remote sensing

MDPI

Article

Multi-Resolution Supervision Network with an Adaptive Weighted Loss for Desert Segmentation

Lexuan Wang [1], Liguo Weng [1,*], Min Xia [1], Jia Liu [1] and Haifeng Lin [2]

[1] Jiangsu Collaborative Innovation Center on Atmospheric Environment and Equipment Technology, Nanjing University of Information Science and Technology, Nanjing 210044, China; 20181223069@nuist.edu.cn (L.W.); xiamin@nuist.edu.cn (M.X.); liujia@nuist.edu.cn (J.L.)

[2] College of Information Science and Technology, Nanjing Forestry University, Nanjing 210037, China; haifeng.lin@njfu.edu.cn

* Correspondence: 002311@nuist.edu.cn

Abstract: Desert segmentation of remote sensing images is the basis of analysis of desert area. Desert images are usually characterized by large image size, large-scale change, and irregular location distribution of surface objects. The multi-scale fusion method is widely used in the existing deep learning segmentation models to solve the above problems. Based on the idea of multi-scale feature extraction, this paper took the segmentation results of each scale as an independent optimization task and proposed a multi-resolution supervision network (MrsSeg) to further improve the desert segmentation result. Due to the different optimization difficulty of each branch task, we also proposed an auxiliary adaptive weighted loss function (AWL) to automatically optimize the training process. MrsSeg first used a lightweight backbone to extract different-resolution features, then adopted a multi-resolution fusion module to fuse the local information and global information, and finally, a multi-level fusion decoder was used to aggregate and merge the features at different levels to get the desert segmentation result. In this method, each branch loss was treated as an independent task, AWL was proposed to calculate and adjust the weight of each branch. By giving priority to the easy tasks, the improved loss function could effectively improve the convergence speed of the model and the desert segmentation result. The experimental results showed that MrsSeg-AWL effectively improved the learning ability of the model and has faster convergence speed, lower parameter complexity, and more accurate segmentation results.

Keywords: multi-resolution supervision; adaptive weighted loss; multi-scale; fusion; deep learning

Citation: Wang, L.; Weng, L.; Xia, M.; Liu, J.; Lin, H. Multi-Resolution Supervision Network with an Adaptive Weighted Loss for Desert Segmentation. *Remote Sens.* **2021**, *13*, 2054. https://doi.org/10.3390/rs13112054

Academic Editors: Carlos López-Martínez, Ramona-Maria Pelich and Minh-Tan Pham

Received: 18 April 2021
Accepted: 20 May 2021
Published: 23 May 2021

1. Introduction

Desertification is a land degradation phenomenon characterized by wind–sand activities in arid and semi-arid areas due to the human–nature imbalance. It is a positive feedback process of environmental instability [1]. A comprehensive, macroscopic, and scientific grasp of the spatial distribution pattern and dynamic change information of desert land types is the basis for preventing and/or controlling desertification [2]. The feature types in desert areas are complex, manual field mapping statistics or visual interpretation consumes time and energy, and the information of dynamic large-scale areas cannot be reflected quickly and accurately [3]. In recent years, satellite remote sensing technology has been developing rapidly, making it possible to obtain remote sensing images in desert areas with low cost, fast speed, and high accuracy [4]. However, due to the complexity of remote sensing image features, there is no universal method for image recognition [5]. Light, water, and other external factors have different effects on the image features of different desert land types, making it difficult to identify land types and distinguish boundaries [6]. Therefore, desert remote sensing image recognition is still a challenging task.

Most of the existing remote sensing image-recognition methods have used sliding windows to extract spectral features and texture features [7,8]. Pi et al. [9] proposed the

desert grassland classification network (DGC) and three-dimensional convolutional neural network (3D-CNN) models to identify desert and grassland. Moghaddam et al. [10] used a multi-layer perceptron (MLP) to classify Isfahan desert images and obtained the land cover map of the Sejzy area. Ge et al. [11] used the artificial intelligence method (ANN), random forest (RF), support vector machine (SVM), and k-nearest neighbor method (KNN) to analyze seven different land cover types in China's dengkou oasis. These methods made full use of the information contained in remote sensing images and effectively improved the land classification accuracy of high-resolution images, but there were still some problems such as time-consuming calculation and inaccurate edge segmentation results. Researches showed that image segmentation methods could better avoid the above problems [5,12].

Traditional desert segmentation methods such as mathematical morphology and threshold segmentation methods were mainly based on remote sensing technology (RS) and geographic and information system (GIS) technologies. These methods' performance depended on many threshold parameters that should be elaborately given. The threshold parameters usually vary in different images, so the traditional methods could only work in a small range of data and cannot be validated in complex circumstances [13,14]. Remote sensing image segmentation methods based on a single path encoder–decoder network to solve pixel-to-pixel prediction have achieved good results [15,16]. Li et al. [17] proposed a land-use segmentation model based on deep learning, which improved the performance of the model by using residuals [18] and multi-scale module ASPP [19]. Ulmas et al. [20] used a deep learning model based on U-Net to identify the land cover type. The features record in desert images usually presents multi-scale characteristics. The extraction and fusion of multi-scale features can help improve the learning ability and the segmentation result [21,22]. The existing single-branch segmentation model did not fully consider the feature information of different scales, and the existing multi-scale feature fusion model requires a lot of computation [23,24]. In order to quickly and accurately segment desert remote sensing images, it is still necessary to further strengthen the multi-scale information fusion effect [25], reduce the number of parameters, and speed up model convergence.

In the field of person re-identification and object detection, the use of deep supervision can effectively improve the network performance [26,27]. When applying this idea with multi-resolution learning to the segmentation task, it is important to achieve balanced loss by considering different contribution of each resolution tasks [28]. Reducing the weight for difficult tasks and increasing the weight for easy tasks can effectively accelerate the convergence speed of training and prevent the model from falling into the local minimum [16]. However, the existing balance loss methods mostly adopted fixed balance parameters or adjust the balance parameters only according to the difficulty of a single task [29].

In view of the above problems, we consider the application of desert remote sensing with the characteristics of large image size, large-scale change, and irregular location distribution of surface objects [30]. This paper regarded the outputs of different branches as different optimization tasks and proposed a multi-resolution supervision network (MrsSeg) with an adaptive weighted loss function (AWL) to automatically segment desert remote sensing images. First, a lightweight backbone was used to extract different-scale features, then a multi-resolution fusion module was adopted to fuse the local and global informations, and finally, a multi-level fusion decoder was used to aggregate and merge the object features at different levels to get the desert segmentation result. An improved adaptive weighted loss function was also designed to automatically optimize the training process. The main contributions of this work are as follows: (1) This paper took the segmentation results of each resolution as an independent optimization task and proposed a multi-resolution supervision network (MrsSeg) to better promote the feature fusion process. (2) According to the characteristics of desert images, a specialized multi-resolution aggregation module was proposed to better recover the detailed information of desert segmentation results by aggregating features from low to high resolution. (3) In order to improve the efficiency of the multi-resolution supervision network, an adaptive weighted loss function (AWL) was designed. By giving priority to the easy tasks, the improved

loss function could effectively improve the convergence speed of tranining and the desert segmentation result. (4) A new desert image dataset was collected, including desert, gobi, oasis, and river. The experimental results on the self-constructed dataset showed that the proposed model obtained better performance in the desert segmentation task compared with existing approaches.

2. Materials and Methods

Desert remote sensing images are usually characterized by large image size, large scale change, and irregular location distribution of surface objects [30]. In order to quickly and accurately segment desert images, this paper proposes a multi-resolution supervision network to effectively fuse local information and global information, so as to improve the desert segmentation effect. According to the characteristics of multi-resolution outputs of the network, an adaptive weighted loss function was proposed to further improve the segmentation performance of the network.

2.1. Multi-Resolution Supervision Network

In the existing remote sensing image segmentation methods, the feature fusion model is often used to extract multi-scale features and preserve spatial details [31]. However, it can be seen from Figure 1 that the mutil-branch model (Figure 1a) was short of dealing with high-level features combination of parallel branches, the lack of feature communication between parallel branches led to insufficient learning ability, and the additional branches on high-resolution images limited the acceleration of training speed. Commonly used pyramid feature map fusion methods include image pyramid [32], feature pyramid [33], and spatial pyramid pool (SPP) [34] module (Figure 1b).The SPP module uses shallow semantic information to enhance high-level features by extracting high-resolution context semantics and enhancing receptive fields. However, the segmentation results of this method are limited to the feature layer where the spatial pooling pyramid is located, and implementing the SPP module is usually time-consuming. The feature pyramid (Figure 1c) fuses the deep semantic information into the shallow network layer by layer through the top-down path. This feature fusion method of aggregating context information not only increases the local information extraction ability of the deep neural network but also makes the shallow network have certain deep-level semantic information.

