International Journal of Geo-Information

MDPI

Article

Multi-Scale Remote Sensing Semantic Analysis Based on a Global Perspective

Wei Cui *, Dongyou Zhang, Xin He, Meng Yao, Ziwei Wang, Yuanjie Hao, Jie Li, Weijie Wu, Wenqi Cui and Jiejun Huang

School of Resources and Environmental Engineering, Wuhan University of Technology, Wuhan 430070, China; gis@whut.edu.cn (D.Z.); 2962575697@whut.edu.cn (X.H.); yaomeng@whut.edu.cn (M.Y.); zwei@whut.edu.cn (Z.W.); haoyuanjie@whut.edu.cn (Y.H.); Ljie@whut.edu.cn (J.L.); wwjie@whut.edu.cn (W.W.); W.Q.Cui@whut.edu.cn (W.C.); hjj@whut.edu.cn (J.H.)

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

Received: 4 July 2019; Accepted: 16 September 2019; Published: 17 September 2019

Abstract: Remote sensing image captioning involves remote sensing objects and their spatial relationships. However, it is still difficult to determine the spatial extent of a remote sensing object and the size of a sample patch. If the patch size is too large, it will include too many remote sensing objects and their complex spatial relationships. This will increase the computational burden of the image captioning network and reduce its precision. If the patch size is too small, it often fails to provide enough environmental and contextual information, which makes the remote sensing object difficult to describe. To address this problem, we propose a multi-scale semantic long short-term memory network (MS-LSTM). The remote sensing images are paired into image patches with different spatial scales. First, the large-scale patches have larger sizes. We use a Visual Geometry Group (VGG) network to extract the features from the large-scale patches and input them into the improved MS-LSTM network as the semantic information, which provides a larger receptive field and more contextual semantic information for small-scale image caption so as to play the role of global perspective, thereby enabling the accurate identification of small-scale samples with the same features. Second, a small-scale patch is used to highlight remote sensing objects and simplify their spatial relations. In addition, the multi-receptive field provides perspectives from local to global. The experimental results demonstrated that compared with the original long short-term memory network (LSTM), the MS-LSTM's Bilingual Evaluation Understudy (BLEU) has been increased by 5.6% to 0.859, thereby reflecting that the MS-LSTM has a more comprehensive receptive field, which provides more abundant semantic information and enhances the remote sensing image captions.

Keywords: LSTM; multi-scale; remote sensing object; image captioning

1. Introduction

According to Tobler's first law of geography, everything is related to everything else, but near things are more related to each other [1]. However, the spatial distance that defines adjacency is a problem. The Academician Li Xiaowen proposed the concept of spatial temporal proximity as follows: the concept of "flow" is applied to understand and express distance and adjacency, thereby resulting in the concept of spatial temporal proximity [2,3].

In remote sensing image captioning [4–9], the above principle shows that the accurate recognition of remote sensing images often requires the spatial relationships between objects. Especially in the cases of different objects with the same spectrum, different types of remote sensing object may have the same spectral, texture and shape features and can only be more accurately interpreted by using adjacent objects and their spatial relations. However, the traditional object recognition methods [10–12]

cannot identify the spatial relationships between objects while recognizing remote sensing objects. Deep neural networks [13,14] can automatically extract and optimize features and have the advantages of high accuracy and universality. They have been widely used in scene classification [15–20] and semantic or instance segmentation [21,22]. In recent years, image captioning [23] based on recurrent neural networks (RNNs) has emerged, and this method can generate natural language descriptions containing the spatial relationships of remote sensing objects. In particular, the attention-based LSTM provides a feasible way to realize the comprehensive interpretation of remote sensing objects and their spatial relationships in remote sensing images [4].

However, when using the LSTM to simultaneously identify remote sensing objects and their spatial relationships, the problem of the uncertainty of the object neighborhood must be solved.

According to Tobler's first law of geography [1], to better understand remote sensing objects, it is often necessary to refer to the related objects in the neighborhood of the objects and the spatial relationships between them, but there is no uniform scale in this relevant neighborhood. Therefore, In the deep learning algorithms, in order to effectively overcome limited GPU memory, the semantic accuracy and computational burden must be balanced. To ensure the effective operation of the algorithm and better identify remote sensing objects, it is often necessary to segment a fixed-size sample to carry out network training. However, the adjacent regions as described by the first law of geography cannot be quantified by using a uniform scale. The current method of segmenting samples according to a single scale will easily lead to incomplete information of the sample scale and thus cannot accurately describe the semantics of remote sensing objects.

In addition, there is another problem of the influence of an uncertain scale range on remote sensing recognition. When recognizing and classifying the samples, professionals who own the domain knowledge often use visual methods on the original image and dynamically transform the observation scope to obtain the corresponding semantic information. This process adopts a remote sensing semantic understanding mechanism that is similar to the global perspective. However, when the labeled remote objects are extracted based on a fixed and unified scale, their surrounding environment is often removed, this may lead to some of the objects with same spectral, texture and shape features having different semantic labels. Thus, we form some samples of "fake conflict" that are semantically contradictory but substantially reasonable. When the samples of "fake conflict" with inconsistent semantics are sent to the model for training, logical chaos occurs in the calculation of the model error and a stable output cannot be learned. Multi-scale research may be a way to conquer the impact of the samples of "fake conflict".

The multi-scale concept in this study refers to the relationship between the local and the global, that is, a small local area and a large area within a certain neighborhood. However, the current multi-scale research focuses on the multi-scale feature extraction and fusion of the same training sample image [24–31]. There are few studies on how to consider the semantic analysis of variable regions in different spatial extents [32–35].

Therefore, the MS-LSTM network, which is a remote sensing image captioning model based on multi-scale semantics, is proposed: In the MS-LSTM network, the multi-scale concept refers to the relationship between the local and the global, that is a small local area and a large area within a certain neighborhood. For restoring the global perspective, a large scene classification input is added to restore the environment and contextual information of samples so that the multi-scale semantic complement is performed to realize error monitoring and correction.

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

1. We propose the following multi-scale segmentation principles. Large-scale objects focus on describing the overall nature of the functional area where remote sensing objects are located and provide the function of the global perspective for remote sensing. Small-scale objects are used for the semantic interpretation of remote sensing images which requires simple and efficient generation of image caption sentence. In this way, the samples of two scales are generated in

pairs, and the multi-receptive field effect between local and global is realized, which realizes the simplicity, efficiency and accuracy of the image caption sentence.

2. To solve the "fake conflict" problems and fuse large-scale scene information, we innovatively design a multi-scale LSTM parallel deep neural network. The original LSTM network structure was modified, and a large-scale semantic memory unit was designed to store the influence of the large-scale semantic information on the currently generated words. At the output-end of the MS-LSTM, a multi-layer perceptron is used to fuse the information of the two scales. The designed parallel network combines the advantages of multi-scale semantic information, deep neural networks and recurrent networks, constructing a multi-scale remote sensing image captioning model.

The remainder of this paper is organized as follows: Section 2 discusses related work. Remote sensing image captioning network based on multi-scale semantics is presented in Section 3. The experiments and analysis are provided in Section 4. Finally, the conclusion is presented in Section 5.

2. Related Work

Remote image captioning describes a certain number of remote sensing objects and the spatial relationship between them by natural language. As the correlative objects within a sentence naturally form a larger-scale remote sensing object, it is necessary to analyze a multi-scale neural network. The captioning of remote sensing images usually uses Recurrent Neural Networks (RNN), especially the Long-Short Term Memory (LSTM) which stores information learnt from experience into memory units and can avoid long-term dependencies through forgetting mechanisms and is suitable for sequence modeling, therefore, this section will describe the issue according to multi-scale remote sensing, RNN and remote image captioning.

2.1. Multi-Scale

Currently, the multi-scale research of remote sensing images based on deep learning focuses on two directions. One is multi-scale feature fusion, which extracts the multi-scale features from a remote sensing image and fuses them for the recognition and classification of remote sensing objects. The other is multi-scale patches, which takes different parts of an image to interpret multi-scale remote sensing images. Multi-scale feature studies often use multi-scale convolution kernels to extract the features of images in parallel. For example, the sliding window covers only one pixel when the 1×1 filter is applied to each channel, which means it focuses more on continuous spectral characteristics, while the 3×3 and 5×5 kernels can focus on different potential local spatial structures due to the different size of receptive fields [26]. Another method is the atrous convolution of an image with a 3×3 kernel and rates of r = 1, 2 and 4 for the target pixel are used for urban land use and land cover classification [27]. It can be seen that higher sampling rates indicate larger fields-of-view (FOVs) compared to the standard convolution filter that requires more parameters to enlarge the FOV. The atrous filter enlarges the FOV without increasing the number of parameters to be calculated, and thus saves computational resources. Additionally, high-level features containing semantic information and low level features containing fine details are fused together through up-sampling and concatenation [28] so that different layers of feature are both considered for detection tasks, especially for small object detection. In a similar model [29], the deep residual network (ResNet) is used as the encoder, and high level features are combined with corresponding low-level features as the up-sampling stage decoder. In an improved fully convolutional network (FCN) [30], the multi-scale contexts are obtained and fused for semantic understanding through a spatial residual inception (SRI) module, which continuously fusing multi-level features. Some others have proposed the feature fusion pyramid network (FFPN) [31] that strengthens the reuse of different scales' features; FFPN can extract the low level locations and high level semantic information that have important impacts on multi-scale ship detection and the precise location of densely parked ships. To better exploit the spatial information, for each pixel, multi-scale image

patches are generated at different spatial scales, which are all centred at the same central spectrum but with smaller spatial scales. Then, the spectral–spatial features from each scale patch are extracted through the multi-scale CNNs [32]. A region-based majority voting CNN is proposed in [33]. First, the images are segmented under multiple resolutions into multiple regions as the basic processing units. Then within each segmented region, the perceptive fields of the generated point voters were classified through the multi-scale CNN. In [34], a three-step region-growing segmentation method was proposed to reduce both under-segmentation and over-segmentation. It uses a feature generation method to transform the three-dimensional neighborhood features of a point into a two-dimensional feature image, which is taken as the input of a multi-scale CNN for training and testing tasks. A novel regional division strategy based on local and non-local decisions [35] was devised to distinguish homogeneous and heterogeneous regions. Then, for homogeneous regions, a multi-scale CNN architecture with larger spatial window inputs was constructed to learn the joint spectral spatial features. For heterogeneous regions, a fine-grained CNN architecture with smaller spatial window inputs was constructed to learn the hierarchical spectral features.

