remote sensing

MDPI

Article

Multi-Scale Semantic Segmentation and Spatial Relationship Recognition of Remote Sensing Images Based on an Attention Model

Wei Cui *, Fei Wang, Xin He, Dongyou Zhang, Xuxiang Xu, Meng Yao, Ziwei Wang and Jiejun Huang

School of Resources and Environmental Engineering, Wuhan University of Technology, Wuhan 430070, China; flyking@whut.edu.cn (F.W.); 2962575697@whut.edu.cn (X.H.); gis@whut.edu.cn (D.Z.); xxx1049731704209@whut.edu.cn (X.X.); yaomeng@whut.edu.cn (M.Y.); zwei@whut.edu.cn (Z.W.); hjj@whut.edu.cn (J.H.)

* Correspondence: cuiwei@whut.edu.cn; Tel.: +86-136-2860-8563

Received: 14 April 2019; Accepted: 24 April 2019; Published: 2 May 2019

Abstract: A comprehensive interpretation of remote sensing images involves not only remote sensing object recognition but also the recognition of spatial relations between objects. Especially in the case of different objects with the same spectrum, the spatial relationship can help interpret remote sensing objects more accurately. Compared with traditional remote sensing object recognition methods, deep learning has the advantages of high accuracy and strong generalizability regarding scene classification and semantic segmentation. However, it is difficult to simultaneously recognize remote sensing objects and their spatial relationship from end-to-end only relying on present deep learning networks. To address this problem, we propose a multi-scale remote sensing image interpretation network, called the MSRIN. The architecture of the MSRIN is a parallel deep neural network based on a fully convolutional network (FCN), a U-Net, and a long short-term memory network (LSTM). The MSRIN recognizes remote sensing objects and their spatial relationship through three processes. First, the MSRIN defines a multi-scale remote sensing image caption strategy and simultaneously segments the same image using the FCN and U-Net on different spatial scales so that a two-scale hierarchy is formed. The output of the FCN and U-Net are masked to obtain the location and boundaries of remote sensing objects. Second, using an attention-based LSTM, the remote sensing image captions include the remote sensing objects (nouns) and their spatial relationships described with natural language. Finally, we designed a remote sensing object recognition and correction mechanism to build the relationship between nouns in captions and object mask graphs using an attention weight matrix to transfer the spatial relationship from captions to objects mask graphs. In other words, the MSRIN simultaneously realizes the semantic segmentation of the remote sensing objects and their spatial relationship identification end-to-end. Experimental results demonstrated that the matching rate between samples and the mask graph increased by 67.37 percentage points, and the matching rate between nouns and the mask graph increased by 41.78 percentage points compared to before correction. The proposed MSRIN has achieved remarkable results.

Keywords: multi-scale; semantic segmentation; image caption; remote sensing; LSTM; U-Net; upscaling; downscaling

1. Introduction

Deep neural networks [1,2] are gradually being applied to high-resolution remote sensing image analysis [3], especially in scene classification [4–9], semantic segmentation [10], or single-scale remote sensing object recognition [11,12], and they all have achieved good results. Unfortunately, most of

the existing studies do not address the interpretation of spatial relationships between remote sensing objects, which limits the understanding of remote sensing objects, especially when the phenomenon of different objects with the same spectrum in remote sensing appears.

The phenomenon of different objects with the same spectrum in remote sensing is quite common. It is difficult to identify objects only by their own textures, spectra, and shape information. Object identification requires multi-scale semantic information and spatially adjacent objects to assist in decision-making. The spatial relationship between remote sensing objects is of great significance to the recognition of remote sensing objects when different objects have the same spectrum, for example, many different types of buildings with similar shapes and spectral features, such as commercial buildings and workshops. The traditional object recognition methods [13–15] can only identify the object by its spectral, texture, and shape features without considering its adjacent objects. Therefore, it is impossible to accurately distinguish the different objects with the same spectrum without additional information. However, commercial buildings are often adjacent to green spaces and squares, and workshops are more adjacent to other factories and warehouses. In this way, it is possible to effectively identify commercial buildings and workshops through adjacent object categories.

According to existing research, scene classification describes the entire patch of the sample but does not involve remote sensing objects. Although semantic segmentation can identify the location and boundaries of remote sensing objects, it does not include the interpretation of complex spatial relationships between remote sensing objects, which leads to a certain degree of incomplete semantic understanding of remote sensing images. How to carry out a comprehensive semantic description of remote sensing objects and their spatial relationships is an issue that still needs further study.

The prosperity of image captions based on recurrent neural networks (RNNs) [16], especially the attention-based LSTM [17], can provide not only image description but also the attention location corresponding to the currently generated word at different time steps, which provides a new way to address the problems above. Chen et al. proposed a novel group-based image captioning scheme (termed GroupCap) [18], which jointly models the structured relevance and diversity among group images towards an optimal collaborative captioning. Previous works only used the global or local image feature. A model with 3-Gated model [19] was proposed to fuse the global and local image features together for the task of image captioning. In recent years, more studies have focused on the relationship between generated words and corresponding regions in the image. An attribute-driven attention model [20] was proposed to focus on training a good attribute-inference model via the RNN for image captioning. The uniqueness of the model lied in the usage of an RNN with the visual attention mechanism to observe the images before generating captions. Khademi et al. presented a novel context-aware, attention-based deep architecture [21] that employed a bidirectional grid LSTM for image captioning. The bidirectional grid LSTM took visual features of an image as the input and learned complex spatial patterns based on two-dimensional context.

In recent years, the application of reinforcement learning [20–23] in image caption has also been a hot topic, which adjusts the generation strategies using the change of the reward functions in the caption generation process to dynamic vocabulary generation.

However, most of the current studies focus on the scene semantic description of ordinary digital images [24,25]. To use the deep RNN or LSTM to execute the semantic analysis [26–30] of remote sensing objects, the following problems must be solved:

Location ambiguity: At different time steps, the attention mechanism is based on 14×14-sized image features and corresponds to 196 spatial locations in remote sensing images. There are some deviations [31], however, that limit the application in remote sensing object recognition.

Boundary ambiguity: the nouns (label of objects) in captions cannot accurately segment the boundaries of remote sensing objects in an image; thus, it is impossible to identify the spatial relationship between the objects.

Spatial scale ambiguity: Everything is related to everything else, but near things are more related to each other [32]. The surroundings of objects are various, which makes it difficult to detect remote

sensing objects using a uniform scale model. Sometimes we need a large scale to contain the neighboring and context information to identify remote sensing objects accurately.

To solve the above problems, we present the MSRIN, which is based on an FCN, a U-net, and an attention-based LSTM. The MSRIN can generate remote sensing image descriptions at multi-spatial scales, segment objects in images, and recognize their spatial relationships from end-to-end. First, a remote sensing image is semantically segmented through an FCN and a U-net on two spatial scales such that each pixel in the original image is labelled with two semantic labels; therefore, a hierarchical relationship of a multi-scale remote sensing object can be formed. Second, the features of the same image obtained using a pre-trained Visual Geometry Group 19 (VGG-19) network are input for the attention-based LSTM, which outputs the captions that describe the two-scale remote sensing objects and their spatial relationships. Finally, the relationship between the nouns in the caption and the object mask graphs is established through the attention weight matrix. In this way, the remote sensing objects from the U-Net get their spatial relationship from captions. To overcome the spatial deviations of the attention weight matrix from the LSTM, the MSRIN designs an attention-based multi-scale remote sensing object identification and correction mechanism. Our method produces a complete semantic interpretation of remote sensing images.

In summary, the main contributions of this paper are as follows:

1. A multi-scale semantic caption strategy is proposed. Based on this strategy, a parallel network (the MSRIN) is designed to completely interpret the semantic information of remote sensing images.
2. We discuss the remote sensing object recognition and correction mechanism based on the attention weight matrix and multi-scale semantic segmentation using the FCN and the U-Net, simultaneously realizing the instance segmentation of the remote sensing images and the spatial relationship identification from end-to-end.

The remainder of this paper is organized as follows: Section 2 discusses related work. Multi-scale semantic segmentation and spatial relationship recognition of remote sensing images based on an attention model is presented in Section 3. Experiments and analysis are listed in Section 4. Discussion is presented in Section 5. Finally, the conclusion is presented in Section 6.

2. Related Work

2.1. Scene Classification

Because of the GPU memory limitations, high-resolution imagery must be segmented into patches for Convolutional Neural Networks (CNN) models, and the label is always attached to a remote sensing image sample [8,10,33–35] in scene classifications. To manage and retrieve patches easily, the previous research studies modified the structure of traditional deep convolution neural networks into two different forms, i.e., cascade and parallel models, according to the characteristics of remote sensing images [36]. In the cascade model, the corresponding layer structure in a traditional CNN is transformed to reduce the total parameters. For example, a global average pooling layer is used to replace the fully connected network as the classifier [10] or to insert a region-based cascade pooling (RBCP) method between the last normal down-sampling layer and the classifier to aggregate convolutional features from both the pre-trained and the fine-tuned convolutional neural networks [36]. Parallel models try to extract more abundant features for scene classification by designing parallel network structures [35]. All of the above methods achieved satisfactory results. However, the interpretation of remote sensing images more than meets the needs of spatial operations of geographical objects; therefore, it is not enough to obtain only the labels of remote sensing image patches.