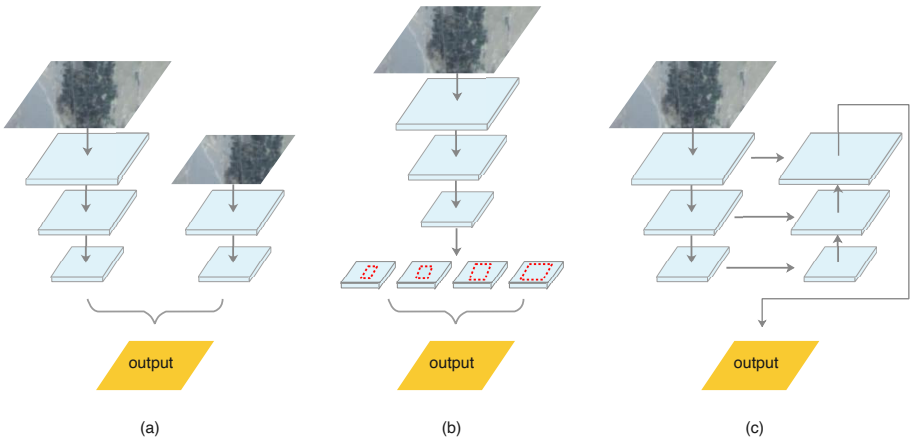

Figure 1. Segmentation model structure comparison: (**a**) multi-branch network; (**b**) spatial pyramid pooling network; (**c**) feature reuse network.

Inspired by the above ideas, the structure of the improved multi-resolution segmentation network in this paper was shown in Figure 2. The structure aimed to better extract and fuse local and global information through supervised training among multiple branches so as to improve the segmentation ability.

The MrsSeg was a lightweight desert image segmentation method that combined multi-resolution semantics to encode features. The whole network could be divided into

three parts, among which the encoder module consisted of a lightweight backbone network and multi-resolution fusion modules and the decoder module was designed as a simple and effective up-sampling module that combined low-level and high-level features.

The overall structure of the MrsSeg-AWL was illustrated in Figure 2. First, we used the pre-trained MobilenetV2 [35] as a lightweight backbone network to obtain different levels features of desert image. Then, we used the multi-resolution fusion module to fuse the multi-level semantic information to improve the feature representation ability of the network, and then adopted the multi-resolution supervised training to improve the feature extraction ability of each branch to promote feature fusion ability. Finally, the segmentation result with the same size as the input image was obtained by up-sampling the feature map of the multi-level fusion decoder.

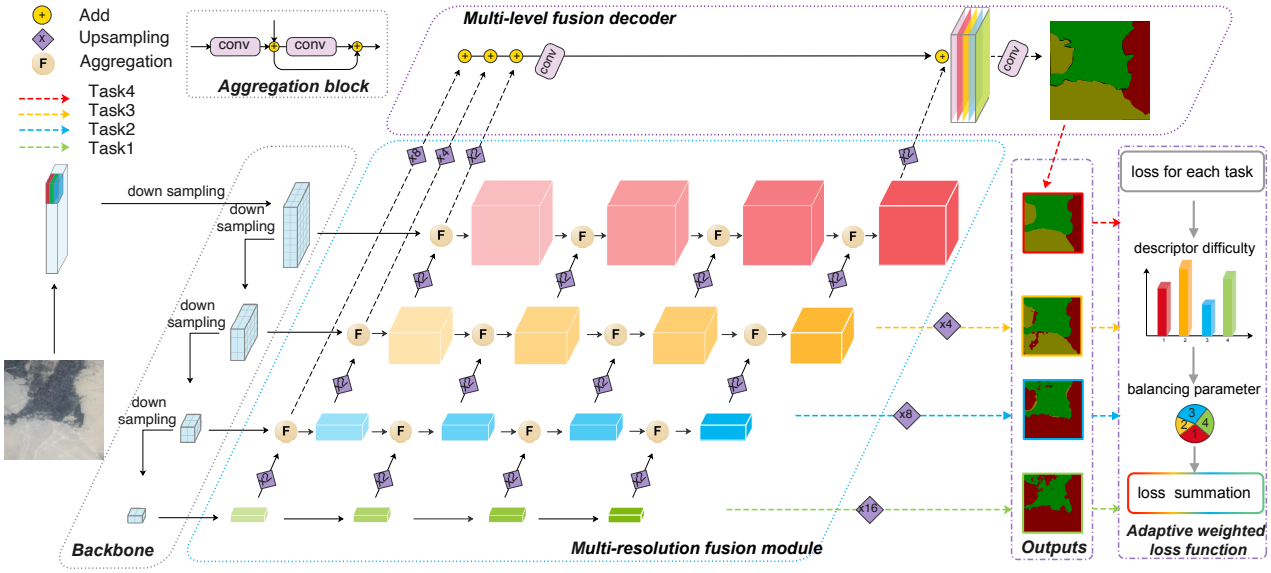

Figure 2. The overall structure of MrsSeg-AWL.

2.1.1. Backbone

Desert images usually have a large image size. In order to improve the model segmentation efficiency, a pre-trained Mobilenetv2 was used in this paper as the lightweight backbone. Inverted residual with a linear bottleneck was adopted in Mobilenetv2, this structure not only ensured the efficiency of feature extraction, but also effectively reduced the number of parameters.

The inverted residual with a linear bottleneck is shown in Figure 3. The inverted structure was designed according to the idea of "expansion–convolution–compression". First, 1×1 point-wise convolution (PW) was used to expand the input F to a high-dimensional embedding space, and then a 3×3 depth-wise separable convolution (DW) was used for filtering. Subsequently, the features were projected back to a low-dimensional representation with a 1×1 linear convolution. Finally, the low-dimensional outputs were added to the inputs by the skip connection to obtain the final output. The inverted residual with a linear bottleneck *Bott* could be computed as follows:

$$Bott(F) = F + PW(DW(PW(F))), \quad (1)$$

where F is the input feature map, PW is 1×1 point-wise convolution layer, DW is 3×3 depth-wise separable convolution layer.

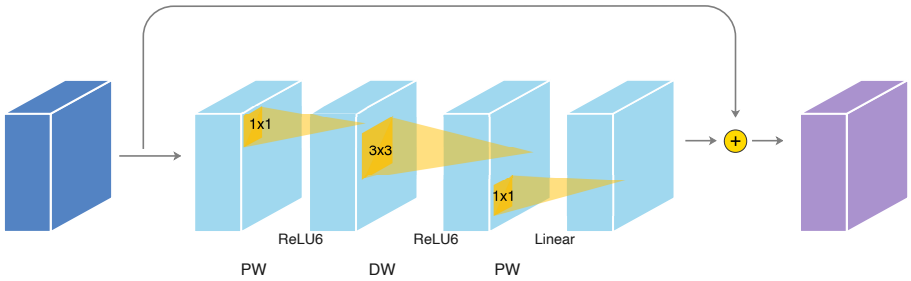

Figure 3. Inverted residual with linear bottleneck: *PW* represents point-wise convolution, *DW* represent depth-wise convolution.

2.1.2. Multi-Resolution Fusion Model

The feature map output from the backbone upper layer had a smaller size and higher semantic information [35]. This kind of high-level information has been experimentally proven to play a key role in the subsequent segmentation task. However, the greater the stride of the downsampling of the network was, the more the spatial details of the image were lost. This led to deep encoder blocks' lack of low-level features and made it difficult for decoders to recover local details. This problem motivated us to propose an aggregation strategy to fuse local detail and global information in different depth positions of feature extraction networks to achieve better performance.

