieved word and the actual content; still some errors in semantic refinement
Zhao et al. [22]	Image segmentation	Optimize attention through segmentation proposals; focus on highly structured semantic content	Only applicable to high-resolution remote sensing images; The number of segmentation proposals is fixed
Cul et al. [147]	Image segmentation	Reduce the accumulated error; construct a new attention-correction mechanism to improve the location accuracy	Used in landslide caption; hard to judge the type and depth of landslide. The landslide covered by vegetation cannot be recognized.
Sumbul et al. [148]	Text summarization	Eliminate redundancy; enrich vocabulary; more suitable for complex scenarios	Introducing complex text consumes an amount of time and space
Murali et al. [150]	VQA	Solve the problem of similarity and indistinguishable in remote sensing images; obtain more information	The VQA model requires additional training and is no less complex than the RSIC task

3.5. Large Visual Language Models

Recent advancements in large visual language models (LVLMs) [151] in the field of natural images have demonstrated their potential to offer users innovative ways to engage with visual content. Moreover, excellent LVLMs such as GPT-4 [152], LLaVA [153] and MiniGPT-v2 [154] have extended this capability by integrating visuals and language, but they still face challenges when applied to remote sensing scenarios due to significant differences between remote sensing images and natural images across various dimensions including resolution, spectral information and image background complexity. These disparities pose difficulties for existing LVLMs in comprehending and processing remote sensing imagery. Despite the remarkable achievements of LVLMs in natural images, further research is necessary to explore their application in remote sensing.

3.5.1. Early Exploration of LVLMs

In the early stage, fine-tuning technology was not mature in terms of LVLMs, and the LVLMs without fine-tuning could only be widely used in detection tasks. Drawing inspiration from template-based image caption methods, researchers proposed an automatic remote sensing image caption (ARSIC) model [155], which instructed a large language model (LLM) [156] to automatically generate a caption by describing its objects. During this process, they first employed the LLM model to identify the names and locations of landmarks, then leveraged the Minimum Spanning Tree (MST) algorithm [157] to calculate the precise locations of various buildings. They also combined geographical words commonly used in the RSICD dataset such as "near", "surrounded by" and "on two sides of" to develop an API to facilitate geographic analysis and describe spatial relationships between objects. Ultimately, the LLM generates captions with assistance from the API.

3.5.2. Fine-Tuned LVLMs

The most recent research in computer vision (CV) has revealed that fine-tuning LVLMs with small, high-quality datasets can lead to impressive performance in both visual and language comprehension. Many state-of-the-art RSIC methods are based on fine-tuning LVLMs.

Wei et al. [158] proposed a visual language pre-training architecture for bilingual RSIC. Word embedding was combined with the image features extracted by CLIP, and then the caption was generated by pre-trained GPT-2 [159]. GPT-2 was built with transformer and trained on a huge dataset. They also used additional language pre-training model fine-tuning to solve the problem of RSIC, which increased the number of training parameters but produced better results. Moreover, they also created Diors-Captions, a new optical RSIC dataset.

In order to improve the robustness of the generated captions, two primary postprocessing methodologies for selecting the best caption from a set of predicted captions are proposed by Ricci et al. [160]; they are selection and generation, respectively. The selection method selects the most similar one among all the generated captions, while the generated method uses BERT to generate semantic embedding of the generated captions and further project them into the VaE [50] potential space. Finally, a fine-tuned GPT-2-based decoder is used to generate the captions.

Hu et al. [161] introduced the RSICap dataset, a high-quality RSIC dataset, which catalyzed the advancement of LVLM in remote sensing. They also developed RSGPT, a dedicated LVLM specifically for remote sensing. To optimize training efficiency, they implemented strategies such as loading pre-training weights and fine-tuning network components for RSGPT. The pre-training weights from instructBLIP [162], designed for visuospatial inference tasks, were integrated into RSGPT and subsequently fine-tuned using the proposed RSICap dataset. The resulting model exhibits versatility and can be effectively leveraged across diverse visual language tasks.

A large language and vision assistant model (LLaVA) [163] can also be used for analyzing and processing remote sensing data. The image encoder and LLM weights are frozen to train the projection network with common image–text pairs, then the projection

network and image encoder are frozen to fine-tune the LLM. While the encoder and decoder were both frozen, low rank adaptive (LoRA) [164] was also used to fine-tune RS-LLaVA, which performed well in both RSIC and RSVQA. The training procedure of RS-LLaVA is shown in Figure 9.

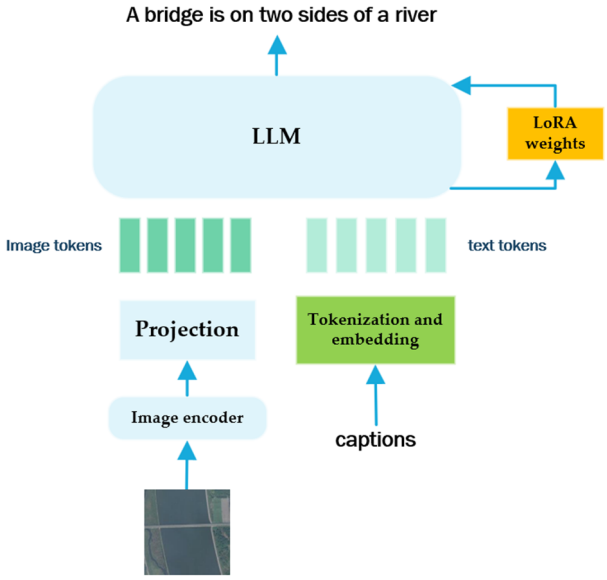


Figure 9. The training procedure of RS-LLaVA.

The RS-CapRet proposed in [165] possessed the powerful capabilities of LLM and could be used as an encoder for RSIC. During the training process, the entire LLM and visual encoder were not fine-tuned. Instead, their parameters were frozen and only a linear layer was trained. Impressive results were obtained in text–image retrieval and RSIC tasks by RS-CapRet, which also offered an easy-to-use training method.

3.5.3. LVLMs with Huge Data

Recent studies [166] have demonstrated impressive results in training multimodal large language models (MLLMs) [167] by instruction tuning to connect LLM and vision, while instruction tuning in the multimodal field of remote sensing remains underexplored. Zhan et al. [168] created a high-quality instruction tracking dataset with 968 k samples, including single-task image–text instructions and multi-task conversation instructions. On the basis of this dataset, SkyEyeGPT was developed specifically to solve visual language-understanding problems of remote sensing. The architecture of SkyEyeGPT was composed of a visual encoder, an alignment layer and an LLM-based decoder. After the remote sensing visual features are projected into the language domain through the alignment layer, they are fed together with task-specific instructions to the LLM-based decoder to forecast the response to the remote sensing open task. SkyEyeGPT is a unified and efficient model since it does not require any extra encoders or external plug-in modules.

Kartik et al. [169] proposed GeoChat, a multifunctional remote sensing VLM, which was obtained by fine-tuning LLaVA-1.5 [170]. They created a new remote sensing multi-modal instruction-tracking dataset consisting of 318 k image–instruction pairs in different multitasking formats, which used the existing object-detection dataset [171] to create a brief description of the image, then used Vicuna v1.5 [172] to create a dialogue by the generated text alone and used VQA datasets and scene-classification datasets to obtain corresponding capabilities. GeoChat is also the first comprehensive large-scale remote sensing model that has shown excellent zero-shot performance across a range of remote sensing tasks.

While RSGPT [161] and GeoChat [169] have been proposed for exploring the multi-modal large language model (MLLM) in remote sensing and addressing multiple tasks, RSGPT necessitates individual model tuning for each specific task, thereby limiting its

generalizability in open set domains. Additionally, although GeoChat was trained on optical RS images, it did not achieve comprehensive understanding in synthetic aperture radar (SAR) and infrared modalities. To address these limitations, Zhang et al. [173] developed a large-scale multi-modal dataset comprising multi-sensor RS images based on existing RS datasets. They introduced a universal MLLM named EarthGPT, designed for multi-sensor RS image understanding. EarthGPT effectively integrates various RSIC tasks and demonstrates superior performance in different tasks.

The existing remote sensing dataset samples generally do not exceed 10 k, which has become a major bottleneck limiting the training of LVLMs. Due to the limited amount of data, the model is susceptible to overfitting during training. Liu et al. [174] combined and expanded remote sensing datasets such as RSITMD [175], RSICD [40] and UCM [39] by a factor of 12, and then integrated part of the unmanned aerial vehicle (UAV) [176] image dataset. This combined large-scale dataset effectively addressed the issue of scarce training data in remote sensing. Their proposed visual language model was named RemoteCLIP and used massive pre-training data to train. RemoteCLIP learned robust visual features that contained rich visual concept semantics of satellite images and was widely applicable to a wide range of relevant downstream tasks.

The change-detection task can accurately locate the change region, but it lacks a deeper understanding of the underlying meaning of the change. A change captioning task can provide rich semantic-level interpretation information, but it may be insufficient in providing accurate pixel-level change location. In order to combine the advantages of the two tasks, Liu et al. [177] proposed a novel multi-level change interpretation (MCI) model and a large language model (LLM) named interactive Change-Agent. The MCI model includes two branches: pixel-level change detection and semantic-level change captioning. Finally, the obtained Change-Agent can not only explain and analyze the changes of remote sensing images comprehensively, but also has an intelligent dialogue function, which opens up a new way for intelligent remote sensing applications.

3.6. Few-Shot Learning

The purpose of few-shot learning [178] is to acquire the capability to solve tasks by learning from limited supervised information. Image caption often necessitates a large amount of labeled data, which are relatively scarce in remote sensing environments. Inadequate data naturally impacts the model's performance. How to achieve comparable performance to other deep learning methods using limited labeled data is a hot point of RSIC.

3.6.1. Meta Caption

Few-shot learning is the application of meta-learning in the field of supervised learning. The meta caption [143] mentioned above can also achieve the effect of few-shot learning, as the model acquired meta-features from natural image classification and remote sensing image classification, so that it requires minimal data to train the RSIC model for generating accurate captions.

3.6.2. Semi-Supervised Learning

Chen et al. [179] proposed a semi-supervised learning method based on ensemble and self-distillation [180]. They divided the existing dataset into three groups, randomly selecting 1% of image–caption pairs as a set of data Dxy, the remaining unpaired images as a set of data Dx and the unpaired captions as a set of data Dy.

The formula for training loss is:

$$L = L_{xy} + \alpha L_x + \beta L_y \tag{5}$$

Among them, L_{xy}, L_x and L_y are calculated on the three datasets D_{xy}, D_x and D_y, respectively, and α and β are the weights to balance the corresponding loss terms.

The framework of their proposed method can be summarized as Figure 10. They initially trained M basic models using paired data and combined these models using mean

teacher learning [181]. Mean teacher was a temporal ensemble technique that averaged model weights rather than their outputs in order to prevent overfitting. This model was subsequently utilized to generate captions for unpaired images, which were referred to as pseudo-labels. However, pseudo-labels are not always reliable, and indiscriminate use of them during training could result in error accumulation and gradual degradation of model performance. Therefore, a confidence-reweighting method was proposed to assign different weights to each training sample based on the confidence level of the pseudo-labels. In this way, unpaired images participated in the training.

