forests

MDPI

Article

Green Space Reverse Pixel Shuffle Network: Urban Green Space Segmentation Using Reverse Pixel Shuffle for Down-Sampling from High-Resolution Remote Sensing Images

Mingyu Jiang [1], Hua Shao [1,*], Xingyu Zhu [1] and Yang Li [2,3,4]

[1] School of Geomatics Science and Technology, Nanjing Tech University, Nanjing 211816, China; 202261223014@njtech.edu.cn (M.J.); 202261223029@njtech.edu.cn (X.Z.)
[2] School of Geography, Nanjing Normal University, Nanjing 210023, China; skys1017@163.com
[3] Key Laboratory of Virtual Geographic Environment, Nanjing Normal University, Ministry of Education, Nanjing 210023, China
[4] Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China
* Correspondence: shaohua@njtech.edu.cn; Tel.: +86-025-5813-9842

Abstract: Urban green spaces (UGS) play a crucial role in the urban environmental system by aiding in mitigating the urban heat island effect, promoting sustainable urban development, and ensuring the physical and mental well-being of residents. The utilization of remote sensing imagery enables the real-time surveying and mapping of UGS. By analyzing the spatial distribution and spectral information of a UGS, it can be found that the UGS constitutes a kind of low-rank feature. Thus, the accuracy of the UGS segmentation model is not heavily dependent on the depth of neural networks. On the contrary, emphasizing the preservation of more surface texture features and color information contributes significantly to enhancing the model's segmentation accuracy. In this paper, we proposed a UGS segmentation model, which was specifically designed according to the unique characteristics of a UGS, named the Green Space Reverse Pixel Shuffle Network (GSRPnet). GSRPnet is a straightforward but effective model, which uses an improved RPS-ResNet as the feature extraction backbone network to enhance its ability to extract UGS features. Experiments conducted on GaoFen-2 remote sensing imagery and the Wuhan Dense Labeling Dataset (WHDLD) demonstrate that, in comparison with other methods, GSRPnet achieves superior results in terms of precision, F1-score, intersection over union, and overall accuracy. It demonstrates smoother edge performance in UGS border regions and excels at identifying discrete small-scale UGS. Meanwhile, the ablation experiments validated the correctness of the hypotheses and methods we proposed in this paper. Additionally, GSRPnet's parameters are merely 17.999 M, and this effectively demonstrates that the improvement in accuracy of GSRPnet is not only determined by an increase in model parameters.

Keywords: urban green space; high-resolution remote sensing imagery; deep learning; semantic segmentation

Citation: Jiang, M.; Shao, H.; Zhu, X.; Li, Y. Green Space Reverse Pixel Shuffle Network: Urban Green Space Segmentation Using Reverse Pixel Shuffle for Down-Sampling from High-Resolution Remote Sensing Images. *Forests* **2024**, *15*, 197. https://doi.org/10.3390/f15010197

Academic Editor: Giorgos Mallinis

Received: 24 November 2023
Revised: 15 January 2024
Accepted: 17 January 2024
Published: 19 January 2024

1. Introduction

According to United Nations projections, the world's population is expected to increase by 2.25 billion, reaching a total of 9.15 billion by the year 2050 [1]. Such rapid population growth imposes significant ecological pressures on cities [2], highlighting issues such as the urban heat island (UHI) effect [3] and various pollution problems in the process of urbanization. Urban green spaces (UGS) represent a crucial component of urban ecosystems, comprising the vegetative entities within urban areas [4]. UGS play a pivotal role in the sustainable development of cities [5]. On one hand, UGS contribute significantly to enhancing the urban ecological environment [6] and addressing environmental issues like air pollution [7] and noise pollution [8]. On the other hand, UGS can improve the

physical and mental well-being of residents [9], reduce stress and anxiety [10], and promote a healthier lifestyle [11].

However, the environmental pollution and land issues resulting from urban expansion and construction pose a significant threat to the urban ecosystem [12]. Therefore, for the sustainable and healthy development of cities and the advancement of the United Nations Sustainable Development Goal No. 11 [13], the rapid and accurate acquisition of UGS information has become increasingly critical. In the past decade, local government authorities conducted surveys and mapping of green spaces using field investigations. Nevertheless, these traditional survey methods were time-consuming, resource-intensive, and often provided incomplete and untimely information, and some small-scale and scattered green spaces were also neglected. This hindered the development and implementation of relevant policies aimed at sustainability. Innovative approaches for UGS survey and statistics methods are still in need of further research.

With the rapid advancement of earth-observing technology, the data acquisition capability of remote sensing has significantly improved, marking the dawn of a new era in multi-platform, multi-angle, multi-sensor, all-weather, and all-time earth observation [14]. Developing a method for the accurate and rapid extraction of UGS information using multispectral or hyperspectral remote sensing imagery is a critical issue. Gandhi et al. [15] proposed a satellite image vegetation change detection method based on the Normalized Difference Vegetation Index (NDVI), utilizing Landsat TM remote sensing data and NDVI and DEM data for multisource vegetation classification. Zhou et al. [16] introduced a city forest type discrimination method based on Linear Spectral Mixture Analysis (LSMA) and Support Vector Machine (SVM), which used LSMA to extract three different vegetation endmembers, including broadleaf forest, coniferous forest, and low vegetation, and their abundances. Zhang et al. [17] generated a fine classification system for the 2015 Global 30 m Land Cover Classification (GLC_FCS30-2015) using Landsat image time series and high-quality training data from the Global Spatial Temporal Spectra Library (GSPECLib) on the Google Earth Engine computing platform. Nevertheless, UGS are characterized by fragmentation and complex backgrounds [18]; thus, there are numerous small-scale UGS, such as roadside trees and independent trees. Small-scale UGS are frequently overlooked because the spatial resolution of multispectral or hyperspectral remote sensing imagery is often limited compared with nature imagery, and small-scale UGS cannot be found even in low-spatial resolution remote sensing imagery. The aforementioned factors will lead to a significant discrepancy between extracted UGS and ground truth. Hence, it is crucial to improve the accuracy of identifying fragmented and scattered UGS.

High-spatial resolution remote sensing imagery plays a pivotal role in addressing the aforementioned challenge. Its spatial resolutions are a few meters or even less than 1 m, offering detailed texture and spectral information of ground objects. This enables a comprehensive understanding of urban environments, supporting decisions for sustainable urban development. However, even with high spatial resolution, remote sensing imagery still lags behind natural imagery in terms of resolution and quality because of the influence of factors like satellite altitude, satellite optical system performance, and manufacturing costs, resulting in the widespread existence of mixed pixels. The edge contours and surface texture information of ground objects are eroded. On the other hand, a UGS encompasses various types of plants with different spectral information, and the spectral information varies greatly between plant species. Consequently, the model needs to store a more extensive range of information to effectively capture the diversity in UGS. At the same time, high-spatial resolution remote sensing imagery contains various objects with similar spectral information, such as UGS, farmlands, forests, and some plant-rich water bodies, for which spectral distinguishment is challenging. These characteristics make the boundary of the spectrum of ground objects in the feature space more “steep” and thus, it is difficult to accurately extract UGS information using threshold methods or shallow learning methods.