The high-level features of desert images contained more global information, while the low-level features contained more local information such as color, texture, and edge. Effective fusion of high-level and low-level features could improve the segmentation effect. Based on this, the multi-resolution fusion module with a top-down fusion mechanism was designed (Figure 2). The module was composed of aggregation blocks (Figure 4). Each aggregation block had two inputs. The low-level feature was from the previous aggregation block at the same branch, and the high-level feature was from aggregation block at low-resolution branch. When the input feature map came from the backbone network, the number of channel (C1) would be adjusted by the 1×1 convolution to C2 to match the dimension of high-level feature. At the same time, the skip connection was used to connect input and output, which could effectively avoid the problems of information loss and gradient disappearance and improve the model's optimization ability. The multi-resolution fusion model made each feature representation from low-resolution to high-resolution continuously receive information from other parallel branches, so as to obtain richer high-resolution representation. This made the final output feature map more accurate.The aggregation block Agg could be computed as follows:

$$Agg(LF, HF) = CBR(CBR(LF) + Up(HF)) + CBR(LF) + Up(HF). \quad (2)$$

where HF is a high-level feature, LF is a low-level feature, CBR represents 3×3 convolution layer followed by one batch normalization layer and relu activation function, and Up represents the bilinear interpolation upsampling layer.

The architectures of MrsSeg are shown in Table 1. The encoder of MrsSeg contained two parts, including the lightweight backbone network and multi-resolution fusion modules. A pre-trained Mobilenetv2 was used in this paper as the lightweight backbone to down-sample the 512×512 training image to $1/16$ of itself. A multi-resolution fusion module took the multi-scale output of backbone as input. Each branch of multi-resolution fusion module had four aggregation blocks. The first aggregation block was used to unify the channel of feature map to 64, and the rest of the aggregation blocks were used to multi-scale information fusion.

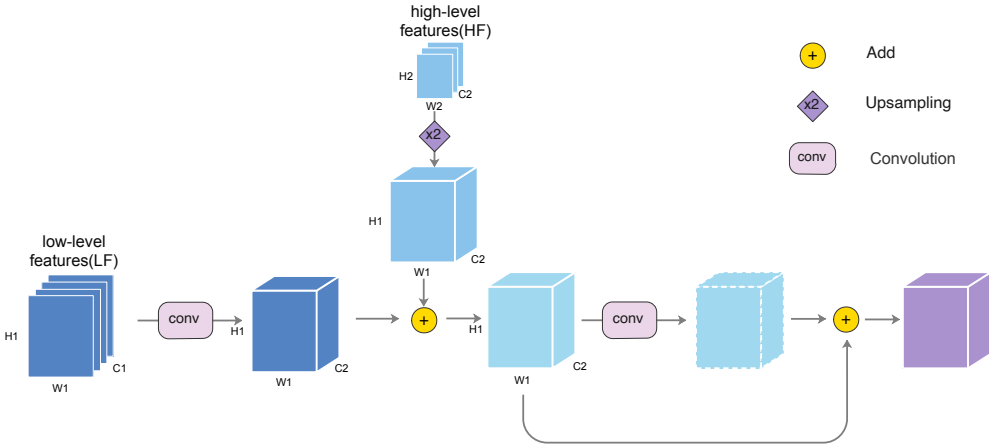

Figure 4. Structure of aggregation block in multi-resolution fusion module.

Table 1. The architectures of MrsSeg Each aggregation branch contains one or more operators. Each operation has output channels c, stride s, repeated n times. The expansion factor is t applied to expand the channel number of the operation. Conv2d means the convolutional layer, followed by one batch normalization layer and relu activation function. Bottleneck indicates inverted residual block. Aggblock represents aggregation block.

	Backbone					Multi-Resolution Fusion			
Branch	**Operator**	**t**	**c**	**n**	**s**	**Outputs Size**	**Operator**	**c**	**n**	**Outputs Size**
Branch 4	Conv2d	-	32	1	2	256^2	aggblock	64	4	256^2
	bottleneck	1	16	1	1	256^2				
Branch 3	bottleneck	6	24	2	2	128^2	aggblock	64	4	128^2
Branch 2	bottleneck	6	32	3	2	64^2	aggblock	64	4	64^2
Branch 1	bottleneck	6	64	4	2	32^2	aggblock	64	4	32^2
	bottleneck	6	96	3	1	32^2				

2.1.3. Multi-Level Fusion Decoder

According to the research of [25], not all the features of the stages were necessary to contribute to the decoder module. This motivated us to find a lightweight method to incorporate multi-level context into encoded features. The decoder in this paper was designed as a simple and effective upsampling module that integrated low-level and high-level features. First, the first feature map of each row (left to right in the multi-resolution module) was upsampled to the original image size through bilinear interpolation and added together, as shown in the black dotted line in Figure 2. Then, the result was followed by convolution operation and added with the output feature maps of the multi-resolution fusion module (upper right feature map in multi-resolution module) so that high-level features and low-level details were further fused. Finally, the fused feature map was subjected to a convolution operation followed by a softmax function to obtain the segmentation result.

2.2. Adaptive Weighted Loss Function

In this paper, the output of different resolution branches of multi-resolution structure (MrsSeg) was regarded as different optimization tasks, and the supervised training method was used to promote multi-resolution fusion. In a multi-output structure, it is usually important to achieve the loss balance by integrating the multi-branches loss. However, the existing loss balancing parameter was determined uniformly or only determined by single task difficulties. In the case that balancing parameters were calculated without considering

task difficulty for each branch, losses that did not match task difficulties of each branch were propagated (Figure 5a), and it seemed to reduce the effect of multitask learning.

In order to solve the above problem, an adaptive weighted loss function (AWL) was proposed to adjust the balancing parameters according to task difficulties for each branch. By reducing the weight of difficult tasks and increasing the weight of easy tasks, it could effectively accelerate the convergence speed of training and help to improve the segmentation result. The improved adaptive weighted loss function is shown in Figure 5b.

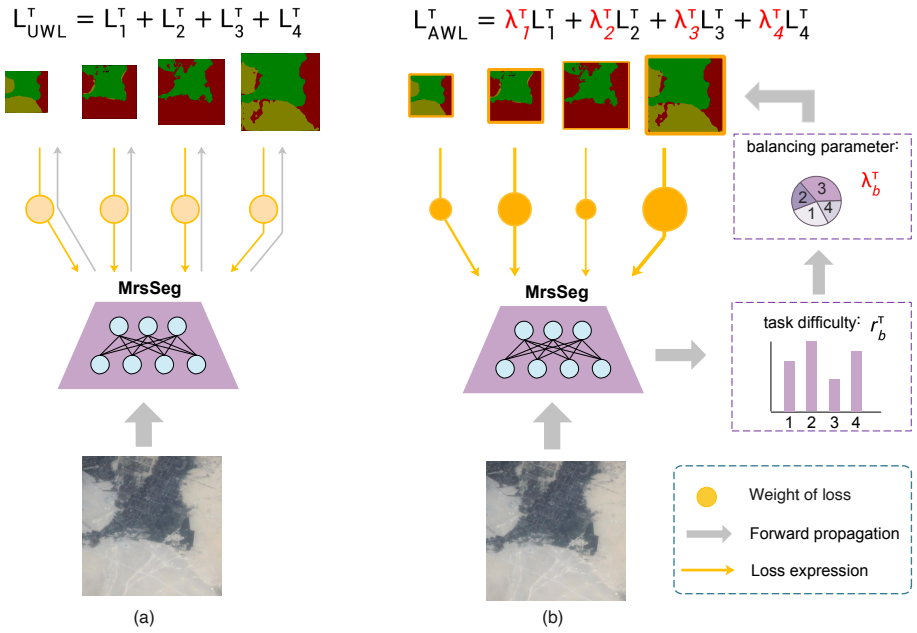

Figure 5. Concept of adaptive loss function: (**a**) Uniform weighted loss function; (**b**) Adaptive weighted loss function. Rectangles of different sizes represent MrsSeg's four branches prediction results. Gray arrows represent the loss expression of each branch. The sizes of the circles on the arrows indicate the balancing weight for multi-branch structure: the larger the radius of the circle, the greater the weight.

In this paper, the method of quantifying branch task difficulty and adjusting balancing parameters was used to achieve the purpose of adaptive loss balance. The branch task difficulty was calculated by loss reduction, so first we calculated the moving average k^{τ}_b of the current loss L^{τ}_b of each branch, as follows:

$$\begin{aligned} k^{\tau}_b &= \frac{k^{\tau-1}_b}{L^{\tau}_b + k^{\tau-1}_b} k^{\tau-1}_b + \frac{L^{\tau}_b}{L^{\tau}_b + k^{\tau-1}_b} L^{\tau}_b \\ &= (1-\alpha) k^{\tau-1}_b + \alpha L^{\tau}_b, \end{aligned} \tag{3}$$

where $\alpha \in [0,1]$ is a discount factor and $\alpha = \frac{L^{\tau}_b}{L^{\tau}_b + k_b{}^{\tau-1}}$, $b \in N$ is the number of each branch task, τ is the current training iteration, k^{τ}_b is the current moving average, and $k^{\tau-1}_b$ is the previous moving average.