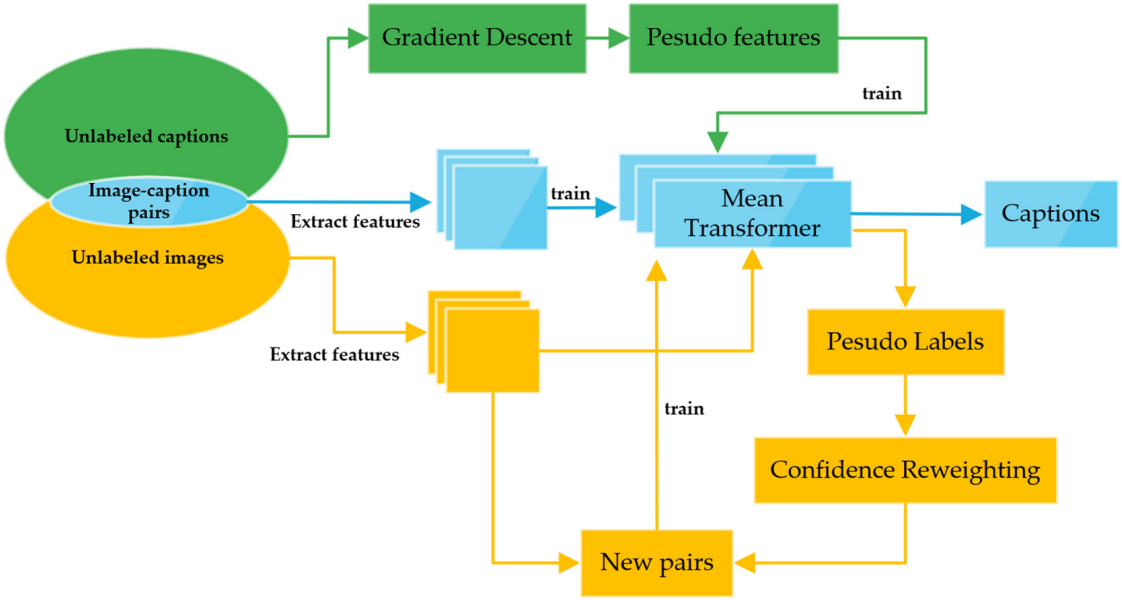


Figure 10. A semi-supervised image caption method.

To enable the participation of unpaired captions in model training, previous studies often used GANs to generate pseudo-image features [182] or model parameters [183], which led to unsatisfactory results with limited data. By applying gradient descent to a trained ensemble, it was possible to effectively generate meaningful pseudo-features without introducing additional parameters. As a result, the risk of overfitting on a small number of training instances decreased.

3.6.3. Few-Shot RSIC

On the basis of the above work, Zhou et al. [184] introduced a few-shot learning method based on ensemble and self-distillation into RSIC to alleviate the problem of dependence on a large number of remote sensing datasets. The difference was that they additionally used a small amount of unlabeled data for self-supervised learning in few-shot scenarios, and they obtained a scene-classification model that could be used on the decoder of RSIC to enhance the model's generalization ability. They also employed SCST to further optimize the model's performance, which achieved satisfactory results with only 80% of the labeled dataset and still maintained acceptable effectiveness with just 60% of the dataset. This demonstrated its efficacy in few-shot training.

The research conducted in [185] employed less than 1.0% of the dataset to replicate a realistic few-shot situation. They adapted their model ensemble strategy by forgoing the training of multiple models and instead integrating a singular model at various time intervals while continuously updating it. When trained using only 0.8% of the RSIC dataset, their proposed method exhibits significant advantages over alternative methods. This study represents the pioneering use of such a minimal amount of data for training an RSIC model.

4. Datasets and Evaluation Metrics of RSIC

Data play a crucial role in deep learning, as a high-quality dataset can substantially enhance model training and prediction accuracy. Image generation, an important task in deep learning, is enough to prove the importance of datasets. Evaluation metrics are served as an intuitive measure of model quality, and enhancing these metrics can also lead to ascensive model performance through reverse optimization. Considering the pivotal role of datasets and evaluation metrics in deep learning, this section will offer a brief introduction of the datasets and evaluation metrics relevant to RSIC.

4.1. Datasets

The datasets in natural image captioning, such as Flickr30k [186] and MSCOCO [187], are usually a large size. In particular, MSCOCO is not only extensive but also widely used in numerous deep learning tasks. The dataset of RSIC was originally inspired by remote sensing image-classification tasks, including Sydney-Captions, UCM-Captions, RSICD, etc. They were all small datasets until the emergence of a larger dataset, NWPU-Captions [83]. This enabled the probability of training on large-scale data sets for RSIC. In addition, LEVIR-CC [188] is specifically designed for the RSICC task. And there are another two small change caption datasets named LEVIR-CCD and Dubai CCD [64], which are also valuable for researchers. Some explanations of a few open resource datasets and some details of different datasets are presented in Table 4.

Table 4. Some detailed information of RSIC datasets.

Dataset	Image	Captions	Category	Characteristics
Sydney-Captions [39]	613	3065	7	Minimum quantity
UCM-Captions [39]	2100	10,500	21	The simplest captions
RSICD [40]	10,921	24,233	30	Less than 5 captions per image, randomly supplemented
NWPU-Captions [83]	31,500	157,500	45	The largest in quantity, variety and vocabulary
LEVIR CC [188]	20,154	50,385	5	Five captions for a pair of images, reflecting changes in the image
LEVIR-CCD [64]	1000	2500	13	A small dataset annotated by annotators with RS background
Dubai CCD [64]	1000	2500	6	The image size is small

4.1.1. Sydney-Captions Dataset

The Sydney-Captions Dataset, expanded by Qu et al. [39], was annotated on the Sydney Dataset [189] that was originally used in remote sensing image classification. Each image is accompanied by five distinct sentences, which are all provided by the same individual. It is comprised of seven categories with varying numbers of images. Each image has a pixel value of 500 × 500 and a resolution of 0.5 m. This dataset includes a total of 613 images and 3065 captions. An example image and its corresponding captions are shown in Figure 11.



Figure 11. Sample image and captions of Sydney-Captions.

4.1.2. UCM-Captions Dataset

The UCM-Captions Dataset was also created by Qu et al. [39], and it was derived from the UC Merced Land Use Dataset [190]. Five descriptive sentences with various comments are included with every remote sensing image. These sentences encompass 21 categories and are supplied by the same person. Each category of ground object has one hundred images with a pixel resolution of 0.3048 m and a pixel size of 256 × 256. The dataset contains 2100 images and 10,500 captions in total. An example image and its corresponding captions are shown in Figure 12.



Figure 12. Sample image and captions of UCM-Captions.

4.1.3. RSICD

RSICD was constructed by Lu et al. [40] in 2017, including 10,921 images across 30 categories, and each image has a resolution of 224 × 224 pixels. Annotators provided 1–2 descriptive sentences for each image, resulting in a total of 24,233 sentences. The annotations for each image are sourced from different individuals. To ensure that there are five sentences for each image, existing sentences are randomly replicated when an image has fewer than five captions. As a result, the five captions for an image may not be entirely distinct within this dataset. An example image and its corresponding captions are shown in Figure 13.



Figure 13. Sample image and captions of RSICD.

4.1.4. NWPU-Captions

The images in NWPU-Captions are sourced from the publicly available benchmark dataset NWPU-RESISC45 [191], which comprises 31,500 images with 256 × 256 pixels and a resolution of 0.228 m. It encompasses 45 scene categories, each with 700 images.

Cheng et al. [83] completed the caption part of NWPU-Captions, consisting of 157,500 sentences. The annotation for the 31,500 images was carried out manually by seven experienced volunteers. The advantage of NWPU-Captions over existing benchmark datasets lies not only in its larger scale but also in its comprehensive coverage of complex

scenes and the diversity and richness of its descriptive vocabulary. An example image and its corresponding captions are shown in Figure 14.

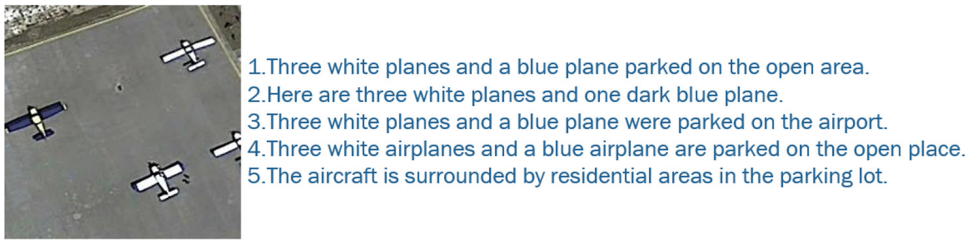


Figure 14. Sample image and captions of NWPU-Captions.

4.1.5. LEVIR-CC

The LEVIR-CC dataset is derived from the original LEVIR-CD dataset [6], which was initially employed for remote sensing change detection. The LEVIR-CD dataset comprises 637 bi-temporal images, each with a size of 1024 × 1024 and a resolution of 0.5 m. These images were collected from 20 different areas in various cities across the United States using the Google Earth API, with a time span ranging from 5 to 15 years.

Liu et al. [188] annotated the LEVIR-CD dataset to create the LEVIR-CC dataset by cropping each image into blocks of 256 × 256 pixels. In total, they annotated 10,077 pairs of images and a total of 50,385 sentences. This dataset focuses on multiple changing scenes and objects such as buildings, roads, playgrounds and rivers, while disregarding light changes, shadow changes, small changes related to trees and other insignificant alterations. The example images and their corresponding captions are shown in Figure 15.



Figure 15. Sample images and captions of LEVIR-CC.

4.1.6. LEVIR-CCD

This dataset was annotated by [64] et al. in accordance with the LEVIR CD dataset, consisting of image pairs that have been cropped into 256 × 256 pixel patches. Each image pair is accompanied by five distinct change captions provided by annotators with remote sensing background.

During the annotation process, the annotator first visually examines the scenes in the two images to determine the nature of the changes occurring between them. They then consider the relationships between various attributes, such as color and size, as well as the different variations present. The captions are annotated based on the location and relative position of objects within the images. This dataset comprises 500 image pairs in total, resulting in 2500 change captions. The example images and their corresponding captions are shown in Figure 16.



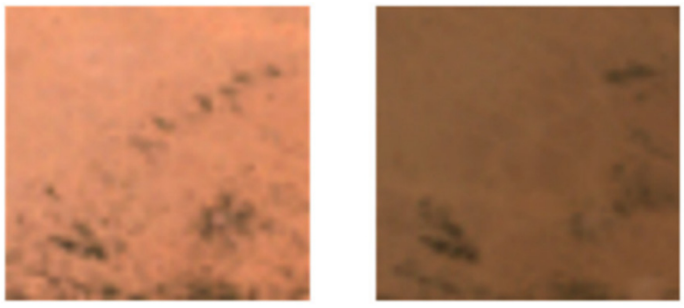
1.The grass appears more green and the trees have increased in size.
2.There is no change except the color of the grass and the size of the trees.
3.Only the trees and the grass appear different.
4.Slight changes related to the color of the grass and the size of the trees.
5.There is no significant change to mention except for the grass that have grown.

Figure 16. Sample images and captions of LEVIR-CCD.

4.1.7. Dubai CCD

This dataset was created by [64] to study the changes in urbanization in the Dubai region from 2000 to 2010, in order to better understand and automatically describe the urbanization phenomenon in Dubai. The size of the captured dual-time image is 2569 × 2468 pixels. It was obtained by an Enhanced Thematic Mapper Plus (ETM+) sensor onboard Landsat7 [192]. They then extracted 500 tiles of sizes 50 × 50 from the bitemporal images.

They also resorted to Google Maps to better inspect the area and understand the physical composition of the area. From the visual examination, six different large categories were identified: roads, houses, buildings, green spaces, lakes and islands. Annotators were asked to explain the properties of spatial distribution and changes in at least three words. Finally, each image in the dataset has 5 different change captions, with 2500 captions in total, a maximum length of 23 words, an average length of 7.35 words and a vocabulary size of 246 unique words. The example images and their corresponding captions are shown in Figure 17.



1.No difference in this area.
2.Nothing has changed.
3.No change occurred in this area.
4.Everything remains the same.
5.No changes to mention.

Figure 17. Sample images and captions of Dubai CCD.

4.2. Evaluation Metrics

The most frequently employed metrics for RSIC originated from natural image captions. In this section, these evaluation metrics will be briefly summarized below. The advantages, disadvantages and application scenarios of various evaluation metrics are presented in Table 5.

Table 5. Some detailed information on the evaluation metrics.

Evaluation Metrics	Advantages	Disadvantages	Application Scenarios
BLEU [193]	Easy to calculate	Does not consider semantic fluency or synonyms	Machine translation; image caption
METEOR [194]	Takes into account word meaning and grammatical structure	Too many hyperparameters; lack of suitability for multiple datasets	Machine translation; image caption
ROUGE [195]	Measures the repeatability and importance of words	Sensitive to length differences	Automatic summarization; image caption
CIDEr [196]	Considers semantic meaning, context and visual content	Complex calculations	Image caption
SPICE [188]	Focuses more on semantics; more consistent with humans	Ignores grammatical correctness	Image caption
S_m [106]	Overall model evaluation	Less frequently utilized	Image caption

4.2.1. BLEU

Bilingual Evaluation Understudy (BLEU) [193] was proposed by IBM in 2002 and it was used to analyze the degree of co-occurrence of n-tuples in the translation to be evaluated and the reference translation. Originally designed for machine translation tasks, it has been adapted to image captions due to their similarity in comparing generated sentences with standard ones. BLEU calculates the proportion of N-grams co-occurring between the generated and reference sentences within sentence length; in other words, it evaluates sentence quality based on the frequency of co-occurrence words. Its value ranges from 0 to 1, with higher values indicating better sentence-generation effects. While BLEU is intuitive for assessing consistency with standard sentences and is easy to calculate, it does not consider semantic fluency or synonyms.

