remote sensing

MDPI

Article

Remote Sensing Monitoring of Grasslands Based on Adaptive Feature Fusion with Multi-Source Data

Weitao Wang [1], Qin Ma [1,*], Jianxi Huang [2], Quanlong Feng [2], Yuanyuan Zhao [2], Hao Guo [2], Boan Chen [2], Chenxi Li [2] and Yuxin Zhang [2]

[1] College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China; wangweitao1998@cau.edu.cn

[2] College of Land Science and Technology, China Agricultural University, Beijing 100083, China; jxhuang@cau.edu.cn (J.H.); fengql@cau.edu.cn (Q.F.); zhaoyuanyuan@cau.edu.cn (Y.Z.); guohaolys@cau.edu.cn (H.G.); SY20203213109@cau.edu.cn (B.C.); s20203213049@cau.edu.cn (C.L.); zyx_geo@cau.edu.cn (Y.Z.)

* Correspondence: maq782003@cau.edu.cn

Abstract: Grasslands, as an important part of terrestrial ecosystems, are facing serious threats of land degradation. Therefore, the remote monitoring of grasslands is an important tool to control degradation and protect grasslands. However, the existing methods are often disturbed by clouds and fog, which makes it difficult to achieve all-weather and all-time grassland remote sensing monitoring. Synthetic aperture radar (SAR) data can penetrate clouds, which is helpful for solving this problem. In this study, we verified the advantages of the fusion of multi-spectral (MS) and SAR data for improving classification accuracy, especially for cloud-covered areas. We also proposed an adaptive feature fusion method (the SK-like method) based on an attention mechanism, and tested two types of patch construction strategies, single-size and multi-size patches. Experiments have shown that the proposed SK-like method with single-size patches obtains the best results, with 93.12% accuracy and a 0.91 average f1-score, which is a 1.02% accuracy improvement and a 0.01 average f1-score improvement compared with the commonly used feature concatenation method. Our results show that the all-weather, all-time remote sensing monitoring of grassland is possible through the fusion of MS and SAR data with suitable feature fusion methods, which will effectively enhance the regulatory capability of grassland resources.

Keywords: grassland remote sensing monitoring; deep learning; multi-spectral and synthetic aperture radar data; convolutional neural network; adaptive feature fusion

Citation: Wang, W.; Ma, Q.; Huang, J.; Feng, Q.; Zhao, Y.; Guo, H.; Chen, B.; Li, C.; Zhang, Y. Remote Sensing Monitoring of Grasslands Based on Adaptive Feature Fusion with Multi-Source Data. *Remote Sens.* **2022**, *14*, 750. https://doi.org/10.3390/rs14030750

Academic Editor: Maria Laura Carranza

Received: 27 December 2021
Accepted: 31 January 2022
Published: 6 February 2022

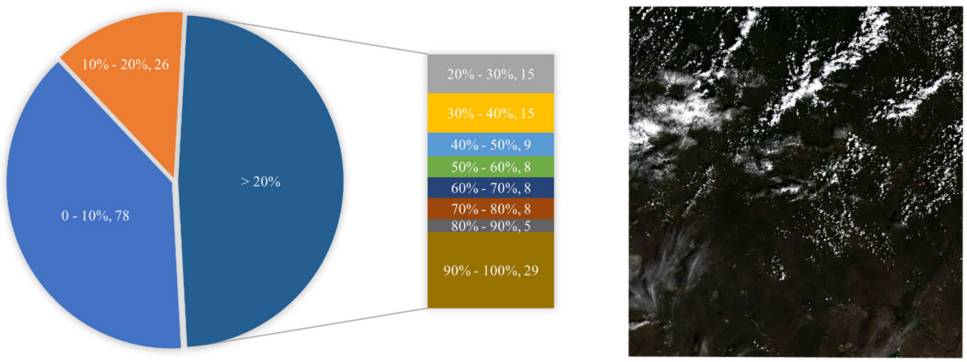
1. Introduction

Grasslands are an important part of terrestrial ecosystems. The Pilot Analysis of Global Ecosystems (PAGE) is a study conducted by the World Resources Institute, which "examines grassland ecosystems of the world using a large collection of spatial and temporal data" [1]. It shows that the world's grasslands cover a total area of 52.5 million km^2, accounting for 40.5% of the Earth's total terrestrial area (excluding Greenland and Antarctica), store 34% of the total carbon in terrestrial ecosystems, maintain 30% of net primary productivity and provide about 30–50% of the world's livestock products. Grasslands also support 25% of the world's population, and not only provide humans with products of direct economic value, such as meat and milk, but also serve extremely important ecological services, such as climate regulation, wind and sand control, water conservation, biodiversity conservation and so on. However, grassland ecosystems are facing a serious problem: land degradation [2,3]. More than 23% of the global terrestrial area is affected by degradation, with grassland being the main type of area affected. The degradation of grassland occurring in arid and semi-arid areas is a serious threat to ecological security, and is one of the main causes of frequent disasters, such as insect infestations and dust storms. Therefore, the

large-scale and high-precision monitoring of grasslands has important ecological value and research significance.

Among the existing land use and land cover (LULC) monitoring methods, remote sensing images, especially satellite multi-spectral (MS) images, are widely used due to their advantages of having wide detection ranges, short monitoring periods and large data volumes. For example, Phinn et al. [4] used multi-spectral images from Landsat-5 TM and Quickbird-2 to monitor seagrass biodiversity on the Eastern Banks in Moreton Bay, Australia, an area containing a range of seagrass species, cover and biomass levels. Lu et al. [5] used Landsat TM/ETM+ data to map and monitor land degradation in areas under human-induced stresses, such as the western Brazilian amazon. Wiesmair et al. [6] calculated the NDVI and $MSAVI_2$ using WorldView-2 multi-spectral images, then used these two indicators separately as predictors for vegetation cover in their random forest regression analyses. Robinson et al. [7] proposed a method for combining multi-resolution multi-spectral images and labels, resulting in a high-resolution (1 m) land cover map of the contiguous US, etc. [8–13]. However, we find that most of the multi-spectral images used in previous studies, including those mentioned above, are carefully processed public datasets or selected satellite images. They are often quite clear, ruling out the interference of clouds and shadows.

However, when our goal shifts to grassland monitoring, especially over large areas and long periods of time, satellite images with clouds and their shadows can be difficult to work around. We counted 201 scenes of Landsat8 OLI data covering the study area (48°25′ N, 116°49′ E–50°03′ N, 118°50′ E) in 2020, and found that nearly half of the scenes contained more than 20% clouds (Figure 1a).



(**a**) (**b**)

Figure 1. Landsat8 OLI data cloud coverage statistics (48°25′ N, 116°49′ E–50°03′ N, 118°50′ E, 2020) obtained using the USGS EarthExplorer system; (**a**) distribution of scenes with different cloud coverage, labeled in the format of "cloud coverage, number"; (**b**) example of a scene in our study area with slightly heavy clouds.

Synthetic aperture radar (SAR) data have all-weather and all-day capability, and thus are well-suited for this situation, which is commonly influenced by cloud cover. Additionally, existing studies have demonstrated that the proper use of VH (vertical–horizontal) and VV (vertical–vertical) polarization SAR data can be helpful for grass classification [14]. All of these indicate that the fusion of MS and SAR data will bring great convenience for remote sensing and grassland monitoring. Therefore, our research proposes a multi-source fused grassland remote sensing monitoring method. Different from existing works [15–17] that fuse two kinds of images at the data level (layer stack, Ehlers fusion, etc.), our method uses feature-level fusion, which fuses features extracted independently from each type of data to improve the ability to resist cloud interference and improve accuracy. Meanwhile, due to the high labor and time costs required for the pixel-by-pixel annotation of high-resolution remote sensing image data, patch-based methods

are widely used in deep learning-based remote sensing image analysis [18,19]. Considering that with single-size patches, it is difficult to accommodate various types of features with complex and different shapes and sizes, multi-size patches are used in our method. Finally, we select a typical experimental area containing various ecological subdivisions at the China–Mongolia–Russia border to verify the effectiveness of the method.