Using k^{τ}_b, we defined the current task difficulty r^{τ}_b of each branch as follows:

$$r^{\tau}_b = \frac{k^{\tau}_b}{k^{\tau-1}_b}. \tag{4}$$

A large r^{τ}_b means that the current optimization step did not reduce the loss much; that is, optimization for the current branch task was difficult. In particular, if $r > 1$, it seemed that the task stepped into a local minimum. Therefore, we introduced the balancing

parameter λ_b^τ to reduce the weight of the branch task with large r_b^τ and increase the weight of the branch task with small r_b^τ . The formula for balancing parameters is:

$$\lambda_b^\tau = \frac{\sum_{i=1}^{N} r_i^\tau - r_b^\tau}{\sum_{i=1}^{N} r_i^\tau}. \tag{5}$$

The overall loss function L_{AWL}^τ was defined as the sum of each adjusted branch loss:

$$L_{AWL}^\tau = \frac{1}{N}\sum_{b=1}^{N} \lambda_b^\tau L_b^\tau. \tag{6}$$

Algorithm 1 shows the application of the adaptive weighted loss algorithm in the training process. First, the cross-entropy loss function was used to calculate the loss of each branch. Then, the moving average of each branch (k_b^τ) could be obtained by Equation (3). The smaller the moving average changed, the more difficult the branch was to optimize, so Equation (4) was used to calculate the optimizing difficulty of each branch (r_b^τ). According to the principle of giving priority to the easy task, the balance parameter of each branch (λ_b^τ) was allocated by Equation (5). Finally, the weight of each branch loss was adjusted by balance parameters to obtain the overall loss, and the network parameters were updated through back-propagation.

Algorithm 1 Application of Adaptive Weighted Loss in the Training Process

Input: batch images I
Output: network parameters W
Initialize network parameters W;
for $\tau = 1$ to max_iter **do**
  Get batch images I;
  Calculate CrossEntropyLoss L_b^τ for each branch;
  if $\tau = 1$ **then**
    Initialize $k_b^{\tau-1}$ with L_b^τ;
  end if
  Update moving average k_b^τ for each branch with Equation (3);
  Calculate task difficulty r_b^τ with Equation (4);
  Calculate balancing parameters λ_b^τ with Equation (5);
  Get final Loss L_{AWL}^τ with Equation (6);
  Using L_{AWL}^τ backpropagation and update network parameters W
end for
return W

Desert imagery has the characteristics of large-scale change and irregular location distribution of surface objects. In view of the above characteristics, MrsSeg adopted multi-resolution feature aggregation modules in order to extract and fuse multi-resolution features of desert image. Aiming at the structural features of multi-resolution outputs of the model, this work designed auxiliary loss function on multiple resolution branches to improve the feature extraction. Because the training difficulty of different branch tasks is different, and the optimization difficulty of each branch is different in different training stages, an adaptive weighted loss function was designed, which could improve the convergence speed of the model and the desert segmentation result by giving priority to the easy tasks.

3. Results

3.1. Data and Pre-Processing

Xinjiang is the region with the largest desertification area, the widest desertification distribution, and the most serious desertification damage in China. The region is deep inland, forming a distinct temperate continental climate. Desert (sandy desert) and gobi (Gobi desert) are the main land types in this area. Unlike desert, gobi is mainly covered with bare gravel and stones. Based on the above reasons, Xinjiang was selected as the sampling object of the desert area segmentation data set.

The images collected in this paper were all from Environment and Disaster Monitoring and Forecasting Small Satellite Constellations A, B (HJ-1A/B). The satellite has unique advantages of autonomous control, medium and high resolution, wide coverage, etc. It can stably obtain the medium resolution remote sensing data covering the whole country every half a year and is the preferred remote sensing data source for carrying out high-dynamic national desert and desertification land monitoring.

The data set contained desert, gobi, oasis, and river categories (Figure 6) and was divided into training set, verification set, and test set by the ratio of 6:2:2. In order to make the model more robust, the data were expanded from 1665 to 6660 pieces by random flipping, rotating, and other image enhancement methods, described in Figure 7.

Figure 6. Categories of desert data set: (**a**) Desert, (**b**) Gobi, (**c**) River, (**d**) Oasis.

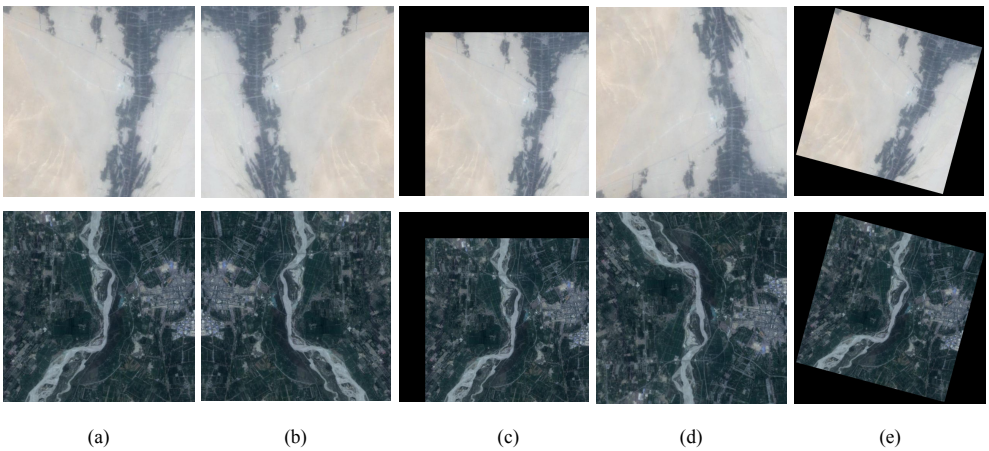

Figure 7. Different data augmentation methods: (**a**) Original image; (**b**) Horizontal flip; (**c**) Random scale; (**d**) Vertical flip; (**e**) Random rotation.

3.2. Experimental Results

The experimental environment was Intel Core i7-8700k (Intel Corporation is an American multinational corporation and technology company headquartered in Santa Clara, CA,

USA, in Silicon Valley) eight-core processor, RTX2070 8G independent graphics card, 32G memory, and 1T hard disk; the software that we used was PyTorch framework (PyTorch is an open source machine learning library based on the Torch library, primarily developed by Facebook. Facebook, Inc., is an American technology conglomerate based in Menlo Park, CA, USA). For a fair comparison, all runs were trained with stochastic gradient descent method. The hyper-parameters were set as follows: batchsize = 4, momentum = 0.9, weight decay = 0.00005. We used cosine decay as learning rate decay strategy. To obtain a quantitative evaluation result, we adopted Frame Per Second (FPS) and mean Intersection over Union (mIoU) as metrics. FPS is the number of images that can be processed per second. The larger its value, the faster the prediction speed of the model. The mIoU calculates the intersection ratio of all classes. This index can better reflect the accuracy and completeness of model segmentation in different terrain type areas in the experiment, as defined below:

$$mIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}p_{ji}-p_{ii}}, \tag{7}$$

where $k+1$ is the number of classes (including background); p_{ii} is the number of pixels that belong to class i and were classifified correctly, p_{ij} is the number of pixels that belong to class i and were classifified as class j.

Due to the large number of data in the desert data set, the mini-batch training method was adopted for model training. In the verification process, the model result will inevitably be biased towards the final iteration of the batch data. When mini-batch randomly extracts batch data from desert data set, batch data samples imbalance may occur. When such a situation occurs in the last iteration, the verification curve will be jittered. However, this does not affect the overall training trend of the model. Hence, average value of mIoU every 20 epochs was used in Figures 8 and 9 as the points of the curve so as to better reflect the overall trend and performance of the model.

In this section, we first analyze the results of ablation experiments and then demonstrate the effect and role of the adaptive weighted loss function. Finally, we compare the results of MrsSeg-AWL and the existing segmentation network in desert land type segmentation task.