With the development of deep learning, image segmentation methods based on deep neural networks have found wide applications in tasks such as pedestrian detection [19],

lane recognition [20], and object identification [21]. They have become vital solutions for urban planning, environmental monitoring, ecological research, and more. Furthermore, these methods offer new possibilities for extracting UGS from high-spatial resolution remote sensing imagery because deep learning models can easily extract complex features without manual design and substantial prior knowledge and can learn nonlinear mapping relationship between inputs and outputs [22]. Consequently, Xu et al. [23] proposed a deep learning classification method for UGS, utilizing phenological characteristics as constraints. This approach takes full advantage of the spectral and spatial information provided by high-resolution remote sensing imagery from different periods. Vegetation phenological features are introduced as auxiliary bands into deep learning networks for training and classification. Wang et al. [21] introduced a multi-level UGS segment architecture based on DeepLab V3+, aimed at extracting urban green space information from high-resolution remote sensing imagery. Shi et al. [24] presented a general deep learning (DL) framework for large-scale urban green space mapping and generated fine-grained UGS maps (UGS-1 m) for 31 major cities in mainland China. Liu et al. [25] introduced a novel hybrid approach, the Multi-Scale Feature Fusion and Transformer Network (MFFTNet), as a fresh deep learning method for extracting urban green spaces from GF-2 high-resolution remote sensing satellites. However, the task of extracting UGS is fundamentally different from natural image semantic segmentation tasks [19–21]. To begin with, UGS exhibit highly irregular and unpredictable edges compared with objects like people, cars, and buildings, resulting in irregular shapes in remote sensing imagery. Additionally, UGS span a wide range of spatial scales, encompassing large areas like parks, as well as small isolated green spaces such as individual trees and garden plots. Moreover, there are similarities in surface texture features among different UGS. Due to these factors, directly applying natural image semantic segmentation models to the task of UGS segmentation will lead to a performance decline.

To address the aforementioned issues, following a thorough investigation and analysis of UGS, in this paper we proposed an end-to-end UGS segmentation model, named the Green Space Reverse Pixel Shuffle Network (GSRPnet), for extracting UGS from GF-2 remote sensing imagery. The main work presented in this paper can be summarized as follows: (1) To minimize the loss of UGS information during the model down-sampling process, and in line with the characteristics of UGS (a low-rank feature, analysis in later section), an enhanced UGS feature extraction backbone network called RPS-ResNet was proposed. RPS-ResNet replaces the large kernel convolutional layer and the max-pooling layer in the origin ResNet-50 with a reverse Pixel Shuffle approach, which is proposed in this paper. Specifically, the last residual convolutional layer in the origin ResNet-50 was also removed to reduce the model's parameters without compromising accuracy. (2) Instead of using cross entropy or binary cross entropy for segmentation tasks, Focal Loss and the Dice coefficient are combined for GSRPnet training, and the effects of these two losses on the segmentation accuracy of a UGS under different weights are discussed. (3) To validate the correctness and effectiveness of the ideas and modules proposed in this paper, ablation studies were conducted. In addition, five segmentation models were introduced to illustrate the superiority of GSRPnet.

2. Materials and Methods

2.1. Data Sources

The UGS segmentation task can be seen as a pixel-level image segmentation task. In this paper, the Green Space 2018 dataset (GS2018) and the Wuhan Dense Labeling Dataset (WHDLD) [26] were utilized for both the training of GSRPnet and the validation of accuracy, alongside other deep neural networks for comparison [27–31].

2.1.1. Green Space 2018 Dataset

Deep neural networks typically require a substantial number of paired images for training and testing. However, the currently available publicly accessible data are insufficient to achieve the requirements. To this end, this paper curated a dataset suitable for the training and accuracy verification of deep neural segment networks, named Green Space 2018 Dataset (GS2018). To provide better coverage of the distribution of UGS in various cities across China, GS2018 was created using Gaofen-2 remote sensing imagery, with a spatial resolution of nearly 1 m, from the main urban areas of six cities in the summer of 2018. These cities include Peking, Shanghai, Guangzhou, Shenzhen, Nanjing, and Wuhan. The global urban boundaries (GUBs) [32] were used to define the main urban areas of these cities. Among these, Peking, Shanghai, Guangzhou, and Shenzhen are the most economically developed cities in China, and studying these four cities helps analyze the impact of economic development on UGS. Nanjing and Wuhan are significant cities in the Yangtze River Basin and were chosen as representative subjects. The relative geographical locations of these six cities are shown in Figure 1.

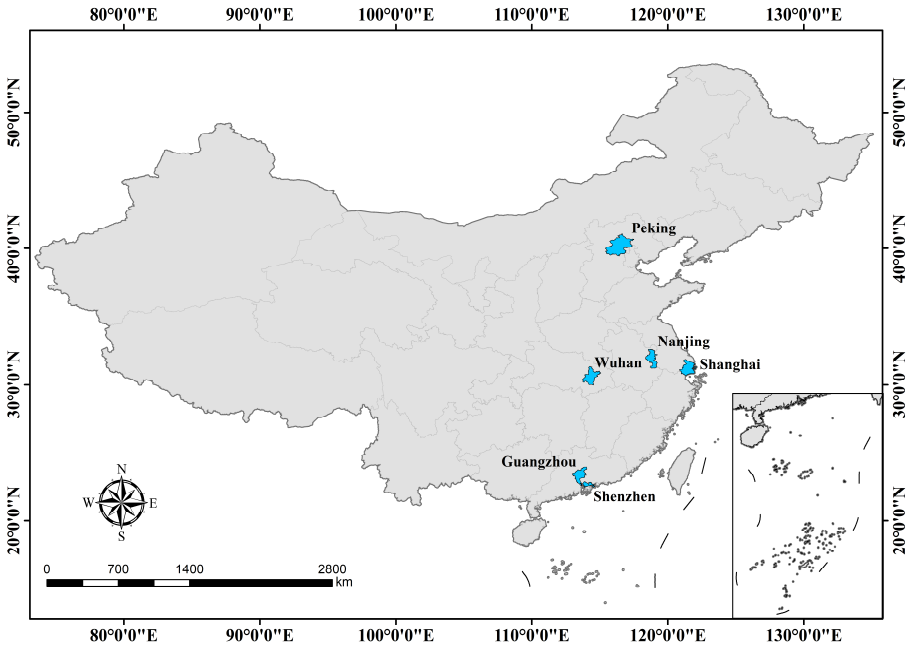

Figure 1. Relative geographical location of the study area.