2.2. Semantic Segmentation

Semantic segmentation algorithms assign a label to every pixel in an image and are the basis of instance segmentation. The research includes two cases: CNN series and RNN series.

2.2.1. CNN Series

A CNN is used not only to label the samples but also to classify the pixels to achieve semantic segmentation [37]. Meanwhile, semantic segmentation has been applied to the remote sensing recognition of buildings and other objects [38,39]. High-resolution imagery must be segmented into patches for CNNs due to Graphics Processing Unit (GPU) memory limitations, thus in a limited area, to make full use of the output features of different convolution layers to achieve a better semantic segmentation effect, the researchers often use a multi-depth network model [40] or design a multiple-feature reuse network in which each layer is connected to all the subsequent layers of the same size, enabling the direct use of the hierarchical features in each layer [41]. Emerging new networks, such as U-Net [42] and DenseNet [43], have also been applied in remote sensing image semantic segmentation [44]. The application scenario also extends from surface geographic objects to continuous phenomena such as highly dynamic clouds [45]. Some studies introduce the attention mechanism [46,47] to achieve an ideal segmentation effect by suppressing low-level features and noise through high-level features.

In general, semantic segmentation of remote sensing images based on a CNN has been developed from a simple transplanting network structure to the design of a creative network structure [48] according to the characteristics of remote sensing and has achieved good results. Application scopes are expanded from building extracting [49], built-up area extracting [48], and mapping impervious surfaces [50] to oil palm tree detection [51].

2.2.2. RNN Series

RNNs [52] are an important branch of the deep learning family, which are widely used for sequence analysis. In hyperspectral remote sensing images, there are tens of hundreds of spectral bands, which can be regarded as a related spectral sequence. Therefore, an RNN is proposed for hyperspectral image classification and achieves excellent classification performance [53,54]. With further research, the RNN model, which takes both spatial and spectral features into account [55], has also been applied to hyperspectral remote sensing semantic segmentation. In this way, the comprehensive utilization of spectral-spatial information is realized and good results are achieved. Another method is to input bands information into an LSTM network as an l-length sequence [56], which also achieves good results. The ability of an RNN to process time series data can also be used to process synthetic aperture radar (SAR) images [57]. The fine structure of SAR images can be retained as much as possible by filtering noise from multi-temporal radar images.

Most of researchers regard the spectral features of each individual pixel as one sequential feature for the RNN input layer. Recently, a novel strategy for constructing sequential features is proposed [58], and similar pixels collected from the entire image are used to construct the respective sequential features and the strategy achieves significant improvements in classification.

2.3. Remote Sensing Image Captioning

Remote sensing image captioning aims to generate comprehensive captions that summarize the image content at a semantic level [26]. Relevant research originated from natural language descriptions of images [16,24] in the field of computing. More recently, attention-based LSTMs [17] have emerged to describe not only the semantic information of images but also the image region corresponding to the words generated at the current moment through a 14×14 weight matrix. Because the image captions and image features are input into the RNN at the same time, there is a debate about whether it focuses on captions or on images at each time step (i.e., at each time step the model decides whether to attend to the image or to the visual sentinel) [25]. Although image captioning of ordinary digital images has achieved good results, it has encountered many difficulties in remote sensing fields. Compared with ordinary digital images, the remote sensing images from satellites or aircraft have a unique "view of God," which makes the remote sensing images have no directional distinction and lack a focused object

or centre. All of these factors increase the difficulty of obtaining natural language descriptions of remote sensing images. Despite these difficulties, some researchers have made useful advances. Qu et al. [26] used an RNN to describe remote sensing images using a natural language. Shi et al. [27] proposed a remote sensing image captioning framework that leverages the techniques of a convolutional neural network (CNN). Both methods used a CNN to represent the image and to generate the corresponding captions from recurrent neural networks or pre-defined templates. To better describe the remote sensing images, and after a comprehensive analysis of scale ambiguity, category ambiguity and rotation ambiguity, a large-scale benchmark dataset of remote sensing images is presented to advance the task of remote sensing image captioning [28]. Wang et al. [29] used semantic embedding to measure the image representation and the caption representation. The captioning performance is based on CNNs, and the authors regarded caption generation task as a latent semantic embedding task, which can be solved via matrix learning. Zang et al. [30] presented a new model with an attribute attention mechanism for the description generation of remote sensing images by introducing the attributes from the fully connected layer of CNN, where the attention mechanism perceives the whole image while knowing the correspondence between regions and words, and the proposed framework achieves robust performance. Then, various image representations and caption generation methods were tested and evaluated. This work made a great step forward in the research on remote sensing image captions.

Although research on remote sensing image captioning has recently made some achievements, there are still some limitations, such as words in image captions that cannot correspond to remote sensing objects one by one, and relatively weak descriptions of the spatial relationships. In particular, the image region corresponding to the attention weight matrix often does not match the remote sensing object corresponding to the word at the same time step [31]. To better understand the semantic information of remote sensing images, further research is still needed.

3. Methodology

The MSRIN was defined first as a multi-scale remote sensing image caption strategy. A parallel network structure was designed to identify multi-scale remote sensing objects and spatial relationships based on attention.

3.1. Strategy of Multi-Scale Caption Design

According to Tobler's first law of geography, everything is related to everything else, but near things are more related to each other [32]. We propose a strategy for multi-scale captioning:

1. Each caption consists of small-scale remote sensing objects and their spatial relationships, which implicitly constitute a large-scale object, as shown in Figure 1 and Table 1.
2. Usually, one object is selected as the main object in a small-scale image caption, while other objects are subordinate to it through spatial relationships. In this way, each class of small-scale objects will not repeat within one large-scale object.
3. If there are two or more large-scale objects, the corresponding number of captions are joined by the word "with."

Table 1 shows the multi-scale classification system in our experiment. Large-scale and small-scale categories are encoded in 1X and 2x respectively, where 1 and 2 represent large-scale and small-scale information respectively, X represents a large-scale category number, and x represents a small-scale category number. In our system, there are 9 large-scale categories and 10 small-scale categories. Each large-scale category has a one-to-many relationship with the small-scale categories.

Following the caption strategy, the multi-scale image caption output from the LSTM looks like this:

$$\text{noun}_1\ \text{R}_{12}\ \text{noun}_2, \ldots, \text{with noun}_i\ \text{R}_{ij}\ \text{noun}_j, \ldots, \text{noun}_n$$

In the example above, two large-scale objects, which are composed of small-scale objects and their spatial relationship, are on both sides of "with." Thus, the caption contains two scales of spatial semantic information. The noun$_i$ describes a remote sensing object O_i, and R_{ij} describes the spatial relationship between the remote sensing objects O_i and O_j (e.g., "road cross residence with road next to service"). Noun$_i$ and noun$_j$ are always different in one clause.

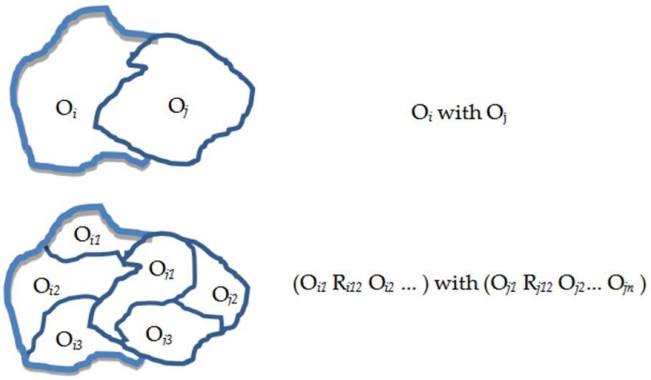

Figure 1. Multi-spatial scale semantic segmentation and image caption. It shows a two-scale hierarchy of one image. An image contains many large-scale objects, each large-scale object contains many small-scale objects, and there are spatial relationships between objects of the same scale. Our strategy of captioning is to describe both information of scale and spatial relationship contained in an image as completely as possible.

Table 1. Classification of multiscale remote sensing objects.

Value	10	11	12	13	14
large-scale	residence region	industry region	service region	village region	forest region
value	15	16	17	18	
large-scale	uncompleted region	road region	other region	Green-space region	
value	20	21	22	23	24
small-scale	residence	industry	service	village	forest
value	25	26	27	28	29
small-scale	uncompleted	road	other	Green-space	waterbody

Our strategy is still valid when generating captions for more complex scenes, as shown in Figure 2. We use small-scale objects and spatial relationships between them to describe the large-scale objects, and then connect each clause with "with." When there are more large-scale objects in an image, we use this method to iterate to form a complete caption containing multi-scale semantic information.

(green_space next_to service) with (road cross waterbody) with (service next_to uncompleted and road) with (road next_to green_space and uncompleted and service)

Figure 2. Sample with complex scenes. It shows an image with more complex scenes that contain four large-scale objects. (**a**) is the input image; (**b**) is the large-scale segmentation map of (**a**); (**c**–**f**) are the small-scale objects contained in each large-scale object. (**c**) corresponds to the clause "green_space next_to service," (**d**) to the clause "road cross waterbody," (**e**) to the clause "service next_to uncompleted and road," and (**f**) to the clause "road next_to green_space and uncompleted and service.".