2.2. RNN Series

As an important branch of the deep learning family, RNNs [36] are widely used for sequence analysis. In remote sensing images, there are multiple spectral bands, especially in hyperspectral remote sensing images which contain hundreds of spectral bands, which can be regarded as a sequence of spectral. Therefore, some RNN related networks are proposed for hyperspectral image classification [37,38], and have achieved excellent performance in classification job. In further research, the improved RNN semantic segmentation model not only utilizes spectral features, but also takes into account spatial features. In particular, a hyperspectral image cube surrounding a central pixel is considered to be a hyperspectral pixel sequence, and an RNN is used to simulate the dependencies between different neighborhood pixels [39]. This achieves comprehensive utilization of spectral spatial information. Also in another image classification method based on spectral spatial information [40], inspired by the widely used convolutional neural network (CNN), convolutional operators across spatial domains are incorporated into the network to extract spatial features, which can also get better results. In recent years, the convolutional end of this method has also improved. By combining the network of convolutional long short-term memory and 3-D convolutional neural networks, a better classification effect is obtained [41]. At the same time, the ability of RNN to process time series data can also be used to process synthetic aperture radar (SAR) images. The two-dimensional image block is converted into a one-dimensional sequence and then imported into the LSTM to learn the potential spatial correlation [42], and the feature can be classified in Softmax classifier [43]. In addition, RNN can also be used in the intelligent interpretation of high resolution remote sensing image. According to our investigation, the RNN network is mainly combined with the CNN network to perform the task of remote sensing image captioning. Specifically, it is described in the next section in detail. Recently, a new strategy for constructing sequential features has been proposed [44]. Different to most studies, in this study, the spectral characteristics of each pixel of the multi-temporal images are considered as a continuous feature of the RNN input layer. This strategy constructs the sequence feature by collecting similar pixels from a single image, and has achieved significant improvements in classification.

2.3. Remote Sensing Image Captioning

Remote sensing image captioning is to understand the given images at a semantic level and generate comprehensive natural sentences [5]. Image captioning [23,45] is an import task in the field of computer vision. In recent years, attention-based LSTMs [46] have been proposed, which not only describes the content of the image, but also obtains a 14×14 weight matrix of a corresponding image region when generating word at the current time. Since the image captions and the image features are simultaneously input into the RNN, a new adaptive attention model that can automatically decide when to rely on visual signals and when to just rely on the language model is proposed [47]. Although the

above method has achieved good results in digital images, its application in the field of remote sensing is still lacking. Remote sensing image captioning is more complicated than natural image description. Because the remote sensing image of satellite or aircraft has a unique the global perspective, it has no directionality and lacks the center of focus. All of these factors make natural language descriptions of remote sensing images more difficult. Despite the above difficulties, researchers have carried out useful study on remote sensing image captioning. Qu et al. [5] used both visual and textual information of remote sensing images to generate natural sentences. Shi et al. [6] propose a novel and promising remote sensing image captioning framework that leverages the techniques of FCN. Both methods use a CNN to extract image features and to generate the corresponding captions from RNNs or pre-defined templates. To fully advance the task of remote sensing captioning, after considering the particularity of remote sensing images, a large-scale aerial image data set of remote sensing images is proposed to better describe remote sensing images [7]. In addition, Wang et al. [8] proposed a novel method which used Latent semantic embedding by metric learning for remote sensing image multi-sentence captioning task. Zhang et al. [9] propose a new model with an attribute attention mechanism for remote sensing image captioning. At the same time, it is explored whether the attributes extracted from remote sensing images have an impact on the attention mechanism. This work has greatly facilitated the application of image captioning in the field of remote sensing. Chen et al. [48] propose an intelligent semantic understanding method for high-spatial-resolution (HSR) remote sensing images via geospatial relation captions, which use an attention-based LSTM to implement geospatial relation captioning for HSR images. Recently, to realize the end-to-end identification of remote sensing objects and their spatial relationships, a multi-scale remote sensing image interpretation network based on a FCN, a U-Net and a LSTM is proposed [4]. In addition, Zhang et al. [49] propose a training mechanism of multi-scale cropping for remote sensing image captioning based on an encoder-decoder model to improve the generalization of feature representation. Although image captioning has been applied to the semantic understanding of remote sensing images in recent years, there are still many problems to be solved. For example, words generated in image captioning cannot correspond to remote sensing objects one by one, and the description of spatial relationships is relatively weak. In particular, the image region that the attention weight matrix corresponds to at each moment does not match the remote sensing object corresponding to the word that generated at the same time step [50]. To better describe remote sensing images and understand the semantic information in them, further exploration is still needed.

In summary, in the multi-scale image captioning research, most of them focused on extracting features of different scales from the same image. However, the multi-scale study of variable regions is relatively weak and needs further exploration.

3. Methodology

Firstly, a multi-scale sample segmentation principle is proposed. On this basis, a multi-scale semantic remote sensing image captioning network is designed and implemented.

3.1. The Multi-Scale Principle

According to the basic principles of urban planning [51] and the semantic segmentation of remote sensing images, we proposed the following multi-scale segmentation and matching principles:

1. Static large-scale segmentation principle: A large scale image should be large enough to restore the global perspective and reconstruct the surrounding environment of remote sensing objects and important contextual information of remote sensing objects. According to the principle of urban planning, the city block is the basic unit of an urban space that is composed of networked urban streets and its enclosed urban construction land, and it possesses certain social significance due to the organizational effect of its form and function. The block therefore contains the sufficient and pure semantic information that is needed for functional classification [52]. According to the basic principle of urban planning, the spatial scale of the block is approximately 150 m × 150 m, so large-scale patch segmentation uses a static, unified spatial scale. First, it contains the core parts

of most blocks, which can support the accurate classification of VGG network. Second, it removes the interference of roads and other buildings around the block, which reduces the difficulty of network training. The specific operation uses the road network to realize the segmentation of the block, and then constructs the large-scale sample of 150 m × 150 m based on the geometric center of the block, which uses ArcGIS as the platform.

2. Dynamic small-scale segmentation principle: A small-scale image should be small enough to be composed of a central object and its immediately adjacent objects. The image caption is prominent in the centre remote sensing object, and the spatial relationship is simple, which is convenient for network training. Therefore, we set the scale of the small patch to approximately 60 m × 60 m.
3. Multi-scale matching principle: Since the two scale samples are extracted separately, it is necessary to design a matching rule. According to the principle that the geometric space distance between geographic objects is often not equal to the geospatial distance [2]. This paper proposes the matching principle of two scale samples based on the inclusion relationship of blocks: Firstly find the block to which the small-scale image belongs, secondly and then construct a large-scale image of 150 m × 150 m through the center of the block, finally form a one-to-many mapping relationship.

We propose the above multi-scale segmentation and matching principles for two main reasons:

(a) As a geographical unit with a single function and category, the block has classification stability for all small-scale objects in the block, while the large-scale image constructed through the geometric center of small-scale image may have a mismatch with the block semantic type. In fact, small-scale samples are more semantically consistent with the large-scale samples constructed by the geometric center of the block. This is what Li Xiaowen said: geometric space distance is often not equal to geospatial distance [2].

(b) Large-scale images constructed by the geometric center of small-scale images are disturbed by the surrounding environment, especially when similar small-scale images match large-scale images of different environments, which increases complexity and leads to an unstable classification of large-scale VGG Network.

As shown in Figure 1, comparing (b) with (c), G can better match f on spectrum and texture than F. Comparing (b) with (d), G can better match e than E. Meanwhile, e and f are the same kind of images, but E and F show great difference in spectrum and texture. These illustrate that large-scale images constructed by the geometric center may fail to match small-scale feature and also bring ambiguous large-scale information that would make large-scale VGG network unable to perform stable classification

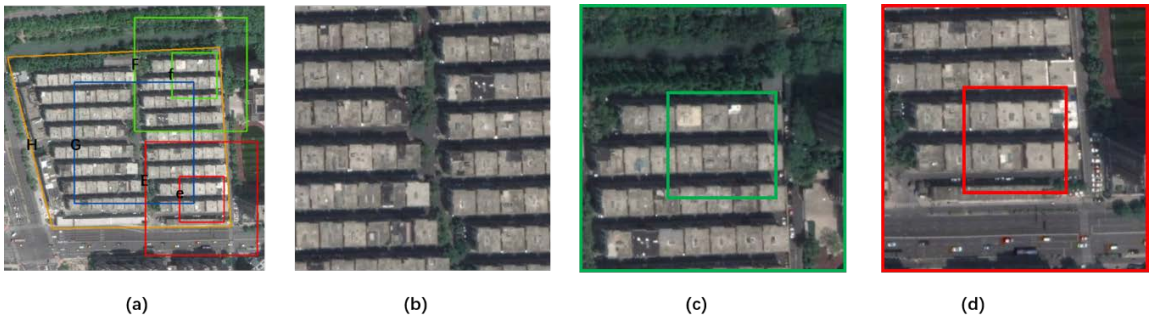

Figure 1. (**a**) Image e and f are small-scale images. Image E and F are large-scale images constructed by the geometric center of e and f, respectively. Large-scale image G is the center of the block. H is the boundary of block; (**b**) correspond to G in (**a**); (**c**) correspond to F, f in (**a**); (**d**) correspond to E, e in (**a**).