GS2018 is a high-precision UGSs semantic segmentation dataset with a unified standard. In GS2018, UGS are categorized based on their functionality into five types: park green spaces, protected green spaces, plaza green spaces, subsidiary green spaces, and ecological green spaces. GS2018 was created using the object-oriented method. The original remote sensing images and the annotation binary label images were uniformly divided into small patches with no overlapping areas. GS2018 comprises a total of 9536 pairs of remote sensing images and binary label images, all with a resolution of 256 × 256 pixels. Among these, 7136 pairs were allocated for the training set, 2000 pairs for the test set, and 400 pairs for the validation set. Some examples of the GS2018 dataset are presented in Figure 2, where the highlighted white areas represent UGS.

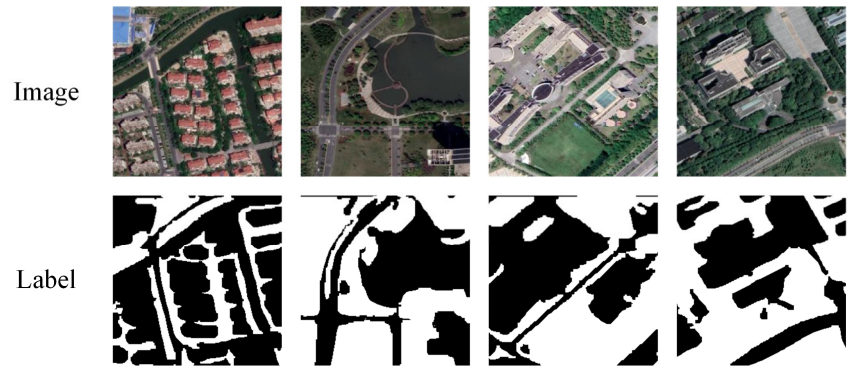

Figure 2. Example of the GS2018 dataset.

2.1.2. Wuhan Dense Labeling Dataset

The Wuhan Dense Labeling Dataset (WHDLD) [26] was produced using the Gaofen-2 remote sensing image of Wuhan, which contained 4940 RGB images with a size of 256 × 256 and a spatial resolution of 2 m. The WHDLD classifies features into six categories, i.e., building, road, pavement, vegetation, bare soil, and water. In order to fit the green space segmentation task, buildings, roads, pavements, bare soil, and water are considered as the background, and vegetation as the target. Some examples of the WHDLD are presented in Figure 3, where the highlighted white areas represent UGS.

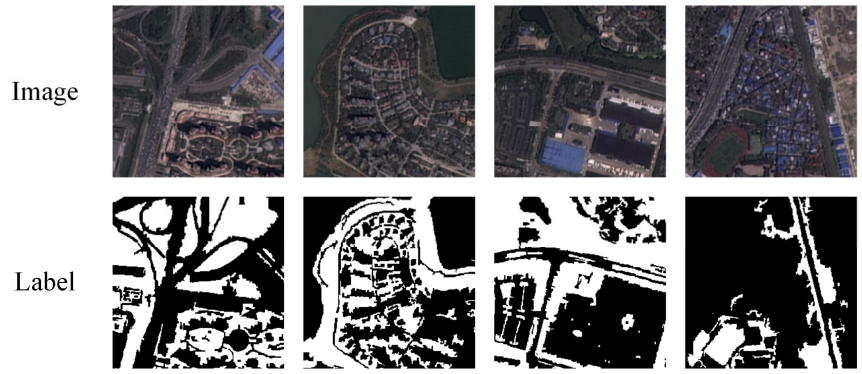

Figure 3. Example of the WHDLD.

2.2. Method

2.2.1. Reverse Pixel Shuffle

Traditional down-sampling methods commonly use convolution operations, compressing pixels within the convolution kernel's range according to specific rules. However, these traditional down-sampling methods will lead to information loss. This becomes more severe, particularly when the segmentation task focuses on fragmented and scattered green spaces. To address the negative impact of the model's initial down-sampling on UGS segmentation, we propose the reverse Pixel Shuffle (RPS) down-sampling method. In contrast to traditional down-sampling methods, RPS transforms the relationships between adjacent pixels into relationships between layers, thus minimizing image information loss while reducing image size. Figure 4 illustrates the principle and process of the RPS method. More specifically, it groups pixels in sets of four, from left to right and top to bottom, combining pixels at the same position within each set into a separate layer, thus halving the image resolution.

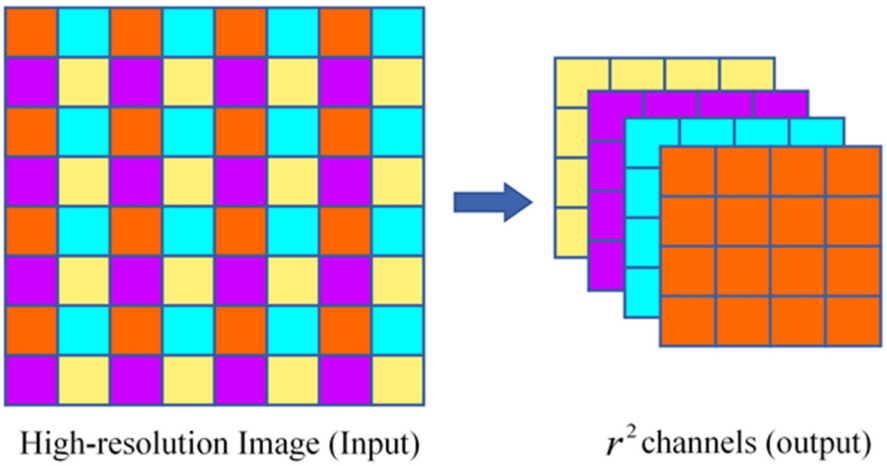

Figure 4. Illustration of the Reverse Pixel Shuffle process.

2.2.2. RPS-ResNet

Neural networks face constraints on their depth due to the issues of gradient vanishing and exploding, hindering model convergence. These problems can be effectively addressed with weight initialization and intermediate normalization layers. However, even with these solutions, the issue of degradation persists as the model becomes deeper. To tackle the degradation problem in deep neural networks, a solution was proposed by [33] in the form of residual networks. The structure of a residual convolutional layer is illustrated in Figure 5.

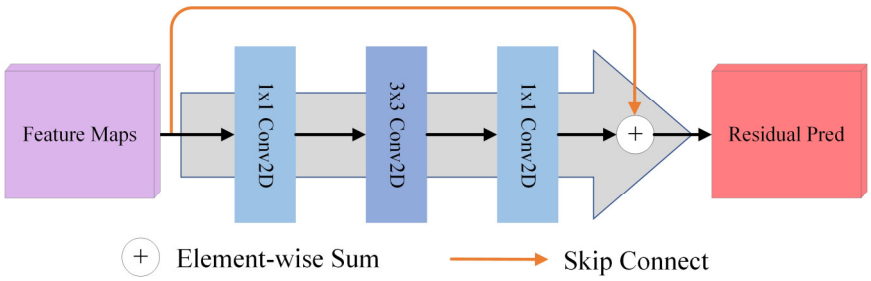

Figure 5. Illustration of the residual structure.