4.2.2. METEOR

Metrics for Evaluation of Translation with Explicit Ordering (METEOR) is also derived from machine translation. It was proposed by Satanjeev et al. in 2002 [194]. In contrast to BLEU, which only matches identical words, METEOR uses exact matching words, stems and semantically similar words to achieve word-level matching. The advantage over BLEU lies in METEOR's aim to produce smooth sentences rather than disordered ones. It takes into account word meaning and grammatical structure, making it more adaptable to language changes. It is widely used in fields such as machine translation, VQA and text summarization. METEOR ranges from 0 to 1, and the higher the METEOR value, the better the sentences generated. However, METEOR also has some shortcomings, such as too many hyperparameters and lack of suitability for multiple datasets.

4.2.3. ROUGE

Recall-oriented Understudy for Gisting Evaluation (ROUGE) was introduced by Lin et al. [195] in 2004 and was extensively used in automatic summarization, where L stood for the longest common sentence. This metric is similar to BLEU because it assesses the similarity of a sentence to its reference sentence. However, the difference is that BLEU calculates accuracy, while ROUGE-L measures recall rate. It can measure the repeatability and importance of words, while being sensitive to length differences between the reference and generated captions. The range of ROUGE-L is between 0 and 1, and the larger the value, the higher the quality of the generated caption. Like BLEU, this metric does not take synonyms into consideration, either.

4.2.4. CIDEr

Different from the aforementioned evaluation metrics, Consensus-based Image Description Evaluation (CIDEr) is specifically designed for image caption and was introduced

by Vedantam [196] at CVPR 2015. CIDEr can be conceptualized as a three-dimensional extension of the BLEU index. It treats each sentence as a "document", which is represented in the form of Term Frequency Inverse Document Frequency (TF-IDF) vector, with the coherence of image captions being assessed through computation of the weight for each n-tuple (TF-IDF) and calculation of cosine similarity between reference captions and captions generated by models. By assigning weights, the model can emphasize key information with higher scores while deemphasizing less important descriptors with lower scores. Its advantages include considering semantic meaning and context, effectively describing visual content and successfully integrating reinforcement learning methods. However, its complexity in calculations is a drawback. When combined with reinforcement learning, the model can focus more on crucial information. Generally speaking, a higher CIDEr value indicates better quality of generated descriptions.

4.2.5. SPICE

In 2016, Anderson et al. [197] developed the Semantic Propositional Image Caption Evaluation (SPICE), which was created specifically to address the issue of image caption. In contrast to earlier evaluation metrics that concentrate on word similarity, SPICE assesses the sentence structure of output captions. The object, attribute and their connection are encoded in the form of a scene diagram, and the F-value of these three items is used to compute the SPICE value. This approach focuses more on semantics than surface structure, and is more consistent with how people actually communicate, but it does not take grammatical correctness into account.

4.2.6. S_m

S_m and S_m^* represent the average performance of a model and are commonly used for overall model evaluation. Given the challenge of improving deep learning methods in specific aspects, these metrics are less frequently utilized as they do not capture the strengths and weaknesses of individual components. But they are more often used in remote sensing image change caption.

5. Discussion

5.1. Analysis of State-of-the-Art Models on Different Datasets

In this section, we will conduct a comparative analysis of the outcomes derived from advanced methodologies across diverse datasets. The information presented in this section is collected and summarized from research findings from peers. To enhance tabular presentation, we appended zeros to results with fewer than four decimal places without altering the raw data. Our objective is to furnish novices with a comprehensive understanding of research advancements within this domain through this review and analysis to expedite their immersion into research endeavors.

5.1.1. Analysis on Sydney-Captions

The Sydney-Captions dataset is the smallest among the three datasets. Generally, a smaller dataset tends to yield superior results. However, its evaluation metrics can only rank second among the three datasets. This is attributed to the fact that compared with UCM-Captions, the captions of Sydney-Captions are lengthier and more informative. The performance of different models on the Sydney-Captions dataset is shown in Table 6.

In terms of metrics, the BLEU score of SkyEyeGPT significantly surpassed that of other models, reaching state-of-the-art levels in all current methods. SkyEyeGPT is the latest breakthrough in remote sensing LVLMs; it was developed on an extensive remote sensing instruction-tracking dataset and has demonstrated robust capabilities in visual language tasks such as VQA. VQA closely resembles RSIC and often encompasses more intricate image information than RSIC does, including number of objects, position, action, etc. SkyEyeGPT can provide detailed answers rather than merely describing the picture alone, so it is no surprise that it achieves remarkable results in BLEU scores. Similarly,

incorporating object counts into RSIC has shown amazing results in METEOR and SPICE metrics, implying numerical accuracy in caption tasks is important.

Table 6. The model performance on Sydney-Captions.

Methods	Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr	SPICE
Encoder–Decoder	Word-sentence [35]	0.7891	0.7094	0.6317	0.5625	0.4181	0.6922	2.0411	
	DMSFF [48]	0.8324	0.7489	0.6591	0.5851		0.7218	3.1898	
	SVMs [62]	0.7787	0.6835	0.6023	0.5305	0.3643	0.6746	2.2222	
	PMM [66]	0.8160			0.5560	0.4150	0.7360	2.3650	0.4300
Attention mechanism	MLA [81]	0.8057	0.7189	0.6448	0.5822	0.4665	0.7472	2.2028	0.4005
	VSCA [82]	0.7908	0.7200	0.6605	0.6088	0.4031	0.7354	2.6788	0.4637
	MC-Net [84]	0.8340	0.7500	0.6780	0.6070	0.4060	0.7390	2.5640	0.4560
	MSISAM [85]	0.7643	0.6919	0.6283	0.5725	0.3946	0.7172	2.8122	
	CHA [87]	0.8170	0.7420	0.6570	0.5910		0.7210	2.2910	
	MSMI [89]	0.8420	0.7570	0.6720	0.6010	0.4210	0.7330	2.8510	0.4450
	MAN [92]	0.7723			0.5617	0.4020	0.7284	2.6137	
	Adaptive [93]	0.8220	0.7410	0.6620	0.5940	0.3970		2.7050	
	SCAMET [95]	0.8072	0.7136	0.6431	0.5846	0.4614	0.7258	2.3570	
	Deformable [96]	0.8373	0.7771	0.7198	0.6659	0.4548	0.7860	3.0369	0.4839
	TrTr-CMR [97]	0.8270	0.6994	0.6002	0.5199	0.3803	0.7220	2.2728	
	PKGT [99]	0.8317	0.7783	0.7248	0.6824	0.4528	0.7706	2.8476	0.4405
	MGT [100]	0.8668	0.8087	0.7588	0.7122	0.4770	**0.7970**	**3.2776**	0.4809
	RAT [101]	0.8610	0.7900	0.7270	0.6700	0.4460	0.7870	2.8580	
	TSFE [102]	0.8279	0.7606	0.6951	0.6290	0.4228	0.7713	2.6561	
	CCT [103]	0.8660	0.7990	0.7470	0.6910	0.4600	0.7910	2.8670	
	AA [107]	0.8143	0.7351	0.6586	0.5806	0.4111	0.7195	2.3021	
	GAF-Net [108]	0.7950	0.7330	0.6590	0.6260	0.4580	0.7300	2.7720	
	RASG [109]	0.8000	0.7217	0.6531	0.5909	0.3908	0.7218	2.6311	0.4301
	GVFGA [110]	0.7681	0.6846	0.6145	0.5504	0.3866	0.7030	2.4522	0.4532
	GLCM [111]	0.8041	0.7305	0.6745	0.6259	0.4421	0.6965	2.4337	
	LAM [112]	0.7365	0.6440	0.5835	0.5348	0.3693	0.6827	2.3513	0.4351
	STA [113]	0.7910	0.7030	0.5920	0.5790			2.3060	
	SAA [114]	0.7155	0.6323	0.5469	0.4660	0.3132	0.6035	1.8027	0.3865
Reinforcement learning	TCE [118]	0.7937	0.7304	0.6717	0.6193	0.4430	0.7130	2.4042	
	SCST [120]	0.8016	0.7095	0.6314	0.5724	0.4183	0.7262	2.5392	
	Mask-CE [123]	0.8338	0.7572	0.6772	0.5980	0.4346	0.7660	2.6982	
Other auxiliary tasks	Object Count [138]	0.8659	0.8532	0.7625	0.7162	**0.5936**	0.7660	2.6887	**0.7355**
	DFEN [139]	0.7980	0.6970	0.6140	0.5420	0.3730	0.7230	2.0090	0.4490
	JTTS [142]				0.6518	0.4327	0.7474	2.8102	0.4729
	Meta capton [143]	0.8531	0.7915	0.7344	0.6800	0.4303	0.7417	2.6660	0.4359
	CRSR [51]	0.7994	0.7440	0.6987	0.6602	0.4150	0.7488	2.8900	0.4845
	SAM [22]	0.7795	0.7019	0.6392	0.5861	0.3954	0.7299	2.3791	
	SD-RSIC [148]	0.7610	0.6660	0.5860	0.5170	0.3660	0.6570	1.6900	
LLM	RSGPT [161]	0.8226	0.7528	0.6857	0.6223	0.4137	0.7477	2.7308	
	RS-CapRet [165]	0.7820			0.5450			2.3900	0.4230
	SkyEyeGPT [168]	**0.9185**	**0.8564**	**0.8088**	**0.7740**	0.4662	0.7774	1.8106	
Few-shot learning	SFRC [184]	0.8256	0.7449	0.6678	0.5939	0.4349	0.7560	2.6388	0.4445
	FRIC [185]	0.8239	0.7347	0.6582	0.5883	0.4241	0.7315	2.5539	0.4481

MGT demonstrated superior performance in ROUGE and CIDEr, highlighting the significant impact of regional features in transformer. The seamless integration of the self-attention mechanism and regional features in transformer was evident, and the crucial role of fine-tuning the image generated by CLIP should not be underestimated. CRSR also achieved a high SPICE score, indicating a noticeable enhancement in semantics through its semantic refinement module. In addition, RAT, CCT, Meta caption and DMSFF also outperformed other models.

5.1.2. Analysis of UCM-Captions

Among the three initial datasets, the UCM-Captions dataset has the greatest overall metrics because of its simplistic captions, which make it the least difficult dataset, so the

effect of machine learning is also closest to the original sentence. However, the disparity in results obtained by various models on the UCM-Captions dataset is minimal, and there are no instances of significant deviation. The performance of different advanced models on UCM-Captions is demonstrated in Table 7.

Table 7. The model performance on UCM-Captions.