For the above two problems, we designed an algorithm based on the multi-scale matching principle. First, we use the road network to realize the segmentation of the block, so that we can get a series of blocks which compose B_List. The size of B_List is n, and n is the number of all blocks. Then

we constructed large-scale images of 150 m × 150 m by the geometric center of each block, and all the large-scale images compose S_List. The size of S_List is also n, and S_List[i] corresponds to a large-scale sample constructed with the geometric center of B_List[i], $i \epsilon [1, n]$. We match a corresponding large-scale image which is denoted as S for one small-scale image which is denoted as s by the following Algorithm 1:

Algorithm 1 For matching small-scale images with large-scale images

```
Input: small-scale image s, a series of blocks B_List, a series of large-scale images S_List, the size of B_List n
Output: large-scale image S.
Begin
    For i = 1 to n; step = 1; do
        If B ∪ s = B_List[i] Then
            S = S_List[i]
            Return S;
        Else
            Continue;
    End For
End
```

The above operation utilizes the spatial analysis function provided by ArcGIS to achieve automatic matching of samples of different scales.

According to the multi-scale segmentation and matching principle that was mentioned above, the small-scale "fake conflict" samples and normal samples and their corresponding large-scale samples are shown in Figure 2. The 'fake conflict' samples mean that the remote sensing objects in the small-scale samples have similar spectral, texture, shape and spatial relationship between the objects, but have different classification information in large scale environment which lead to network training process conflict. The normal samples mean samples that will not cause above-mentioned problems.

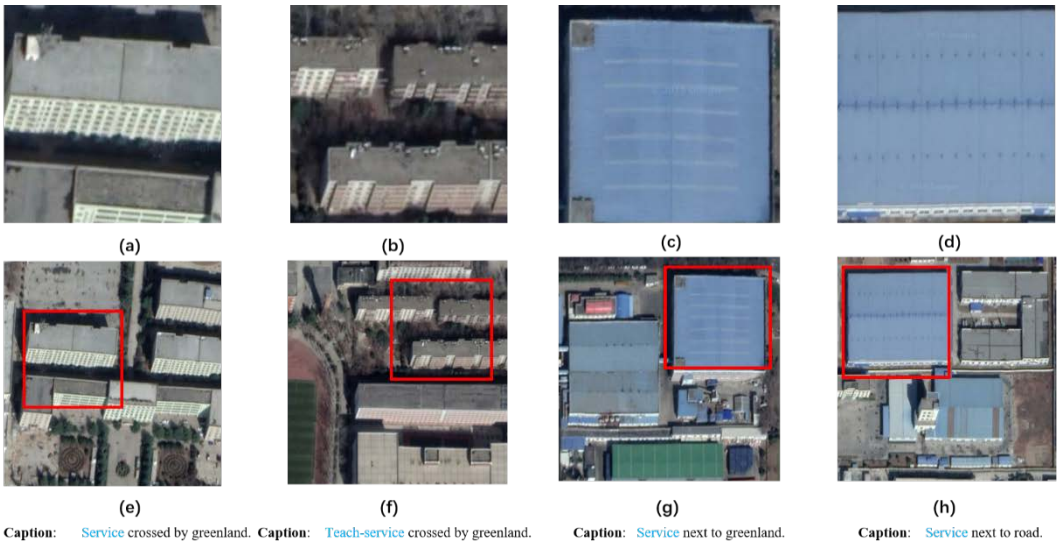

Figure 2. A pair of "fake conflict" samples and a pair of normal samples. (**a**–**d**) are four small-scale samples; (**e**–**h**) are corresponding large-scale samples of (**a**–**d**); (**a**,**b**) are a pair of 'fake conflict' samples; (**c**,**d**) are a pair of normal samples. We demonstrate small scale samples in large scale samples by drawing red boxes.

3.2. Multi-scale Image Captioning Network Structure

The proposed network structure adopts two CNNs and a multi-scale LSTM. In the encoding phase, the CNN is used to extract the remote sensing image features, and the VGG19 network is applied to the research. The large-scale VGG is used for large-scale scene classification which requires retraining, and the small-scale VGG adopts the migration learning method only for extracting small-scale image features, so a well-trained VGG is used. In the decoding stage, an innovative design was implemented based on the original attention-based LSTM, which integrates the scene semantic information that is provided by the large-scale VGG network with the small-scale remote sensing image features that are extracted by the CNN to form a multi-scale LSTM. The network structure is shown in Figure 3.

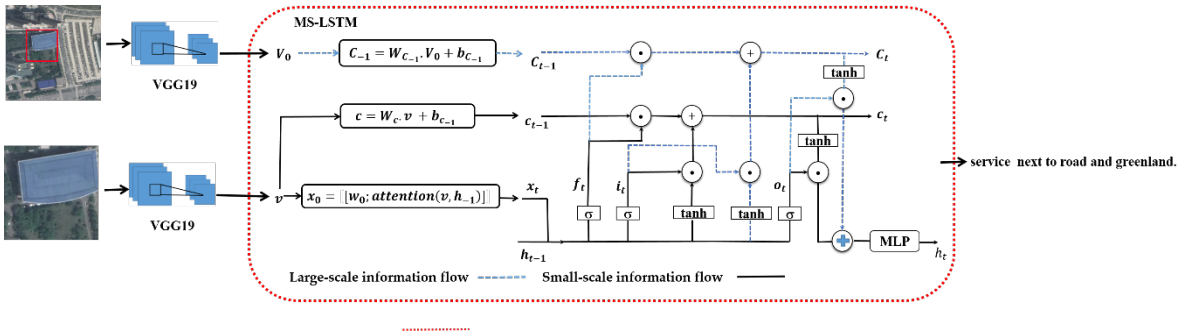

Figure 3. Network structure. It shows the overall network structure of the MS-LSTM. In our network, a small-scale remote sensing image and its corresponding large-scale image are input into two VGG19 networks respectively to extract two scale semantic information. At the output-end, a 1024-dimensional large-scale semantic memory unit and a 1024-dimensional small-scale memory unit are concatenated to form a 2048-dimensional vector, which is input into a MLP. And then the MLP outputs a 1024-dimensional vector, which is activated by tanh and then updates the hidden layer information h_t through the output gate.

3.3. MS-LSTM

Two scales of features extracted from two VGG networks are input into the MS-LSTM. The specific design features of the MS-LSTM are shown as follows. At time t_0, the semantic information V_0 that is obtained by the large-scale VGG network is added to the original network such that the network simultaneously incorporates image information from two spatial scales, and the new large-scale information V_0 is derived from the classification results of the VGG19. The network adds a large-scale semantic memory unit to store the semantic information of the large-scale remote sensing images. The large-scale semantic information memory unit is updated by using an input gate and a forget gate. At the output-end, a 1024-dimensional large-scale semantic memory unit and a 1024-dimensional small-scale memory unit are concatenated to form a 2048-dimensional vector, which is input into a MLP. And then the MLP outputs a 1024-dimensional vector, which is activated by tanh and then updates the hidden layer information h_t through the output gate. And the effects of the two semantic scales are combined to obtain a better remote sensing image caption.

The improved network is calculated as follows at time t_0.

The initial value of the original memory unit can be computed using:

$$c_0 = f_0 \cdot c_{-1} + i_0 \cdot \sigma_h(W_c[h_{-1}, x_0] + b_c) \tag{1}$$

The initial value of the large-scale semantic memory unit can be computed using:

$$C_0 = f_0 \cdot C_{-1} + i_0 \cdot \sigma_h(W_C[h_{-1}, x_0] + b_C) \tag{2}$$

The initial values of the input gate and forget gate can be computed using:

$$i_0 = \sigma(W_i[h_{-1}, x_0] + b_i) \tag{3}$$

$$f_0 = \sigma\Big(W_f[h_{-1}, x_0] + b_f\Big) \tag{4}$$

x_0, c_{-1}, h_{-1} and C_{-1} can be computed using:

$$x_0 = [w_0; attention(v, h_{-1})] \tag{5}$$

$$c_{-1} = W_{c_{-1}} \cdot v + b_{c_{-1}} \tag{6}$$

$$h_{-1} = W_{h_{-1}} \cdot v + b_{h_{-1}} \tag{7}$$

$$C_{-1} = W_{C_{-1}} \cdot V_0 + b_{C_{-1}} \tag{8}$$

where v is the small-scale feature with a dimension of $14 \times 14 \times 512$, V_0 is the large-scale semantic information, and its dimension is the number of large-scale categories. w_0 is the initial word embedding vector with a dimension of 512.

The improved network is calculated as follows at time t:

The values of the input gate and forget gate can be computed using:

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{9}$$

$$f_t = \sigma\Big(W_f[h_{t-1}, x_t] + b_f\Big) \tag{10}$$

where the value of input x_t can be computed using:

$$x_t = [w_t; attention(v, h_{t-1})] \tag{11}$$

In addition, the values of the original memory unit and large-scale semantic memory can be computed using:

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \sigma_h(W_c[h_{t-1}, x_t] + b_c) \tag{12}$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \sigma_h(W_C[h_{t-1}, x_t] + b_C) \tag{13}$$

The large-scale semantic memory unit integrates the output V_0 from the large-scale VGG classification network and combines the attention-based image information with the information of the generated words at the previous moment in each update, thus combining the semantics of the two scales.

The hidden layer information h_t at time t can be computed using:

$$h_t = o_t \cdot \sigma_h(MLP(c_t + C_t)) \tag{14}$$

where σ is the sigmoid activation function, σ_h is the tanh activation function, each W is a weight, b is the bias, h_{t-1} is the hidden layer information of the previous moment, x_t is the current small-scale network input information, and C_t is the value of the large-scale semantic memory unit at the current time.