Instead of directly fitting the mapping of the underlying layers, the model aims to fit the residual relative to the input. The residual predict process can be represented as (1):

$$X_{t+1} = X_t + F(X_t) \quad (1)$$

where X_{t+1} and X_t are the output and input of the residual network, and F represents the residual network, respectively.

ResNet [33] has been widely used in various tasks [34–36]. To better adapt ResNet to the green space segmentation task, we offer targeted optimizations to ResNet-50 based on the characteristics of UGS. The enhanced backbone network is referred to as Reverse Pixel Shuffle ResNet (RPS-ResNet).

The network configuration of ResNet-50 is depicted in Figure 6a. Firstly, we replace the large kernel convolution and max-pooling layers used for down-sampling with the proposed RPS. This transformation converts relationships between adjacent pixels into relationships between layers, mitigating the impact of mixed pixels in remote sensing imagery on the accuracy of UGS segmentation. Additionally, we simplified the network structure by removing the final residual convolution layer, Conv5, to accelerate the convergence speed and improve the accuracy of UGS segmentation in RPS-ResNet. Specifically, the network input of RPS-ResNet first undergoes a layer of Reverse Pixel Shuffle, down-sampling the

feature map by a factor of two. Subsequently, it passes through a 3 × 3 convolutional layer and a Silu activation function to expand the feature map dimensions. It is noteworthy that this differs from the original ResNet-50, which uses large kernel convolutions and max-pooling layers, performing two times down-sampling operations in total. RPS-ResNet only reduces the resolution of the input by half. RPS-ResNet's R2–R4 are designed to be residual convolutional layers similar to ResNet-50's Conv2–Conv4. The specific network configuration of RPS-ResNet is illustrated in Figure 6b.

ResNet-50	
Layer name	
Conv1	7×7, 64, stride 2
Conv2	3 × 3 max pool, stride 2, 64 $\begin{bmatrix} 1\times1.\ 64 \\ 3\times3.\ 64 \\ 1\times1.\ 256 \end{bmatrix}\times3$
Conv3	$\begin{bmatrix} 1\times1.\ 128 \\ 3\times3.\ 128 \\ 1\times1.\ 512 \end{bmatrix}\times4$
Conv4	$\begin{bmatrix} 1\times1.\ 256 \\ 3\times3.\ 256 \\ 1\times1.\ 1024 \end{bmatrix}\times6$
Conv5	$\begin{bmatrix} 1\times1.\ 512 \\ 3\times3.\ 512 \\ 1\times1.\ 2048 \end{bmatrix}\times3$

(**a**)

RPS-ResNet	
Layer name	
R1	Reverse Pixel Shuffle 3×3 Conv2D, stride 2, 64
R2	$\begin{bmatrix} 1\times1.\ 64 \\ 3\times3.\ 64 \\ 1\times1.256 \end{bmatrix}\times3$
R3	$\begin{bmatrix} 1\times1.128 \\ 3\times3.128 \\ 1\times1.512 \end{bmatrix}\times4$
R4	$\begin{bmatrix} 1\times1.\ 256 \\ 3\times3.\ 256 \\ 1\times1.1024 \end{bmatrix}\times6$

(**b**)

Figure 6. Comparison of different feature extraction backbones: (**a**) the original ResNet-50 and (**b**) the proposed RPS-ResNet.

2.2.3. GSRPnet Network Structure

In this paper, we proposed a deep neural network for UGS segmentation, named the Green Space Reverse Pixel Shuffle Network (GSRPnet). To expedite the model fitting speed and improve the accuracy of the UGS segmentation task, GSRPnet utilize the proposed RPS-ResNet as the feature extraction backbone. GSRPnet concatenates the feature map output by each residual convolutional layer to connect contextual information and high–low level features, enhancing the accuracy of UGS recognition. As shown in Figure 7, GSRPnet's network structure is straightforward, yet subsequent experiments indicate its effectiveness. GSRPnet can be divided into three components: the down-sampling module, the feature extraction module, and the feature compression filtering module.

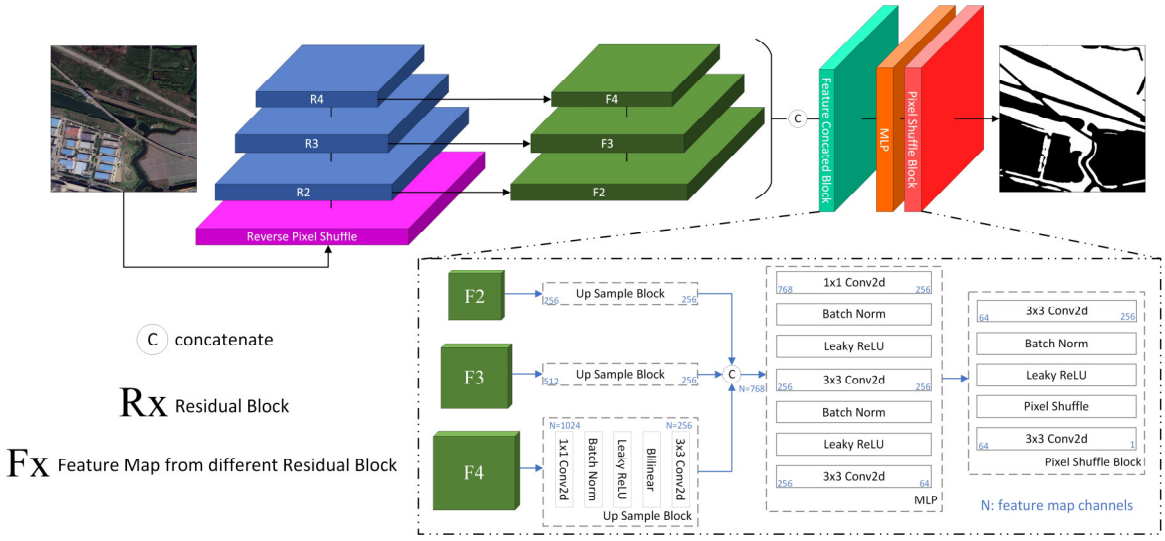

Figure 7. Overall architecture of the proposed GSRPnet.

In the down-sampling module, the input remote sensing images undergo a half-resolution reduction using RPS to conserve computational resources, along with a convolutional layer to expand the feature map dimension. In the feature extraction module, features are extracted through R2, R3, and R4, and the resulting feature maps (designated as F2, F3, and F4) have their resolution increased to half of the original resolution using bilinear interpolation. Subsequently, these feature maps are concatenated along the channel dimension to form a global feature block named the Feature Concatenated Block. In the feature compression filtering module, the global feature block undergoes compression and filtering of green space features through an MLP layer. Finally, the Pixel Shuffle method [37] is used to raise the feature map resolution to the target space, obtaining the binary image of a UGS.