Methods	Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr	SPICE
Encoder–Decoder	Word-sentence [35]	0.7931	0.7237	0.6671	0.6202	0.4395	0.7132	2.7871	
	DMSFF [48]	0.8306	0.7589	0.6972	0.6345		0.7318	3.2956	
	SVMs [65]	0.7653	0.6947	0.6417	0.5942	0.3702	0.6877	2.9228	
	PMM [66]	0.8620			0.7120	0.4820	0.8250	3.6540	0.5360
Attention mechanism	MLA [81]	0.8864	0.8233	0.7735	0.7271	0.5222	0.8441	3.3074	0.5021
	VSCA [82]	0.8900	0.8416	0.7987	0.7575	0.4931	0.8478	3.8314	0.5227
	MLCA-Net [83]	0.8260	0.7700	0.7170	0.6680	0.4350	0.7720	3.2400	0.4730
	MC-Net [84]	0.8450	0.7840	0.7320	0.6790	0.4490	0.7860	3.3550	0.5200
	MSISAM [85]	0.8727	0.8096	0.7551	0.7039	0.4652	0.8258	3.7129	
	CHA [87]	0.8230	0.7680	0.7100	0.6590		0.7560	3.1920	
	MSMI [89]	0.8430	0.7750	0.7110	0.6510	0.4530	0.7850	3.3810	0.5240
	MAN [92]	0.8480			0.6642	0.4562	0.8065	3.3172	
	Adaptive [93]	0.8390	0.7690	0.7150	0.6750	0.4460		3.2310	
	SCAMET [95]	0.8460	0.7772	0.7262	0.6812	0.5257	0.8166	3.3773	
	Deformable [96]	0.8230	0.7700	0.7228	0.6792	0.4439	0.7839	3.4629	0.4825
	TrTr-CMR [97]	0.8156	0.7091	0.6220	0.5469	0.3978	0.7442	2.4742	
	PKGT [99]	0.9048	0.8704	0.8410	0.8139	0.5466	0.8657	4.2749	0.5701
	MGT [100]	0.9128	0.8811	0.8537	0.8295	0.5601	**0.8877**	4.4839	0.5839
	RAT [101]	0.8950	0.8530	0.8110	0.7660	0.5030	0.8580	3.7930	
	TSFE [102]	0.8985	0.8546	0.8126	0.7645	0.4893	0.8800	3.6701	
	CCT [103]	0.9220	0.8900	0.8640	0.8330	0.5730	0.8830	4.1560	
	GAF-Net [108]	0.8690	0.8380	0.7870	0.7530	0.4790	0.8100	3.7760	
	RASG [109]	0.8518	0.7925	0.7432	0.6976	0.4571	0.8072	3.3887	
	GVFGA [110]	0.8319	0.7657	0.7102	0.6596	0.4436	0.7845	3.3270	0.4853
	GLCM [111]	0.8182	0.7540	0.6986	0.6468	0.4619	0.7524	3.0279	
	LAM [112]	0.8570	0.8120	0.7750	0.7430	0.5100	0.8260	3.7580	0.5350
	STA [113]	0.8170	0.7790	0.6130	0.6060			3.1990	
	SAA [114]	0.7828	0.7276	0.6759	0.6333	0.3803	0.6864	2.9057	0.4194
Reinforcement learning	TCE [118]	0.8210	0.7622	0.7140	0.6700	0.4775	0.7567	2.8547	
	SCST [120]	0.8382	0.7901	0.7443	0.7011	0.4464	0.7763	3.5653	
	Mask-CE [123]	0.8936	0.8517	0.8057	0.7650	0.5138	0.8585	3.8992	
	ADC [129]	0.8533	0.7568	0.6785	0.6117	**0.8324**	0.8087	**4.8650**	
Other auxiliary tasks	Object Count [138]	**0.9233**	**0.9152**	**0.8863**	**0.8596**	0.6235	0.8875	4.3622	**0.6126**
	DFEN [139]	0.8510	0.7840	0.7280	0.6770	0.4590	0.8050	3.1770	0.5010
	JTTS [142]				0.7376	0.4906	0.8364	3.7102	0.5231
	Meta capton [143]	0.9185	0.8876	0.8603	0.8347	0.5401	0.8758	4.3311	0.5710
	CRSR [51]	0.9060	0.8561	0.8122	0.7681	0.4956	0.8586	3.8069	0.5201
	SAM [22]	0.8538	0.8035	0.7572	0.7149	0.4632	0.8141	3.3489	
	SD-RSIC [148]	0.7480	0.6640	0.5980	0.5380	0.3900	0.6950	2.1320	
LLM	RSGPT [161]	0.8612	0.7914	0.7231	0.6574	0.4221	0.7834	3.3323	
	RS-LLaVA [165]	0.9000	0.8488	0.8030	0.7603	0.4921	0.8578	3.5561	
	RS-CapRet [168]	0.8330			0.6450			3.4290	0.5250
Few-shot learning	SFRC [184]	0.8856	0.8143	0.7778	0.7149	0.4706	0.8167	3.7595	0.5098
	FRIC [185]	0.8665	0.8162	0.7683	0.7226	0.4771	0.8163	3.5589	0.5128

We can see from the table that the object-count method had a strong performance in BLEU scores and SPICE scores, and achieved the highest results in all BLEU 1–4. This may be attributed to the UCM-Captions containing numerous elements that require quantification, such as aircrafts and so on. The model trained using the original method can only be described with terms like "many", which lacks precision. Employing an object-counting approach enhances accuracy significantly. Except for the object-count method, CCT also has a good performance in BLEU scores. Grid features and regional features were merged in CCT, and the excellent performance of CCT might be attributed to the novel exploration of regional features in RSIC.

Surprisingly, the highest METEOR metric came from the ADC model, which was essentially a reinforcement learning model. Probably because the ADC model used the method of reverse excitation of the generated caption back to the image features, it generated many captions in a reverse manner that are semantically more accurate than the ground truth. MGT performed best on ROUGE, and similarly CCT and MGT both used region-level features. In addition to the models mentioned above, there are still some models that achieved leading levels on UCM-Captions, such as meta caption, which closely matched CCT performance on UCM-Captions and achieved an excellent performance in BLEU scores. Not only that, as a large visual language model, RS-LLaVA also achieved good results on this dataset.

5.1.3. Analysis on RSICD

RSICD has the lowest overall score among the three datasets, while it contains the largest amount of data and the most diverse types of objects. Its repeated captions demand a high level of precision in assessment. The scores obtained by various advanced models on RSICD are evident in Table 8.

Table 8. The model performance on RSICD.

Methods	Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr	SPICE
Encoder–Decoder	Word-sentence [35]	0.7240	0.5861	0.4923	0.4250	0.3197	0.6260	2.0629	
	SA-FWC [43]	0.7200	0.6000	0.5100	0.4300	0.3700	0.6500	2.8300	
	VRTMM [50]	0.7813	0.6721	0.5645	0.5123	0.3737	0.5713	2.7150	
	SVMs [65]	0.6112	0.4277	0.3153	0.2411	0.2303	0.4588	0.6825	
	PMM [66]	0.7360			0.4540	0.3730	0.6600	2.6340	0.4770
Attention mechanism	MLA [81]	0.8058	0.6778	0.5866	0.5163	0.4718	0.7237	2.7716	0.4786
	VSCA [82]	0.7965	0.6856	0.5964	0.5224	0.3745	0.6833	2.9343	0.4914
	MLCA-Net [83]	0.7570	0.6430	0.5390	0.4610	0.3510	0.6460	2.3560	0.4440
	MC-Net [84]	0.7280	0.6060	0.5110	0.4330	0.3600	0.6410	2.4540	0.4630
	MSISAM [85]	0.7836	0.6679	0.5774	0.5042	0.3672	0.6730	2.8436	
	CHA [87]	0.7700	0.6490	0.5320	0.4710		0.6510	2.3630	
	MSMI [89]	0.7930	0.6810	0.5770	0.4980	0.3740	0.6820	2.7930	0.4910
	MAN [92]	0.7272			0.4436	0.3690	0.6575	2.5276	
	Adaptive [93]	0.7980	0.6470	0.5690	0.4890	0.2850		2.4040	
	SCAMET [95]	0.7681	0.6309	0.5352	0.4611	0.4572	0.6979	2.4681	
	Deformable [96]	0.7581	0.6416	0.5585	0.4923	0.3550	0.6523	2.5814	0.4579
	TrTr-CMR [97]	0.6201	0.3937	0.2671	0.1932	0.2399	0.4895	0.7518	
	PKGT [99]	0.6967	0.5830	0.5045	0.4431	0.3332	0.6078	2.7401	0.4691
	MGT [100]	0.7027	0.5980	0.5208	0.4601	0.3417	0.6227	2.8510	0.4807
	TSFE [102]	0.8054	0.7064	0.6208	0.5486	0.3793	0.7020	3.0570	
	CCT [103]	0.7980	0.6930	0.6080	0.5330	0.3830	0.6920	2.8810	
	GAF-Net [108]	0.7920	0.6760	0.5820	0.5100	0.4020	0.7100	2.7920	
	RASG [109]	0.7729	0.6651	0.5782	0.5062	0.3626	0.6691	2.7549	0.4719
	GVFGA [110]	0.6779	0.5600	0.4781	0.4165	0.3285	0.5929	2.6012	0.4683
	GLCM [111]	0.7767	0.6492	0.5642	0.4937	0.3627	0.6779	2.5491	
	LAM [112]	0.6756	0.5549	0.4714	0.4077	0.3261	0.5848	2.6285	0.4671
	STA [113]	0.8130	0.6810	0.5410	0.4960			2.6950	
	SAA [114]	0.6196	0.4819	0.3902	0.3195	0.2733	0.5143	1.6386	0.3598
Reinforcement learning	TCE [118]	0.7608	0.6358	0.5471	0.4791	0.3425	0.6687	2.4665	
	SCST [120]	0.7704	0.6549	0.5632	0.4882	0.3698	0.6696	2.6840	
	Mask-CE [123]	0.8042	0.6996	0.6136	0.5414	0.3937	0.7058	2.9839	
	ADC [129]	0.7397	0.5526	0.4635	0.4102	0.2213	0.7131	2.2430	
Other auxiliary tasks	Object Count [138]	0.8221	0.7480	0.6999	0.6285	**0.5289**	**0.7408**	2.7522	**0.5429**
	DFEN [139]	0.7660	0.6360	0.5380	0.4630	0.3730	0.6850	2.6050	0.4770
	JTTS [142]				0.5135	0.3773	0.6823	2.7958	0.4877
	Meta capton [143]	0.7067	0.5936	0.5119	0.4471	0.3394	0.6082	2.7425	0.4772
	CRSR [51]	0.8192	0.7171	0.6307	0.5574	0.4015	0.7134	**3.0687**	0.5276
	SAM [22]	0.7016	0.5614	0.4648	0.3934	0.3291	0.5706	1.7031	
	SD-RSIC [148]	0.6440	0.4740	0.3690	0.3000	0.2490	0.5230	0.7940	
VLLM	RSGPT [161]	0.7032	0.5423	0.4402	0.3683	0.3010	0.5334	1.0294	
	RS-CapRet [165]	0.7410			0.4550			2.6050	0.4840
	SkyEyeGPT [168]	**0.8773**	**0.7770**	**0.6890**	**0.6199**	0.3623	0.6354	0.8937	
Few-shot learning	SFRC [184]	0.8009	0.6952	0.6084	0.5345	0.3882	0.6974	2.8727	0.5096
	FRIC [185]	0.7925	0.6853	0.5977	0.5238	0.3912	0.6981	2.8830	0.5133

Most mainstream models achieve a BLEU-1 score around 0.75–0.8, while BLEU-2 is approximately 0.6–0.7, BLEU-3 is about 0.5–0.6 and BLEU-4 is about 0.4–0.5. However, SkyEyeGPT also achieved superior performance, with all four BLEU metrics surpassing those of mainstream models. Strong performance on this dataset is also achieved with the object-count method, with the highest score on three metrics, METEOR, ROUGE and SPICE. MLA also outperformed others in terms of METEOR, thanks to its multi-level attention to regional, linguistic, visual and semantic aspects which not only enhanced feature attention but also improved the accuracy of word usage and semantic coherence in sentences. The high METEOR score was attributed to its emphasis on semantic coherence rather than repetition rate with ground truth. Additionally, dataset corrections reduced the likelihood of incorrect sentences during training and generation. These corrections may also explain why MLA achieved high results on ROUGE by reducing the number of wrong words in ground truth and increasing the probability of words appearing in generated sentences. Additionally, SCAMET also achieves a high METEOR score while CRSR should lead in ROUGE. The success of CRSR in ROUGE could be credited to its cross-modal retrieval for supplementing semantic information and a semantic refinement module for filtering error information. This model was able to only focus on useful information and could optimize sentence structure with supplementary semantic information. The performance of both few-shot learning methods was also noteworthy.

5.2. Pros and Cons of Different Methods

Based on the previous research of the principles of various methods, the understanding of the models, and the performance on different datasets, we summarized the advantages and disadvantages of the previous seven methods in RSIC. We only highlighted some common issues, which are not absolute. This approach allows researchers to select more suitable methods for their own research and also encourages them to further enhance the strengths or address the shortcomings for a new breakthrough within each method. Pros and cons of different methods are shown in Table 9.

Table 9. Pros and cons of different methods.