The network adds an MLP at the output-end and adds semantic information C_t from the large-scale semantic memory unit to the input of h_t to restore the comprehensive semantic information of samples and improve the accuracy of the semantic analysis of remote sensing objects. The large scale plays the role of the global perspective and directly participates in the output of small scale image captioning at different times t, thus ensuring that each generated word integrates the semantic information of the two scales. Therefore, the improved network organically combines different scales of remote sensing image features, which can ensure the scale effect and accuracy of the output multi-scale remote sensing image captioning.

4. Experiment and Analysis

4.1. Introduction of Test Areas and Samples

This study involves three areas: (1) the latitude range is 30°28′41″ to 30°32′29″ and the longitude range is 114°22′42″ to 114°28′11″ in the Wuhan Guanggu area; (2) the latitude range is 30°14′27″ to 30°17′22″ and the longitude range is 120°9′19″ to 120°6′33″ in the Hangzhou Xihu district, Shangcheng district, Xicheng district and Gongshu district; and (3) the latitude ranges from 34°43′2″ to 34°48′35″ and the longitude ranges from 113° 41′5″ to 113°46′37″ in the Huiji district, Guancheng district and Jinshui district of Zhengzhou. This study aims to optimize the recognition effect of small-scale samples by using large scale semantic information. Large scale samples contain information on small scale samples and their surrounding fields. Therefore, in this study, each small-scale sample corresponds to one large scale sample according to a 1:1 ratio between the small scale and large scale. The two scales of the samples are selected from three research areas. Each scale has 2814 patches, respectively, and there are a total of 5628 patches. To validate our proposal, the samples dataset is composed of 'fake conflict' samples and normal samples. There are 1510 'fake conflict' samples and 1304 normal samples in our dataset. To ensure sufficient samples in the validation set and training set and reasonable results, 1970 of samples were used as the training set and 844 of samples were used as the verification set for each scale. For each small-scale sample image in the training set, we manually add an image caption sentence.

4.2. Network Parameters

The networks involved in the experiment include two VGG networks, and an MS-LSTM network integrating large-scale semantic information. In addition, the original attention-based LSTM is used as a baseline. The training of the large-scale VGG and MS-LSTM are separate. The large-scale information input into MS-LSTM is the block scene category of the image, not the feature of it, so we need to train large-scale VGG in the block scene to strengthen its ability to classify and identify neighborhoods. When the training of large scale VGG is completed, we use the classification result of large-scale VGG together with the features extracted by small-scale VGG as the input of MS-LSTM, then we can train MS-LSTM.

(1) The small-scale image was input into the pre-trained VGG without training, and the Conv5_3 features were extracted. The size of the feature map was $14 \times 14 \times 512$, which was used as a part of inputs of the two LSTMs.

(2) The large-scale image was input into the large-scale VGG network for scene classification which requires training. When large-scale VGG training is optimal, the output of the large-scale VGG network is input into the MS-LSTM network as semantic information. Based on our previous experimental experience, the learning rate was 0.005, the batch_size was 40, and the number of iterations was 3000. In addition, the loss function of the large-scale VGG training is the Cross Entropy Loss, which can be computed using:

$$\text{loss} = -\sum_{i=1}^{n} y_i \log(\hat{y}_i) \tag{15}$$

where n is the batch_size, y_i is the ground truth of input, $\hat{y}_i$ is the output from network. In the training phase, we use the multi-scale segmentation principle and matching algorithm to match large-scale images to small-scale images, and manually label large-scale images, while no manual labeling is required in the prediction phase.

(3) The number of hidden layer neurons of both LSTM networks was 1024, the embedding dimension of the word vector was 512, the batch_size was 40, and the number of iterations was 60. The softmax function was used as the nonlinear activation function. As seen from Figure 4, learning curve keeps fluctuating without converging when learning rate is 0.01. When learning rate is 0.001 or 0.0001, the learning curves tend to be smooth after 1500 times of iteration. However, the learning

curve converge faster when learning rate changes into 0.001 than 0.0001. Briefly speaking, it is inappropriate to set learning rate too high or too low and we choose to set learning rate as 0.001.

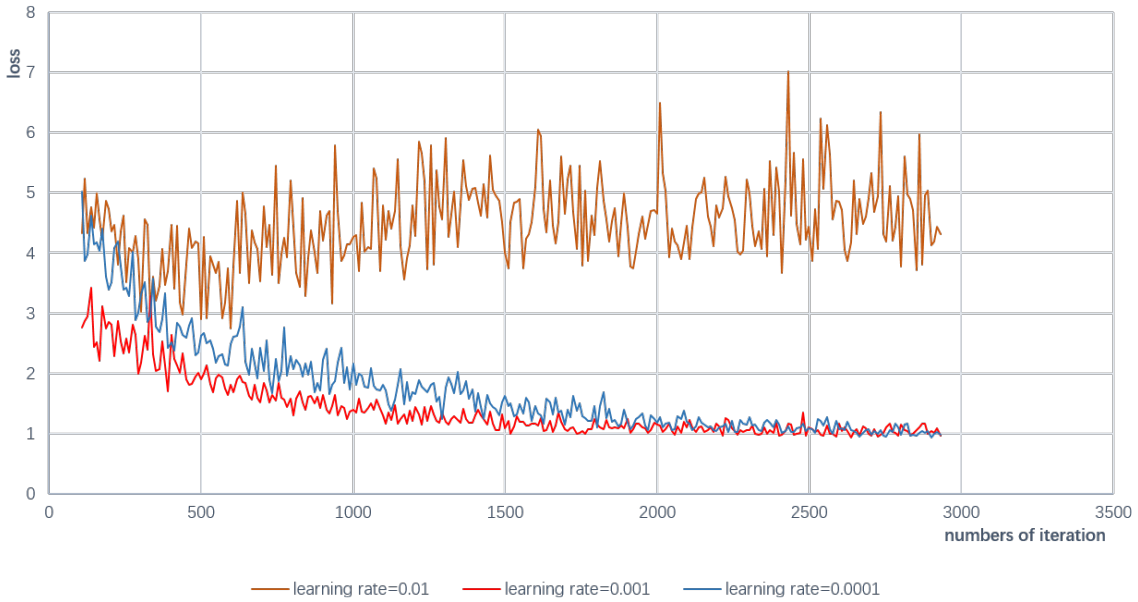

Figure 4. Learning curves demonstration when training in different learning rate. The *x*-axis represents the number of network iteration, and the *y*-axis represents the loss during network training.

4.3. Analysis of Experimental Results

To study the effect of large-scale information on recognition, the LSTM, MS-LSTM+ Ground Truth(GT) and MS-LSTM+VGG were used for comparison purposes. The LSTM uses the original LSTM network and only uses small-scale samples. The MS-LSTM+GT adds large-scale GT as the large-scale semantic information in the MS-LSTM network to show the role of large-scale semantic information in extreme cases. The data that are used are the small-scale samples and corresponding large-scale samples. The MS-LSTM+VGG has the same structure as the MS-LSTM+GT network, but the large-scale GT is replaced by the large-scale features that are extracted by the VGG as large-scale semantic information to verify whether the actual large-scale semantic information output by the VGG19 network can play the role of the global perspective; the data that are used are small-scale samples and large-scale samples. Since the LSTM network itself has the ability to learn sample image features, in order to verify the importance of large-scale semantics, the sample set contains 'fake conflict' samples and normal samples.

The general overview of the small-scale validation set samples is shown Table 1.

Table 1. Validation set of small-scale samples.

Samples	Fake Conflict Samples	Normal Samples	Total
Number	452	392	844

Among the 844 samples in the validation set, there are 452 'fake conflict' samples and 392 normal samples.

The large-scale samples were classified by using the VGG network, and the results are shown in Table 2:

Table 2. VGG classification confusion matrix.

Category	Residence	Service	School	Forest	Greenland	Total	Accuary
residence	238	6	5	0	7	256	92.97%
service	7	192	3	2	5	209	91.87%
school	1	9	112	0	9	131	85.50%
forest	0	3	0	110	2	115	95.65%
greenland	2	7	1	5	118	133	88.72%
total	248	217	121	117	141	844	
accuacy	95.97%	88.48%	92.56%	94.02%	83.69%		

It can be seen from the VGG classification confusion matrix that forest have the best effects with recall rates exceeding 92%. The residence and service results were followed with recall rates exceeding 90%. The recall rates of greenland and school are low at 88.72% and 85.50%, respectively.

We compared the evaluation indexes of the original LSTM network and the MS-LSTM network, and the results were shown in Table 3.

Table 3. Metrics of the LSTM and MS-LSTM for the validation set.

Model	BLEU_1	BLEU_2	BLEU_3	BLEU_4	METEOR	ROUGE_L	CIDEr
LSTM	0.80269	0.71196	0.64893	0.60019	0.43378	0.79564	5.17350
MS-LSTM	0.85907	0.77514	0.72409	0.68121	0.48250	0.85400	5.58904

According to the comparison, it can be seen that the scores of the various metrics of the MS-LSTM network that integrates the large-scale semantic information are all higher than those of the original LSTM network, but there is little difference. This result shows that if we only use the image captioning metrics, large-scale semantic information will play a limited role in the MS-LSTM network.

Next, we analyse the output image caption sentences from the network: misrecognition denotes that the centre object cannot be recognized, while recognition means that the centre object can be recognized; the analysis results are shown in Table 4:

Table 4. Performances of the 3 networks with the validation set.

Model	Recognition	Misrecognition	Total	Recognition Accuracy
LSTM	702	142	844	83.18%
MS-LSTM+GT	842	2	844	99.76%
MS-LSTM+VGG	805	39	844	95.38%

As seen from the above table, 702 samples can be correctly recognized from the captions that are generated by the original LSTM network, and 142 samples cannot be recognized, thereby resulting in an 83.18% accuracy rate.

In the captions that are generated by the MS-LSTM+GT network, only 2 samples could not be recognized, resulting in a 99.76% accuracy rate, which is a 16.58% increase compared to the original LSTM.