2.2.4. Loss Function

In this paper, following previous works [24–26], we combined Focal Loss with the Dice coefficient to selectively enhance segmentation accuracy. Focal Loss, introduced by [38] and designed to address the issue of imbalanced positive and negative samples in dense object detection tasks, is defined as follows:

$$FL(p_t) = -\alpha_t(1-p_t)^{\mu}\log(p_t) \tag{2}$$

$$p_t = \begin{cases} p & y = 1 \\ 1-p & otherwise \end{cases} \tag{3}$$

where p represents the probability value of a model's output pixel being classified as green space, while α_t and μ are weighting factors and focusing parameters. These are used to adjust imbalances between samples, and $\alpha_t \in [0,1]$, $\mu > 0$.

The Dice coefficient is a function that measures the similarity between two sets, with values ranging from 0 to 1. The Dice coefficient is defined as (4):

$$Dice = \frac{2 \times |X \cap Y|}{|X| + |Y|} \tag{4}$$

where $|\,|$ and $\cap$ represent the summation and intersection operators, and X and Y represent the model's output and the ground truth labels, respectively.

During the training process, the total loss of GSRPnet can be expressed as (5):

$$L_{total} = \lambda \times FL(P_t) + \gamma \times Dice \tag{5}$$

where γ and λ are hyperparameters that control the respective weights of the two different losses in the overall loss.

3. Experiment

3.1. Accuracy Is Not Heavily Reliant on the Depth of the Neural Networks

By analyzing the spatial distribution and spectral information of a UGS, it can be found that the UGS constitutes a kind of low-rank feature. This implies that the accurate determination of whether a location is green space can be achieved with just one or a few surrounding pixels. Excessive down-sampling (which means a larger receptive field) does not significantly enhance the accuracy of UGS segmentation task. On the other hand, using conventional methods (convolution, filtering) for down-sampling during the model's input phase leads to a loss of UGS information. To validate this hypothesis, experiments were conducted, as described in this section.

As shown in Table 1, we used ResNet-50 as the backbone for Model-1 and consider it as the baseline. The other four models are modifications built on Model-1. In comparison with Model-1, which utilizes the complete ResNet-50 as the feature extraction backbone, Model-2 removes the last residual convolutional layer (Conv5) only. This adjustment achieves nearly

identical accuracy as Model-1 while reducing device requirements and computational resources. This indicates that the feature maps with a larger receptive field output by the residual convolutional layer Conv5 in Model-1 do not contribute to the accuracy improvement in the UGS segmentation task. To further investigate the characteristics of UGS, we built Model-3 based on Model-2 by removing the max-pooling layer. Compared with Model-2, Model-3, with one fewer down-sampling operation, has larger-sized feature map output by the residual convolutional layers Conv2-4. Although the receptive field is smaller, Model-3 exhibits further improvement in the IoU and OA accuracy evaluation metrics. This suggests that, compared with models with deeper network structures, a wider model can achieve better results in UGS segmentation tasks.

Table 1. Ablation study for the urban green space segmentation model's structure on the GS2018 validation dataset. The best and second-best results are highlighted in bold and underlined, respectively.

Model	Conv1	Max Pooling	Conv2	Conv3	Conv4	Conv5	IoU ↑ (%)	OA ↑ (%)	Params ↓ (M)	FLOPs ↓ (G)
Model-1	√	√	√	√	√	√	84.090	93.221	51.406	339.666
Model-2	√	√	√	√	√		84.067	93.382	<u>18.008</u>	**179.203**
Model-3	√		√	√	√		<u>84.274</u>	<u>93.463</u>	<u>18.008</u>	216.042
Model-4	Reverse Pixel Shuffle		√	√	√		**86.702**	**94.393**	**17.999**	<u>215.751</u>

The number of FLOPs were tested using an open-source package for a single image with a resolution of 256 × 256. The ↑ and ↓ indicate the larger the better and the smaller the better, respectively.

To minimize the loss of UGS information caused by down-sampling during the model's input stage, we proposed a reverse Pixel Shuffle (RPS) method. RPS is applied to Model-4. Compared with the other models, Model-4 achieves the best results with IoU and OA accuracy evaluation metrics of 86.702% and 94.393%, respectively.

3.2. Effectiveness of RPS-ResNet

In Section 2.2.2, we introduced RPS-ResNet, a ResNet-50 improvement designed based on the distribution characteristics of UGS. To investigate the effectiveness of RPS-ResNet, we conducted the following ablation study on the GS2018 dataset. GSPRnet and GSRSnet represent models using RPS-ResNet and ResNet-50 as the feature extraction backbone, respectively. Both models were trained and validated under identical conditions. Figure 8 illustrates the performance curves of the two models. GSPRnet exhibits significant advantages over GSRSnet both in the initial and highest performance. Moreover, GSPRnet's IoU accuracy improvement curve is smoother, indicating a more stable training process, with less fluctuation in IoU metrics during training. This suggests that the enhanced RPS-ResNet outperforms the original ResNet-50 in UGS segmentation tasks.

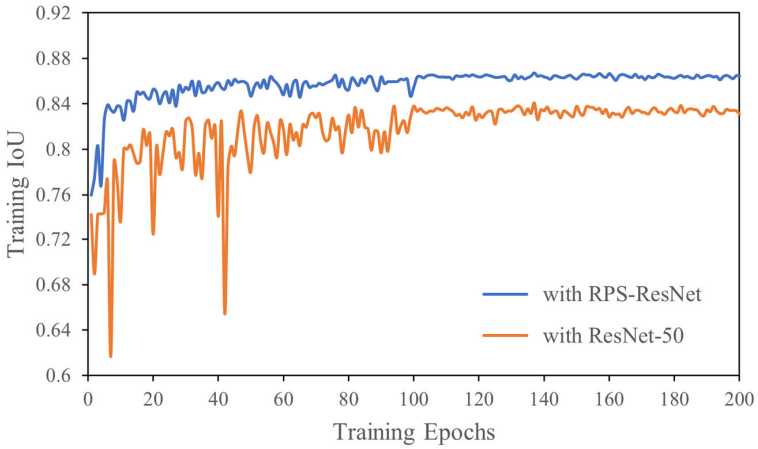

Figure 8. Ablation study for the different feature extraction backbones on the GS2018 validation dataset.

3.3. The Weight for Two Losses

GSPRnet, which is proposed in this paper, uses the combination of Focal Loss and the Dice coefficient as the overall loss function. In order to illustrate the influence of the weights of Focal Loss and the Dice coefficient in the overall loss on the model performance, we conducted an ablation study.

Firstly, keeping the weight γ of the Dice coefficient fixed at 1, we varied the weight λ of Focal Loss from 1, increasing by 0.1 each time, to explore the impact of the Focal Loss's proportion in the total loss on the experimental results. As shown in Figure 9a, the model's IoU initially experiences a continuous decline until λ reaches around 1.5, after which it gradually increases. When λ reaches 2, the training IoU reaches a maximum value, but the further increase in λ leads to a subsequent decline. When λ reaches 2.5, the training IoU is even lower than initial level.