In the first section, a detailed ablation experiment was performed on MrsSeg-AWL to better understand the gain effect of each improved component. The ablation experiments results are shown in Figure 8 and Table 2. Backbone network and the number of branches remain unchanged (Net1), the introduction of skip-connection into the aggregation blocks(Net2) could effectively avoid the network degradation caused by the increase of network layers. By changing the convolution before the upsampling mode to the upsampling before convolution mode in multi-level fusion decoder's final stage (Net3), the decoding capability of the decoder was enhanced to better recover the detailed features of desert images. Compared with the model that only used the cross-entropy loss function, the MrsSeg-AWL using the adaptive weighted loss function improved mIoU by 3.8%. It could be clearly seen from Figure 8 that the mIoU of MrsSeg-AWL rapidly improved between 0 and 40 epochs, and the mIoU curve did not oscillate after 140 epochs, indicating that adaptive weighted loss function effectively improved the convergence speed of the model.

The training curves of MrsSeg with different loss strategies compared with the single-branch(Baseline) are shown in Figure 9 and Table 3. It can be seen from Table 3 that when fixed balancing parameters were used, increasing the number of integrated branches could effectively increase the mIoU value of the model, which shows that the use of additional branch loss has a positive effect on the final result. Compared with the model only trained with cross-entropy loss, the model trained with four-branch loss improved the mIoU by 2.2%, and the model that used adaptive weighted loss function improved the mIoU by 3.9%. It can be seen from Figure 9d that the training curve of MrsSeg with adaptive weighted loss was steeper between 0 and 40 epochs, and the curve did not oscillate

after 140 epochs. Compared with other loss strategies, MrsSeg-AWL training curves also achieved the highest mIoU. The experimental result shows that adaptive weighted loss function effectively improved the convergence speed and the mIoU of the model.

Table 2. Ablation study of MrsSeg-AWL.

Method	Extra Skip-Connection	Upsampling Before Convolution	Adaptive Weighted Loss	mIoU(%)
Net1				44.4
Net2	✓			47.3
Net3	✓	✓		54.2
MrsSeg-AWL	✓	✓	✓	58.0

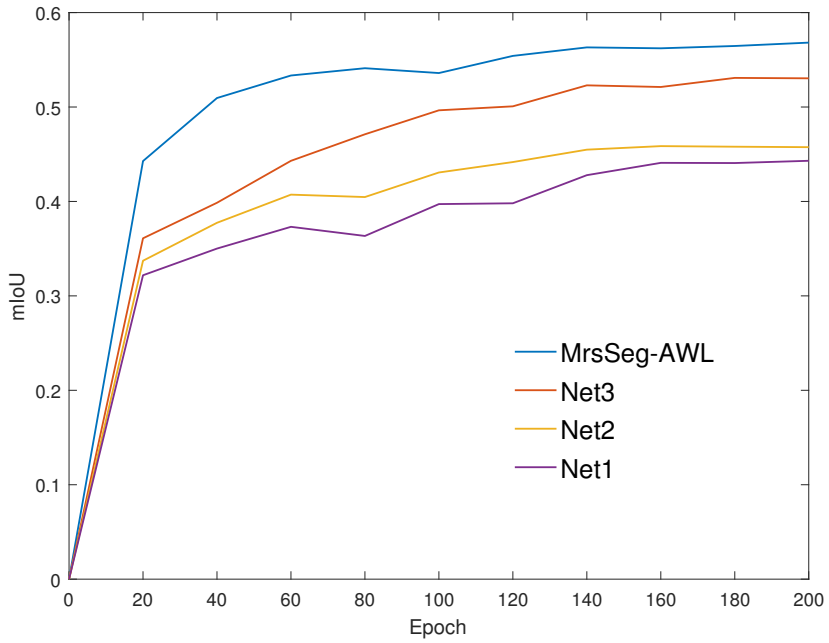

Figure 8. Training curve of ablation experience. The curves represent the mIoU change of the validation dataset during the training process. Each point of the curves was the average value of mIoU every 20 epochs.

Table 3. Different loss strategies results.

Loss Strategies	Branch 1	Branch 2	Branch 3	Branch 4	mIoU(%)
Single-branch	✓				54.2
Two-branches	✓	✓			55.3
Three-branches	✓	✓	✓		56.1
Four-branches	✓	✓	✓	✓	56.3
Adaptive weighted	✓	✓	✓	✓	58.0

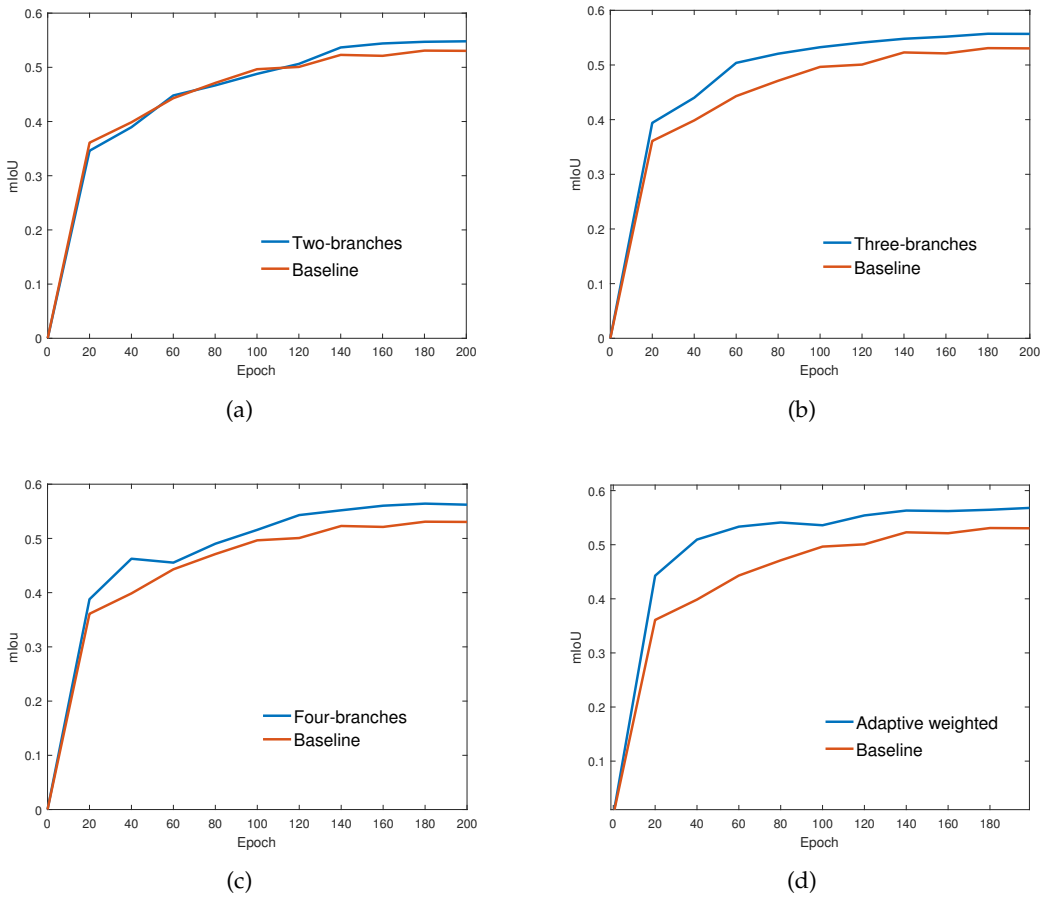

Figure 9. Training curves comparison for different loss strategies: Single-branch loss strategy was used as the baseline for the experiment.The curves represent the mIoU change of the validation dataset during the training process. Each point of the curves was the average value of mIoU every 20 epochs. (**a**) Two-branches losss strategy, (**b**) Three-branches loss strategy, (**c**) Four-branches loss strategy, (**d**) Adaptive weighted loss strategy.

Figure 10 shows the balancing parameters curve of MrsSeg-AWL in different training stages. The balancing parameters represent the optimization degree of each branch. The larger the balance parameter was, the easier the branch was to be optimized. Each line in Figure 10 corresponds to the four branch outputs in the same color as in Figure 2. On the one hand, the top-down comparison of Figure 10a–d shows that the branch task with higher resolution has larger balance parameter values in the whole training process, indicating that the integration of global information and local information could better optimize the training process. On the other hand, by observing the curve, it could be found that in different training stages, the optimization difficulty of each branch loss in the multi-resolution supervision network is different. If fixed balancing parameters were adopted, the proportion of each branch loss cannot be dynamically adjusted, such that the model cannot be further optimized. The experimental results further demonstrated that the adaptive weighted loss function was helpful to adjust the influence of each branch loss on the total loss in different training stages, and gave priority to training the branch with large optimization space, so as to accelerate the convergence speed and improve the accuracy.