The advantages of a multi-scale semantic caption strategy are as follows:

Hierarchically describing the spatial relationship between objects according to scale effect can simplify the type of spatial relationship (including "next_to", "near", "cross", "surround" and "surround_by"). Due to the scope limitation of sample patch, only the spatial neighbourhood relationship between large-scale objects is considered and described using "with" such that the network training is facilitated.

3.2. Multi-Scale Network Structure

Corresponding to the multi-scale semantics caption strategy, the MSRIN consists of three different deep neural networks: an FCN and a U-Net for multi-scale semantic segmentation, and an attention-based LSTM for image caption, both the FCN and the U-Net are used to semantically segment the same remote sensing image at two different spatial scales. The output of FCN and U-Net are masked to obtain the location and boundaries of remote sensing objects.

Meanwhile, the same sample is input into one attention-based LSTM network, with the output captions following the principles of multi-scale remote sensing captioning. To match noun_t with the remote sensing object via attention and to overcome the location deviation of attention, we designed a multi-scale remote sensing object recognition and correction mechanism. The structure of the MSRIN is shown in Figure 3.

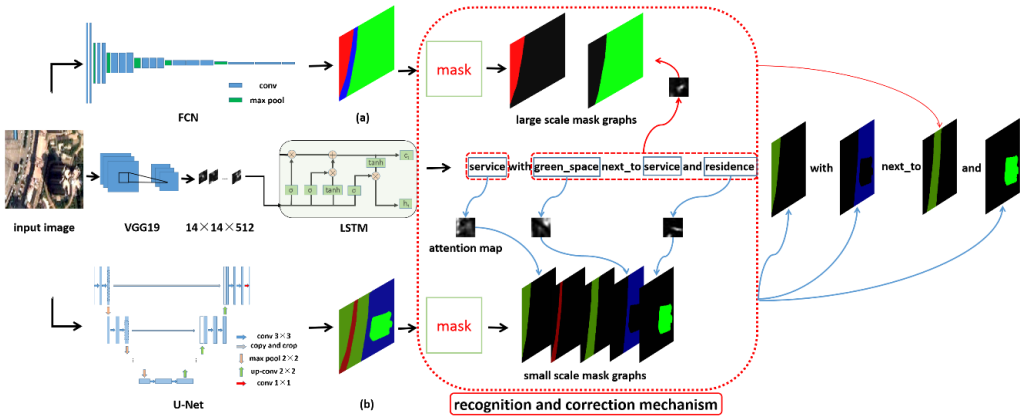

Figure 3. Network structure. It shows the overall network structure of the MSRIN. In our network, one remote sensing image is input into three branch networks. (**a**) is the large-scale segmentation map of the FCN output, (**b**) is the small-scale segmentation map of the U-Net output, and they are masked to obtain the location and boundaries of remote sensing objects. The LSTM outputs image captions and attention areas. The process of identification and correction is given in Section 3.3. The multi-scale objects recognition and correction mechanism attaches the object (the mask graphs from U-Net) to nount through the weight matrix at time step t.

The FCN [37] can achieve pixel-to-pixel classification by using full convolution, up-sampling, and jump structure. The U-Net [42] follows the idea of FCN for image semantic segmentation and combines the features of coding–decoding structures and jumping networks. From the encoder to the decoder, there is usually a direct information connection to help the decoder recover the target details better. Considering the features of small-scale objects are more complex and large-scale objects are more abstract, we used a U-Net to segment small-scale objects and a FCN to segment large-scale objects, and the segmentation effect is shown in Figure 4.

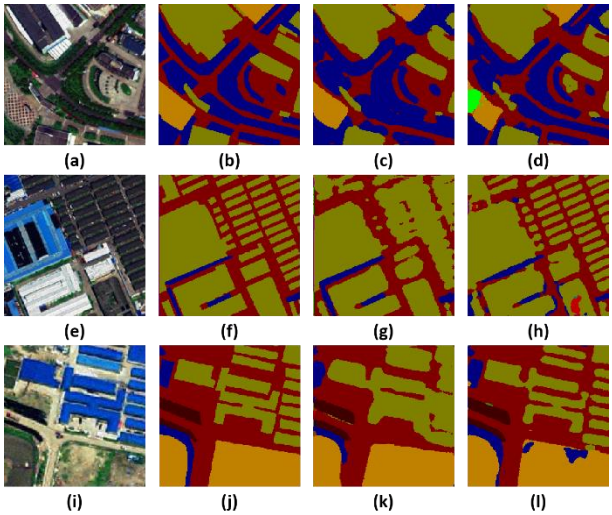

Figure 4. Segmentation effect of FCN and U-Net. It shows three examples of FCN and U-Net segmentation effect comparison. (**a**,**e**,**i**) are the input images. (**b**,**f**,**j**) are corresponding ground truth of the images. (**c**,**g**,**k**) are segmentation maps of FCN. (**d**,**h**,**l**) are segmentation maps of U-Net. It was found that FCN performed better at segmenting large objects, while smaller objects were easier to aggregate into blocks, so FCN was more suitable for large-scale segmentation. U-Net worked well when smaller objects were segmented but tended to misclassify some small fragments of the large-scale objects when it was being segmented, so U-Net was more suitable for small-scale segmentation.

The core of the networks is the multi-scale objects recognition and correction mechanism, which attaches the object (the mask graphs from U-Net) to noun_t through the weight matrix at time step t. In this way, the remote sensing objects get their spatial relationship from captions. In other words, the MSRIN will output not only a series of the remote sensing objects (the mask graphs) but also the spatial relationships between them from image captions.

Unfortunately, the attention weights were computed from a 14×14 size feature map. Thus, the spatial location accuracy was relatively low, leading to a mismatch between noun_t in the caption and objects in the image at some time step t, as shown in Figure 5. To solve this problem, our paper proposes a multi-scale remote sensing object recognition and correction mechanism.

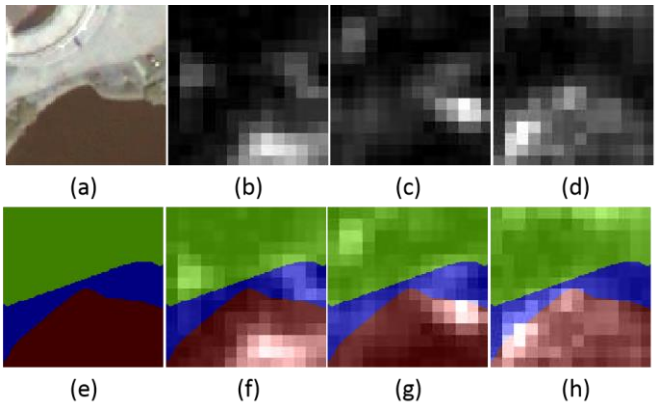

Figure 5. Attention weight matrix error. It shows mismatches between nouns in the caption and objects in the image. (**a**) is the input image; (**b**–**d**) are the attention maps for generating "green_space," "service," and "waterbody," respectively; (**e**) is a small-scale segmentation map of (**a**); (**f**–**h**) are overlaid maps of (**b**–**d**), respectively, with (**e**). As shown in the figure, the attention area of the first generated noun "green_space" corresponds to the object "waterbody" in the image, which resulted in mismatch. Of the three nouns contained in the generated image caption, only the third noun matched the right object.

3.3. Remote Sensing Objects Recognition and Correction Mechanism

Attention-based LSTM provides a 14×14 weight matrix at different time steps, which is the basis for implementing the remote sensing object recognition and correction.

3.3.1. Remote Sensing Object Recognition

The remote sensing object recognition is based on the attention weight matrix and U-Net mask graphs. First, we resample the 14×14 attention weight matrix to a 210×210 size. Then, we denote the attention map (weight matrix) at location $(i, j) \in L \times L (L = 210)$ at time step t as a_{ij}^t, and the U-Net mask graphs at location $(i, j) \in L \times L$ as m_{ij}. In the mask graphs, the area where the object is located has a pixel value of the class label C and the rest is 0:

$$m = \begin{cases} C, (i,j) \in object \\ 0, (i,j) \notin object \end{cases} \tag{1}$$

The values of intersect areas can be computed using:

$$v_{ij} = \frac{1}{C} \alpha_{ij}^t . m_{ij} \tag{2}$$

where C is the normalization factor such that v_{ij} sums to 1. The mean value of the intersect areas (weight mean value) can be computed using:

$$v_{mean} = \frac{1}{n} \sum_{ij} v_{ij} \tag{3}$$

where n is the total number of pixels of the remote sensing object. Then, the mask graph with the largest mean value will be selected. If the class label of the selected mask graph (object) is consistent with the noun_t in the caption at current time step t, it means the mask graph represents the noun_t of current time t, and the location and boundary of the remote sensing object will be identified using the selected mask graph.

However, at time step t, the label of the selected mask graph often does not match noun_t in the captions, as shown in Figure 5. To solve this problem, we propose a multi-scale remote sensing object correction algorithm.

3.3.2. Remote Sensing Correction Algorithm

If the mismatch happens, the correction algorithm needs to upscale and search for the large-scale object, which the current weight matrix pays attention to first. The detailed method is shown below:

The MSRIN first scales up to the large-scale objects region that are large-scale mask graphs output from the FCN, then calculates the weights mean value in each large-scale object and takes the maximum one as a candidate object. In the candidate object, the MSRIN downscales to small-scale mask graphs and selects the remote sensing object whose class label corresponds to noun_t using a one-to-one relationship, thus completing the correction.