In the captions that are generated by the MS-LSTM+VGG network, 39 samples could not be recognized, and the accuracy rate is 95.38%, which is slightly lower than that of the MS-LSTM+GT.

To analyse the effect of the network on 'fake conflict' samples, we further analyse the results. The results from the 452 'fake conflict' samples were statistically validated, as shown in the Table 5.

Table 5. Performances of the 3 networks with respect to the 'fake conflict' samples in validation set.

Model	Recognition	Misrecognition	Total	Recognition Accuracy
LSTM	324	128	452	71.68%
MS-LSTM+GT	450	2	452	99.56%
MS-LSTM+VGG	413	39	452	91.37%

According to Tables 5 and 6, we can evaluate the performances of the three methods. Using the original LSTM network, 128 of the 452 'fake conflict' samples could not be recognized and 14 of 392 normal samples could not be recognized. After used the modified MS-LSTM+GT network, 2 of the 452 'fake conflict' samples could not be recognized and all normal samples could be recognized. With the improved MS-LSTM+VGG network, 39 of the 452 conflict samples could not be recognized and all normal samples could be recognized.

Table 6. Performances of the 3 networks with respect to normal samples in validation set.

Model	Recognition	Misrecognition	Total	Recognition Accuracy
LSTM	378	14	392	96.43%
MS-LSTM+GT	392	0	392	100%
MS-LSTM+VGG	392	0	392	100%

The experiments above show that recognition accuracy of the normal samples is relatively consistent in the original LSTM, MS-LSTM+GT and MS-LSTM+VGG. Meanwhile, the original LSTM network with the 'fake conflict' samples cannot obtain better accuracy. The improved MS-LSTM network can withstand the impacts of 'fake conflict' samples due to the addition of large-scale semantic information and still maintains the ideal performance. According to the results of the MS-LSTM+VGG network, the influence of the VGG classification error on large-scale semantic information needs to be further analysed. We counted 39 samples from the MS-LSTM+VGG network that incorrectly recognized the "fake conflict" samples of the validation set, and the results by category are shown in Table 7.

Table 7. Contrast between the VGG accuracy and the MS-LSTM+VGG accuracy.

Category	Residence	Service	School	Forest	Greenland	Total
number of fake conflict samples	85	122	91	77	77	452
Number of misrecognitions	7	8	11	3	10	39
recognition accuracy	91.76%	93.44%	87.91%	96.10%	87.01%	91.37%
VGG accuracy	92.97%	91.87%	85.50%	95.65%	88.72%	

It can be found that the classification accuracy of greenland and school is 87.01%, 87.91% while VGG accuracy of greenland and school is 88.72%, 85.50%. Meanwhile, the VGG has a good classification effect for categories such as forests, service, and the MS-LSTM+VGG has high recognition accuracy in those categories. The above results indicate that the quality of the large-scale information will affect the performance of the network. The better the effect of the VGG is, the more accurate the extraction of the large-scale information and the better performance of the improved MS-LSTM +VGG network, and vice versa.

As shown in Figure 5, the LSTM network might misrecognize center objects of small scale images. Although only one word different from manual caption, the semantic information could be totally changed. The classic metric BLEU is an index to evaluate the similarity of the whole sentence. According to Figure 4, the semantic information changes greatly when the central objects are different while the other parts of the sentence remain unchanged. However, the BLEU_1 index changes by 0.056 in that condition according to Table 3. Therefore, there are some limitations when using BLEU to assess the recognition of target objects.

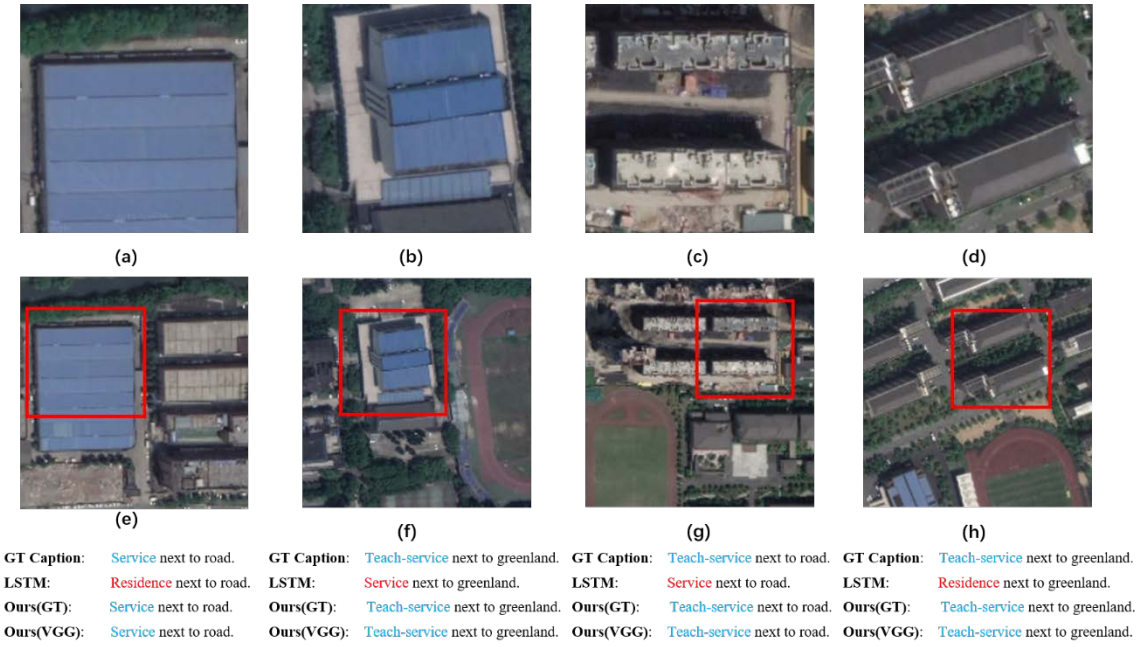

Figure 5. Three network generated image caption sentences and GT caption sentence. (**a**–**d**) are four small-scale samples; (**e**–**h**) are corresponding large-scale samples of (**a**–**d**).

4.4. Impact Analysis of Fake Conflict Samples

In order to further analyse the impact of the 'fake conflict' samples on the network, we redesigned the sample training set and the verification set. First, the samples are divided into two groups, which are the 'fake conflict' samples set and the normal samples set, and samples sets include small-scale samples and large-scale samples. In the 'fake conflict' samples set, there are 1510 small-scale 'fake conflict' samples, of which 1058 are allocated to the training set and 452 are allocated to the verification set. Small-scale normal samples set is composed of 1304 normal samples, of which 912 are allocated to the training set and 392 are allocated to the verification set. Certainly, there is one-to-one match between each large-scale sample and each small-scale sample. The two groups of samples were respectively input into the original LSTM for training, and the variation of the loss curve with the training times was obtained, as shown in Figure 6.

As shown in Figure 6, the loss decreases as the number of iterations increases. It can be seen that the deceleration of the loss of the normal samples is obviously faster than that of the 'fake conflict' samples. We conclude that 'false conflict' samples will make it difficult to accurately recognize various features, which retard the network's convergence.

To further learn the influence of the 'fake conflict' samples on the experimental results, we conducted four contrasting experiments. Two experiments include inputting only the 'fake conflict' samples into the original LSTM and MS-LSTM networks for training. The others include inputting only the normal samples into the above two networks for their respective training.

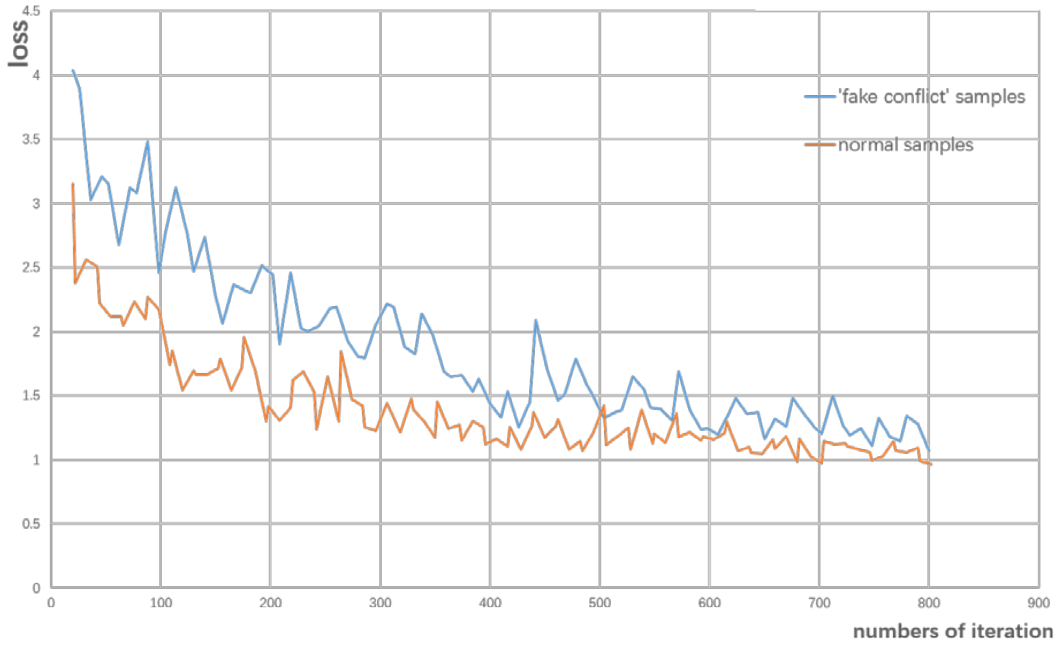

Figure 6. Loss curves when training network only with 'fake conflict' samples or normal samples. The x-axis represents the number of network iteration, and the y-axis represents the loss during network training.

The metrics of the four control experiments were compared, and the results are shown in the following table. The errors between the two groups of samples in different networks are also calculated in Table 8.