Methods	Advantages	Disadvantages
Encoder–Decoder	It serves as the foundation for all models and is highly susceptible enhancement	Limited capacity for improvement
Attention mechanism	It can help us more accurately focus on and utilize image features	High sensitivity to the resolution of images, and traditional dataset images tend to be blurry, which may lead to overfitting
Reinforcement learning	Enhanced model performance without the need for additional datasets	Increased demand for computing power
Learning with auxiliary tasks	More accurate object and attribute information, fewer errors	Most of them rely on pre-trained models, and double model training
LVLMs	It can be used in multi-tasks without training	High cost
Few-shot learning	It reduces reliance on datasets	Hard to achieve sota performance, and the effect is not obvious on small datasets.

The encoder–decoder structure was the foundation for all RSIC models. Most research is based on it, but simply changing the structure of it may only yield limited improvements in model effectiveness. An attention mechanism can assist the model in focusing more on specific regions and details of interest. However, it is sensitive to the resolution of images. In low-resolution remote sensing images, excessive enhancement of the attention mechanism may not necessarily improve the effect. Reinforcement learning offers the advantage of not requiring additional datasets. Given that data resources are scarce in remote sensing, introducing reinforcement learning strategies can enhance model performance without in-

creasing resource consumption. However, this approach also leads to increased complexity and computing power requirements for the model. To incorporate more object features and attribute features into the model, several other deep learning tasks are typically used as auxiliary tasks to support RSIC. With guidance from multi-modal and multi-attribute knowledge, RSIC models can sometimes achieve better results. At the same time, they usually require pre-trained models for other tasks or double training amounts. LLMs can simultaneously handle multiple tasks without needing retraining; some tasks only need fine-tuning as necessary. Nevertheless, training LLMs is costly and demands substantial infrastructure. Lastly, few-shot learning reduces an RSIC model's reliance on extensive datasets; fine effects can be achieved with only a portion of the available datasets, but this method cannot lead to optimal or even superior model effects. And when working with very limited datasets such as less than 10 images, the resulting performance remains subpar.

6. Future Directions

The increasing advancement of CV and NLP has led to a wider range of applications in remote sensing. As one of them, RSIC has been able to achieve increasingly remarkable results. Nevertheless, there still exist numerous challenges and difficulties that need to be addressed, which are complex and will serve as the primary focus of RSIC research in the future.

(1) Large and high-quality datasets

At present, the three most commonly used datasets in RSIC are all small. Concerning NWPU-Captions, the slightly larger dataset has too few users due to its being released only recently, and the quantity is still too small compared with the COCO dataset in natural image caption. With the significant success of LVLMs in CV and NLP, the need for large and high-quality datasets has become even more urgent. The lack of datasets greatly limits a lot of research. In addition to the small number of datasets, there are many errors in the existing datasets. Although some scholars have corrected them [81], some sentences are too simple and the problem of insufficient sentences has not been solved. This highlights the need to balance large datasets with high-quality annotations. Generally, as the dataset grows larger, the annotation workload increases, which may lead to a decline in annotation quality. To tackle this challenge, active learning [198] can be a viable option and can select the most relevant data for annotation and maximizing the balance between dataset accuracy and annotation effort. In a word, this domain currently lacks a common dataset like COCO, for experimenting with three datasets for each study costs time and resources, relatively.

(2) Opportunities for Regional Features

Region-level features have historically driven significant advancements in natural image caption [199–201] thanks to the fact that the COCO dataset is both an image caption dataset and an object-detection dataset. By utilizing Faster-RCNN, researchers extracted regions from the COCO dataset to acquire bottom-up features, which resulted in a substantial enhancement in transformer-based image caption. Conversely, as the RSIC dataset originates from remote sensing image classification, manual region annotation is required for extracting regional features. Some scholars have begun to work on this aspect [101] and have yielded promising outcomes in transformer-based RSIC models [103]. A prominent direction for future research involves further optimizing RSIC models through leveraging region-level features instead of CNN pre-extracted features within specific remote sensing environments.

(3) Opportunities and Challenges of LVLMs

The introduction of LVLMs presents both opportunities and challenges for RSIC. The opportunities lie in the robust object-recognition capability of remote sensing AI large models, which can swiftly and accurately identify various features on the surface such as farmland, crops and buildings. This significantly enhances the efficiency of analyzing remote sensing images and is suitable for processing multi-dimensional data including visi-

ble light, multispectral, hyperspectral and LiDAR data. SkyEyeGPT [168] has demonstrated outstanding results in RSIC, surpassing all other models using different deep learning methods during the same period, and showcasing the potential of LVLMs in RSIC. However, LVLMs also pose numerous challenges such as requiring substantial data support for their construction as mentioned above, along with issues related to data security and privacy [202]. Many recent RSIC missions involve LVLM with several excellent articles published [169,203]. Although there have been proposals for LVLMs applicable to visual language tasks, few studies have compared and analyzed their performance under different tasks against other contemporary models. Enhancing the adaptability of LVLMs and their effectiveness under specific tasks is crucial for future research.

(4) Remote Sensing Image Change Caption

Remote sensing image change caption is employed for the analysis of geomorphic changes in remote sensing imagery. Despite being a relatively recent undertaking, substantial progress has been achieved through advanced research outcomes [204,205]. In this area, there is a unified dataset known as LEVIR-CC, and numerous deep learning methodologies still remain unexplored within the remote sensing image change caption framework. One particularly promising avenue for future work involves applying various deep learning techniques to address challenges encountered in remote sensing image change caption. For example, existing change caption models primarily focus on dual-time images and have not explored the change caption of sequence images [206]. However, in certain cases, sequence images are more meaningful for accurately describing changes in remote sensing images. Additionally, current remote sensing image change caption methods often focus solely on satellite remote sensing images, neglecting datasets of other modalities, such as radar and infrared images. Investigating multimodal remote sensing image change caption methods [207,208] is crucial for broadening the application scope of remote sensing image change caption and enhancing its universality.

(5) Diffusion model

The current RSIC tasks primarily rely on autoregressive decoding to generate coherent sentences. However, this approach is hindered by high latency and substantial computational costs. In contrast, non-autoregressive methods enhance inference speed by predicting multiple tokens simultaneously, but their performance may be reduced due to the absence of sequential dependencies. Recently, non-autoregressive decoding based on diffusion models has shown promise in the iterative refinement of natural image captions [209]. Nevertheless, the unique characteristics of remote sensing images restrict its effectiveness, complicating the creation of robust inputs and, consequently, impacting the accuracy of the generated captions.

(6) Deeper alignment between image and text

Image and text convey distinct dimensions of information. Aligning them enables models to gain a more comprehensive understanding of the image's content, thereby assisting in interpreting the meaning of text within specific contexts. For instance, visual features can identify the presence of objects in an image, while textual features can explain an object's function or purpose. By aligning visual and textual information, we can enhance the accuracy and relevance of generated captions, ensuring that the text accurately reflects what is depicted in the image. Numerous scholars have proposed methods for vision–text alignment. For example, Yang et al. [210] matched specific objects in remote sensing images with corresponding proper nouns in generated sentences. Additionally, they introduced a vision–text alignment mechanism [211] in the decoder to produce precise descriptive sentences. Research into the deeper alignment between image and text will significantly contribute to advancements in remote sensing image caption.

(7) More application scenarios

Although RSIC has made significant progress, there is limited research on the application level. It is crucial to integrate RSIC technology with specific application scenarios. For

instance, the study in [147] has produced captions for landslides, and we can also generate image captions for earthquakes [212] and floods [213], which is essential for disaster warning. In the future, there are numerous remote sensing scenarios that can be analyzed with RSIC such as vegetation distribution [214] that provide guidance for reclamation, afforestation and greening work. It can also be applied to urban building layout [215], which facilitates overall urban-construction planning. We can also introduce RSIC tasks into high-dimensional datasets or datasets of other modalities, such as LiDAR point cloud images; attempts made on this basis have already appeared in this area [216], but due to the different imaging mechanisms [217], transferring models trained on conventional remote sensing images to datasets of other modalities will be a challenge. Additionally, given the development of multimodal data-fusion technology [218], incorporating datasets of different modalities to obtain more comprehensive information and generate more detailed captions might be a good option. In summary, there are various remote sensing fields where image caption can be applied. Future work should focus on further developing image caption in different remote sensing environments to unlock new possibilities for RSIC technology.

7. Conclusions

This article provides a review of RSIC based on deep learning. First of all, we outlined the methods we used to obtain and filter the papers we discussed. We then categorized techniques in RSIC into six categories according to methods, which were encoder–decoder framework, attention mechanism, reinforcement learning, auxiliary task learning, large visual language models and few-shot learning. Later, we reviewed the latest advancement of these six parts in recent years, and briefly sorted out the application of different models under different methods. By comparing with natural image caption, we obtained new insights for the development of RSIC in the future. Then, the datasets and evaluation metrics for RSIC were briefly introduced, which lays a foundation for evaluating and analyzing different models. Using index comparison, an analysis of the outstanding results achieved in recent years was conducted on different datasets to identify which methods have excellent performance and examine their reasons for success briefly. Finally, according to the previous research in RSIC and the analysis of their results, the future directions of RSIC were forecasted and summarized. Through this review, researchers are able to gain a more comprehensive understanding of the current methods, development, opportunities, challenges and future trends of RSIC.

Author Contributions: All of the authors made significant contributions to the manuscript. The review was created and written by P.L. guided by the oversight of K.Z. and J.W. helped with the review of the related literature. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (NSFC) under grant numbers 62076093, 61871182, 62206095, by the Fundamental Research Funds for the Central Universities under grant numbers 2022MS078, 2023JG002, 2023JC006.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: The authors express their sincere gratitude to the academic editors and reviewers for their valuable comments and constructive suggestions.

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

References

1. Wang, A.; Rauf, A.; Ozturk, I.; Wu, J.; Zhao, X.; Du, H. The key to sustainability: In-depth investigation of environmental quality in G20 countries through the lens of renewable energy, economic complexity and geopolitical risk resilience. *J. Environ. Manag.* **2024**, *352*, 120045. [CrossRef]
2. Zhang, B.; Wu, Y.; Zhao, B.; Chanussot, J.; Hong, D.; Yao, J.; Gao, L. Progress and challenges in intelligent remote sensing satellite systems. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 1814–1822. [CrossRef]