The key to the above process is that the strategy of multi-scale captions made of small-scale objects of each class will not be repeated within each large-scale object such that in the large-scale object region, the small-scale object that matches to noun_t can be selected. The Algorithm 1 is shown as follows:

Algorithm 1. For Multi-Scale Remote Sensing Objects Recognition and Correction

Input: noun, weight matrix at time step t.
Small scale object set $o = \{o_i\}$, $i \in 1, n]$, n is the number of mask graphs from U-Net in one sample patch,
Large scale object set $O = \{O_j\}$, $j \in [1, m]$, m is the number of mask graphs from FCN in the same sample patch with the U-Net.
weight_graph is a visual graph of the weight matrix generated at the current moment.
Output: $\mathbf{o_{selected}}$.

For i = 1 to n; step = 1; do //search the small-scale object that the current weights graph pay attention to
{ weight_graph intersect with o_i; //determine the area of attention on a small-scale object
Calculate mean value of intersect area; //basis for selecting small-scale candidate
Update o_i to small_candidate when weights mean value is the current maximum mean; //update candidate based on the mean value
}
If the class label of the small_candidate is equal to $noun_t$; //the generated noun matches the object
Then $o_{selected}$ = small_candidate; //the object was recognized.
Else //there is a mismatch between $noun_t$ and the candidate, so a correction process will start
{//upscale, search the candidate large-scale object that the current weights graph pay attention to
For j = 1 to m; step = 1; do //search the large-scale object that the current weights graph pay attention to
{weight_graph intersect with O_j; //determine the area of attention on a large-scale object
Calculate mean value of intersect area; //basis for selecting large-scale candidate
Update O_j to large_candidate when weights mean value is the current maximum mean; //update candidate based on the mean value
}
Downscaling in large_candidate; //downscale, determine small-scale object based on the large-scale candidate
search the small-scale object oi which class label is corresponded to $noun_t$ in large_candidate; //the target small scale object in the large-scale candidate
$o_{selected} = o_i$; //thus the object was recognized and corrected.
}

3.3.3. Case Analysis

The following example analyzes the process of multi-scale remote sensing object recognition and correction, as shown in Figures 6–8. The generated caption is "service with green_space next_to service and surround residence." The reference caption is "service with green_space next_to service and surround residence." The mean value of remote sensing objects at each time is shown in Table 2.

Table 2. Mean value of remote sensing objects.

	The Mean Value of Weight at Every Object When Generating the First "Service"	The Mean Value of Weight at Every Object When Generating the Second "Service"
service_0	0.000061702 (correct)	0.000016412
service_1	0.000015618	0.000019851
road_0 (with)	0.000029001	0.000027496
green_space _0	0.000015094	0.000023904
residence_0	0.000021753	0.000046018 (incorrect)
service_region	0.000061702	0.000016288
residence_region	0.0000164134	0.0000265598

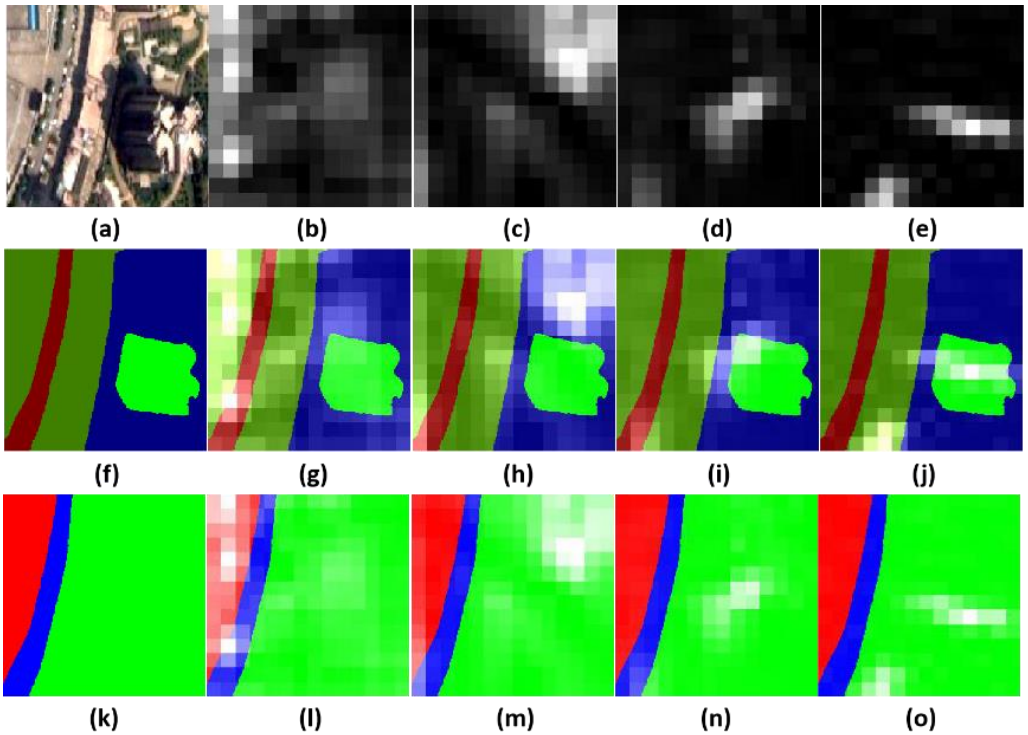

Figure 6. Remote sensing object recognition and correction. It shows the process of multi-scale remote sensing object recognition. (**a**) is the input image; (**b**–**e**) are the attention maps for generating "service," "green_space," "service," and "residence," respectively; (**f**) is a small-scale segmentation map of (**a**); (**g**–**j**) are overlaid maps of (**b**–**e**), respectively, with (**f**); (**k**) is a large-scale segmentation map of (**a**); (**l**–**o**) are overlaid maps of (**b**–**e**), respectively, with (**k**). As shown in (**I**,**n**), when generating the second "service," the spatial location of attention weights is incorrect at the small scale, but it is correct at the large scale.

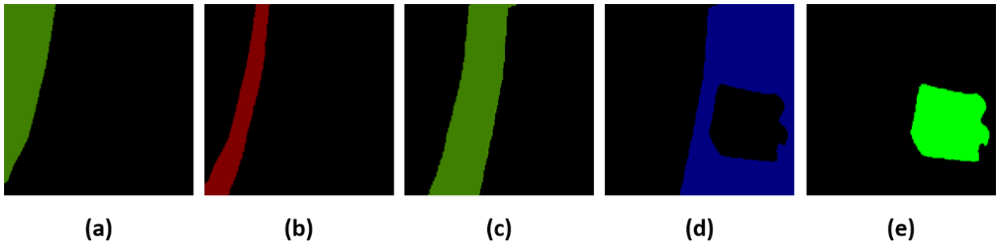

Figure 7. Small-scale objects. It shows the small-scale objects of the image. (**a**) is service_0 (in order to distinguish between different objects of the same class, we number each object); (**b**) is road_0; (**c**) is service_1; (**d**) is green_space _0; and (**e**) is residence_0.

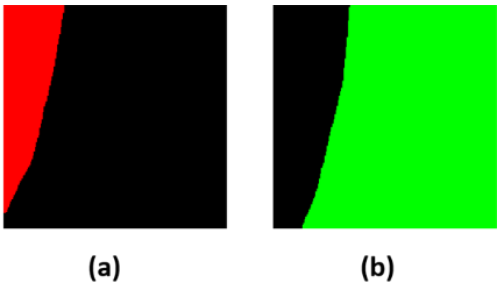

Figure 8. Large-scale objects. It shows the large-scale objects of the image. We divided the image into two large-scale objects by the road. (**a**) is service_region, which contains small-scale object service_0; (**b**) is residence_region, which contains small-scale object service_1, green_space_0, and residence_0.

Table 2. The mean value of weight at every object when two "service" words are generated. The object of concern on the small scale was correct when generating the first "service" and it was incorrect when generating the second "service".

The caption "service with green space next to next_to service and surround residence" was divided into two parts using "with" (representing a road in the image). "Service" can describe both remote sensing objects "service_0" and "service_1." The optimal case is that the attention object of the first generated "service" corresponds to the object "service_0" and the second corresponds to "service_1." The sub-optimal case is that the two attention regions of the "service" are both aimed at "service_1." By comparing the mean value of small-scale objects, it was found that the object of concern was correct when the first generated "service" aimed at object "service_0" and it was incorrect when the second "service" aimed at "residence_0." Thus, the correction algorithm upscaled, comparing the mean value of large-scale objects; the attention region of the second generated "service" was "residence_region," and it contained the object "service_1." It is noted that the attention was correct in the case of the large scale. At this time, the original incorrect object of concern was corrected, and the optimal situation was achieved.