2. Materials and Methods

2.1. Study Area

In this study, the area between 48°25′–50°03′ N and 116°49′–118°50′ E at the border of China–Mongolia–Russia was selected as a typical study area, as shown in Figure 2. Specifically, the red boxes in Figure 2 represent the location of the study area. The study area has a semi-dry, humid climate, with relatively serious soil desertification, which is a critical area for wind and sand control, and a typical area for the remote sensing monitoring of grassland. Meanwhile, the study area is rich in ecological subdivisions, with a large area of typical grassland in the north, and Hulun Lake and its surrounding grassland and some sandy areas in the middle and south, which are representative of various grassland ecosystems. Therefore, the study of this region has considerable reference value and can be used to judge the effectiveness of our method.

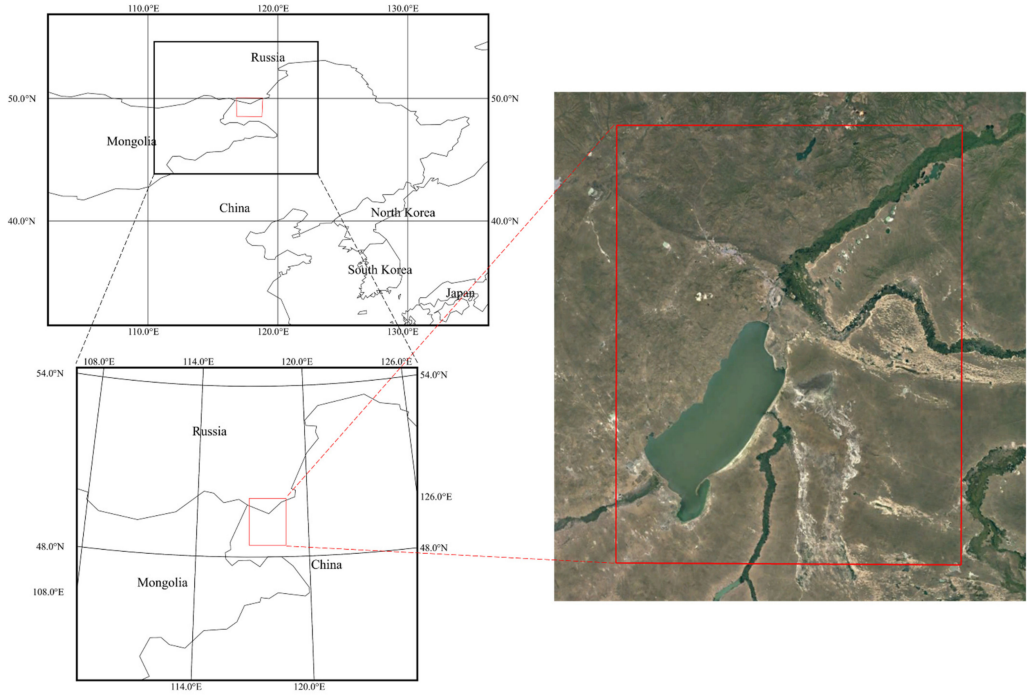


Figure 2. Location of the study area.

2.2. Dataset

2.2.1. Data Introduction and Pre-Processing

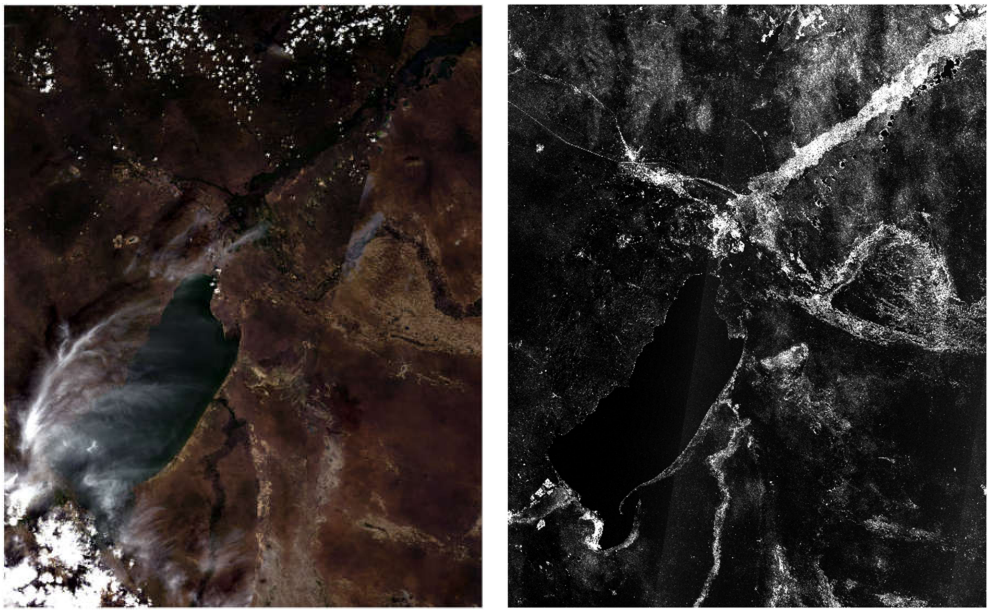
In this study, multi-spectral remote sensing images and SAR remote sensing images are used. The data collection time is selected from July to August, when grass growth is at its peak. In particular, MS data were collected on 9 July and 1 August 2020, and SAR data were collected on 26 July and 12 August 2020.

For the MS data, we use Landsat-8 Collection 1 Level-2 surface reflectance products provided by the United States Geological Survey (USGS), which are pre-processed officially. These data contain 7 different bands, including three RGB bands, one NIR band, two SWIR bands and one coastal band, with a 30m resolution.

For the SAR data, we use Sentinal-1B Level-1 Ground Range Detected (GRD) high-resolution products provided by the European Space Agency (ESA), which have a 10 m

resolution. These data are then processed using SNAP Desktop, which includes steps such as apply orbit file, GRD border noise removal, thermal noise removal, speckle filtering, radiometric calibration, terrain correction and geocoding to obtain VV- and VH-polarized data. After that, the data are resampled to 30m resolution to match the Landsat-8 data.

Finally, both data are cropped, mosaicked and layer overlaid using ENVI, referring to the study area. The images were also normalized using the 2% linear contrast stretch method. The final results of data pre-processing are shown in Figure 3.



(**a**) (**b**)

Figure 3. (**a**) MS image after data pre-processing (Landsat-8, 4, 3 and 2 bands); (**b**) MS image after data pre-processing (Sentinal-1, VH polarized).

2.2.2. Data Labeling

In the data labeling part, if the data are directly labeled as grassland and non-grassland, it is easy to annotate only simple samples, which means that the model cannot account for all kinds of ground objects in the non-grassland samples, and this eventually leads to the problem of a false high model evaluation index and poor robustness. Therefore, in this study, we adopt multi-category labeling, and label the features into five categories: grassland, farmland, water, man-made buildings and bare land.