Table 4 shows the desert segmentation result by four mainstream lightweight backbones, which were all pre-trained on Imagenet classification inn the case that the feature fusion module and the adaptive loss function did not change. The experimental results show that MobilenetV2 achieved the best desert segmentation result. Although its segmentation time was a little bit slower than ResNet-18, its mIoU was 2.2% higher than

ResNet-18. Therefore, in the next comparative experiment, we used MobileV2 as the backbone network.

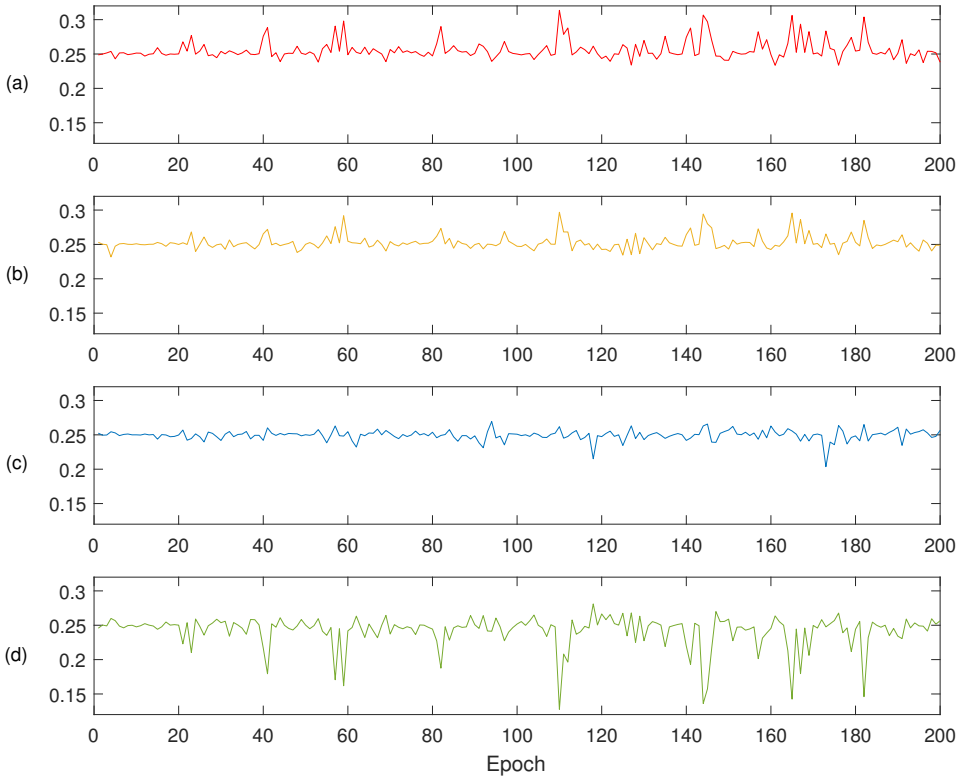

Figure 10. Balancing parameters curve of MrsSeg-AWL: (**a**) Task4; (**b**) Task3; (**c**) Task2; (**d**) Task1.

Table 4. Comparsion of different backbones.

Backbone	mIoU(%)	FPS	Paramteter (M)
MobilenetV2	58.0	77.0	3.3
ResNet-34	57.0	69.0	22.4
ResNet-18	55.8	82.0	12.23
SuffleNet	54.6	70.0	2.3

The performances of different models in the desert segmentation task are shown in Table 5. It can be seen from the table that the improved MrsSeg-AWL achieved the highest mIoU, and adaptive weighted loss function improved the mIoU by 1.7% without increasing the prediction time of MrsSeg. In the experiment, we found that FCN based on Vgg was prone to the problem of hard convergence. FPN achieved the fastest prediction speed, but its accuracy was unsatisfactory. DeepLabV3+ was better than MrsSeg-AWL with respect to speed and comparable to MrsSeg-AWL with respect to accuracy. However, the last one requires less parameter tuning. Experimental results showed that the proposed MrsSeg-AWL with multi-resolution fusion network and adaptive weighted loss function has better performance in desert segmentation task than the existing segmentation network.

Table 6 shows he land type segmentation results of each model. It can be seen that the IoU of desert and oasis categories was generally high, indicating that that land type was easier to identify when the sample was sufficient. The IoU of MrsSeg-AWL segmentation result reached 84% and 86%, respectively. Due to the small number size and the large-scale change in river samples, the IoU of this category was generally low. The MrsSeg-AWL's IoU of the river category reached 23.1%, which is 3.4% higher than DeepLabV3+. It showed that the multi-resolution supervision network could better learn the characteristics of river

samples and obtain more accurate segmentation results when the number of samples was small and the sample scale changed greatly.

Table 5. Segmentation performance of different models.

Method	Backbone	mIoU (%)	FPS	Parameters (M)
MrsSeg-AWL	MobilenetV2	58.0	77.0	3.3
DeepLabV3+	MobilenetV2	56.6	93.0	5.8
MrsSeg	MobilenetV2	56.3	77.0	3.3
DenseASPP	MobilenetV2	54.3	85.0	2.48
UNet	-	52.6	33.0	13.3
ENet	-	49.4	69.0	-
FPN	MobilenetV2	48.0	128.0	4.4
DFANet	-	45.7	69.0	1.97
FCN32s	Vgg	20.1	27.1	134.4

Table 6. Land types segmentation IoU of different models.

Method	Desert	Oasis	Gobi	River	Background
MrsSeg-AWL	84.0	86.0	39.6	**23.1**	46.4
DeepLabV3+	83.9	82.0	41.1	19.7	46.8
DenseASPP	81.5	81.9	38.9	15.1	44.7
UNet	75.2	86.3	35.1	14.1	43.0
ENet	78.5	80.2	37.3	11.8	41.9
FPN	76.1	78.1	27.3	10.9	39.6
DFANet	71.1	76.6	14.6	20.8	25.1

3.3. Desert Segmentation Results

The segmentation results of desert remote sensing images are shown in Figure 11. It can be seen that the images segmented by FPN have a large area of mis-segmentation, indicating that the feature fusion network with a single branch cannot make full use of local and global semantic information, leading to pixel-level classification errors. Compared with FPN, Unet significantly reduces he false detection area in desert and gobi land types, but the classification result of river samples was still not accurate enough, and the segmentation edge was rough. MrsSeg-AWL and the state-of-art segmentation network Deeplabv3+ performed well on desert and oasis segmentation task. While extracting the multi-resolution features of desert images, MrsSeg-AWL used adaptive weighted loss function to promote multi-resolution feature fusion through supervised training. This enabled MrsSeg-AWL to better learn the characteristics of river samples for river types with a small number of samples and large-scale change and obtained more accurate and clear desert segmentation maps.

In order to test the desert segmentation ability in the non-sampling region, we randomly selected some desert images in the Nile valley and carried out the segmention test on these images using the model trained on the desert dataset. The segmentation results are shown in Figure 12. It can be seen from the figure that FPN and Unet's segmentation results showed large areas of desert and gobi false detection areas, indicating that these methods' feature extraction ability needed to be improve. MrsSeg-AWL and the state-of-art segmentation network Deeplabv3+ had a better overall classification results. MrsSeg-AWL using multi-resolution supervising network had more accurate segmentation edges of the river category. The test result shows that MrsSeg-AWL had good desert segmentation application potential.