4. Data and Experiments

4.1. Introduction of Experiment Area and Sample

To better integrate professional and research ideas, we selected 1835 patches with the longitudes ranging from $114^{\circ}23'50''E$ to $114^{\circ}25'7''E$, the latitudes ranging from $30^{\circ}27'50''N$ to $30^{\circ}30'37''N$, and a total area of 9.06 km^2 of remote sensing images in Guanggu in 2009. To make the number of samples in the verification set and training set sufficient and the results reasonable, we allocated 1167 samples to the training set and 668 samples to the verification set. For each sample image, we gave three captions that were as different as possible.

4.2. Network Parameters and Experiment

The basic functions of the MSRIN include image segmentation and image caption. The function of semantic segmentation of the original images is obtained based on a pre-trained FCN and a pre-trained U-Net.

In general, when fine-tuning network parameters, in order to reduce the learning cost, we first adjusted the number of iterations of the network. For example, in FCN, we first set a larger number of iterations and observed the loss function change during the iterations, where the trend is shown in Figure 9. Then, according to this, we selected an appropriate number of iterations required for the network to reach stability and kept it unchanged during the subsequent tuning process. When adjusting other parameters, we followed the principle of single factor experiments: fine-tune a certain parameter while keeping other parameters unchanged until the parameter is optimal, and then adjust other parameters one by one. For example, when adjusting the batch size of LSTM, we set the initial value to 25 according to experience and gradually adjusted the value according to the trend of Bleu_1, where the change process of Bleu_1 is shown in Figure 10. Finally, we selected the batch size when Bleu_1 was the highest.

After adjusting the sub-networks of the MSRIN one by one, we determined the basic parameters of each network. In FCN, we set the learning rate to 1×10^{-5}, the batch size to 1, and the number of iterations to 60,000. In U-Net, we set the learning rate to 1×10^{-4}, the batch size to 20, and the number of iterations to 120. The function of image caption was obtained based on an attention-based LSTM. The original image was input into a pre-trained VGG-19 and the features of conv5_3 were extracted. The size of the feature map was $14 \times 14 \times 512$, which was used as a part of the input of the LSTM. In the LSTM, we set the hidden layers to 1024, the embedding dimension of the word vector to 512, the learning rate to 0.001, the batch size to 20, the number of iterations to 120, and we used the softmax function as the nonlinear activation function.

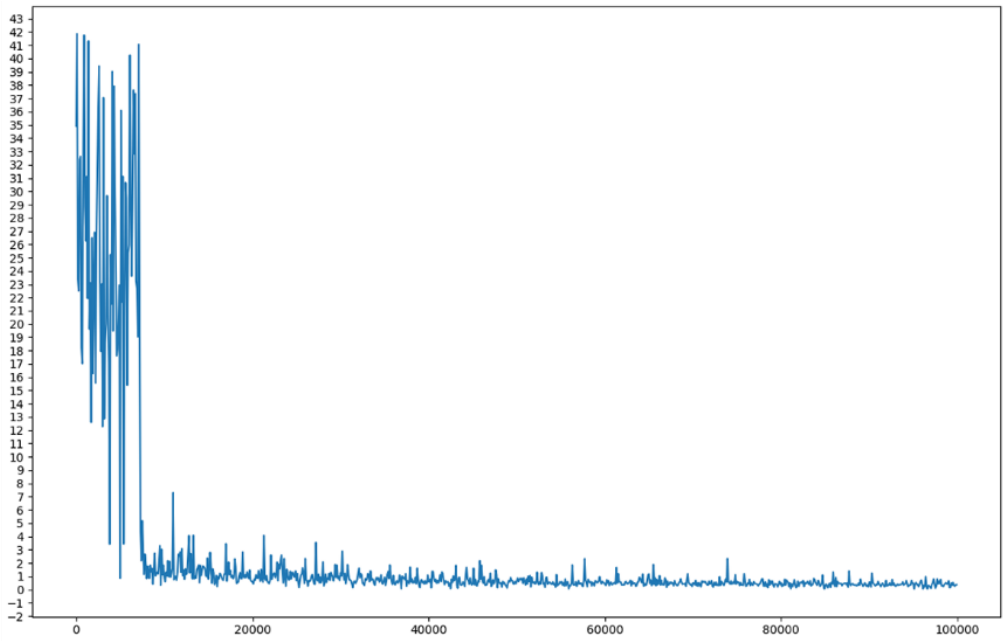

Figure 9. The loss value of FCN during training. It shows the trend of loss values during training. From the figure, we can see that in the early period of the iteration (about before 5000 times), the loss value violently oscillated and then dropped sharply. In the medium term (around 5000–50,000), the loss value decreased slightly and tended to be stable. In order to ensure that the network has stabilized, we chose 60,000 as the number of iterations.

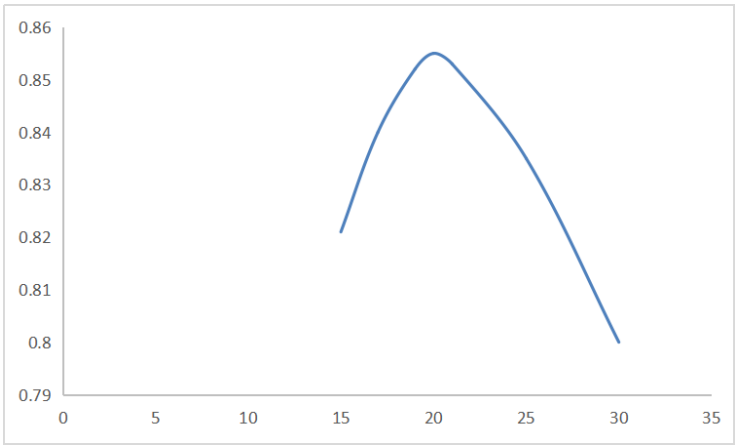

Figure 10. Bleu_1 of different batch sizes. It shows the trend of Bleu_1 when the other parameters were constant and only the batch size was changed. As the batch size increased, Bleu_1 increased first and then decreased, and the effect of batch size on Bleu_1 was obvious, so a suitable batch size was necessary.

4.3. Experiment Evaluation

We randomly allocated the total samples to the training set and validation set according to the sample set sizes in Section 4.1. After an image caption experiment, we obtained a set of satisfactory experimental results, in which Bleu_1 was 0.893, Bleu_2 was 0.744, Bleu_3 was 0.655, and Bleu_4 was 0.587. Then, we kept the number of training sets and validation sets unchanged, randomly allocated samples, and performed nine independent Monte Carlo runs. We compared the Bleu_1, Bleu_2, Bleu_3, and Bleu_4 of those 10 experiments, where the trend of Bleu is shown in Figure 11. In these ten experiments, the mean values of Bleu_1, Bleu_2, Bleu_3, and Bleu_4 was 0.8982, 0.7466, 0.6521, and 0.5866, respectively, and the standard deviations were 0.004, 0.009, 0.011, and 0.012, respectively, which

proved the stability and reliability of the experimental results. We selected the results of the first experiment as the basis for the subsequent experiments and analyses.

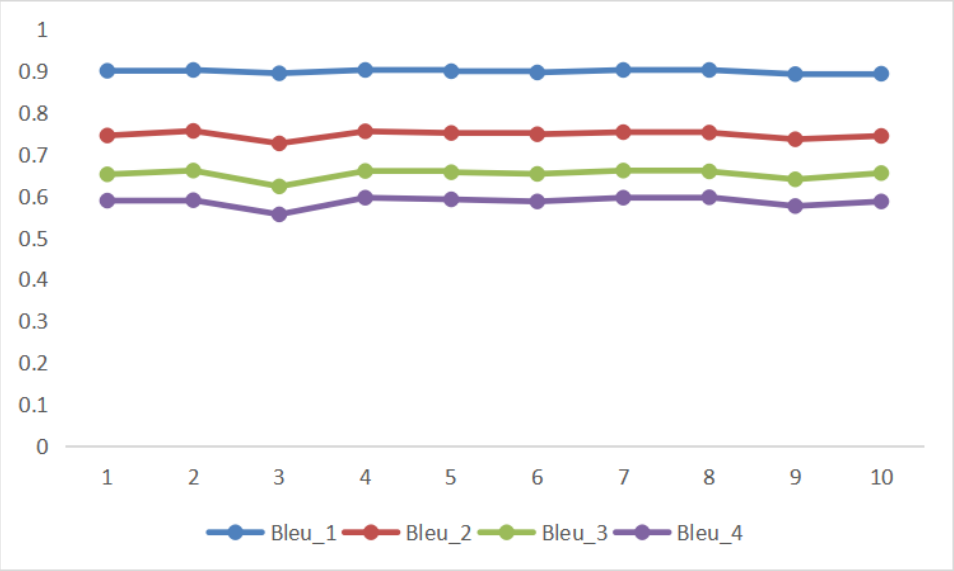

Figure 11. Bleu trend of ten experiments. It shows the trend of Bleu. From the figure, we can see that in ten experiments, the variation amplitudes of Bleu_1, Bleu_2, Bleu_3, and Bleu_4 are small, which can prove the randomness of data distribution and the robustness of the algorithm.