During the labeling process, the remote sensing images were quantitatively examined according to the multi-temporal NDVI index curves and spectral curves in the Google Earth Engine [20]; meanwhile, the Google Earth super-high-resolution RGB images were combined as references to improve the accuracy of the labeling (Figure 4).

There are 1075 samples in the dataset, which contains 435 grassland samples, 109 farmland samples, 238 water samples, 126 man-made building samples and 167 bare land samples. The whole samples are randomly split into a training set, validation set, and test set according to the ratio of 6:1.5:2.5. The specific sample distribution is shown in Figure 5.

2.3. Our Method

2.3.1. Overview

In this paper, we propose an adaptive feature fusion method based on the attention mechanism in order to fuse MS and SAR images better and improve the classification accuracy and resistance against cloud interference. We trained the feature extraction and fusion network by extracting multi-size patches in multi-source remote sensing images as

training samples, and finally achieved the classification and remote sensing monitoring of grassland. Our method includes the following main components: (1) multi-source and multi-size patch extraction from selected sample points; (2) a feature extraction network based on a CNN (Convolutional Neural Network); (3) adaptive feature fusion and a classification module, as shown in Figure 6. It is necessary to clarify that in the "Adaptive feature fusion and classification" part of Figure 6, concatenation is the commonly used feature fusion method, and the SE-like and SK-like methods are the feature fusion methods we proposed. These three methods are used to fuse features separately.

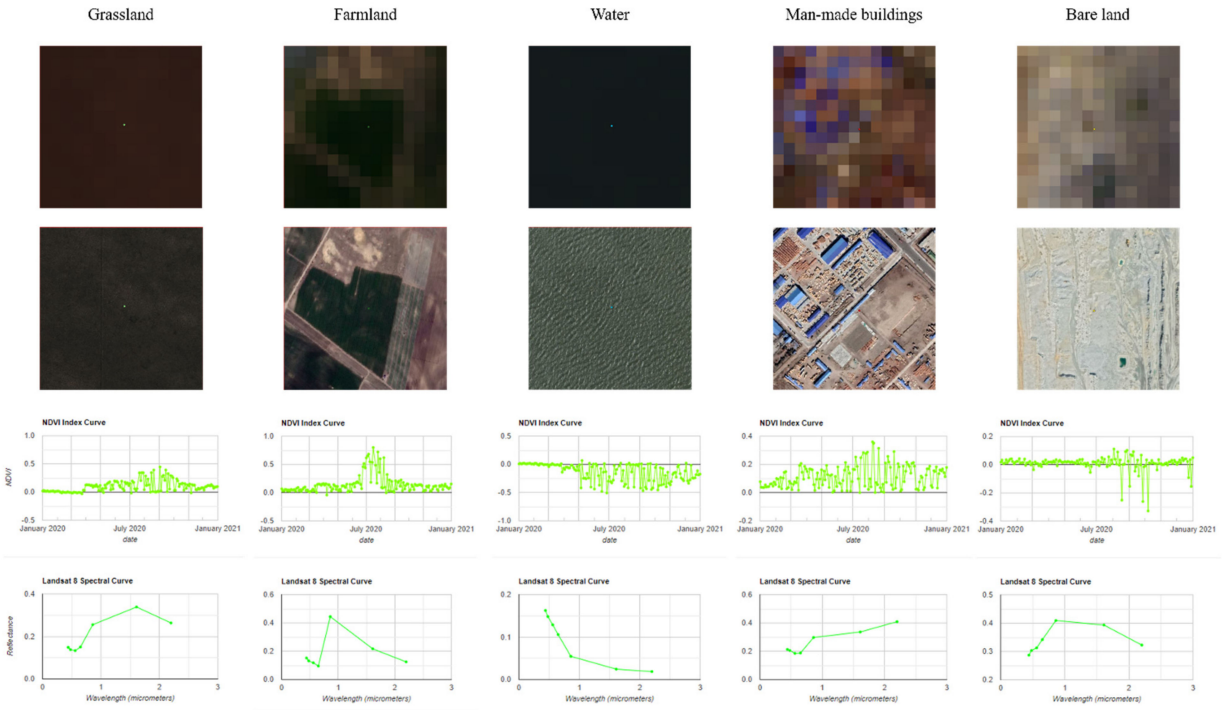


Figure 4. Examples of typical samples and methods to assist with labeling. (**Top**: MS remote sensing image from Landsat-8. **Middle**: super-high-resolution RGB image from Google Earth. **Bottom**: multi-temporal NDVI index curves and spectral curves generated by Google Earth Engine).

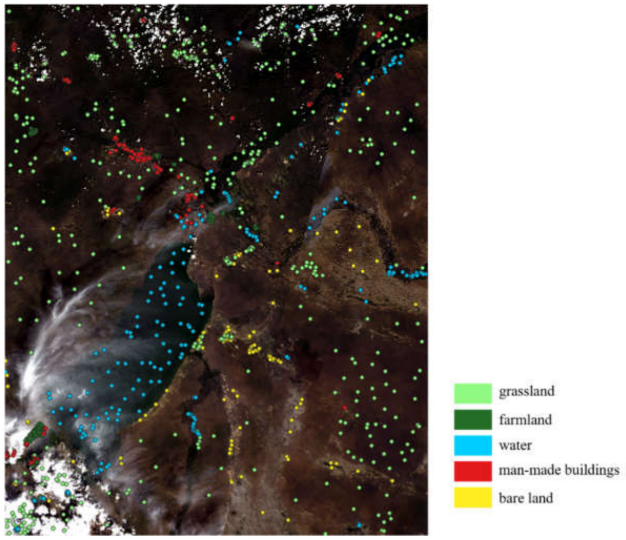


Figure 5. Landsat-8 image (4, 3 and 2 bands) and samples used in our study.

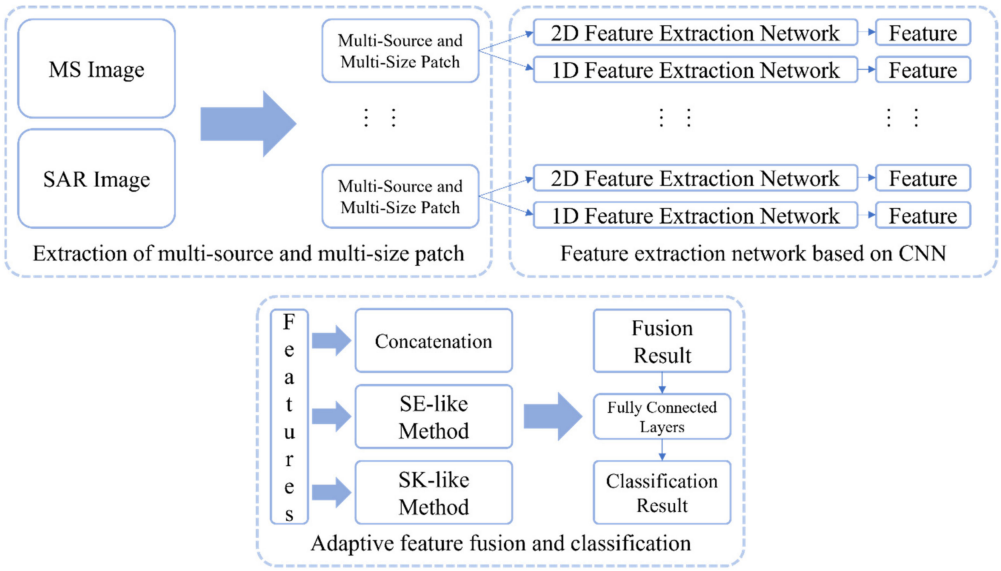


Figure 6. Methods and processes proposed in this article.