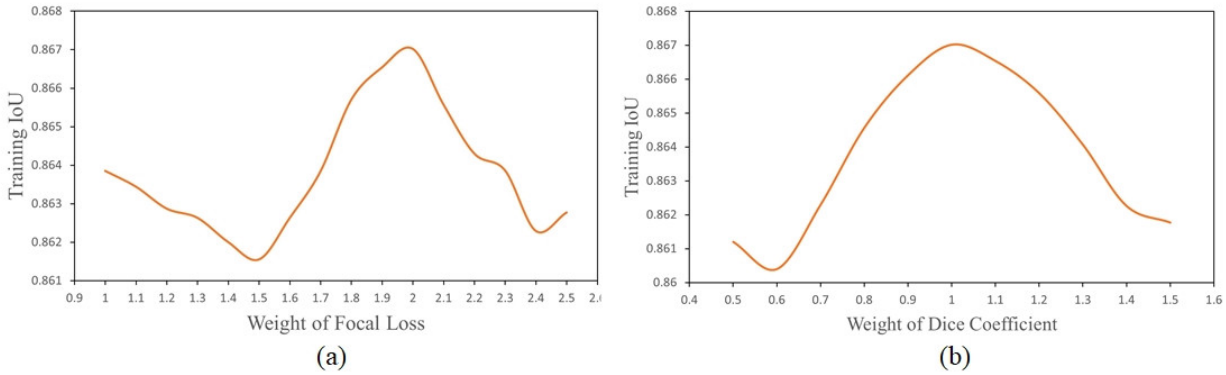

Figure 9. Ablation study for different loss weight settings on the GS2018 validation dataset. (**a**) The change trend in the training IoU with the increase in the Focal Loss weight. (**b**) The change trend in the training IoU with the increase in the Dice coefficient weight.

For the Dice coefficient, we fixed the weight λ of Focal Loss at 2, which achieved the best results in experiment, while the weight γ of the Dice coefficient started from 0.5, increasing by 0.1 each time. Figure 9b illustrates the variation trend in the training IoU with the increase in the weight γ, with the best result being achieved when γ is 1.

The experimental results indicate that the model achieves the best performance when the weights of the hyperparameters λ and γ are set to 2 and 1, respectively.

3.4. The Effect of Different Loss Functions on the Urban Green Space Segmentation Task

Traditional segmentation models often utilize cross entropy or binary cross entropy as the loss function. In contrast to nature image segmentation tasks [16–18], the UGS segmentation task exhibits a relatively balanced proportion between the foreground (UGS) and background (non-UGS). In some specific areas (such as forest parks), the foreground may even outnumber the background. Simultaneously, the spatial distribution of a UGS is fragmented and scattered, with isolated green spaces (such as individual trees) appearing frequently. In such cases, cross entropy performs poorly, especially in the boundary regions of a UGS.

In the previous section's experiments, we only discussed the use of hybrid loss (the combination of Focal Loss and the Dice coefficient). In this section, we explore the impact of different loss functions on UGS segmentation task accuracy. Keeping the other conditions consistent, we train the model using cross entropy, binary cross entropy, and hybrid loss, respectively.

Figure 10 illustrates the variation trend in the IoU as the epoch increases when using different loss functions. In comparison with models using cross entropy or binary cross entropy, the model using hybrid loss achieves the highest IoU metric at the beginning of the training and exhibits the highest accuracy improvement rate over the next 20 epochs. On the other hand, models using cross entropy or binary cross entropy loss experience

significant fluctuations in the IoU metric during the first 60 epochs, whereas the model using hybrid loss demonstrates a smoother IoU improvement. Furthermore, over the entire training period of 200 epochs, the model using hybrid loss maintains a performance advantage in the IoU compared with models using the other two loss functions.

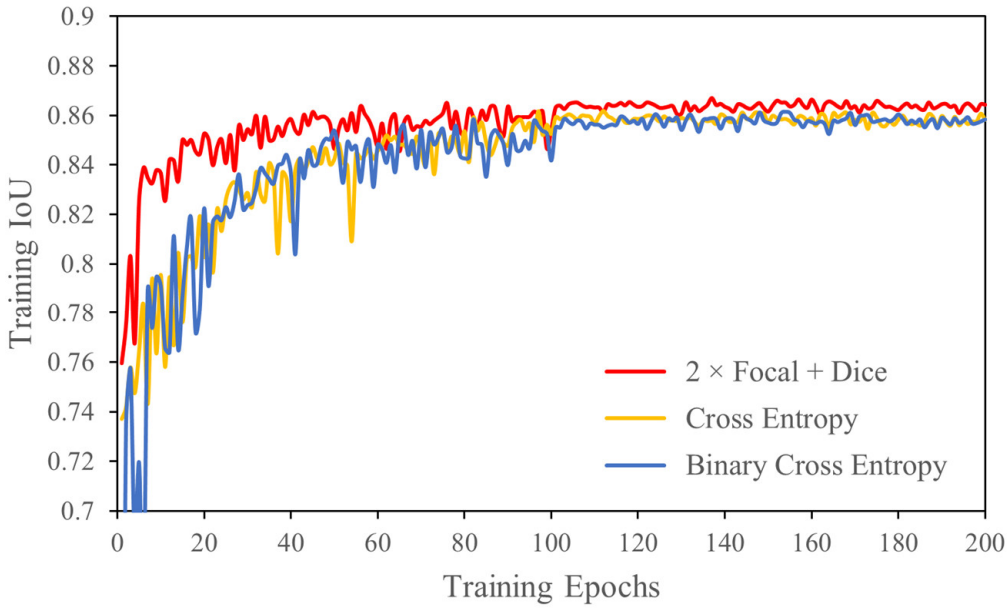

Figure 10. Ablation study for different loss functions on the GS2018 validation dataset.

As shown in Table 2, the model using hybrid loss achieved the highest values of 93.665%, 86.702%, and 94.393% for precision (Pre), intersection over union (IoU), and overall accuracy (OA), respectively. Notably, the model using hybrid loss reached its peak accuracy at the 136th epoch, while the other models attained their best accuracy after the 150th epoch. This indicates that the model using hybrid loss exhibits a faster convergence speed. The results above suggest that the use of hybrid loss can more effectively guide model training, enabling a model to store more features of a UGS and enhancing the performance of the model.

Table 2. The UGS semantic segmentation model obtained the best accuracy and epoch on the validation set with different loss functions. The best and second-best results are highlighted in bold and underlined, respectively.

Loss Type	Pre ↑ (%)	IoU ↑ (%)	OA ↑ (%)	Best Epoch ↓
Cross entropy	92.845	<u>86.178</u>	94.113	160
Binary cross entropy	<u>93.408</u>	86.145	<u>94.145</u>	<u>154</u>
2 × Focal + Dice	**93.665**	**86.702**	**94.393**	**136**

The ↑ and ↓ indicate the larger the better and the smaller the better, respectively.