3. Selva, M.; Santurri, L.; Baronti, S. Improving hypersharpening for WorldView-3 data. *IEEE Geosci. Remote Sens. Lett.* **2018**, *16*, 987–991. [CrossRef]
4. Sefercik, U.G.; Alkan, M.; Jacobsen, K.; Atalay, C.; Buyuksalih, G. Quality analysis of Worldview-4 DSMs generated by least squares matching and semiglobal matching. *J. Appl. Remote Sens.* **2021**, *15*, 034515. [CrossRef]
5. Hestrio, Y.F.; Soleh, M.; Hidayat, A.; Afida, H.; Gunawan, H.; Maryanto, A. Satellite data receiving antenna system for pleiades neo observation satellite. *J. Phys. Conf. Ser.* **2021**, *1763*, 012019. [CrossRef]
6. Cheng, G.; Han, J.; Lu, X. Remote sensing image scene classification: Benchmark and state of the art. *Proc. IEEE* **2017**, *105*, 1865–1883. [CrossRef]
7. Zhao, Z.; Li, J.; Luo, Z.; Li, J.; Chen, C. Remote sensing image scene classification based on an enhanced attention module. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 1926–1930. [CrossRef]
8. Chen, H.; Qi, Z.; Shi, Z. Remote sensing image change detection with transformers. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5607514. [CrossRef]
9. Shi, Q.; Liu, M.; Li, S.; Liu, X.; Wang, F.; Zhang, L. A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5604816. [CrossRef]
10. Wang, Y.; Bashir, S.M.A.; Khan, M.; Ullah, Q.; Wang, R.; Song, Y.; Guo, Z.; Niu, Y. Remote sensing image super-resolution and object detection: Benchmark and state of the art. *Expert Syst. Appl.* **2022**, *197*, 116793. [CrossRef]
11. Sharma, M.; Dhanaraj, M.; Karnam, S.; Chachlakis, D.G.; Ptucha, R.; Markopoulos, P.P.; Saber, E. YOLOrs: Object detection in multimodal remote sensing imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *14*, 1497–1508. [CrossRef]
12. Li, R.; Zheng, S.; Zhang, C.; Duan, C.; Su, J.; Wang, L.; Atkinson, P.M. Multiattention network for semantic segmentation of fine-resolution remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5607713. [CrossRef]
13. Zhao, Q.; Liu, J.; Li, Y.; Zhang, H. Semantic segmentation with attention mechanism for remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5403913. [CrossRef]
14. Liu, Y.; Ding, L.; Chen, C.; Liu, Y. Similarity-based unsupervised deep transfer learning for remote sensing image retrieval. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 7872–7889. [CrossRef]
15. Cao, R.; Zhang, Q.; Zhu, J.; Li, Q.; Li, Q.; Liu, B.; Qiu, G. Enhancing remote sensing image retrieval using a triplet deep metric learning network. *Int. J. Remote Sens.* **2020**, *41*, 740–751. [CrossRef]
16. 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]
17. 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]
18. Liu, Q.; Ruan, C.; Zhong, S.; Li, J.; Yin, Z.; Lian, X. Risk assessment of storm surge disaster based on numerical models and remote sensing. *Int. J. Appl. Earth Obs. Geoinf.* **2018**, *68*, 20–30. [CrossRef]
19. Gao, L.; Wang, X.; Johnson, B.A.; Tian, Q.; Wang, Y.; Verrelst, J.; Mu, X.; Gu, X. Remote sensing algorithms for estimation of fractional vegetation cover using pure vegetation index values: A review. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 364–377. [CrossRef]
20. Karthikeyan, L.; Chawla, I.; Mishra, A.K. A review of remote sensing applications in agriculture for food security: Crop growth and yield, irrigation, and crop losses. *J. Hydrol.* **2020**, *586*, 124905. [CrossRef]
21. Wentz, E.A.; Anderson, S.; Fragkias, M.; Netzband, M.; Mesev, V.; Myint, S.W.; Quattrochi, D.; Rahman, A.; Seto, K.C. Supporting global environmental change research: A review of trends and knowledge gaps in urban remote sensing. *Remote Sens.* **2014**, *6*, 3879–3905. [CrossRef]
22. 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]
23. Li, M.; Zang, S.; Zhang, B.; Li, S.; Wu, C. A review of remote sensing image classification techniques: The role of spatio-contextual information. *Eur. J. Remote Sens.* **2014**, *47*, 389–411. [CrossRef]
24. Song, J.; Gao, S.; Zhu, Y.; Ma, C. A survey of remote sensing image classification based on CNNs. *Big Earth Data* **2019**, *3*, 232–254. [CrossRef]
25. Tuia, D.; Volpi, M.; Copa, L.; Kanevski, M.; Munoz-Mari, J. A survey of active learning algorithms for supervised remote sensing image classification. *IEEE J. Sel. Top. Signal Process.* **2011**, *5*, 606–617. [CrossRef]
26. Kotaridis, I.; Lazaridou, M. Remote sensing image segmentation advances: A meta-analysis. *ISPRS J. Photogramm. Remote Sens.* **2021**, *173*, 309–322. [CrossRef]
27. Wang, Y.; Lv, H.; Deng, R.; Zhuang, S. A comprehensive survey of optical remote sensing image segmentation methods. *Can. J. Remote Sens.* **2020**, *46*, 501–531. [CrossRef]
28. Kaur, H. Review of remote sensing image segmentation techniques. *Int. J. Adv. Res. Comput. Eng. Technol. (IJARCET)* **2015**, *4*, 1667–1674.
29. Khelifi, L.; Mignotte, M. Deep learning for change detection in remote sensing images: Comprehensive review and meta-analysis. *IEEE Access* **2020**, *8*, 126385–126400. [CrossRef]
30. Afaq, Y.; Manocha, A. Analysis on change detection techniques for remote sensing applications: A review. *Ecol. Inform.* **2021**, *63*, 101310. [CrossRef]

31. Bai, T.; Wang, L.; Yin, D.; Sun, K.; Chen, Y.; Li, W.; Li, D. Deep learning for change detection in remote sensing: A review. *Geo-Spat. Inf. Sci.* **2023**, *26*, 262–288. [CrossRef]
32. Zhao, B. A systematic survey of remote sensing image captioning. *IEEE Access* **2021**, *9*, 154086–154111. [CrossRef]
33. Bashmal, L.; Bazi, Y.; Melgani, F.; Al Rahhal, M.M.; Al Zuair, M.A. Language Integration in Remote Sensing: Tasks, datasets, and future directions. *IEEE Geosci. Remote Sens. Magazine* **2023**, *11*, 63–93. [CrossRef]
34. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3156–3164.
35. Albawi, S.; Mohammed, T.A.; Al-Zawi, S. Understanding of a convolutional neural network. In Proceedings of the 2017 International Conference on Engineering and Technology (ICET), Antalya, Turkey, 21–23 August 2017.
36. Salehinejad, H.; Sankar, S.; Barfett, J.; Colak, E.; Valaee, S. Recent advances in recurrent neural networks. *arXiv* **2017**, arXiv:1801.01078.
37. Rubinstein, R.Y.; Kroese, D.P. *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation, and Machine Learning*; Springer: New York, NY, USA, 2004; Volume 133.
38. 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]
39. 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, IEEE.
40. 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]
41. Nanal, W.; Hajiarbabi, M. Captioning remote sensing images using transformer architecture. In Proceedings of the 2023 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Bali, Indonesia, 20–23 February 2023; pp. 413–418.
42. Shen, X.; Liu, B.; Zhou, Y.; Zhao, J. Remote sensing image caption generation via transformer and reinforcement learning. *Multimed. Tools Appl.* **2020**, *79*, 26661–26682. [CrossRef]
43. Wang, J.; Wang, B.; Xi, J.; Bai, X.; Ersoy, O.K.; Cong, M.; Gao, S.; Zhao, Z. Remote Sensing Image Captioning with Sequential Attention and Flexible Word Correlation. *IEEE Geosci. Remote. Sens. Lett.* **2024**, *21*, 6004505. [CrossRef]
44. Geetha, G.; Kirthigadevi, T.; Ponsam, G.; Karthik, T.; Safa, M. Image Captioning Using Deep Convolutional Neural Networks (CNNs). *J. Phys. Conf. Ser.* **2020**, *1712*, 012015. [CrossRef]
45. Zhu, Y.; Newsam, S. Densenet for dense flow. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 790–794.
46. Jastrzębski, S.; Arpit, D.; Ballas, N.; Verma, V.; Che, T.; Bengio, Y. Residual connections encourage iterative inference. *arXiv* **2017**, arXiv:1710.04773.
47. Badhe, N.B.; Bharadi, V.A.; Giri, N.; Alegavi, S. Deep Attention Based DenseNet with Visual Switch Added BiLSTM for Caption Generation from Remote Sensing Images. *Int. J. Intell. Eng. Syst.* **2023**, *16*, 677–687.
48. 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]
49. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20 June 2009; pp. 248–255.
50. 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]
51. 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. [CrossRef]
52. Graves, A. *Supervised Sequence Labelling with Recurrent Neural Networks*; Springer Nature: Dordrecht, The Netherlands, 2012; pp. 37–45.
53. Dey, R.; Salem, F.M. Gate-variants of gated recurrent unit (GRU) neural networks. In Proceedings of the 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS), Boston, MA, USA, 6–9 August 2017; pp. 1597–1600.
54. Chouaf, S.; Hoxha, G.; Smara, Y.; Melgani, F. Captioning changes in bi-temporal remote sensing images. In Proceedings of the 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium, 11–16 July 2021; pp. 2891–2894.
55. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems 30*; Guyon, I., Von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Neural Information Processing Systems Foundation, Inc.: La Jolla, CA, USA, 2017.
56. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27*; Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberg, K.Q., Eds.; Neural Information Processing Systems Foundation, Inc.: La Jolla, CA, USA, 2014.
57. Wang, Q.; Li, B.; Xiao, T.; Zhu, J.; Li, C.; Wong, D.F.; Chao, L.S. Learning deep transformer models for machine translation. *arXiv* **2019**, arXiv:1906.01787.
58. Liu, Y.; Lapata, M. Text summarization with pretrained encoders. *arXiv* **2019**, arXiv:1908.08345.

59. Li, G.; Zhu, L.; Liu, P.; Yang, Y. Entangled transformer for image captioning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8928–8937.
60. Li, J.; Yao, P.; Guo, L.; Zhang, W. Boosted transformer for image captioning. *Appl. Sci.* **2019**, *9*, 3260. [CrossRef]
61. Suthaharan, S.; Suthaharan, S. Support vector machine. In *Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning*; Springer Publishing: New York, NY, USA, 2016; pp. 207–235.
62. Rosset, S.; Zhu, J.; Hastie, T. Margin maximizing loss functions. In *Advances in Neural Information Processing Systems 16*; Thrun, S., Saul, L., Schölkopf, B., Eds.; Neural Information Processing Systems Foundation, Inc.: La Jolla, CA, USA, 2003.
63. Belgiu, M.; Drăguţ, L. Random forest in remote sensing: A review of applications and future directions. *ISPRS J. Photogramm. Remote Sens.* **2016**, *114*, 24–31. [CrossRef]
64. Hoxha, G.; Chouaf, S.; Melgani, F.; Smara, Y. Change captioning: A new paradigm for multitemporal remote sensing image analysis. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5627414. [CrossRef]
65. Hoxha, G.; Melgani, F. A novel SVM-based decoder for remote sensing image captioning. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5404514. [CrossRef]
66. 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]
67. Wang, J.; Chen, Z.; Ma, A.; Zhong, Y. Capformer: Pure transformer for remote sensing image caption. In Proceedings of the 2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022; pp. 7996–7999.
68. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y. A survey on vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 87–110. [CrossRef]
69. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *aXiv* **2018**, arXiv:1810.04805.
70. Liu, C.; Zhao, R.; Shi, Z. Remote-sensing image captioning based on multilayer aggregated transformer. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6506605. [CrossRef]
71. Zhang, X.; Wang, Q.; Chen, S.; Li, X. Multi-scale cropping mechanism for remote sensing image captioning. In Proceedings of the 2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 10039–10042.
72. Ramos, R.; Martins, B. Using neural encoder-decoder models with continuous outputs for remote sensing image captioning. *IEEE Access* **2022**, *10*, 24852–24863. [CrossRef]
73. Chen, Z.; Wang, J.; Ma, A.; Zhong, Y. TypeFormer: Multiscale transformer with type controller for remote sensing image caption. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6514005. [CrossRef]
74. Hoxha, G.; Scuccato, G.; Melgani, F. Improving image captioning systems with postprocessing strategies. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5612013. [CrossRef]
75. Galassi, A.; Lippi, M.; Torroni, P. Attention in natural language processing. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 4291–4308. [CrossRef]
76. 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.
77. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6077–6086.
78. Huang, L.; Wang, W.; Chen, J.; Wei, X.Y. Attention on attention for image captioning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4634–4643.
79. Lu, J.; Xiong, C.; Parikh, D.; Socher, R. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Venice, Italy, 22–29 October 2017; pp. 375–383.
80. Zhang, Z.; Zhang, W.; Diao, W.; Yan, M.; Gao, X.; Sun, X. VAA: Visual aligning attention model for remote sensing image captioning. *IEEE Access* **2019**, *7*, 137355–137364. [CrossRef]
81. 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]
82. 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]
83. Cheng, Q.; Huang, H.; Xu, Y.; Zhou, Y.; Li, H.; Wang, Z. NWPU-captions dataset and MLCA-net for remote sensing image captioning. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5629419. [CrossRef]
84. Huang, H.; Shao, Z.; Cheng, Q.; Huang, X.; Wu, X.; Li, G.; Tan, L. MC-Net: Multi-scale contextual information aggregation network for image captioning on remote sensing images. *Int. J. Digit. Earth* **2023**, *16*, 4848–4866. [CrossRef]
85. 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]
86. Li, Y.; Qi, H.; Dai, J.; Ji, X.; Wei, Y. Fully convolutional instance-aware semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision And Pattern Recognition, Venice, Italy, 22–29 October 2017; pp. 2359–2367.