Next, we selected 200 samples from the remote sensing image captioning data set (RSICD) [28] as a validation set to test our experimental model. As shown in our test result, compared with using the VGG-19+LSTM model from the original paper [28], our model outperformed in all the evaluation metrics, where the comparison result of metrics is shown in Table 3. Bilingual Evaluation Understudy of n gram (Bleu_n) calculates the matching degree between n-dimensional phrases and reference captions (GT) [59]; Metric for Evaluation of Translation with Explicit Ordering (METEOR) adds synonym matching on the basis of Bleu to make it more strongly correlated with manual discrimination [60]; Recall-Oriented Understudy for Gisting Evaluation of Longest Common Subsequence (ROUGE_L) and Consensus-based Image Description Evaluation (CIDEr) evaluate the similarity between the generated caption and the GT using the recall rate and cosine similarity, respectively [61,62]. In general, these metrics calculate the matching degree between the generated caption and the GT in different ways, and the larger the metrics values are, the better the generated caption is.

Table 3. Results comparison using VGG-19+LSTM.

	Bleu_1	Bleu_2	Bleu_3	Bleu_4	METEOR	ROUGE_L	CIDEr
Ours	0.774	0.535	0.406	0.314	0.359	0.663	2.745
RSICD	0.583	0.423	0.331	0.270	0.261	0.519	2.033

Table 3 shows the comparison of the results of our model and the model (VGG-19+LSTM) from Reference [28]. Our metrics are higher than that from Reference [28].

There are two reasons for our network performing better on the RSICD:

1. The multi-scale caption design strategy makes the spatial relationship more concise such that the vocabulary size is relatively small, which makes it easier for the network training.
2. The RSICD is a shared test dataset such that the purpose of their experiment is to verify the credibility of the dataset. Therefore, in their test experiment, accuracy is not the main indicator.

We made statistical comparisons between the evaluation metrics and similar work in existing studies, and the results are shown in Table 4:

Table 4. Comparing evaluation metrics.

	Bleu_1	Bleu_2	Bleu_3	Bleu_4	METEOR	ROUGE_L	CIDEr
Mean	0.670	0.509	0.399	0.310	0.235	0.560	0.978
Ours	0.893	0.744	0.655	0.587	0.455	0.779	5.044

Table 4 shows the comparison of the evaluation metrics between the experiment and the mean values from 18 experiments in several related papers. Our evaluation metric values are higher than the mean values.

Comparing our experimental results with the mean values of 18 experimental evaluation metrics from related papers [17,25,26,28,29,63], it is obvious that all of our evaluation metrics scores were better than the mean values. Moreover, in our experience, the pixel accuracy of FCN and U-net for semantic segmentation were 0.89 and 0.93, respectively, which also reached a good level. Therefore, our experimental results are credible and can support subsequent recognition and correction experiments.

We analyzed the reasons for the better experimental results. By comparing the generated captions with the GT, we found that among 668 samples in the validation set, 300 samples contained the word "with" in GT. A total of 256 samples of generated captions containing the word "with," accounting for 85%. The analysis shows that the multi-scale labeling strategy we proposed is feasible and can be used in experiments. This multi-scale spatial relationship description strategy not only greatly reduces the sum of spatial relationship vocabulary in reference captions and the difficulty of network learning but also accurately describes the complex spatial relationship between remote sensing objects. Both reasons are important for increasing the evaluation metric values of the experiment.

Combining words in all sample generation captions, we analyzed the reliability of caption descriptions. The analysis results are shown in Table 5.

Table 5. Reliability analysis for generated captions.

Samples	Total Number of Samples: 668		
	Correct	Incorrect	Total
Words	3417	368	3785
Proportion	90.28%	9.72%	100%
S.R. Word	1396	170	1566
Proportion	89.14%	10.86%	100%
Nouns	2021	198	2219
Proportion	91.08%	8.92%	100%

Table 5 shows the results of our analysis of the generated captions of 668 samples in the validation set. In the table, "word" is the sum of "S.R. word" and "nouns" in the captions, "S.R. word" was the spatial relationship word, and "nouns" were the category nouns. "Correct" means that the word existed in the reference captions (GT), and "incorrect" means that it did not exist.

From Table 5, it is obvious that a total of 3785 words existed in the 668 generated captions, of which 3417 were correct, accounting for 90.28%, and 368 were incorrect, accounting for 9.72%. The Bleu_1 value in our experiment was slightly lower than 0.9028. There are two possible reasons: (1) the influence of other words (such as "and" and "with"), and (2) the introduction of a penalty mechanism in the calculation of the Bleu scores. We divided the 3785 words into nouns and relatives and calculated the statistics. Among them, there were 2219 nouns in total and 2021 were correct, accounting for 91.08%. Moreover, there were 1566 relative words, and 1396 were correct, accounting for 89.14%. Because of the high accuracy of nouns and relative words, it was possible to recognize and correct image objects.

After the recognition and correction of objects, the effect is shown in Table 6.

Table 6. Number of matched nouns before and after correction.

Samples	Total Number of Samples: 668			
	Pre-Correction Matched	**Post-Correction Matched**	**Unmatched**	**Total**
Nouns	929	1856	363	2219
Proportion	41.87%	83.64%	16.36%	100%

Table 6 shows the matching of the nouns and the object mask graphs before and after recognition and correction. "Pre-corrected matching" means that the attention area of the noun was matched with the object mask graphs before correction. "Post-correction matched" means that the attention area of the noun was matched after correction. "Unmatched" means that the attention area of the noun was unmatched with the object mask graphs before and after correction.

5. Discussion

The object recognition and correction mechanism proposed in this paper performs object recognition based on the noun_t generated at time step t and the corresponding attention weight matrix, and multi-scale correction for the mismatched object concerned by the attention weight matrix. Therefore, in order to better analyze the object recognition and correction effects, we divided the 668 samples into two subsets: Sample Set 1 and Sample Set 2. The nouns contained in the generated captions of the samples in Sample Set 1 were all correct, and the captions generated by the samples in Sample Set 2 contained error nouns. In this way, Sample Set 1 could fully realize object recognition and correction, and Sample Set 2 could only identify and correct the remote objects corresponding to the correct nouns. We further analyzed the two sample sets. The overall results before and after correction are shown in Table 7.

Table 7. The results of the overall analysis of the subsets.

Samples	Total Number of Samples: 668							
	All Nouns are Correct: 477				Not All Nouns are Correct: 191			
	Correct	**Incorrect**	**Total**		**Correct**	**Incorrect**	**Total**	
S.R. Word	1013	76	1089		383	94	477	
Proportion	93.02%	6.98%	100%		80.29%	19.71%	100%	
	Pre-Correction Matched	**Post-Correction Matched**	**Unmatched**	**Total**	**Pre-Correction Matched**	**Post-Correction Matched**	**Unmatched**	**Total**
Nouns	781	1541	26	1567	148	315	337	652
Proportion	49.84%	98.34%	1.66%	100%	22.70%	48.31%	51.69%	100%

Table 7 shows the overall analysis of Sample Set 1 and Sample Set 2. From the table, we can see the words contained in the captions generated using the two sample sets and the comparison results of the number of nouns matched with the object before and after correction.

The 477 samples in Sample Set 1 generated a total of 2656 words, of which 2580 words are correct, and 76 words were incorrect. In the generated 2656 words, there were 1567 object nouns, and all of them were correct; 1089 words were relational words, of which 1013 words were correct, and 76 words were incorrect. There were a total of 1129 words contained in the 191 generated captions of Sample Set 2, of which 837 words were correct, and 292 words were incorrect. Among the 1129 words, there were 652 object nouns, of which 454 words were correct, 198 words were incorrect; 477 words were relational words, of which 383 words were correct, and 94 words were incorrect.

In Sample Set 1, we first performed an analysis based on the samples, and the results are shown in Table 8. There were 477 samples in the validation set (i.e., the sample set to be analyzed). Among them, 87 samples, accounting for 18.24%, did not need to be corrected because all nouns were matched with objects. There were 337 samples with incorrect recognition objects before correction, but nouns were all matched with objects after correction, accounting for 70.65%. Generally, after the implementation of our correction method, there were a total of 424 samples in which each noun in the generated captions were

matched with the corresponding objects, accounting for 88.89%, an increase of 70.65 percentage points, which was a remarkable effect. In addition, there were 37 partially corrected samples, accounting for 7.76%; only 16 samples were not corrected, accounting for 3.35%, basically achieving the purpose of recognizing remote sensing objects and the spatial relationship between them.

Table 8. Sample-based analysis of Sample Set 1 before and after correction.

Sample Class	Completely Corrected	Partial Correction	No Corrective Effect	No Need for Correction	Total
Number	337	37	16	87	477
Proportion	70.65%	7.76%	3.35%	18.24%	100%

Table 8 shows the sample-based analysis of Sample Set 1. As shown in the table, more than 78% of the samples were completely or partially corrected.

Next, we conducted an analysis based on category nouns included in the 477 generated captions of Sample Set 1. There were 1567 nouns in all generated captions from the 477 samples. There were 781 nouns whose attention areas were matched with objects before the correction, the proportion was 49.84%, and the matched ones increased to 1541 after correction, accounting for 98.3%, an accuracy increase of 48.50 percentage points. The effect was greatly improved, which means that the method we proposed can meet the demand of remote sensing interpretations.

We performed a similar analysis of 191 samples in Sample Set 2. The results of the sample-based analysis are shown in Table 9. There were multiple objects in one or more classes of some samples, and in the generated captions of these samples, nouns of these classes appeared more than once, and a many-to-many relationship could therefore be constructed. However, the matched objects' judgement of the generated correct nouns in the image will be affected by the incorrect nouns. These samples were classified into having no corrective effect on the statistics, totaling 55 samples. One sample whose words in the generated captions were all incorrect and was classified into no corrective effect.