After the neural network model in the above method is trained, the classification effectiveness of the method is evaluated by a pre-obtained test set. Then, the multi-source remote sensing images of the whole study area are input line by line to obtain the classification results at pixel level, which is also the result of the remote sensing monitoring of the grassland in the study area.

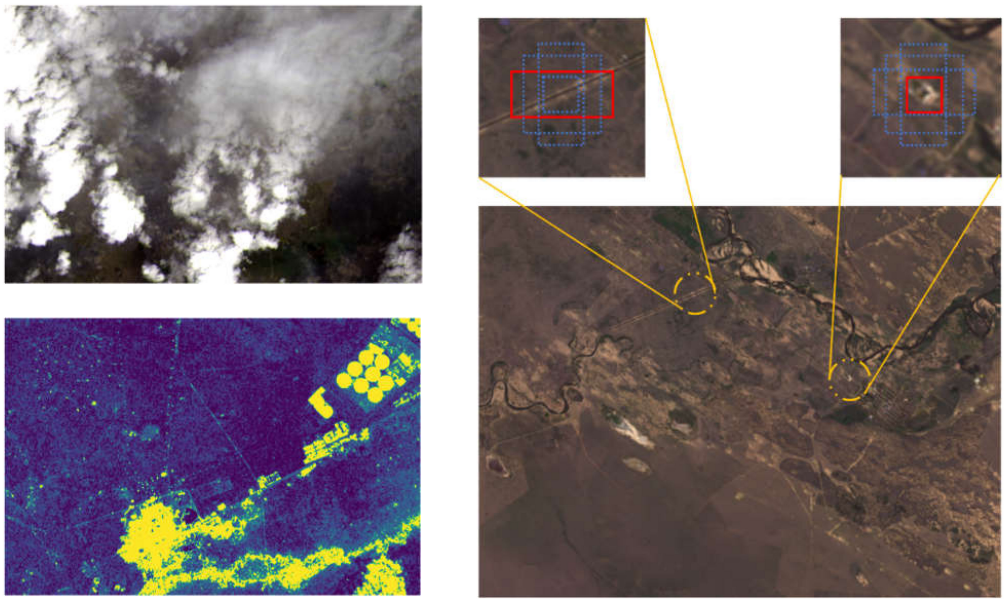
2.3.2. Extraction of Multi-Source and Multi-Size Patch

The patch-based method is widely used in the classification of remote sensing images. Compared with the pixel-based method, the patch-based method combines the sample point and its surrounding pixels within a certain range to a patch, and takes this patch as a complete sample. This method not only considers the spectral information of the sample point itself, but also supplements the texture information around the sample point, making full use of both the spectral and spatial information of remote sensing images, and it has achieved good results in the existing studies.

However, we found that the traditional patch-based methods tend to use a single fixed-size patch, such as 5×5, 11×11, etc. Considering that the actual ground objects vary in sizes and shapes, a single fixed-shape patch often cannot effectively extract the texture features of various types of ground objects at the same time. For example, rivers are less effective in square patches, larger patches tend to cause the ground object in it to be less correlated with the central sample points and smaller patches contain insufficient texture information. Therefore, we propose the method of multi-size patches. As shown in Figure 7, this method constructs multiple patches at the same time for a single sample point, and can generate rectangular patches of different shapes and square patches of different sizes according to the predefined parameters. For all types of ground objects, our method can generate at least one patch that can effectively characterize the central sample point (Figure 7b), which is convenient for subsequent feature extraction and fusion.

To simultaneously utilize the rich color information of hyperspectral images and the ability of SAR images to penetrate cloud, we also propose a multi-source patch extraction method. Considering that MS and SAR images are from two different data sources, the image coordinates corresponding to a certain geographic coordinate are not the same on these two types of images. To solve this problem, we achieved the direct conversion of geographic coordinates to image coordinates on any image containing geographic information using functions from the Geospatial Data Abstraction Library (GDAL). This method gives us the ability to directly extract patches from different data sources to ensure

that they represent the same sample point. At the same time, we used the above multi-size patch extraction method on different data.



(a) **(b)**

Figure 7. (**a**) Multi-source patch extraction and its application scenarios (**top**: MS image influenced seriously by cloud; **bottom**: SAR image); (**b**) multi-size patch extraction and its application scenarios (**top left**: linear ground objects with rectangular patch; **top right**: small ground objects with small square patch).

2.3.3. CNN-Based Feature Extraction Network

We constructed a convolutional neural network as a feature extraction sub-network to extract the features of each patch obtained in Section 2.3.2. Considering that the size of each patch does not exceed 32 × 32 at most, we controlled the depth of the convolutional neural network to prevent overfitting. Additionally, a one-dimensional convolutional neural network is proposed specifically for extracting the spectral information of the sample points themselves. Specifically, each feature extraction sub-network consists of two convolutional layers and one pooling layer using the Leaky-ReLU activation function. Each feature extraction sub-network outputs feature vectors of the same size, and the specific network structure is shown in Figure 8.

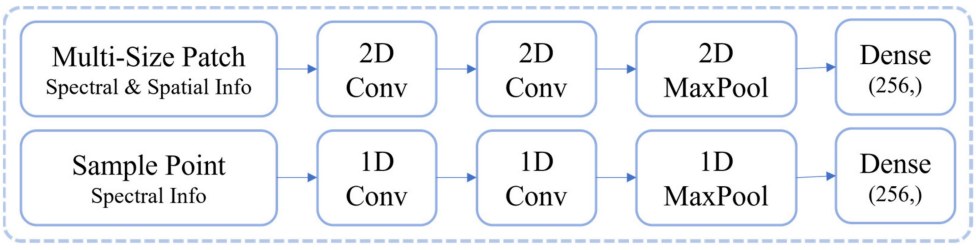


Figure 8. Structure of feature extraction network.

2.3.4. Adaptive Feature Fusion and Classification

As mentioned earlier, in our approach, multiple features extracted from different patches with different data sources and shapes are obtained for any sample point, so a data fusion method is needed to combine these features for the final grassland classification. In particular, for any given sample point, there are only a few sources and sizes of patches

that can effectively characterize the central sample point (Figure 7). This makes it necessary for the features to be given corresponding weights during the fusion process to enhance the influence of the effectively characterized patches and accordingly diminish the influence of the others. In this regard, we proposed two adaptive feature fusion and classification methods, the SE-like method and SK-like method, based on the attention mechanism.

Before presenting our methods, we will explain the traditional feature fusion method, feature concatenation. It obtains the new feature vector by concatenating the features of each patch, as shown in the following equation:

$$U = \left[X^1 X^2 \dots X^N\right],$$

where $X = \left[X^1, X^2, \dots, X^N\right]$ is the input feature vector and U is the result of feature concatenation.

The SE-like adaptive feature fusion method references the idea of a Squeeze-and-Excitation Block [21], in which the features from different patches are used as each channel of the fused feature vector, and squeeze and excitation processes are performed on this feature vector. The SE-like structure can train a one-dimensional weight vector with the same length as the original input feature, and multiply it with the fused feature vector by the channel scale to achieve adaptive feature fusion. Similar to the original structure, the adaptive feature fusion of the SE-like method consists of two major steps, F_{sq} and F_{ex}, which are mathematically expressed as follows:

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j),$$
$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \delta(W_1 z)),$$

where the F_{sq} process is a global average pooling (GAP) process, while the F_{ex} process contains two fully connected layers and the ReLU activation function. The structure of SE-like adaptive feature fusion method is shown in Figure 9, in which 4 is used as an example of the number of features extracted from patches.