87. Wang, C.; Jiang, Z.; Yuan, Y. Instance-aware remote sensing image captioning with cross-hierarchy attention. In Proceedings of the IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium, Online, 26 September–2 October 2020; pp. 980–983.
88. Ma, X.; Zhao, R.; Shi, Z. Multiscale methods for optical remote-sensing image captioning. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 2001–2005. [CrossRef]
89. Wang, Y.; Zhang, W.; Zhang, Z.; Gao, X.; Sun, X. Multiscale multiinteraction network for remote sensing image captioning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 2154–2165. [CrossRef]
90. He, Y.; Carass, A.; Zuo, L.; Dewey, B.E.; Prince, J.L. Self domain adapted network. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*; Springer International Publishing: Berlin, Germany, 2020; pp. 437–446.
91. 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]
92. Meng, Y.; Gu, Y.; Ye, X.; Tian, J.; Wang, S.; Zhang, H.; Hou, B.; Jiao, L. Multi-view attention network for remote sensing image captioning. In Proceedings of the 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium, 11–16 July 2021; pp. 2349–2352.
93. Zia, U.; Riaz, M.M.; Ghafoor, A. Transforming remote sensing images to textual descriptions. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *108*, 102741. [CrossRef]
94. Cornia, M.; Stefanini, M.; Baraldi, L.; Cucchiara, R. Meshed-memory transformer for image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10578–10587.
95. 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]
96. 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]
97. Wu, Y.; Li, L.; Jiao, L.; Liu, F.; Liu, X.; Yang, S. TrTr-CMR: Cross-Modal Reasoning Dual Transformer for Remote Sensing Image Captioning. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5643912. [CrossRef]
98. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
99. Meng, L.; Wang, J.; Yang, Y.; Xiao, L. Prior Knowledge-Guided Transformer for Remote Sensing Image Captioning. *IEEE Trans. Geosci. Remote. Sens.* **2023**, *61*, 4706213. [CrossRef]
100. Meng, L.; Wang, J.; Meng, R.; Yang, Y.; Xiao, L. A Multiscale Grouping Transformer With CLIP Latents for Remote Sensing Image Captioning. *IEEE Trans. Geosci. Remote. Sens.* **2024**, *62*, 4703515. [CrossRef]
101. Zhao, K.; Xiong, W. Exploring region features in remote sensing image captioning. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *127*, 103672. [CrossRef]
102. Guo, J.; Li, Z.; Song, B.; Chi, Y. TSFE: Two-Stage Feature Enhancement for Remote Sensing Image Captioning. *Remote Sens.* **2024**, *16*, 1843. [CrossRef]
103. Zhao, K.; Xiong, W. Cooperative Connection Transformer for Remote Sensing Image Captioning. *IEEE Trans. Geosci. Remote. Sens.* **2024**, *62*, 5607314. [CrossRef]
104. Cai, C.; Wang, Y.; Yap, K.H. Interactive change-aware transformer network for remote sensing image change captioning. *Remote Sens.* **2023**, *15*, 5611. [CrossRef]
105. Zhou, Q.; Gao, J.; Yuan, Y.; Wang, Q. Single-Stream Extractor Network With Contrastive Pre-Training for Remote-Sensing Change Captioning. *IEEE Trans. Geosci. Remote. Sens.* **2024**, *62*, 5624514. [CrossRef]
106. Liu, C.; Yang, J.; Qi, Z.; Zou, Z.; Shi, Z. Progressive scale-aware network for remote sensing image change captioning. In Proceedings of the IGARSS 2023–2023 IEEE International Geoscience and Remote Sensing Symposium, Pasadena, CA, USA, 16–21 July 2023; pp. 6668–6671.
107. 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]
108. Peng, Y.; Jia, Y.; Chen, J.; Ji, X. GAF-Net: Global view guided attribute fusion network for remote sensing image captioning. *Multimed. Tools Appl.* **2024**, *83*, 22409–22431. [CrossRef]
109. 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*, 5608816. [CrossRef]
110. 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*, 5615216. [CrossRef]
111. 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]
112. 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]

113. Cheng, K.; Wu, Z.; Jin, H.; Li, X. Remote Sensing Image Captioning with Multi-Scale Feature and Small Target Attention. In Proceedings of the IGARSS 2024–2024 IEEE International Geoscience and Remote Sensing Symposium, Athens, Greece, 7–12 July 2024; pp. 7436–7439.
114. 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]
115. Abdul, Z.K.; Al-Talabani, A.K. Mel frequency cepstral coefficient and its applications: A review. *IEEE Access* **2022**, *10*, 122136–122158. [CrossRef]
116. Zhang, H.; Parkes, D.C.; Chen, Y. Policy teaching through reward function learning. In Proceedings of the 10th ACM Conference on Electronic Commerce, Stanford, CA, USA, 6–10 July 2009; pp. 295–304.
117. Arulkumaran, K.; Deisenroth, M.P.; Brundage, M.; Bharath, A.A. Deep reinforcement learning: A brief survey. *IEEE Signal Process. Mag.* **2017**, *34*, 26–38. [CrossRef]
118. 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]
119. 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.
120. Rennie, S.J.; Marcheret, E.; Mroueh, Y.; Ross, J.; Goel, V. Self-critical sequence training for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7008–7024.
121. Ranzato, M.A.; Chopra, S.; Auli, M.; Zaremba, W. Sequence level training with recurrent neural networks. *arXiv* **2015**, arXiv:1511.06732.
122. Luo, R. A better variant of self-critical sequence training. *arXiv* **2020**, arXiv:2003.09971.
123. 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]
124. Drenkow, N.; Sani, N.; Shpitser, I.; Unberath, M. A systematic review of robustness in deep learning for computer vision: Mind the gap? *arXiv* **2021**, arXiv:2112.00639.
125. Zhang, L.; Sung, F.; Liu, F.; Xiang, T.; Gong, S.; Yang, Y.; Hospedales, T.M. Actor-critic sequence training for image captioning. *arXiv* **2017**, arXiv:1706.09601.
126. Creswell, A.; White, T.; Dumoulin, V.; Arulkumaran, K.; Sengupta, B.; Bharath, A.A. Generative adversarial networks: An overview. *IEEE Signal Process. Mag.* **2018**, *35*, 53–65. [CrossRef]
127. Rui, X.; Cao, Y.; Yuan, X.; Kang, Y.; Song, W. Disastergan: Generative adversarial networks for remote sensing disaster image generation. *Remote Sens.* **2021**, *13*, 4284. [CrossRef]
128. Pfau, D.; Vinyals, O. Connecting generative adversarial networks and actor-critic methods. *arXiv* **2016**, arXiv:1610.01945.
129. Chavhan, R.; Banerjee, B.; Zhu, X.X.; Chaudhuri, S. A novel actor dual-critic model for remote sensing image captioning. In Proceedings of the 2020 25th International Conference on Pattern Recognition, Milan, Italy, 10–15 January 2020; pp. 4918–4925.
130. Tong, Y.; Chen, Y.; Shi, X. A multi-task approach for improving biomedical named entity recognition by incorporating multi-granularity information. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*; Association for Computational Linguistics (ACL): Stroudsburg, PA, USA, 2021; pp. 4804–4813.
131. Ruder, S. An overview of multi-task learning in deep neural networks. *arXiv* **2017**, arXiv:1706.05098.
132. Toshniwal, S.; Tang, H.; Lu, L.; Livescu, K. Multitask learning with low-level auxiliary tasks for encoder-decoder based speech recognition. *arXiv* **2017**, arXiv:1704.01631.
133. Yang, Q. Three challenges in data mining. *Front. Comput. Sci. China* **2010**, *4*, 324–333. [CrossRef]
134. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
135. Liu, C.; Chen, K.; Qi, Z.; Liu, Z.; Zhang, H.; Zou, Z.; Shi, Z. Pixel-level change detection pseudo-label learning for remote sensing change captioning. In Proceedings of the 2024 IEEE International Geoscience and Remote Sensing Symposium, Athens, Greece, 7–12 July 2024; pp. 8405–8408.
136. Lin, M.; Yang, G.; Zhang, H. Transition is a process: Pair-to-video change detection networks for very high resolution remote sensing images. *IEEE Trans. Image Process.* **2022**, *32*, 57–71. [CrossRef]
137. Li, X.; Sun, B.; Li, S. Detection Assisted Change Captioning for Remote Sensing Image. In Proceedings of the 2024 IEEE International Geoscience and Remote Sensing Symposium, Athens, Greece, 7–12 July 2024; pp. 10454–10458.
138. Ni, Z.; Zong, Z.; Ren, P. Incorporating object counts into remote sensing image captioning. *Int. J. Digit. Earth* **2024**, *17*, 2392847. [CrossRef]
139. Zhao, W.; Yang, W.; Chen, D.; Wei, F. DFEN: Dual feature enhancement network for remote sensing image caption. *Electron.* **2023**, *12*, 1547. [CrossRef]
140. Kumar, S.C.; Hemalatha, M.; Narayan, S.B.; Nandhini, P. Region driven remote sensing image captioning. *Procedia Comput. Sci.* **2019**, *165*, 32–40. [CrossRef]
141. Kandala, H.; Saha, S.; Banerjee, B.; Zhu, X.X. Exploring transformer and multilabel classification for remote sensing image captioning. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6514905. [CrossRef]
142. 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]

143. 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]
144. Hoxha, G.; Melgani, F.; Slaghenauffi, J. A new CNN-RNN framework for remote sensing image captioning. In Proceedings of the 2020 Mediterranean and Middle-East Geoscience and Remote Sensing Symposium (M2GARSS), Tunis, Tunisia, 9–11 March 2020; pp. 1–4.
145. Chowdhary, C.L.; Goyal, A.; Vasnani, B.K. Experimental assessment of beam search algorithm for improvement in image caption generation. *J. Appl. Sci. Eng.* **2019**, *22*, 691–698.
146. Wang, B.; Zheng, X.; Qu, B.; Lu, X. Retrieval topic recurrent memory network for remote sensing image captioning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 256–270. [CrossRef]
147. Cui, W.; He, X.; Yao, M.; Wang, Z.; Li, J.; Hao, Y.; Wu, W.; Zhao, H.; Chen, X.; Cui, W. Landslide image captioning method based on semantic gate and bi-temporal LSTM. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 194. [CrossRef]
148. Sumbul, G.; Nayak, S.; Demir, B. SD-RSIC: Summarization-driven deep remote sensing image captioning. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 6922–6934. [CrossRef]
149. Lobry, S.; Marcos, D.; Murray, J.; Tuia, D. RSVQA: Visual question answering for remote sensing data. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 8555–8566. [CrossRef]
150. Murali, N.; Shanthi, A.P. Remote sensing image captioning via multilevel attention-based visual question answering. In *Innovations in Computational Intelligence and Computer Vision: Proceedings of ICICV 2021*; Springer Nature: Singapore, 2022; pp. 465–475.
151. Zhou, K.; Yang, J.; Loy, C.C.; Liu, Z. Learning to prompt for vision-language models. *Int. J. Comput. Vis.* **2022**, *130*, 2337–2348. [CrossRef]
152. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. Gpt-4 technical report. *arXiv* **2023**, arXiv:2303.08774.
153. Liu, H.; Li, C.; Li, Y.; Lee, Y.J. Improved Baselines with Visual Instruction Tuning. *arXiv* **2023**, arXiv:2310.03744.
154. Chen, J.; Zhu, D.; Shen, X.; Li, X.; Liu, Z.; Zhang, P.; Krishnamoorthi, R.; Chandra, V.; Xiong, Y.; Elhoseiny, M. Minigpt-v2: Large language model as a unified interface for vision-language multi-task learning. *arXiv* **2023**, arXiv:2310.09478.
155. He, Y.; Sun, Q. Towards Automatic Satellite Images Captions Generation Using Large Language Models. *arXiv* **2023**, arXiv:2310.11392.
156. Chen, M.; Tworek, J.; Jun, H.; Yuan, Q.; Pinto, H.P.D.O.; Kaplan, J.; Edwards, H.; Burda, Y.; Joseph, N.; Brockman, G.; et al. Evaluating large language models trained on code. *arXiv* **2021**, arXiv:2107.03374.
157. Pettie, S.; Ramachandran, V. An optimal minimum spanning tree algorithm. *JACM* **2002**, *49*, 16–34. [CrossRef]
158. Wei, T.; Yuan, W.; Luo, J.; Zhang, W.; Lu, L. VLCA: Vision-language aligning model with cross-modal attention for bilingual remote sensing image captioning. *J. Syst. Eng. Electron.* **2023**, *34*, 9–18. [CrossRef]
159. Lee, J.S.; Hsiang, J. Patent claim generation by fine-tuning OpenAI GPT-2. *World Pat. Inf.* **2020**, *62*, 101983. [CrossRef]
160. Ricci, R.; Melgani, F.; Junior, J.M.; Gonçalves, W.N. NLP-Based Fusion Approach to Robust Image Captioning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 11809–11822. [CrossRef]
161. Hu, Y.; Yuan, J.; Wen, C.; Lu, X.; Li, X. Rsgpt: A remote sensing vision language model and benchmark. *arXiv* **2023**, arXiv:2307.15266.
162. Dai, W.; Li, J.; Li, D.; Tiong, A.M.H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.N.; Hoi, S. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Adv. Neural Inf. Process. Syst.* **2024**, *36*, 49250–49267.
163. Bazi, Y.; Bashmal, L.; Al Rahhal, M.M.; Ricci, R.; Melgani, F. RS-LLaVA: A Large Vision-Language Model for Joint Captioning and Question Answering in Remote Sensing Imagery. *Remote Sens.* **2024**, *16*, 1477. [CrossRef]
164. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv* **2021**, arXiv:2106.09685.
165. Silva, J.D.; Magalhães, J.; Tuia, D.; Martins, B. Large Language Models for Captioning and Retrieving Remote Sensing Images. *arXiv* **2024**, arXiv:2402.06475.
166. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36*; Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S., Eds.; Neural Information Processing Systems Foundation, Inc.: La Jolla, CA, USA, 2024.
167. Yin, S.; Fu, C.; Zhao, S.; Li, K.; Sun, X.; Xu, T.; Chen, E. A survey on multimodal large language models. *arXiv* **2023**, arXiv:2306.13549.
168. Zhan, Y.; Xiong, Z.; Yuan, Y. Skyeyegpt: Unifying remote sensing vision-language tasks via instruction tuning with large language model. *arXiv* **2024**, arXiv:2401.09712.
169. Kuckreja, K.; Danish, M.S.; Naseer, M.; Das, A.; Khan, S.; Khan, F.S. Geochat: Grounded large vision-language model for remote sensing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 27831–27840.
170. Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; Gao, J. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing 36*; Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S., Eds.; Neural Information Processing Systems Foundation, Inc.: La Jolla, CA, USA, 2024.