4. Results

In this section, GSRPnet was compared with other five segmentation models, including U-net [27], PSPnet [28], Segnet [29], DeeepLab V3+ [30], and Separable Graph Convolutional Network (SGCN-Net) [31]. A comparison of these models is listed in Table 3.

Table 3. A comparison of the models used in this paper.

Model	Description
U-net [27]	Unet's distinctive U-shaped structure incorporates contracting and expansive pathways, enabling the model to capture contextual information.
PSPnet [28]	PSPnet utilizes a pyramid pooling module to capture contextual information at multiple scales.
Segnet [29]	Segnet uses an encoder–decoder structure with a skip architecture to capture and reconstruct spatial information effectively.
DeepLab V3+ [30]	DeepLab V3+ incorporates atrous convolution [39] and a feature pyramid to capture fine details and context in images.
SGCNnet [31]	SGCNNet applies graph convolutional networks to segmentation tasks to capture contextual information of channel and spatial features.

4.1. Training Details

GSPRnet was implemented using Python 3.7 on the PyTorch 1.12.1 framework, was trained and tested in an environment consisting of Ubuntu 20.04.5 LTS, CUDA 11.3, and CUDNN 8.3.2, and utilized an RTX 2080Ti GPU with 11 GB VRAM. The model was trained with a learning rate initialized at 1×10^{-3}, and the Adam optimizer [40] was used to update the model parameters, with momentum values β_1 and β_2 set to 0.9 and 0.999, respectively. To enhance the data during the model training process, random horizontal and vertical flips were applied. The model was trained for a total of 200 epochs, with a linear decrease in the learning rate at the 100th epoch. The batch size was set to 10. Based on the experimental results in Section 3.3, the loss weight hyperparameters λ and γ were set to 2 and 1, respectively.

4.2. Evaluation Metrics

UGS segmentation can be regarded as a binary classification instance segmentation task. Following the practices outlined in [41–43], we used five metrics to evaluate model performance, including precision (Pre), the F1-score (F1), intersection over union (IoU), overall accuracy (OA), and the number of model parameters (Params). These metrics are defined as follows:

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FP} \tag{7}$$

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{8}$$

$$IoU = \frac{TP}{FP + TP + FN} \tag{9}$$

$$OA = \frac{TP + TN}{FP + TP + FN + TN} \tag{10}$$

where TP, FP, TN, and FN represent True Positive, False Positive, True Negative, and False Negative, respectively. Params describes the total number of trainable parameters in the model.

4.3. Comparison Experiment of the GS2018 Dataset

In this section, we describe experiments conducted on the GS2018 dataset. As shown in Table 4, compared with other models, GSRPnet achieved the best Pre, F1-score, IoU, and OA, with values of 89.328%, 89.427%, 81.344%, and 92.256%, respectively. Overall, when compared with the worst-performing Segnet, our model exhibited remarkable improvements in Pre (5.069%), F1-score (4.207%), IoU (6.576%), and OA (3.188%). Moreover, compared with the second-best-performing SGCNnet, GSRPnet still demonstrated improve-

ments of 1.306% in Pre, 0.431% in the F1-score, 0.725% in IoU, and 0.4% in OA. Particularly noteworthy is the minimal cost of increased model parameters for GSRPnet compared with the U-net model, with a slight increase of 3.21 M. Simultaneously, considering GSRPnet's model parameters compared with the other four comparison models, it has great advantages, even falling below one-third of DeepLab V3+. These results quantitatively reflect the effectiveness of GSRPnet in UGS segmentation tasks and demonstrate that the improved performance of the GSRPnet model is achieved not only through an increase in parameters.

Table 4. Quantitative comparison of the results on the GS2018 test dataset. The best and second-best results are highlighted in bold and underlined, respectively.

Model	Pre ↑ (%)	F1-Score ↑ (%)	IoU ↑ (%)	OA ↑ (%)	Params ↓ (M)
U-net	<u>88.449</u>	88.770	80.273	91.749	**14.789**
PSPnet	86.884	87.965	79.096	91.166	52.495
Segnet	84.259	85.220	74.768	89.068	29.444
DeepLab V3+	85.977	86.250	76.380	89.995	59.339
SGCNnet	88.022	<u>88.996</u>	<u>80.619</u>	<u>91.856</u>	43.908
GSRPnet	**89.328**	**89.427**	**81.344**	**92.256**	<u>17.999</u>

The ↑ and ↓ indicate the larger the better and the smaller the better, respectively.

Figure 11 illustrates the visualized UGS segmentation results of a typical UGS across all comparison models. These areas include urban outskirts, squares, parks with associated water bodies, areas around railways, airport surroundings, and transportation hubs. Because of the special encoder–decoder and skip-connected structure, the U-net model retains more low-rank features; thus, the UGS edges of the recognition results are smoother. However, in the case of a scattered UGS, the recognition accuracy of U-net is reduced. PSPnet provides relatively accurate segmentation of the UGS, but errors in recognition occur when small non-UGS patches appear within the UGS. SegNet and DeepLab V3+ exhibit a jagged pattern in regions where the UGS and non-UGS intersect each other. In particular, the two models do not perform well when UGS are in a strip-like pattern. Overall, DeepLab V3+ outperforms Segnet. SGCNnet produces UGS segmentation results with smooth edges, which aligns better with ground truth. However, it does not perform well in areas where a UGS and non-UGS are in contact. The proposed GSRPnet in this paper demonstrates a smoother edge performance compared with the other models, accurately identifying discrete small UGS, especially in areas where UGS and non-UGS are interspersed with each other, and the output results closely resemble the ground truth.

4.4. Comparison Experiment of the WHDLD

In order to further verify the robustness and generalization of GSRPnet and the correctness of our proposed hypothesis, we conducted experiments on the WHDLD. In this section, we follow the methods of [26], who randomly selected 80% of the WHDLD as the training set and 20% as the testing set. We set the learning rate, number of epochs, and batch size to 0.001, 120, and 10, respectively.

As shown in Table 5, the advantage of GSRPnet is further expanded. Compared with SGCNnet, which achieved the second-best result, GSRPnet achieved 0.116%, 1.458%, 1.989%, and 0.791% improvements in Pre, the F1-score, IoU, and OA, respectively, and the number of model parameters is less than half of that of SGCNnet. Compared with the other four models, GSRPnet exhibits varying degrees of improvement in all indicators.

Figure 11. UGS segment results using the GS2018 test dataset. (**a**) Satellite imagery, (**b**) ground truth, (**c**) GSRPnet, (**d**) Unet, (**e**) PSPnet, (**f**) Segnet, (**g**) DeepLab V3+, and (**h**) SGCNnet.

Table 5. Quantitative comparison results on the WHDLD test dataset. The best and second-best results are highlighted in bold and underlined, respectively.