Table 8. Metrics of the LSTM and MS-LSTM in the validation set when using only the 'fake conflict' samples or normal samples.

Sample Set	Model	BLEU_1	BLEU_2	BLEU_3	BLEU_4	METEOR	ROUGE_L	CIDEr
'Fake conflict' samples	LSTM	0.7605	0.6594	0.5900	0.5338	0.4055	0.7562	4.3598
	MS-LSTM	0.8293	0.7262	0.6632	0.6116	0.4540	0.8224	5.5890
Normal samples	LSTM	0.8566	0.7814	0.7312	0.6949	0.4812	0.8508	6.2104
	MS-LSTM	0.8802	0.8145	0.7731	0.7372	0.5124	0.8851	6.8703

As seen in Table 8, it is obviously that metrics of MS-LSTM is better than LSTM no matter in normal samples or 'fake conflict' samples.

The mean absolute errors (MAEs) of the four BLEU_n indexes in the two types of samples were compared between the two networks. For 'fake conflict' samples, the MAE between the original LSTM network and the MS-LSTM in the four BLEU_n groups is 0.0717. For normal samples, the MAE between the original LSTM network and MS-LSTM in the four BLEU_n groups is only 0.0352. The performance of MS-LSTM is better than LSTM in recognition, the improvement is more obvious in 'fake conflict' samples.

As seen in Figure 7, all three networks can recognize center objects of small scale samples that means all network achieve good performance in normal samples. Then, we demonstrate the performance of three network in 'fake conflict' samples in Figure 7.

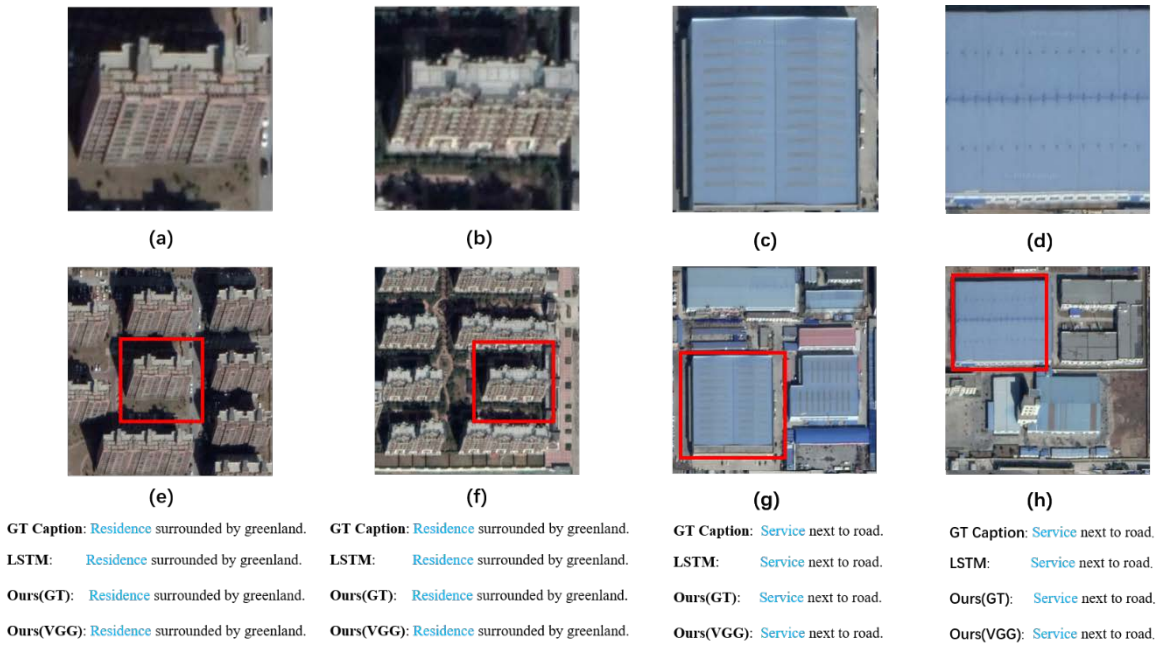

Figure 7. Demonstration of three networks' performance on normal samples. (**a**–**d**) are four small-scale samples; (**e**–**h**) are corresponding large-scale samples of (**a**–**d**); (**a**,**b**) are a pair of normal sample; (**c**,**d**) are a pair of normal sample. The following sentences are GT caption sentence for small scale samples, caption sentence generated by LSTM, caption sentence generated by MS-LSTM+GT and caption sentence generated by MS-LSTM+VGG as well.

We can see from Figure 8 that LSTM fail to recognize center objects of 'fake conflict' samples while MS-LSTM can correctly recognize them which shows the advantage of MS-LSTM in recognition.

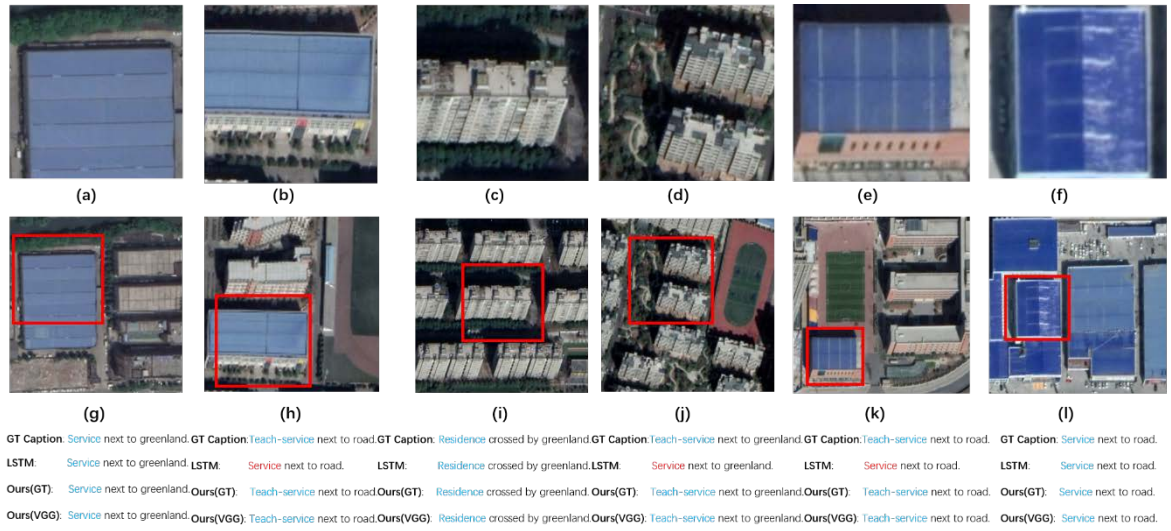

Figure 8. Demonstration of three networks' performance on 'fake conflict' samples. (**a**–**f**) are four small-scale samples; (**g**–**l**) are corresponding large-scale samples of (**a**–**f**); (**a**,**b**) are a pair of 'fake conflict' sample; (**c**,**d**) are a pair of 'fake conflict' sample; (**e**,**f**) are a pair of 'fake conflict' sample. The following sentences are GT caption sentence for small scale samples, caption sentence generated by LSTM, caption sentence generated by MS-LSTM+GT and caption sentence generated by MS-LSTM+VGG as well.

In the dataset that was composed entirely of 'fake conflict' samples, we trained the LSTM and MS-LSTM networks separately. The performances of the two networks for the validation set that only

includes 'fake conflict' samples were assessed. The results are shown in Table 9 based on the number of captions with errors.

Table 9. Performances of the 3 networks only using fake conflict samples.

Model	Recognition	Misrecognition	Total	Recognition Accuracy
LSTM	341	111	452	75.44%
MS-LSTM+GT	448	4	452	99.56%
MS-LSTM+VGG	412	40	452	91.15%

It can be seen from the table that 448 small-scale captions generated by MS-LSTM network can be correctly recognized among the 452 'fake conflict' samples in the validation set, accounting for 99.56% of the total. Meanwhile, recognition number of MS-LSTM+VGG is 412 and recognition accuracy is 91.15%. However, 341 of the small-scale caption sentences that are generated by the original LSTM network can be correctly recognized, accounting for 75.44% of the total. Therefore, MS-LSTM+GT and MS-LSTM+VGG excel in recognize 'fake conflict' samples.

To show 'fake conflict' samples' effect on networks' performance, we compare performances of two networks only uses 'fake conflict' samples or normal samples.

We further compare the quality of original LSTM with our MS-LSTM by training them only uses fake conflict samples or normal samples respectively. As shown in Table 10 and Figure 9, we can see that there is no significant difference between MS-LSTM and original LSTM in normal samples. However, MS-LSTM can recognize 448 of 452 fake conflict samples when original can only recognize 341 of 452 fake conflict samples. So MS-LSTM show great advantage in dealing with fake conflict samples.

Table 10. Validation performance when training in fake conflict samples or normal samples.