171. Wang, D.; Zhang, J.; Du, B.; Xu, M.; Liu, L.; Tao, D.; Zhang, L. Samrs: Scaling-up remote sensing segmentation dataset with segment anything model. In *Advances in Neural Information Processing Systems 36*; Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S., Eds.; Neural Information Processing Systems Foundation, Inc.: La Jolla, CA, USA, 2024.
172. Chiang, W.L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J.E. Vicuna: An Open-source Chatbot Impressing Gpt-4 with 90%* Chatgpt Quality. Available online: https://vicuna.lmsys.org (accessed on 14 April 2023).
173. Zhang, W.; Cai, M.; Zhang, T.; Zhuang, Y.; Mao, X. Earthgpt: A universal multi-modal large language model for multi-sensor image comprehension in remote sensing domain. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5917820.
174. Liu, F.; Chen, D.; Guan, Z.; Zhou, X.; Zhu, J.; Ye, Q.; Fu, L.; Zhou, J. Remoteclip: A vision language foundation model for remote sensing. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5917820. [CrossRef]
175. Yuan, Z.; Zhang, W.; Fu, K.; Li, X.; Deng, C.; Wang, H.; Sun, X. Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval. *arXiv* **2022**, arXiv:2204.09868. [CrossRef]
176. Mittal, P.; Singh, R.; Sharma, A. Deep learning-based object detection in low-altitude UAV datasets: A survey. *Image Vis. Comput.* **2020**, *104*, 104046. [CrossRef]
177. Liu, C.; Chen, K.; Zhang, H.; Qi, Z.; Zou, Z.; Shi, Z. Change-Agent: Toward Interactive Comprehensive Remote Sensing Change Interpretation and Analysis. *IEEE Trans. Geosci. Remote. Sens.* **2024**, *62*, 5635616. [CrossRef]
178. Wang, Y.; Yao, Q.; Kwok, J.T.; Ni, L.M. Generalizing from a few examples: A survey on few-shot learning. *ACM Comput. Surv.* **2020**, *53*, 63. [CrossRef]
179. Chen, X.; Jiang, M.; Zhao, Q. Self-distillation for few-shot image captioning. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Online, 5–9 January 2021; pp. 545–555.
180. Allen-Zhu, Z.; Li, Y. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *arXiv* **2020**, arXiv:2012.09816.
181. Barraco, M.; Stefanini, M.; Cornia, M.; Cascianelli, S.; Baraldi, L.; Cucchiara, R. CaMEL: Mean teacher learning for image captioning. In Proceedings of the 26th International Conference on Pattern Recognition, Montreal, QC, Canada, 21–25 August 2022; pp. 4087–4094.
182. Laina, I.; Rupprecht, C.; Navab, N. Towards unsupervised image captioning with shared multimodal embeddings. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 7414–7424.
183. Rusu, A.A.; Rao, D.; Sygnowski, J.; Vinyals, O.; Pascanu, R.; Osindero, S.; Hadsell, R. Meta-learning with latent embedding optimization. *arXiv* **2018**, arXiv:1807.05960.
184. Zhou, H.; Du, X.; Xia, L.; Li, S. Self-learning for few-shot remote sensing image captioning. *Remote Sens.* **2022**, *14*, 4606. [CrossRef]
185. Zhou, H.; Xia, L.; Du, X.; Li, S. FRIC: A framework for few-shot remote sensing image captioning. *Int. J. Digit. Earth* **2024**, *17*, 2337240. [CrossRef]
186. Plummer, B.A.; Wang, L.; Cervantes, C.M.; Caicedo, J.C.; Hockenmaier, J.; Lazebnik, S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2641–2649.
187. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V*; Springer International Publishing: Berlin, Germany, 2014; pp. 740–755.
188. Liu, C.; Zhao, R.; Chen, H.; Zou, Z.; Shi, Z. Remote sensing image change captioning with dual-branch transformers: A new method and a large scale dataset. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5633520. [CrossRef]
189. Masek, J.G.; Honzak, M.; Goward, S.N.; Liu, P.; Pak, E. Landsat-7 ETM+ as an observatory for land cover: Initial radiometric and geometric comparisons with Landsat-5 Thematic Mapper. *Remote Sens. Environ.* **2001**, *78*, 118–130. [CrossRef]
190. Yang, Y.; Newsam, S. Bag-of-visual-words and spatial extensions for land-use classification. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, CA, USA, 2–5 November 2010; pp. 270–279.
191. Zhang, F.; Du, B.; Zhang, L. Saliency-guided unsupervised feature learning for scene classification. *IEEE Trans. Geosci. Remote Sens.* **2014**, *53*, 2175–2184. [CrossRef]
192. Chen, H.; Shi, Z. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote Sens.* **2020**, *12*, 1662. [CrossRef]
193. 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.
194. 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.
195. Lin, C.Y.; Och, F.J. Looking for a few good metrics: ROUGE and its evaluation. In Proceedings of the 4th NTCIR Workshop, Tokyo, Japan, 2–4 June 2004.
196. Vedantam, R.; Lawrence Zitnick, C.; Parikh, D. Cider: Consensus-based image description evaluation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4566–4575.

197. Anderson, P.; Fernando, B.; Johnson, M.; Gould, S. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part V*; Springer International Publishing: Berlin, Germany, 2016; pp. 382–398.
198. Miller, B.; Linder, F.; Mebane, W.R., Jr. Active learning approaches for labeling text: Review and assessment of the performance of active learning approaches. *Political Anal.* **2020**, *28*, 532–551. [CrossRef]
199. De Silva, V.; Sumanathilaka, T.G.D.K. A Survey on Image Captioning Using Object Detection and NLP. In Proceedings of the 4th International Conference on Advanced Research in Computing, Belihuloya, Sri Lanka, 21–24 February 2024; pp. 270–275.
200. Zhao, F.; Yu, Z.; Zhao, H.; Wang, T.; Bai, T. Integrating grid features and geometric coordinates for enhanced image captioning. *Appl. Intell.* **2024**, *54*, 231–245. [CrossRef]
201. Guo, L.; Liu, J.; Zhu, X.; Yao, P.; Lu, S.; Lu, H. Normalized and geometry-aware self-attention network for image captioning. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10327–10336.
202. Yao, Y.; Duan, J.; Xu, K.; Cai, Y.; Sun, Z.; Zhang, Y. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confid. Comput.* **2024**, *4*, 100211. [CrossRef]
203. Hong, D.; Zhang, B.; Li, X.; Li, Y.; Li, C.; Yao, J.; Yokoya, N.; Li, H.; Ghamisi, P.; Jia, X.; et al. SpectralGPT: Spectral remote sensing foundation model. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 5227–5244. [CrossRef]
204. Chang, S.; Ghamisi, P. Changes to captions: An attentive network for remote sensing change captioning. *IEEE Trans. Image Processing* **2023**, *32*, 6047–6060. [CrossRef]
205. Liu, C.; Zhao, R.; Chen, J.; Qi, Z.; Zou, Z.; Shi, Z. A decoupling paradigm with prompt learning for remote sensing image change captioning. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5622018. [CrossRef]
206. Peng, W.; Jian, P.; Mao, Z.; Zhao, Y. Change Captioning for Satellite Images Time Series. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 6006905. [CrossRef]
207. Sun, Y.; Lei, L.; Guan, D.; Kuang, G.; Li, Z.; Liu, L. Locality Preservation for Unsupervised Multimodal Change Detection in Remote Sensing Imagery. In *IEEE Transactions on Neural Networks and Learning Systems*; IEEE: New York, NY, USA, 2024; pp. 1–15.
208. Zheng, X.; Guan, D.; Li, B.; Chen, Z.; Li, X. Change smoothness-based signal decomposition method for multimodal change detection. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 2507605. [CrossRef]
209. Cheng, Q.; Xu, Y.; Huang, Z. VCC-DiffNet: Visual Conditional Control Diffusion Network for Remote Sensing Image Captioning. *Remote Sens.* **2024**, *16*, 2961. [CrossRef]
210. Yang, C.; Li, Z.; Zhang, L. Bootstrapping interactive image-text alignment for remote sensing image captioning. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5607512. [CrossRef]
211. 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*, 5624711. [CrossRef]
212. Asadi, A.; Baise, L.G.; Koch, M.; Moaveni, B.; Chatterjee, S.; Aimaiti, Y. Pixel-based classification method for earthquake-induced landslide mapping using remotely sensed imagery, geospatial data and temporal change information. *Nat. Hazards* **2024**, *120*, 5163–5200. [CrossRef]
213. Amitrano, D.; Di Martino, G.; Di Simone, A.; Imperatore, P. Flood Detection with SAR: A Review of Techniques and Datasets. *Remote Sens.* **2024**, *16*, 656. [CrossRef]
214. Wang, B.; Yao, Y. Mountain Vegetation Classification Method Based on Multi-Channel Semantic Segmentation Model. *Remote Sens.* **2024**, *16*, 256. [CrossRef]
215. Jia, P.; Chen, C.; Zhang, D.; Sang, Y.; Zhang, L. Semantic segmentation of deep learning remote sensing images based on band combination principle: Application in urban planning and land use. *Comput. Commun.* **2024**, *217*, 97–106. [CrossRef]
216. Hess, G.; Tonderski, A.; Petersson, C.; Åström, K.; Svensson, L. Lidarclip or: How I learned to talk to point clouds. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2024; pp. 7438–7447.
217. Sun, Y.; Lei, L.; Li, Z.; Kuang, G. Similarity and dissimilarity relationships based graphs for multimodal change detection. *ISPRS J. Photogramm. Remote Sens.* **2024**, *208*, 70–88. [CrossRef]
218. Gao, J.; Li, P.; Chen, Z.; Zhang, J. A survey on deep learning for multimodal data fusion. *Neural Comput.* **2020**, *32*, 829–864. [CrossRef]

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.