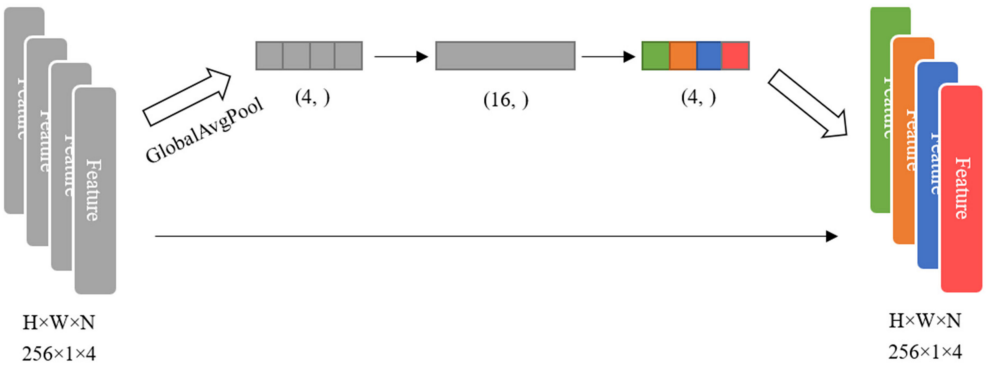


Figure 9. Structure of SE-like adaptive feature fusion method.

The SK-like adaptive feature fusion method is derived from the Selective Kernel Convolution method [22]. Unlike the original work, we replace the convolution results of differently sized convolution kernels with the feature vectors extracted by each feature extraction network as the input to the SE-Block. For the input feature vector $X = \left[X^1, X^2, \dots, X^N\right]$, the SK-Block first performs element-level summation as follows:

$$U = \sum_{s=1}^{N} X^s,$$

and the following steps are similar to the SE-like method, and contain a global average pooling F_{gp} procedure and an F_{fc} procedure consisting of a fully connected layer:

$$s_c = F_{gp}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j),$$
$$z = F_{fc}(s) = \delta(\mathcal{B}(Ws)),$$

where δ is the ReLU activation function and $\mathcal{B}$ represents the batch regularization. After that, further, we construct multiple attention vectors based on the number of input feature vectors and apply the Softmax operator among the attention vectors to ensure that the sum is 1 in the channel direction:

$$w_c^i = \frac{e^{W_c^i z}}{\sum_{j=1}^{N} e^{W_c^j z}}, i = 1, 2 \ldots N.$$

The final feature vector $V = [V_1, V_2, \ldots, V_c]$ is obtained by computing the original feature vector with the attention vector as follows:

$$V_c = \sum_{i=1}^{N} w_c^i \cdot X_c^i, \sum_{i=1}^{N} w_c^i = 1.$$

The structure of SK-like adaptive feature fusion method is shown in Figure 10, in which 4 is used as an example of the number of features extracted from patches.

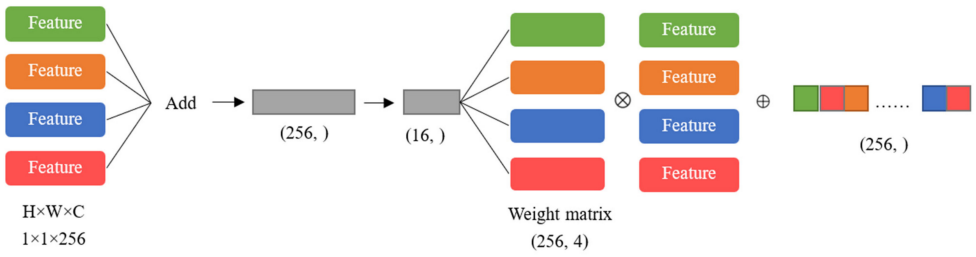


Figure 10. Structure of SK-like adaptive feature fusion method.

2.4. Evaluation Metrics

During training and testing, several evaluation metrics were used to compare the performance differences between the proposed methods. Specifically, overall accuracy and average F1-score were calculated for the training, validation and test set. The overall accuracy is the ratio of correctly classified samples to the total number of samples:

$$\text{overall accuracy} = \frac{1}{K} \sum_{k=1}^{K} \frac{TP + TN}{TP + TN + FP + FN}$$

where K stands for total number of classes, k is for each class, TP is the true positives, TN is the true negatives, FP is for false positives and FN is for false negatives. The F1-score is calculated from precision and recall, and then averaged across the categories to obtain the average F1-score:

$$\text{avg } F_1 \text{ score} = \frac{1}{K} \sum_{k=1}^{K} 2 \frac{pr}{p + r}$$

where K and k stand for the same as before, p stands for precision and r stands for recall. Additionally, the precision and recall can be calculated as follows:

$$p = {TP}/{TP + FP}$$

$$r = {TP}/{TP + FN}$$

where TP, FP and FN also stand for the same as before.

3. Results

3.1. Model Training

3.1.1. Training Process

Our model is trained and validated on a server equipped with an Intel(R) Xeon(R) Gold 6226R @ 2.90GHz processor and a dual NVIDIA(R) Tesla(R) V100 graphics processing unit (GPU).

For each of the methods proposed above, we trained and tested them on the dataset to find the most suitable method for conducting the remote sensing monitoring of grasslands. Figure 11 shows the variation of accuracy and loss with the training process, using the SK-like method as an example, while other methods behave similarly.

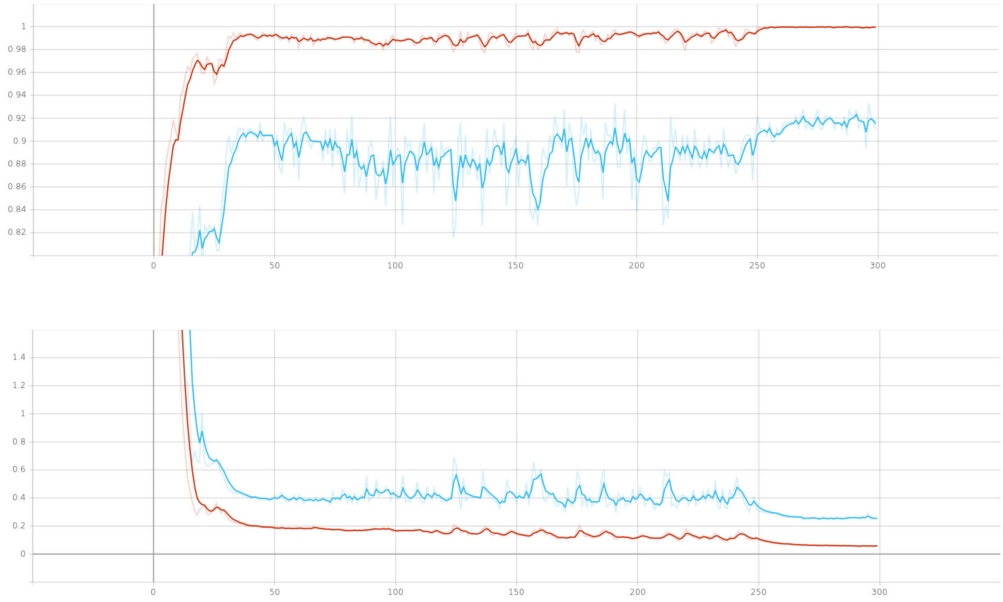


Figure 11. Training curves of accuracy and loss against number of epochs (red: training set; blue: validation set).