Table 9. Sample-based analysis of Sample Set 2 before and after correction.

Sample Class	Completely Corrected	Partial Correction	No Corrective Effect	No Need for Correction	Total
Number	113	0	56	22	191
Proportion	59.16%	0.00%	29.32%	11.52%	100%

Table 9 shows the sample-based analysis of Sample Set 2. As shown in the table, only approximately 59% of the samples were corrected. The correction effect was worse than that of Sample Set 1.

The nouns-based analysis of Sample Set 2 showed that 148 (22.70%) of the nouns matched with the mask graphs before the correction, and it increased to 315 after correction, accounting for 48.31%. The proportion increased by 25.61 percentage points, which was less of an effect than for Sample Set 1. However, 337 nouns still could not be corrected, accounting for 51.69%, which indicated that the correction algorithm could only solve the mismatching problem between the attention weight matrix and the object but could not correct the incorrect words generated by LSTM. In addition, 315 of 341 nouns could not be corrected in Sample Set 2, indicating that for a sample, the higher the Bleu scores, the better the recognition and correction mechanism performs.

The sample-based and nouns-based overall correction effect analysis of Sample Set 1 and Sample Set 2 are shown in Figure 12. We conducted a comprehensive analysis of the corrective effect with a combined Sample Set 1 and Sample Set 2. Before the correction, the number of each noun in the captions generated by the samples matching with the mask graph was 109. When the correction finished, the number rose to 559, equivalent to a proportion increase of 67.37 percentage points. Before and after correction, the number of nouns matching with the mask graph increased from 929 to 1856, a proportion increase of 41.78 percentage points.

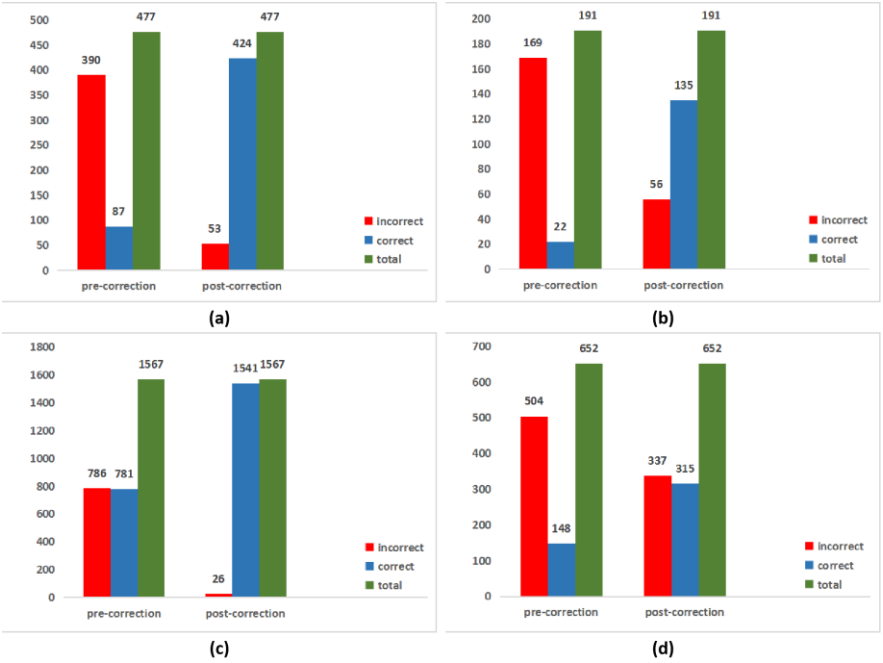

Figure 12. Correction effect analysis. It shows the corrective effect of Sample Set 1 and Sample Set 2. (**a**,**b**) are the sample-based overall correction effect for Sample Set 1 and Sample Set 2, respectively; (**c**,**d**) are the noun-based overall correction effect for Sample Set 1 and Sample Set 2, respectively. As shown in the figure, whether from the perspective of samples or nouns, the correction algorithm proposed in this paper achieved good results. The correction effect of Sample Set 1 was better than that of Sample Set 2.

From the above analysis, the following conclusions can be drawn:

- When the noun in the generated caption was correct and the spatial relationship was incorrect, the remote sensing object could still be recognized, but the spatial relationship could not be corrected.
- When both the nouns and spatial relationship were incorrect, the proposed method was ineffective. This requires further research.

6. Conclusions

In this paper, a multi-scale remote sensing image interpretation network (the MSRIN) was proposed for identifying remote sensing objects and their spatial relationships from end-to-end. First, a remote sensing image was semantically segmented through an FCN and a U-net on two spatial scales such that each pixel in the original image was labelled with two semantic labels; therefore, a hierarchical relationship of a multi-scale remote sensing object could be formed. Second, the features of the same image obtained using a pre-trained VGG-19 network were input for the attention-based LSTM, which outputted the captions that described the two-scale remote sensing objects and their spatial relationships. Finally, the relationship between the nouns in the caption and the object mask graphs was established through the attention weight matrix. In this way, the remote sensing objects from the U-Net got their spatial relationship from the caption. To overcome the spatial deviations of the attention weight matrix from the LSTM, the MSRIN designed an attention-based, multi-scale remote sensing object identification and correction mechanism. Our method produced a complete semantic interpretation of remote sensing images.

Identifying remote sensing objects and their spatial relations was based on the attention weight matrix. In the future, we will improve the attention weight calculation method to achieve more accurate positioning.

The objectivity of the evaluation method based on Bleu depends on the completeness of the captions, which greatly aggravates the burden of sample preparation. Therefore, we will concentrate on exploring the establishment of evaluation metrics and methods that are suitable for remote sensing image captioning.""

The experimental data in this paper was mainly optical remote sensing images. In the future, we will further verify the recognition and correction mechanism proposed in this paper by using other types of data, such as SAR.

Author Contributions: W.C. contributed toward creating the original ideas of the paper. W.C. conceived and designed the experiments. F.W. and X.H. prepared the original data, performed the experiments and analyzed the experimental data with the help of D.Z., X.X., and M.Y. and W.C. wrote and edited the manuscript. F.W., X.H., and Z.W carefully revised the manuscript. J.H. contributed constructive suggestions on modifying the manuscript.

Funding: This research was funded by National Key R & D Program of China (Grant No. 2018YFC0810600, 2018YFC0810605).

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

References

1. Lecun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436. [CrossRef]
2. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Netw.* **2015**, *61*, 85–117. [CrossRef]
3. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
4. Cheng, G.; Han, J.; Guo, L.; Liu, Z.; Bu, S.; Ren, J. Effective and efficient midlevel visual elements-oriented land-use classification using vhr remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 4238–4249. [CrossRef]
5. Cheng, G.; Zhou, P.; Han, J. Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 7405–7415. [CrossRef]
6. Han, J.; Zhang, D.; Cheng, G.; Guo, L.; Ren, J. Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 3325–3337. [CrossRef]
7. Han, X.; Zhong, Y.; Zhao, B.; Zhang, L. Scene classification based on a hierarchical convolutional sparse auto-encoder for high spatial resolution imagery. *Int. J. Remote Sens.* **2017**, *38*, 514–536. [CrossRef]
8. Hu, F.; Xia, G.-S.; Hu, J.; Zhang, L. Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. *Remote Sens.* **2015**, *7*, 14680–14707. [CrossRef]
9. Hu, W.; Huang, Y.; Wei, L.; Zhang, F.; Li, H. Deep convolutional neural networks for hyperspectral image classification. *J. Sens.* **2015**, *2015*, 1–12. [CrossRef]
10. Zhong, Y.; Fei, F.; Zhang, L. Large patch convolutional neural networks for the scene classification of high spatial resolution imagery. *J. Appl. Remote Sens.* **2016**, *10*, 025006. [CrossRef]
11. Cui, W.; Zheng, Z.; Zhou, Q.; Huang, J.; Yuan, Y. Application of a parallel spectral–spatial convolution neural network in object-oriented remote sensing land use classification. *Remote Sens. Lett.* **2018**, *9*, 334–342. [CrossRef]
12. Cui, W.; Zhou, Q.; Zheng, Z. Application of a hybrid model based on a convolutional auto-encoder and convolutional neural network in object-oriented remote sensing classification. *Algorithms* **2018**, *11*, 9. [CrossRef]
13. Cannon, R.; Dave, J.; Bezdek, J.; Trivedi, M. Segmentation of a thematic mapper image using the fuzzy c-means clusterng algorthm. *IEEE Trans. Geosci. Remote Sens.* **1986**, *GE-24*, 400–408. [CrossRef]
14. Jeon, B.; Landgrebe, D.A. Classification with spatio-temporal interpixel class dependency contexts. *IEEE Trans. Geosci. Remote Sens.* **1992**, *30*, 663–672. [CrossRef]
15. Baatz, M.; Schape, A.; Multiresolution segmentation: An optimization approach for high quality multi-scale image segmentation. Angew. *Geogr. Inf. 2000, XII, 12–23.* Available online: https://pdfs.semanticscholar.org/364c/c1ff514a2e11d21a101dc072575e5487d17e.pdf?_ga=2.55340014.416308819.1554177081-320853791.1554177081 (accessed on 2 April 2019).

16. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3156–3164. [CrossRef]
17. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhutdinov, 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.
18. Chen, F.; Ji, R.; Sun, X.; Wu, Y.; Su, J. GroupCap: Group-based image captioning with structured relevance and diversity constraints. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 1345–1353. [CrossRef]
19. Yuan, A.; Li, X.; Lu, X. 3G structure for image caption generation. *Neurocomputing* **2019**, *330*, 17–28. [CrossRef]
20. Chen, H.; Ding, G.; Lin, Z.; Zhao, S.; Han, J. Show, observe and tell: Attribute-driven attention model for image captioning. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 13–19 July 2018; pp. 606–612. [CrossRef]
21. Khademi, M.; Schulte, O. Image caption generation with hierarchical contextual visual spatial attention. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018; pp. 2024–2028. [CrossRef]
22. Ranzato, M.; Chopra, S.; Auli, M.; Zaremba, W. Sequence Level Training with Recurrent Neural Networks. Available online: https://arxiv.org/abs/1511.06732 (accessed on 2 April 2019).
23. Shi, H.; Li, P.; Wang, B.; Wang, Z. Image captioning based on deep reinforcement learning. In Proceedings of the 10th International Conference on Internet Multimedia Computing and Service(ICIMCS), Nanjing, China, 17–19 August 2018; pp. 45–49. Available online: https://arxiv.org/10.1145/3240876.3240900 (accessed on 2 April 2019).
24. Karpathy, A.; Li, F.-F. Deep visual-semantic alignments for generating image descriptions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3128–3137. [CrossRef]
25. 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 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3242–3250. [CrossRef]
26. Qu, B.; Li, X.; Tao, D.; Lu, X. Deep semantic understanding of high resolution remote sensing image. In Proceedings of the 2016 International Conference on Computer, Information and Telecommunication Systems (CITS), Kunming, China, 6–8 July 2016; pp. 1–5. [CrossRef]
27. 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]
28. Lu, X.; Wang, B.; Zheng, X.; Li, X. Exploring models and data for remote sensing image caption generation. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 2183–2195. [CrossRef]
29. Wang, B.; Lu, X.; Zheng, X.; Liu, W. Semantic descriptions of high-resolution remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2019**, *99*, 1–5. [CrossRef]
30. 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]
31. Wang, Y.; Lin, Z.; Shen, X.; Cohen, S.; Cottrell, G.W. Skeleton key: Image captioning by skeleton-attribute decomposition. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 7378–7387. [CrossRef]
32. Tobler, W.R. A computer movie simulating urban growth in the Detroit region. *Econ. Geogr.* **1970**, *46*, 234. [CrossRef]
33. Nogueira, K.; Penatti, O.A.B.; dos Santos, J.A. Towards better exploiting convolutional neural networks for remote sensing scene classification. *Pattern Recognit.* **2017**, *61*, 539–556. [CrossRef]
34. Zhang, F.; Du, B.; Zhang, L. Scene classification via a gradient boosting random convolutional network framework. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 1793–1802. [CrossRef]
35. Li, E.; Xia, J.; Du, P.; Lin, C.; Samat, A. Integrating multilayer features of convolutional neural networks for remote sensing scene classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 5653–5665. [CrossRef]
36. Ge, Y.; Tang, Y.; Jiang, S.; Leng, L.; Xu, S.; Ye, F. Region-based cascade pooling of convolutional features for HRRS image retrieval. *Remote Sens. Lett.* **2018**, *9*, 1002–1010. [CrossRef]

37. Shelhamer, E.; Long, J.; Darrell, T. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651. [CrossRef]
38. Huang, Z.; Cheng, G.; Wang, H.; Li, H.; Shi, L.; Pan, C. Building extraction from multi-source remote sensing images via deep deconvolution neural networks. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; pp. 1835–1838. [CrossRef]
39. Cui, W.; Li, R.; Yao, Z.; Chen, S.; Tang, S.; Li, Q. Study on the optimal segmentation scale based on fractual dimension of remote sensing images. *J. Wuhan Univ. Technol.* **2011**, *33*, 463–467. [CrossRef]
40. Xia, J.; Yang, X.; Jia, L. A multi-depth convolutional neural network for SAR image classification. *Remote Sens. Lett.* **2018**, *9*, 1138–1147. [CrossRef]
41. Li, L.; Liang, J.; Weng, M.; Zhu, H. A multiple-feature reuse network to extract buildings from remote sensing imagery. *Remote Sens.* **2018**, *10*, 1350. [CrossRef]
42. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Springer International Publishing: Cham, Switzerland, 2015; Volume 9351, pp. 234–241, ISBN 978-3-319-24573-7.
43. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269. [CrossRef]
44. Tao, Y.; Xu, M.; Lu, Z.; Zhong, Y. DenseNet-based depth-width double reinforced deep learning neural network for high-resolution remote sensing image per-pixel classification. *Remote Sens.* **2018**, *10*, 779. [CrossRef]
45. Drönner, J.; Korfhage, N.; Egli, S.; Mühling, M.; Thies, B.; Bendix, J.; Freisleben, B.; Seeger, B. Fast cloud segmentation using convolutional neural networks. *Remote Sens.* **2018**, *10*, 1782. [CrossRef]
46. Yang, H.; Wu, P.; Yao, X.; Wu, Y.; Wang, B.; Xu, Y. Building extraction in very high resolution imagery by dense-attention networks. *Remote Sens.* **2018**, *10*, 1768. [CrossRef]
47. Panboonyuen, T.; Jitkajornwanich, K.; Lawawirojwong, S.; Srestasathiern, P.; Vateekul, P. Semantic segmentation on remotely sensed images using an enhanced global convolutional network with channel attention and domain specific transfer learning. *Remote Sens.* **2019**, *11*, 83. [CrossRef]
48. Zhang, T.; Tang, H. A comprehensive evaluation of approaches for built-up area extraction from landsat oli images using massive samples. *Remote Sens.* **2019**, *11*, 2. [CrossRef]
49. Sun, G.; Huang, H.; Zhang, A.; Li, F.; Zhao, H.; Fu, H. Fusion of multiscale convolutional neural networks for building extraction in very high-resolution images. *Remote Sens.* **2019**, *11*, 227. [CrossRef]
50. Fu, Y.; Liu, K.; Shen, Z.; Deng, J.; Gan, M.; Liu, X.; Lu, D.; Wang, K. Mapping impervious surfaces in town–rural transition belts using China's GF-2 imagery and object-based deep CNNs. *Remote Sens.* **2019**, *11*, 280. [CrossRef]
51. Li, W.; Dong, R.; Fu, H.; Yu, L. Large-scale oil palm tree detection from high-resolution satellite images using two-stage convolutional neural networks. *Remote Sens.* **2019**, *11*, 11. [CrossRef]
52. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
53. Mou, L.; Ghamisi, P.; Zhu, X.X. Deep recurrent neural networks for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3639–3655. [CrossRef]
54. Wu, H.; Prasad, S. Convolutional recurrent neural networks forhyperspectral data classification. *Remote Sens.* **2017**, *9*, 298. [CrossRef]
55. Ndikumana, E.; Minh, D.H.T.; Baghdadi, N.; Courault, D.; Hossard, L. Deep recurrent neural network for agricultural classification using multitemporal sar sentinel-1 for Camargue, France. *Remote Sens.* **2018**, *10*, 1217. [CrossRef]
56. Liu, B.; Yu, X.; Yu, A.; Zhang, P.; Wan, G. Spectral-spatial classification of hyperspectral imagery based on recurrent neural networks. *Remote Sens. Lett.* **2018**, *9*, 1118–1127. [CrossRef]
57. Liu, Q.; Zhou, F.; Hang, R.; Yuan, X. Bidirectional-convolutional LSTM based spectral-spatial feature learning for hyperspectral image classification. *Remote Sens.* **2017**, *9*, 1330. [CrossRef]
58. Ma, A.; Filippi, A.M.; Wang, Z.; Yin, Z. Hyperspectral image classification using similarity measurements-based deep recurrent neural networks. *Remote Sens.* **2019**, *11*, 194. [CrossRef]

59. 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 on Association for Computational Linguistics—ACL'02, Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318. [CrossRef]
60. Lavie, A.; Agarwal, A. Meteor: An automatic metric for MT evaluation with high levels of correlation with human judgments. In Proceedings of the Second Workshop on Statistical Machine Translation—StatMT'07, Prague, Czech Republic, 23 June 2007; pp. 228–231. [CrossRef]
61. Lin, C. ROUGE: A package for automatic evaluation of summaries. In Proceedings of the Workshop on Text Summarization Branches Out—ACL'05, Barcelona, Spain, 25–26 July 2004; pp. 74–81. Available online: https://www.aclweb.org/anthology/W04-1013 (accessed on 2 April 2019).
62. Vedantam, R.; Zitnick, C.L.; Parikh, D. CIDEr: Consensus-based image description evaluation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 4566–4575. [CrossRef]
63. You, Q.; Jin, H.; Wang, Z.; Fang, C.; Luo, J. Image captioning with semantic attention. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 4651–4659. [CrossRef]