Samples	Only 'Fake Conflict' Samples	Only Normal Samples
Total	452	392
LSTM recognition	448	387
MS-LSTM recognition	341	392

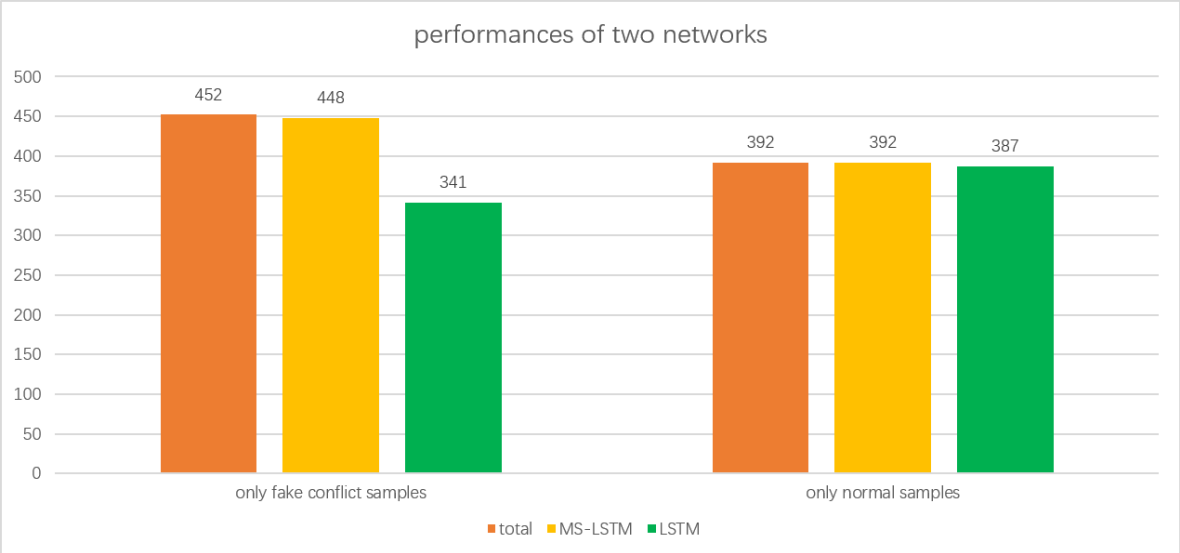

Figure 9. Validation performance when the network only uses fake conflict samples or normal samples. This figure demonstrates total number of two kinds of samples and the recognition number of network.

The experimental results are as follows:

(1) The fake conflict samples will interfere with the networks' training during the learning process. As a result, it is difficult for the network to accurately identify various features, and it is ultimately unable to effectively converge.
(2) The MS-LSTM network based on multi-scale semantics can restore the environmental and contextual information of samples by integrating large-scale semantic information into the LSTM, and it produces better results.
(3) Comparing the LSTM with the MS-LSTM, we can conclude that the improved network will obtain better image captioning performance if the accuracy of VGG is higher. Meanwhile, it also verifies the influence of large-scale semantic information on object recognition.

4.5. Model Comparison and Model Stability Analysis

We added two other remote sensing image captioning models for experimental comparison. All four models are the LSTM model without the attention [23], the SCA-CNN model [53], the Attention-based LSTM [46] model and our MS-LSTM model. We performed three sets of experiments on each of the four models. The data of the three experiments were the entire sample set, the 'fake conflict' samples set and the normal samples set. The experimental results are shown in Table 11, we denoted the LSTM model without the attention as No-Att, and the Attention-based LSTM model as LSTM:

Table 11. Results of each model in comparison experiments.

Sample Set	Model	BLEU_1	BLEU_2	BLEU_3	BLEU_4	METEOR	ROUGE_L	CIDEr
'Fake conflict' samples	No-Att	0.7679	0.6643	0.5924	0.5317	0.3975	0.7532	4.2327
	SCA-CNN	0.7705	0.6858	0.6131	0.5335	0.4012	0.7711	4.5674
	LSTM	0.7605	0.6594	0.5900	0.5338	0.4055	0.7562	4.3598
	MS-LSTM	**0.8293**	**0.7262**	**0.6632**	**0.6116**	**0.4540**	**0.8224**	**5.5890**
Normal samples	No-Att	0.8570	0.7803	0.7281	0.6882	0.4819	0.8528	6.1788
	SCA-CNN	0.8699	0.8105	0.7583	0.7068	0.5016	0.8665	6.4345
	LSTM	0.8566	0.7814	0.7312	0.6949	0.4812	0.8508	6.2104
	MS-LSTM	**0.8802**	**0.8145**	**0.7731**	**0.7372**	**0.5124**	**0.8851**	**6.8703**
All samples	No-Att	0.7955	0.7036	0.6423	0.5943	0.4303	0.7885	5.1077
	SCA-CNN	0.8102	0.7239	0.6545	0.5841	0.4407	0.8112	5.4706
	LSTM	0.8027	0.7120	0.6489	0.6002	0.4338	0.7956	5.1735
	MS-LSTM	**0.8591**	**0.7751**	**0.7241**	**0.6812**	**0.4825**	**0.8540**	**5.5890**

In order to compare the efficiency of the models, we also recorded the time spent on 60 iterations of the four models. The results are shown in Table 12:

Table 12. Comparison of training time of each model.

Model	Training Time
No-Att	24.5 min/60 iteration
SCA-CNN	26.1 min/60 iteration
LSTM	24.6 min/60 iteration
MS-LSTM	**25.5 min/60 iteration**

Comparing the other three models with our MS-LSTM network, we can find that the Bleu index of the MS-LSTM network are higher than the other three models, especially on the 'fake conflict' samples, our model's Bleu index is the most improved. At the same time, by comparing the training time of each model, it can be seen that the training time of our model has not increased significantly, indicating that our network has no additional burden

Then, in order to verify the stability of our MS-LSTM network, we randomly allocate the total samples to the training and validation sets in the same proportions as the previous experiments, and performed 10 independent Monte Carlo runs. We compared the Bleu_1, Bleu_2, Bleu_3 and Bleu_4 of the 10 experiments, where the trend is shown below (Figure 10). In the 10 experiments, the mean values of the Bleu_1, Bleu_2, Bleu_3 and Bleu_4 were 0.85609, 0.77091, 0.71764, 0.67160, respectively, and the standard deviations were 0.00220, 0.00382, 0.00495, and 0.00647, respectively, which proved the stability and reliability of the experimental results.

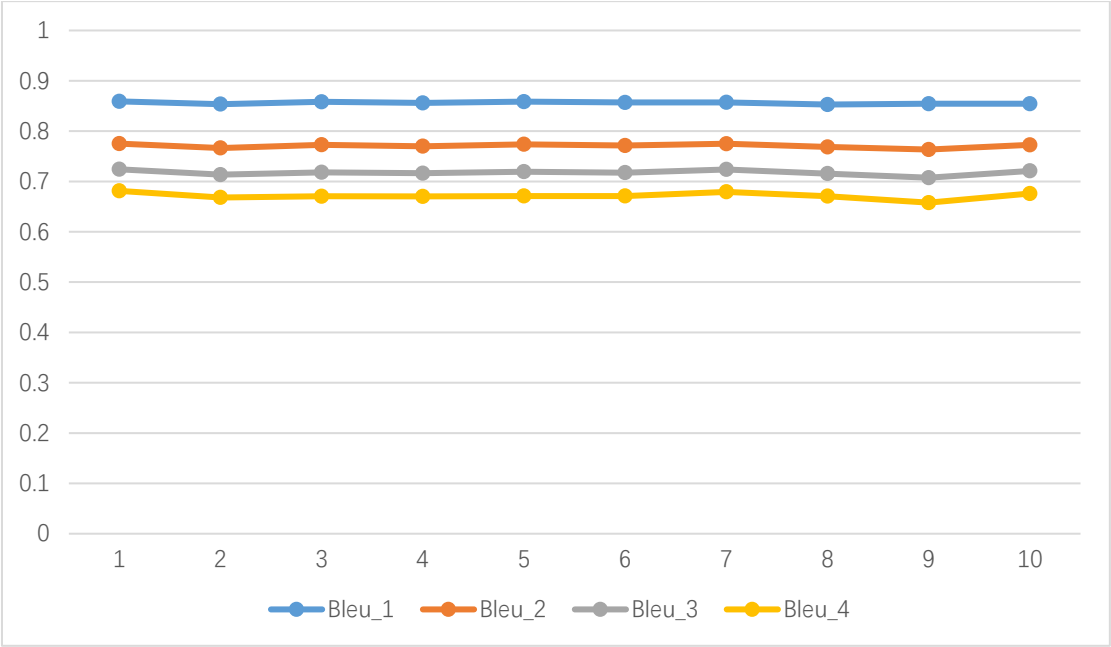

Figure 10. Bleu trend of the 10 experiments. It can be seen from the figure that in these experiments, the variation amplitudes of Bleu_1, Bleu_2, Bleu_3 and Bleu_4 are subtle, which can prove the randomness of the data distribution and the robustness of the algorithm.

5. Conclusions

In this paper, a multi-scale semantic long-short term memory network (the MS-LSTM) was proposed for solving the problem that the single-scale patch cannot achieve high efficiency and comprehensively and accurately caption images. We improved the original LSTM network by adding a large-scale semantic memory unit to store the impact of large-scale semantic information on currently generated words. At the output of the MS-LSTM, a multi-layer perceptron is used to fuse the two scales of information, which effectively resolves the impacts of the 'fake conflict' samples on the training process of the original LSTM. Relative to the original LSTM, the MS-LSTM abstracts and generalizes the semantic information of large-scale scenes by using the VGG network and then inputs it into the LSTM, which is more efficient and accurate than directly inputting the original large-scale image. The improved parallel network provides a global perspective for original LSTM network and also organically combines the advantages of multi-scale semantic information, deep neural networks and recurrent networks.

Our future work will focus on the following directions.

(1) The proposed work lacks universality and does not quantitatively analyses the choice of the spatial scale for remote sensing. We will analyse the consistency of the multi-scale region and multi-scale semantics and establish a quantitative spatial scale segmentation model.

(2) Exploring the combination mode of recurrent neural network and the extreme learning machines [54,55]. LSTM needs a large number of samples and iterations to finish the

training, the speed of convergence is sensitive to hyperparameters. Extreme learning machines, by contrast, have good generalization and extreme fast learning [56,57], and achieves the ability of semi-supervised and unsupervised learning [58]. Besides, in recent years, the extreme learning machines are also applied in remote sensing for the Spectral–Spatial Classification of Hyperspectral Imagery [59–61]. Therefore, our future work will focus on involving the extreme learning machines into remote sensing image captioning to achieve higher efficiency and stronger generalization.