We can see that the model gradually converges in the early stage and stabilizes after 250 epochs. Both training and validation accuracy increase as the loss decreases, and eventually the accuracy of the training set stabilizes around 1, while the accuracy of the validation set remains around 90%.

3.1.2. Training Strategies

Learning rate decay and weight decay are used to enhance the training performance. Among them, learning rate decay helps to reduce the training time without affecting the training result, and weight decay can alleviate the overfitting problem. We set the initial learning rate and weight decay coefficients to 0.001 and 0.0005, respectively, and set them to one tenth of the original values in the first 30 and last 50 epochs, as shown in Figure 12.

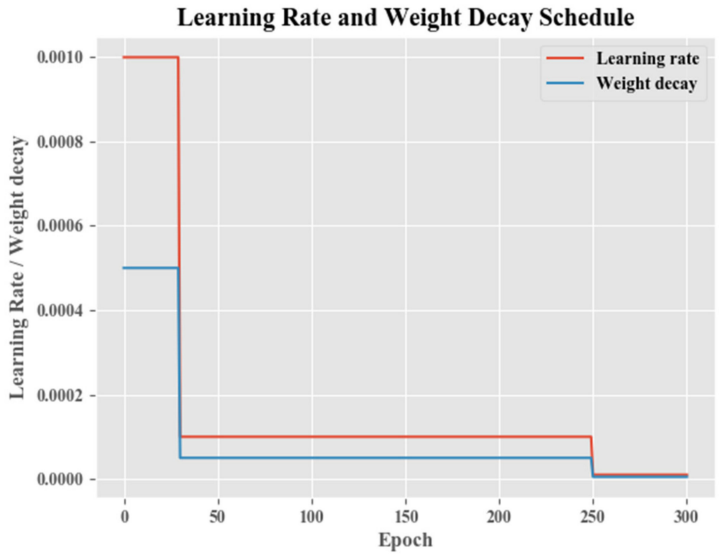
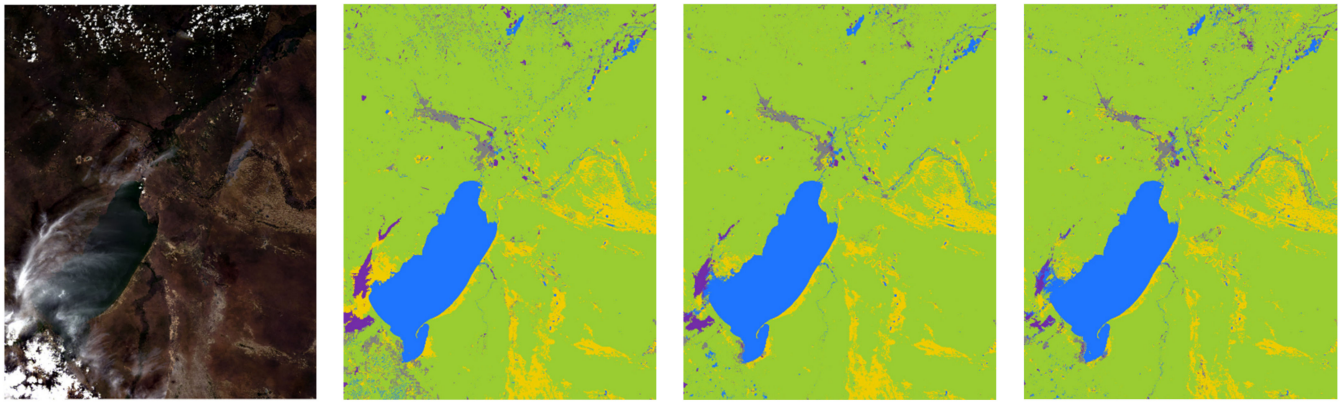


Figure 12. Learning rate and weight decay schedule curves.

3.2. Test Result

Each method and the data combination are evaluated via the test set, and the models trained by each method and data combination are also used to generate the classification results for the whole study area. The test results are shown in Figure 13.



(**a**) (**b**) (**c**) (**d**)

Figure 13. Test results of the whole study area using different data sources and methods. (**a**) Source MS image; (**b**) results without using SAR data; (**c**) results using concatenation; (**d**) results using SK-like model.

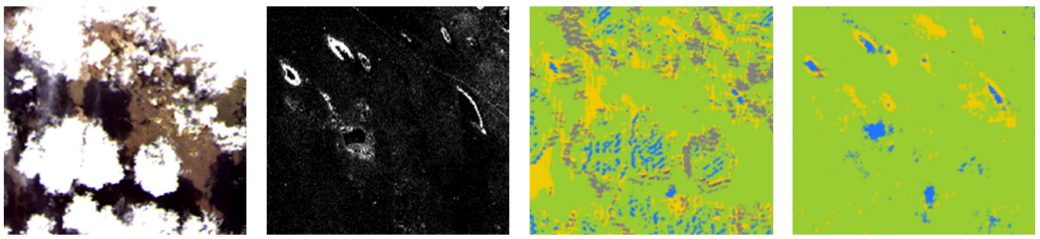
3.2.1. Comparison between Different Data Sources

We use the SK-like model proposed in the previous section to compare the effect of different data on the classification results. As shown in Table 1, neither the MS nor SAR data alone can achieve a satisfactory result. For the MS data only, the accuracy is 85.51% and the average F1-score is 0.80; for the SAR data only, the accuracy is 70.10% and the average F1-score is 0.63. At the same time, we can see that the integration of MS and SAR has a good effect, achieving 93.12% and 0.91 on accuracy and average F1-score, respectively, with an 7.61% and 0.11 improvement compared with using MS only.

Table 1. Overall accuracy and average F1-score for different data sources in test set.

Data Source	Accuracy	Average F1-Score
MS only	85.51%	0.80
SAR only	70.10%	0.63
MS + SAR	**93.12%**	**0.91**

The introduction of the SAR image provides the ability to correctly classify sections in cloud-covered areas. As shown in Figure 14, when using the MS image alone for remote sensing monitoring, the classification results are severely affected by clouds and fog, creating confusion in them. In contrast, the method integrating the MS and SAR data leverages the all-weather capability of SAR data, effectively solving this problem, and successfully distinguished areas such as grassland, farmland and water in cloud-covered areas, bringing considerable accuracy improvement.



(**a**) (**b**) (**c**) (**d**)

Figure 14. Different classification results using different data sources in an example cloud-covered area. (**a**) Source MS image; (**b**) source SAR image; (**c**) results without using SAR data; (**d**) results using SAR data.

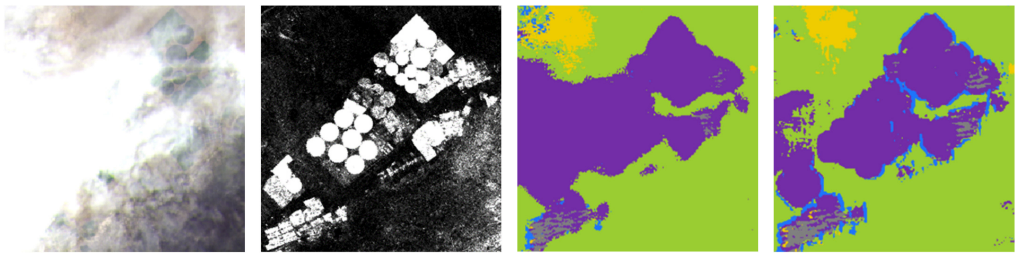
3.2.2. Comparison between Different Methods

Based on the results above, subsequent research will be conducted using the MS+SAR data to investigate the fusion ability of different adaptive feature fusion methods for features extracted from different data sources. The detailed results are shown in Table 2.