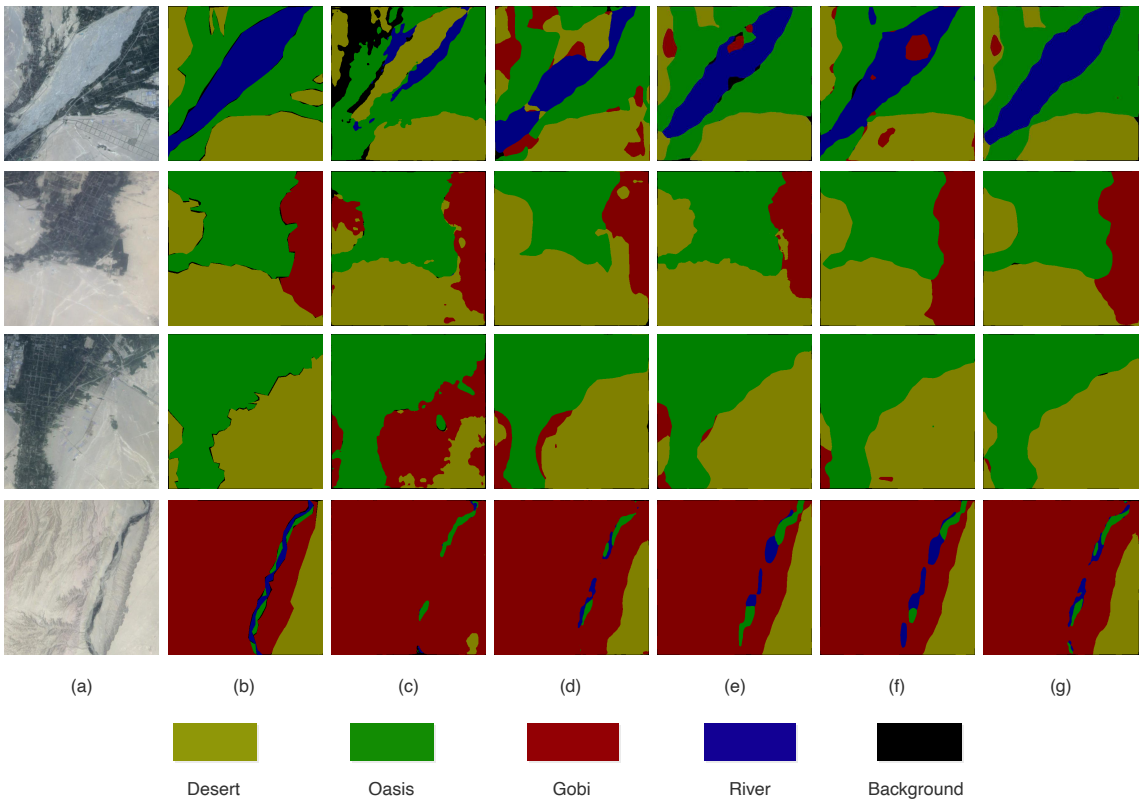

Figure 11. Comparison of different models segmentation results on desert dataset: (**a**) Original image, (**b**) Ground truth, (**c**) FPN, (**d**) Unet,(**e**) DenseASPP, (**f**) Deeplab v3+, (**g**) MrsSeg-AWL.

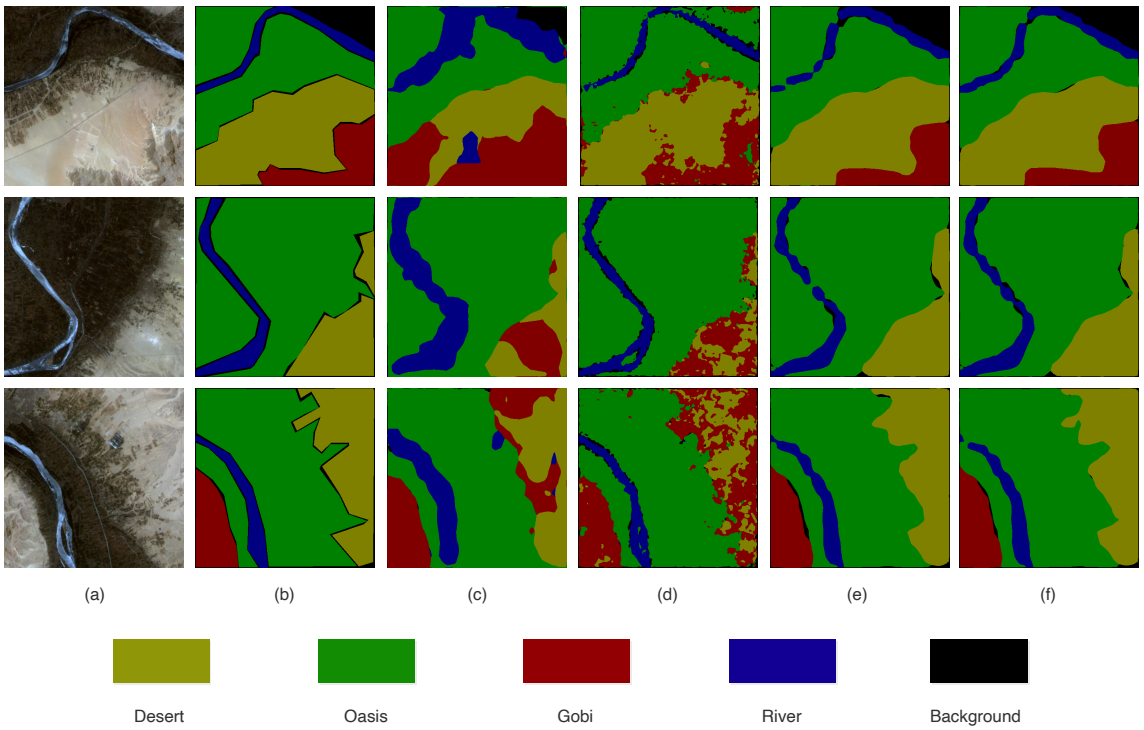

Figure 12. Comparison of different models segmentation results on the Nile valley: (**a**) Original image, (**b**) Ground truth, (**c**) FPN, (**d**) Unet, (**e**) Deeplab v3+, (**f**) MrsSeg-AWL.

4. Discussion

Desert remote sensing images are usually characterized by large image size, large-scale change, and irregular location distribution of surface objects. In the desert data, the gobi and river samples accounted for only 15% and 1% of the total number, respectively. It can be seen from Table 6 that MrsSeg-AWL achieved the highest IoU value in the river category and also reached comparable gobi segmentation result as that of Deeplab V3+. It can be also seen from Figures 11 and 12 that MrsSeg-AWL achieved good segmentation results, especially for complex images such as the small area of the oasis and rivers. MrsSeg-AWL used adaptive weighted loss function to promote multi-resolution feature fusion through supervised training. This enabled MrsSeg-AWL to better learn the characteristics of samples with a small number and large-scale changes and obtained more accurate and clear desert segmentation maps.

In the experiment, we found that there was excessive exposure in some areas, and sthe narrow rivers in the desert area had seasonal flow interruption, which had an impact on the accuracy of desert segmentation. Therefore, in future research, we will focus on the problem of bad effects of image noise, so as to further improve the segmentation results.

5. Conclusions

Accurate desert segmentation results could provide a basis for the timely understanding of the status, extent, and evolution of desert areas. Desert images are usually characterized by large image size, large-scale change, and irregular location distribution of surface objects. The multi-scale fusion method is widely used in the existing deep learning segmentation models to solve the above problems. Based on the idea of multi-scale feature extraction, this paper took the segmentation results of each scale as an independent optimization task and proposed a multi-resolution supervision network (MrsSeg) to further improve the desert segmentation result. Due to the different optimization difficulties of different scale tasks, we also proposed an adaptive weighted loss function (AWL) to automatically optimize the training process. First, we collected remote sensing images of the Xinjiang region from the Environment and Disaster Monitoring and Forecasting Small Satellite Constellations A and B satellites (HJ-1A/B) and used these images to create a desert segmentation data set. Then, a multi-resolution segmentation method based on the adaptive weighted loss function was proposed. Finally, the image segmentation experiments and results analysis were carried out on remote sensing images of the Xinjiang and Nile valleys. The experimental results showed that the proposed multi-resolution supervision network could effectively improve the desert segmentation accuracy under the condition of low parameter complexity. The adaptive weighted loss function accelerated the convergence of the model and further improved the segmentation results. MrsSeg-AWL also showed a certain improvement in the gobi and river categories with few samples and was difficult to segment. To sum up, the improved network is an effective automatic desert remote sensing image segmentation method.

Author Contributions: Conceptualization, L.W. (Liguo Weng) and M.X.; methodology, L.W. (Lexuan Wang) and M.X.; software, L.W. (Lexuan Wang).; validation, L.W. (Lexuan Wang), M.X., and L.W. (Liguo Weng).; formal analysis, L.W. (Lexuan Wang), J.L., and M.X. and H.L.; investigation, L.W. (Lexuan Wang) and M.X.; resources, M.X. and L.W. (Liguo Weng); data curation, L.W. (Liguo Weng) and M.X.; writing—original draft preparation, L.W. (Lexuan Wang); writing—review and editing, M.X. and H.L.; visualization, L.W. (Lexuan Wang); supervision, M.X.; project administration, M.X. and J.L.; funding acquisition, M.X. and J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported in part by the National Natural Science Foundation of PR China (42075130, 61773219).