Author Contributions: Wei Cui contributed toward creating the original ideas of the paper. Wei Cui conceived and designed the experiments. Dongyou Zhang and Xin He prepared the original data, performed the experiments and analysed the experimental data with the help of Meng Yao, Ziwei Wang and Yuanjie Hao and Wei Cui wrote and edited the manuscript. Dongyou Zhang, Xin He, Meng Yao, Jie Li, and Weijie Wu carefully revised the manuscript. Wenqi Cui and Jiejun Huang 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. Tobler, W.R. A Computer Movie Simulating Urban Growth in the Detroit Region. *Econ. Geogr.* **1970**, *46*, 234. [CrossRef]
2. Li, X.; Cao, C.; Chang, C. The First Law of Geography and Spatial-Temporal Proximity. *Chin. J. Nat.* **2007**, *29*, 69–71.
3. Sun, J.; Pan, Y.; He, R.; Liu, H. The enlightenment of geographical theories construction from the First Law of Geography and its debate. *Geogr. Res.* **2012**, *31*, 1749–1763.
4. Cui, W.; Wang, F.; He, X.; Zhang, D.; Xu, X.; Yao, M.; Wang, Z.; Huang, J. Multi-Scale Semantic Segmentation and Spatial Relationship Recognition of Remote Sensing Images Based on an Attention Model. *Remote Sens.* **2019**, *11*, 1044. [CrossRef]
5. 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 2016), Kunming, China, 6–8 August 2016; pp. 1–5.
6. 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]
7. 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]
8. Wang, B.; Lu, X.; Zheng, X.; Li, X. Semantic Descriptions of High-Resolution Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 1274–1278. [CrossRef]
9. 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]
10. 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]
11. Jeon, B.; Landgrebe, D.A. Classification with spatio-temporal interpixel class dependency contexts. *IEEE Trans. Geosci. Remote Sens.* **1992**, *30*, 663–672. [CrossRef]
12. Baatz, M.; Schape, A. An optimization approach for high quality multi-scale image segmentation. *Angew. Geogr. Inf.* **2000**, *12*, 12–23.
13. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
14. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Netw.* **2015**, *61*, 85–117. [CrossRef] [PubMed]
15. 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]
16. 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]

17. 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]
18. 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]
19. 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]
20. 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]
21. 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]
22. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. *arXiv* **2017**, arXiv:1703.06870.
23. 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.
24. Shao, Z.; Fua, H.; Lia, D.; Altanb, O.; Cheng, T. Remote sensing monitoring of multi-scale watersheds impermeability for urban hydrological evaluation. *Remote Sens. Environ.* **2019**, *232*, 111338. [CrossRef]
25. Shao, Z.; Pan, Y.; Diao, C.; Cai, J. Cloud Detection in Remote Sensing Image on Multiscale Features-Convolution Neural Network. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 4062–4076. [CrossRef]
26. 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]
27. Zhang, P.; Ke, Y.; Zhang, Z.; Wang, M.; Li, P.; Zhang, S. Urban Land Use and Land Cover Classification Using Novel Deep Learning Models Based on High Spatial Resolution Satellite Imagery. *Sensors* **2018**, *18*, 3717. [CrossRef] [PubMed]
28. Zhuang, S.; Wang, P.; Jiang, B.; Wang, G.; Wang, C. A Single Shot Framework with Multi-Scale Feature Fusion for Geospatial Object Detection. *Remote Sens.* **2019**, *11*, 594. [CrossRef]
29. Wang, Y.; Liang, B.; Ding, M.; Li, J. Dense Semantic Labeling with Atrous Spatial Pyramid Pooling and Decoder for High-Resolution Remote Sensing Imagery. *Remote Sens.* **2018**, *11*, 20. [CrossRef]
30. Liu, P.; Liu, X.; Liu, M.; Shi, Q.; Yang, J.; Xu, X.; Zhang, Y. Building Footprint Extraction from High-Resolution Images via Spatial Residual Inception Convolutional Neural Network. *Remote Sens.* **2019**, *11*, 830. [CrossRef]
31. Fu, K.; Li, Y.; Sun, H.; Yang, X.; Xu, G.; Li, Y.; Sun, X. A Ship Rotation Detection Model in Remote Sensing Images Based on Feature Fusion Pyramid Network and Deep Reinforcement Learning. *Remote Sens.* **2018**, *10*, 1922. [CrossRef]
32. Li, S.; Zhu, X.; Bao, J. Hierarchical Multi-Scale Convolutional Neural Networks for Hyperspectral Image Classification. *Sensors* **2019**, *19*, 1714. [CrossRef]
33. Lv, X.; Ming, D.; Lu, T.; Zhou, K.; Wang, M.; Bao, H. A New Method for Region-Based Majority Voting CNNs for Very High Resolution Image Classification. *Remote Sens.* **2018**, *10*, 1946. [CrossRef]
34. Yang, Z.; Tan, B.; Pei, H.; Jiang, W. Segmentation and Multi-Scale Convolutional Neural Network-Based Classification of Airborne Laser Scanner Data. *Sensors* **2018**, *18*, 3347. [CrossRef] [PubMed]
35. Feng, J.; Wang, L.; Yu, H.; Jiao, L.; Zhang, X. Divide-and-Conquer Dual-Architecture Convolutional Neural Network for Classification of Hyperspectral Images. *Remote Sens.* **2019**, *11*, 484. [CrossRef]
36. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
37. 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]
38. Wu, H.; Prasad, S. Convolutional Recurrent Neural Networks forHyperspectral Data Classification. *Remote Sens.* **2017**, *9*, 298. [CrossRef]
39. 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]
40. 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]
41. Seydgar, M.; Alizadeh Naeini, A.; Zhang, M.; Li, W.; Satari, M. 3-D Convolution-Recurrent Networks for Spectral-Spatial Classification of Hyperspectral Images. *Remote Sens.* **2019**, *11*, 883. [CrossRef]

42. Geng, J.; Wang, H.; Fan, J.; Ma, X. SAR Image Classification via Deep Recurrent Encoding Neural Networks. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 2255–2269. [CrossRef]
43. Ndikumana, E.; Ho Tong Minh, D.; 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]
44. Ma, A.; Filippi, A.; Wang, Z.; Yin, Z. Hyperspectral Image Classification Using Similarity Measurements-Based Deep Recurrent Neural Networks. *Remote Sens.* **2019**, *11*, 194. [CrossRef]
45. Karpathy, A.; Li, F. Deep Visual-Semantic Alignments for Generating Image Descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 664–676. [CrossRef] [PubMed]
46. 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. *arXiv* **2015**, arXiv:1502.03044.
47. Lu, J.; Xiong, C.; Parikh, D.; Socher, R. Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning. *arXiv* **2016**, arXiv:1612.01887.
48. Chen, J.; Han, Y.; Wan, L.; Zhou, X.; Deng, M. Geospatial relation captioning for high-spatial-resolution images by using an attention-based neural network. *Int. J. Remote Sens.* **2019**, *40*, 6482–6498. [CrossRef]
49. Zhang, X.; Wang, Q.; Li, X. Multi-Scale Cropping Mechanism for Remote Sensing Image Captioning. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium(IGARSS), Yokohama, Japan, 28 July–2 August 2019.
50. Wang, Y.; Lin, Z.; Shen, X.; Cohen, S.; Cottrell, G.W. Skeleton Key: Image Captioning by Skeleton-Attribute Decomposition. *arXiv* **2017**, arXiv:1704.06972.
51. Huang, Y.; Sun, Y. Judgement Characteristics and Quantitative Index of Suitable Block Scale. *J. South China Univ. Technol. (Nat. Sci. Ed.)* **2012**, *40*, 131–138.
52. Wang, R.; Zhang, Y. Taking history as a Lesson: Research on the evoiution of block Sizes from the perspective of typomorphoiogy. *Plan. Des.* **2018**, *10*, 81–85.
53. Chen, L.; Zhang, H.; Xiao, J.; Nie, L.; Shao, J.; Liu, W.; Chua, T.S. SCA-CNN: Spatial and Channel-wise Attention in Convolutional Networks for Image Captioning. *arXiv* **2016**, arXiv:1611.05594.
54. Huang, G.B.; Chen, L.; Siew, C.K. Universal approximation using incremental constructive feedforward networks with random hidden nodes. *IEEE Trans. Neural Netw.* **2006**, *17*, 879–892. [CrossRef] [PubMed]
55. Huang, G.B.; Zhu, Q.Y.; Siew, C.K. Extreme learning machine: A new learning scheme of feedforward neural networks. In Proceedings of the 2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541), Budapest, Hungary, 25–29 July 2004.
56. Huang, G.; Huang, G.; Song, S.; You, K. Trends in extreme learning machines: A review. *Neural Netw.* **2015**, *61*, 32–48. [CrossRef] [PubMed]
57. Mohammed, A.A.; Minhas, R.; Wu, Q.J.; Sid-Ahmed, M.A. Human face recognition based on multidimensional PCA and extreme learning machine. *Pattern Recognit.* **2011**, *44*, 2588–2597. [CrossRef]
58. Huang, G.; Song, S.; Gupta, J.N.D.; Wu, C. Semi-Supervised and Unsupervised Extreme Learning Machines. *IEEE Trans. Cybern.* **2014**, *44*, 2405–2417. [CrossRef] [PubMed]
59. Chen, C.; Li, W.; Su, H.; Liu, K. Spectral-Spatial Classification of Hyperspectral Image Based on Kernel Extreme Learning Machine. *Remote Sens.* **2014**, *6*, 5795–5814. [CrossRef]
60. Li, J.; Xi, B.; Du, Q.; Song, R.; Li, Y.; Ren, G. Deep Kernel Extreme-Learning Machine for the Spectral–Spatial Classification of Hyperspectral Imagery. *Remote Sens.* **2018**, *10*, 1–22. [CrossRef]
61. Salerno, V.M.; Rabbeni, G. An Extreme Learning Machine Approach to Effective Energy Disaggregation. *Electronics* **2018**, *7*, 235. [CrossRef]