Table 2. Overall accuracy and average F1-score for different methods in test set.

Method	Single-Size Patch		Multi-Size Patch	
	Accuracy	Average F1-Score	Accuracy	Average F1-Score
Concatenation	92.10%	0.90	**92.44%**	**0.90**
SE-like model	91.07%	0.88	91.75%	0.89
SK-like model	**93.12%**	**0.91**	90.03%	0.87

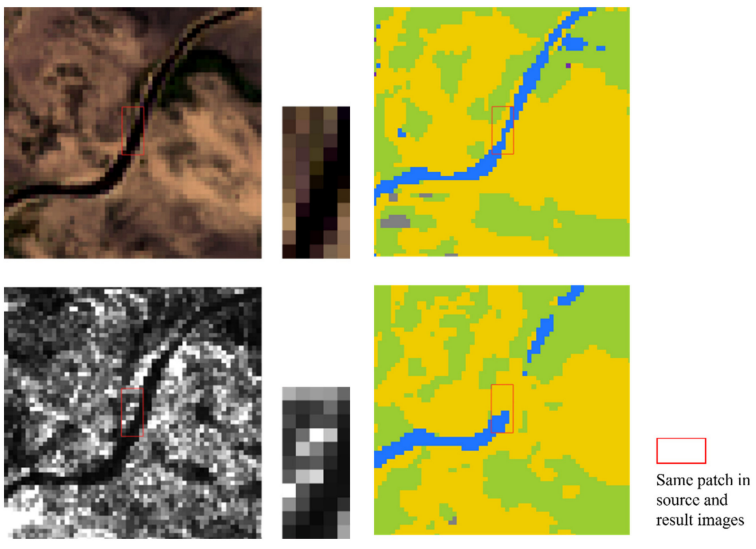
Firstly, we tested and compared the proposed method while using single-size patches. By analyzing the size of the ground objects and conducting several experiments, we selected 15×15 as the size of the patch. As we can see, under such conditions, the SK-like method obtains the best results, with 93.12% accuracy and a 0.91 average f1-score, which has a 1.02% accuracy improvement and a 0.01 average f1-score improvement compared with the commonly used feature concatenation method (Table 2). The difference of effectiveness between the SK-like method and the feature concatenation in fusing two types of data from different sources is also visually evident (Figure 15). The SK-like method we proposed can fuse MS and SAR data better with its capability to adaptively assign weights to features when classifying in scenarios where SAR data have advantages, such as with cloud-covered areas, water and buildings. The SE-like method, on the other hand, performs poorly, and ends up being less effective than traditional methods, and will not be discussed later in the paper.



(**a**) (**b**) (**c**) (**d**)

Figure 15. Different classification results using SK-like method and feature concatenation in an example area mainly contains farmland. (**a**) Source MS image; (**b**) source SAR image; (**c**) results using feature concatenation; (**d**) results using SK-like method.

Secondly, multi-size patches are used to test and compare each method. As introduced in Section 2.3.2, we need a set of sizes to generate patches that can fit more ground targets. After some preliminary experiments, this set of sizes were determined to be 5×11, 11×5, 15×15 and 31×31. Under this setting, we tested and found that multi-size patches showed a slight improvement in both feature concatenation and SE-like methods compare to single-size patches, with a maximum accuracy improvement of 0.68% (Table 2). However, surprisingly, the SK-like method used in conjunction with multi-size patches brought a serious performance degradation—a 3.09% loss in accuracy. By analyzing the generated multi-size patches and the final classification results, we found that one possible reason was that the MS and SAR data we used could not be matched precisely in spatial terms. Although we have performed image registration for these two types of data in the preprocessing stage, they still have a deviation of up to three to five pixels. It has no effect on larger size patches, such as the 15×15 patches used in the single-size patch method, but has a significant effect on smaller size patches, such as the 5×11 patches used in the multi-size patch method. In the case of using smaller size patches, small or linear ground objects may be included in the patches generated from the MS data, but disappear completely in the patches generated from the SAR data (Figure 16a).



(**a**) (**b**)

Figure 16. (**a**) Patches of size 5×11 extracted from the same sample point (same geographic location). (**Top**: MS image; **bottom**: SAR image; **right**: zoomed-in view of the patch). (**b**) Classification result of the example region. (**Top**: results using single-size patches with SK-like method; **bottom**: results using multi-size patches with SK-like method).

For the feature concatenation method, the features extracted from these patches containing misinformation only constitute a minority of all the features in the sample, and have little impact on the classification results. The method still benefits from the additional information brought in by multi-size patches and achieves a small performance gain. However, for the SK-like method, once the features extracted from these patches containing wrong information are given higher weights, the classification results of sample points will be wrong, bringing a serious performance degradation (Figure 16b).

To verify this conjecture, an additional experiment was carried out. Nine scenes of Sentinal-2 MSI Level 2A data were mosaiced and cropped to obtain a Sentinal-2 MS image of the study area. We used this image to replace the originally used Landsat-8 image, and used the original training and test sets for training and testing. The results show that without adjusting the samples based on the cloud coverage of Sentinal-2 image, the combination of multi-size patches and SK-like method can already achieve the same accuracy as single-size patches with the SK-like method, which is a significant improvement compared to the original 3.09% performance loss.

4. Discussion

In this study, we used MS and SAR data for multi-source data fusion and proposed an adaptive feature fusion method based on the attention mechanism. The method proposed in this study has achieved good results for LULC classification on medium-resolution remote sensing images.

4.1. Dataset and Methods Selection

Firstly, our results clearly showed the advantages of using multi-source data compare to single-source data. As shown in Table 1, the integration of MS and SAR data obtained 93.12% and 0.91 on accuracy and average F1-score, respectively, with a 7.61% and 0.11 improvement compared with using MS only, which means that the additional information from SAR data can be of great help in improving the LULC classification accuracy.

Secondly, the combination of the SK-like method that we proposed and single-size patches achieved the most excellent results, with the highest accuracy of 93.12% and an average F1-score of 0.91, demonstrating the advantages of the adaptive feature fusion methods in exploiting the additional information provided by SAR data. Meanwhile, the traditional feature concatenation method with multi-size patches also achieved good results, with the second highest accuracy of 92.44% and an average F1-score of 0.90, illustrating the potential of improving the classification accuracy with the additional information provided by multi-size patches.

In summary, we believe that satisfactory results can be obtained using the SK-like method we proposed in the case of the remote sensing monitoring of grasslands using Landsat-8 MS data and Sentinal-1 SAR data.

4.2. Problems Analysis

Although our method brings a decent performance improvement, there are still some problems. As we can see in Figure 13, the classification accuracy of cloud-covered areas is significantly lower than other regions. After analyzing the samples and the results, we believe that there are two main factors. The first reason is the lack of samples from cloud-covered regions. In order to allow the neural network to conduct the classification properly, similar to normal regions, samples from cloud-covered regions need to cover all five categories, including farmland, buildings and water. Since the percentage of cloud-covered areas is relatively small, the requirement for the density of samples increases accordingly. Therefore, we believe that it is necessary to pay attention to sample selection in future studies. The second reason is the interference of MS images in cloud-covered areas. MS images will introduce false feature information in cloud-covered areas, especially in thinner cloud areas, which can have a significant impact on classification accuracy.

Although the SAR image we used can mitigate this phenomenon, clouds and shadows still lead to a decrease in class probability, which has been supported by past studies [15].