Data Availability Statement: The data and the code of this study are available from the corresponding author upon request. (xiamin@nuist.edu.cn).

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

References

1. Huang, J.; Zhang, G.; Zhang, Y.; Guan, X.; Guo, R. Global Desertification Vulnerability to Climate Change and Human Activities. *Land Degrad. Dev.* **2020**, *10*, 1380–1391. [CrossRef]
2. Yue, Y.; Ye, X.; Zou, X.; Wang, J.; Gao, L. Research on Land Use Optimization for Reducing Wind Erosion in Sandy Desertified Area: A Case Study of Yuyang County in Mu Us Desert, China. *Stoch Envrion. Res. Risk Assess* **2017**, *31*, 1371–1387. [CrossRef]
3. Zhang, L.; Yue, L.; Xia, B. The Study of Land Desertification in Transitional Zones between the MU US Desert and the Loess Plateau Using RS and GIS-A Case Study of the Yulin Region. *Environ. Geol.* **2003**, *44*, 530–534. [CrossRef]
4. Chen, B.; Xia, M.; Huang, J. MFANet: A Multi-Level Feature Aggregation Network for Semantic Segmentation of Land Cover. *Remote Sens*. **2021**, *13*, 731. [CrossRef]
5. Xia, M.; Cui, Y.; Zhang, Y.; Liu, J.; Xu, Y. DAU-Net: A Novel Water Areas Segmentation Structure for Remote Sensing Image. *Int. J. Remote Sens*. **2021**, *42*, 2594–2621. [CrossRef]
6. Zhang, F.; Tiyip, T.; Johnson, V.C.; Kung, H.; Ding, J.; Zhou, M.; Fan, Y.; Kelimu, A.; Nurmuhammat, I. Evaluation of Land Desertification from 1990 to 2010 and Its Causes in Ebinur Lake Region, Xinjiang China. *Environ. Earth Sci.* **2015**, *73*, 5731–5745. [CrossRef]
7. Weng, L.; Wang, L.; Xia, M.; Shen, H.; Liu, J.; Xu, Y. Desert classification based on a multi-scale residual network with an attention mechanism. *Geosci. J.* **2020**, *25*, 387–399. [CrossRef]
8. Xia, M.; Tian, N.; Zhang, Y.; Xu, Y.; Zhang, X. Dilated multi-scale cascade forest for satellite image classification. *Int. J. Remote Sens*. **2020**, *41*, 7779–7800. [CrossRef]
9. Pi, W.; Du, J.; Liu, H.; Zhu, X. Desertification Glassland Classification and Three-Dimensional Convolution Neural Network Model for Identifying Desert Grassland Landforms with Unmanned Aerial Vehicle Hyperspectral Remote Sensing Images. *J. Appl. Spectrosc.* **2020**, *87*, 309–318. [CrossRef]
10. Moghaddam, M.H.R.; Sedighi, A.; Fasihi, S.; Firozjaei, M.K. Effect of Environmental Policies in Combating Aeolian Desertification over Sejzy Plain of Iran. *Aeolian Res.* **2018**, *35*, 19–28. [CrossRef]
11. Ge, G.; Shi, Z.; Zhu, Y.; Yang, X.; Hao, Y. Land Use/Cover Classification in an Arid Desert-Oasis Mosaic Landscape of China Using Remote Sensed Imagery: Performance Assessment of Four Machine Learning Algorithms. *Glob. Ecol. Conserv.* **2020**, *22*, e00971. [CrossRef]
12. Xia, M.; Wang, T.; Zhang, Y.; Liu, J.; Xu, Y. Cloud/shadow segmentation based on global attention feature fusion residual network for remote sensing imagery. *Int. J. Remote Sens*. **2021**, *42*, 2022–2045. [CrossRef]
13. Moustafa, O.R.M.; Cressman, K. Using the Enhanced Vegetation Index for Deriving Risk Maps of Desert Locust (*Schistocerca Gregaria*, Forskal) Breeding Areas in Egypt. *J. Appl. Remote Sens.* **2015**, *8*, 084897. [CrossRef]
14. Wang, S.; Mu, X.; Yang, D.; He, H.; Zhao, P. Road Extraction from Remote Sensing Images Using the Inner Convolution Integrated Encoder-Decoder Network and Directional Conditional Random Fields. *Remote Sens*. **2021**, *13*, 465. [CrossRef]
15. Shelhamer, E.; Long, J.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*. **2017**, *39*, 640–651. [CrossRef] [PubMed]
16. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *Computer Vision—ECCV 2018*; Lecture Notes in Computer Science; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer International Publishing: Cham, Switzerland, 2018; Volume 11211, pp. 833–851.
17. Li, L. Deep Residual Autoencoder with Multiscaling for Semantic Segmentation of Land-Use Images. *Remote Sens.* **2019**, *11*, 2142. [CrossRef]
18. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26–30 June 2016; IEEE: Las Vegas, NV, USA, 2016; pp. 770–778.
19. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell*. **2018**, *40*, 834–848. [CrossRef]
20. Ulmas, P.; Liiv, I. Segmentation of Satellite Imagery Using U-Net Models for Land Cover Classification. *arXiv* **2020**, arXiv:2003.02899.
21. Duarte, D.; Nex, F.; Kerle, N.; Vosselman, G. Multi-Resolution Feature Fusion for Image Classification of Building Damages with Convolutional Neural Networks. *Remote Sens*. **2018**, *10*, 1636. [CrossRef]
22. Song, X.; Jiang, S.; Herranz, L. Multi-scale multi-feature context modeling for scene recognition in the semantic manifold. *IEEE Trans. Image Process.* **2017**, *26*, 2721–2735. [CrossRef]
23. Xia, M.; Liu, W.; Wang, K.; Song, W.; Chen, C.; Li, Y. Non-intrusive load disaggregation based on composite deep long short-term memory network. *Expert Syst. Appl*. **2020**, *160*, 113669. [CrossRef]
24. Xia, M.; Zhang, X.; Weng, L.; Xu, Y. Multi-Stage Feature Constraints Learning for Age Estimation. *IEEE Trans. Inf. Forensics Secur*. **2020**, *15*, 2417–2428. [CrossRef]
25. Shahrezaei, I.H.; Kim, H.C. Fractal analysis and texture classification of high-frequency multiplicative noise in sar sea-ice images based on a transform-domain image decomposition method. *IEEE Access* **2020**, *8*, 40198–40223. [CrossRef]
26. Wu, D.; Wang, C.; Wu, Y.; Wang, Q.C.; Huang, D.S. Attention deep model with multi-scale deep supervision for person re-identification. *IEEE Trans. Emerg. Top. Comput. Intell.* **2021**, *5*, 70–78. [CrossRef]

27. Shen, Z.; Liu, Z.; Li, J.; Jiang, Y.G.; Xue, X. Object detection from scratch with deep supervision. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *42*, 99. [CrossRef] [PubMed]
28. Hosono, T.; Hoshi, Y.; Shimamura, J.; Sagata, A. Adaptive Loss Balancing for Multitask Learning of Object Instance Recognition and 3D Pose Estimation. In Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 3–8 November 2019; pp. 2587–2592.
29. Chen, Z.; Badrinarayanan, V.; Lee, C.-Y.; Rabinovich, A. GradNorm: Gradient Normalization for Adaptive Loss Balancing in Deep Multitask Networks. In Proceedings of the International Conference on Machine Learning Research, Stockholmsmässan, Stockholm, Sweden, 10–15 July 2018; Volume 80, pp. 794–803.
30. Kampffmeyer, M.; Salberg, A.-B.; Jenssen, R. Semantic Segmentation of Small Objects and Modeling of Uncertainty in Urban Remote Sensing Images Using Deep Convolutional Neural Networks. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 680–688.
31. Zhang, W.; Huang, H.; Schmitz, M.; Sun, X.; Mayer, H. A multi-resolution fusion model incorporating color and elevation for semantic segmentation. *ISPRS Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2017**, *42*, 513–517. [CrossRef]
32. Lowe, D.G. Object Recognition from Local Scale-Invariant Features. In Proceedings of the Seventh IEEE International Conference on Computer Vision, Kerkyra, Greece, 20–27 September 1999; Volume 2, pp. 1150–1157.
33. Lin, T.-Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944.
34. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. In *Computer Vision—ECCV 2014*; Lecture Notes in Computer Science; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer International Publishing: Cham, Switzerland, 2014; Volume 8691, pp. 346–361.
35. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520.