Additionally, the combination of SK-like methods with multi-size patches is limited by the alignment problems that exist in the data itself, meaning their combination cannot achieve the expected results. A possible solution is to use same-source data, such as Sentinel-1 (SAR data) and Sentinel-2 (MS data) to replace the combination of Sentinel-1 (SAR data) and Landsat-8 (MS data) data used in this paper. The effect of this solution has been proven by an additional experiment, which can be tried in future studies.

5. Conclusions

In conclusion, the method proposed in this paper adaptively fuses multispectral data and synthetic aperture radar data using the attention mechanism, effectively solves the problem of cloud and shadow interference when using remote sensing images, such as Landsat, for grassland classification, and provides a technical tool for achieving all-time, all-weather and large-scale grassland remote sensing monitoring, which helps to improve the capability of remote sensing-based grassland monitoring. This means that we can further detect and mitigate problems, such as grassland degradation, all over the world, and protect the Earth's ecological environment.

Author Contributions: Conceptualization, validation and writing—review and editing, W.W. and Q.M.; methodology, software, formal analysis, investigation, resources and writing—original draft preparation and visualization, W.W.; data curation, W.W., C.L., B.C. and Y.Z. (Yuxin Zhang); supervision, Q.M., J.H., Q.F., Y.Z. (Yuanyuan Zhao) and H.G.; project administration, J.H.; funding acquisition, J.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China, grant number 2018YFE0122700. This research was also funded by the Provincial Natural Science Foundation Project, grant number ZR2021MC099.

Data Availability Statement: Data available on request due to privacy.

Acknowledgments: The authors acknowledge support from the National Key Research and Development Program of China (Grant number 2018YFE0122700). The authors also acknowledge support from the Provincial Natural Science Foundation Project (Grant number ZR2021MC099).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. White, R. *Pilot Analysis of Global Ecosystems: Grassland Ecosystems*; World Resources Institute: Washington, DC, USA, 2000; ISBN 9781569734612.
2. Liu, X.; Wang, Z.; Zheng, K.; Han, C.; Li, L.; Sheng, H.; Ma, Z. Changes in Soil Carbon and Nitrogen Stocks Following Degradation of Alpine Grasslands on the Qinghai-Tibetan Plateau: A Meta-analysis. *Land Degrad. Dev.* **2021**, *32*, 1262–1273. [CrossRef]
3. Li, X.; Perry, G.L.W.; Brierley, G.; Gao, J.; Zhang, J.; Yang, Y. Restoration Prospects for Heitutan Degraded Grassland in the Sanjiangyuan. *J. Mt. Sci.* **2013**, *10*, 687–698. [CrossRef]
4. Phinn, S.; Roelfsema, C.; Dekker, A.; Brando, V.; Anstee, J. Mapping Seagrass Species, Cover and Biomass in Shallow Waters: An Assessment of Satellite Multi-Spectral and Airborne Hyper-Spectral Imaging Systems in Moreton Bay (Australia). *Remote Sens. Environ.* **2008**, *112*, 3413–3425. [CrossRef]
5. Lu, D.; Batistella, M.; Mausel, P.; Moran, E. Mapping and Monitoring Land Degradation Risks in the Western Brazilian Amazon Using Multitemporal Landsat TM/ETM+ Images. *Land Degrad. Dev.* **2007**, *18*, 41–54. [CrossRef]
6. Wiesmair, M.; Feilhauer, H.; Magiera, A.; Otte, A.; Waldhardt, R. Estimating Vegetation Cover from High-Resolution Satellite Data to Assess Grassland Degradation in the Georgian Caucasus. *Mred* **2016**, *36*, 56–65. [CrossRef]
7. Robinson, C.; Hou, L.; Malkin, K.; Soobitsky, R.; Czawlytko, J.; Dilkina, B.; Jojic, N. Large Scale High-Resolution Land Cover Mapping with Multi-Resolution Data. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 12718–12727.
8. Selvaraj, S.; Case, B.S.; White, W.L. Discrimination of Common New Zealand Native Seaweeds from the Invasive Undaria Pinnatifida Using Hyperspectral Data. *J. Appl. Remote Sens.* **2021**, *15*, 024501. [CrossRef]

9. Pan, Y.; Pi, D.; Chen, J.; Chen, Y. Remote Sensing Image Fusion with Multistream Deep ResCNN. *J. Appl. Remote Sens.* **2021**, *15*, 032203. [CrossRef]
10. Su, H.; Peng, Y.; Xu, C.; Feng, A.; Liu, T. Using Improved DeepLabv3+network Integrated with Normalized Difference Water Index to Extract Water Bodies in Sentinel-2A Urban Remote Sensing Images. *J. Appl. Remote Sens.* **2021**, *15*, 018504. [CrossRef]
11. Yuan, X.; Shi, J.; Gu, L. A Review of Deep Learning Methods for Semantic Segmentation of Remote Sensing Imagery. *Expert Syst. Appl.* **2021**, *169*, 114417. [CrossRef]
12. Ball, J.E.; Anderson, D.T.; Chan, C.S. Comprehensive Survey of Deep Learning in Remote Sensing: Theories, Tools, and Challenges for the Community. *J. Appl. Remote Sens.* **2017**, *11*, 2609. [CrossRef]
13. Zhang, L.; Zhang, L.; Du, B. Deep Learning for Remote Sensing Data A Technical Tutorial on the State of the Art. *IEEE Geosci. Remote Sens. Mag.* **2016**, *4*, 22–40. [CrossRef]
14. Nicolau, A.P.; Flores-Anderson, A.; Griffin, R.; Herndon, K.; Meyer, F.J. Assessing SAR C-Band Data to Effectively Distinguish Modified Land Uses in a Heavily Disturbed Amazon Forest. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *94*, 102214. [CrossRef]
15. Solórzano, J.V.; Mas, J.F.; Gao, Y.; Gallardo-Cruz, J.A. Land Use Land Cover Classification with U-Net: Advantages of Combining Sentinel-1 and Sentinel-2 Imagery. *Remote Sens.* **2021**, *13*, 3600. [CrossRef]
16. Pereira, L.O.; Freitas, C.C.; Sant'Anna, S.J.S.; Reis, M.S. Evaluation of Optical and Radar Images Integration Methods for LULC Classification in Amazon Region. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 3062–3074. [CrossRef]
17. Khan, A.; Govil, H.; Kumar, G.; Dave, R. Synergistic Use of Sentinel-1 and Sentinel-2 for Improved LULC Mapping with Special Reference to Bad Land Class: A Case Study for Yamuna River Floodplain, India. *Spat. Inf. Res.* **2020**, *28*, 669–681. [CrossRef]
18. Sharma, A.; Liu, X.; Yang, X.; Shi, D. A Patch-Based Convolutional Neural Network for Remote Sensing Image Classification. *Neural Netw.* **2017**, *95*, 19–28. [CrossRef] [PubMed]
19. Liu, B.; Du, S.; Du, S.; Zhang, X. Incorporating Deep Features into GEOBIA Paradigm for Remote Sensing Imagery Classification: A Patch-Based Approach. *Remote Sens.* **2020**, *12*, 3007. [CrossRef]
20. Gorelick, N.; Hancher, M.; Dixon, M.; Ilyushchenko, S.; Thau, D.; Moore, R. Google Earth Engine: Planetary-Scale Geospatial Analysis for Everyone. *Remote Sens. Environ.* **2017**, *202*, 18–27. [CrossRef]
21. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [CrossRef] [PubMed]
22. Li, X.; Wang, W.; Hu, X.; Yang, J. Selective Kernel Networks. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 510–519.