Model	Pre ↑ (%)	F1-Score ↑ (%)	IoU ↑ (%)	OA ↑ (%)	Params ↓ (M)
U-net	85.845	84.491	74.801	89.213	**14.789**
PSPnet	84.087	83.438	73.391	88.412	52.495
Segnet	85.258	81.948	71.445	87.802	29.444
DeepLab V3+	83.348	83.835	72.736	87.152	58.036
SGCNnet	<u>87.664</u>	<u>84.567</u>	<u>75.003</u>	<u>89.582</u>	43.908
GSRPnet	**87.830**	**86.025**	**76.992**	**90.373**	<u>17.999</u>

The ↑ and ↓ indicate the larger the better and the smaller the better, respectively.

Figure 12 shows the visualized results of UGS segmentation for all models on the WHDLD testing set. Similar to the results of GS2018, when faced with scattered UGS, U-net still struggles to discern effectively, and there is a phenomenon of misclassification. PSPnet is insensitive to continuous strip-like UGS. Segnet is the least effective model, and it cannot distinguish well when the spectral information of adjacent objects is similar. DeepLab V3+ shows a serious misclassification when dividing scattered UGS, particularly in residential areas. SGCNnet does not produce fine UGS well in the area of interleaving UGS and non-UGS. The GSRPnet proposed in this paper still shows the best results on the WHDLD, which is the closest to the ground truth, and further proves the advanced ability of GSRPnet in the UGS segmentation task.

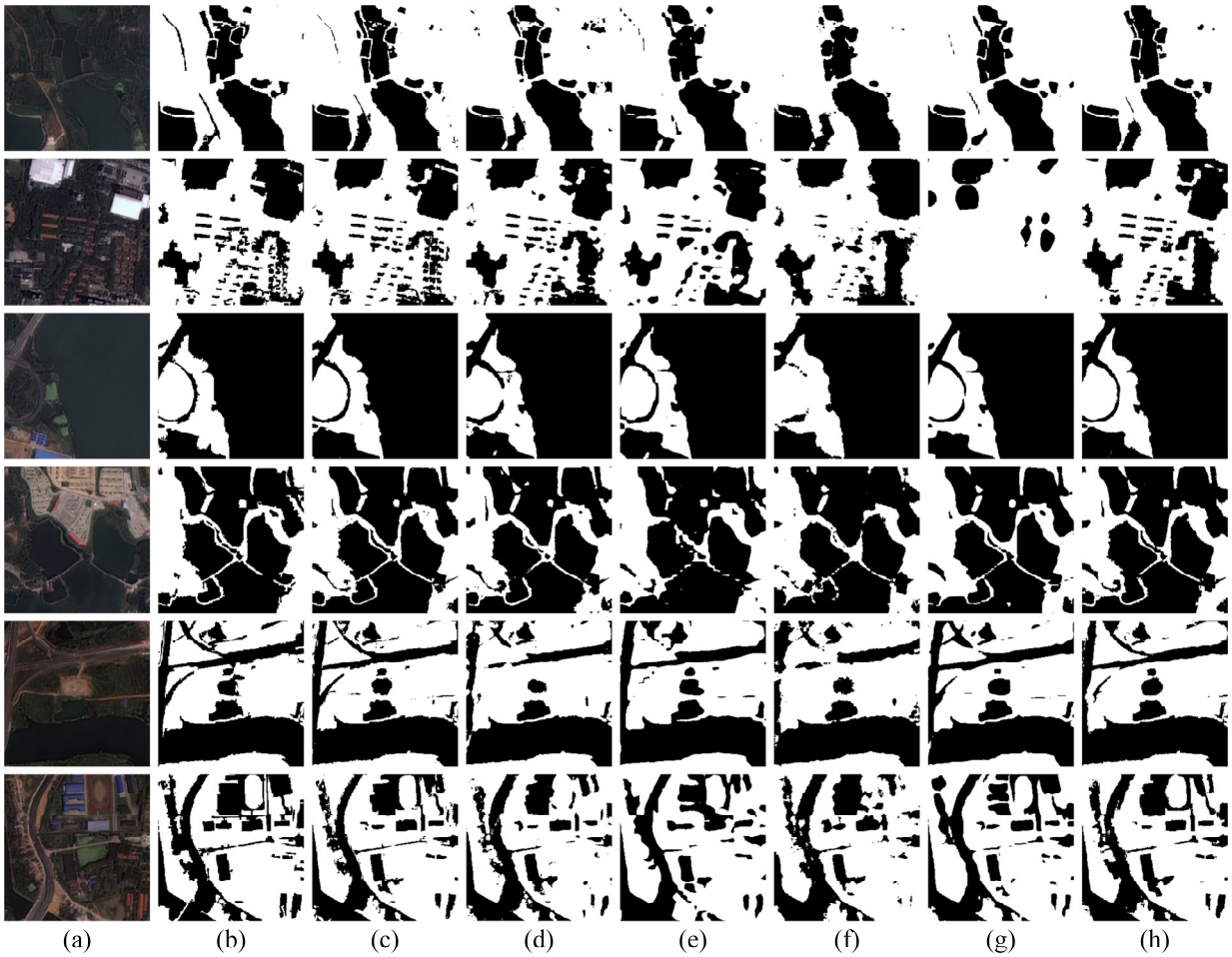

Figure 12. UGS segment results under the WHDLD test dataset. (**a**) Satellite imagery, (**b**) ground truth, (**c**) GSRPnet, (**d**) Unet, (**e**) PSPnet, (**f**) Segnet, (**g**) DeepLab V3+, and (**h**) SGCNnet.

5. Discussion

Benefiting from the high-spatial resolution and relatively short revisit period of remote sensing imagery, extracting UGS information from such imagery proves to be a quick and effective approach. Furthermore, with the development of deep learning technology, semantic deep segmentation neural networks can achieve higher accuracy compared with traditional methods. However, the exiting deep segmentation neural networks still encounter challenges when applied directly to the UGS segmentation task. After studying the characteristics of UGS, we found that a UGS is a kind of low-rank features. Because of the characteristics of low-level features, the detection of a UGS is not highly dependent on the model's ability to understand semantic features, which means that the recognition ability of UGS surface texture and color features is more important. In order to further research on the UGS segmentation task with deep learning technology, we proposed an end-to-end UGS segmentation model named GSRPnet. In comparison with five other segmentation models [27–31], GSRPnet exhibits several advantages as follows:

1. In addressing the challenges posed by small-scale and fragmented UGS, GSRPnet outperforms the other five models. For instance, in scenarios where roads and buildings are surrounded by trees or other types of UGS (Figure 11, sixth row and Figure 12, second row), GSRPnet excels in clearly distinguishing their outlines without interruptions or omissions. In contrast, Segnet [29] and DeepLab V3+ [30] struggle to discern surface objects distinctly and may even misclassify non-UGSs as UGS. This discrepancy arises from the fact that these models use regular down-sampling methods (such as convolution, resampling, etc.) at the model input stage to conserve

computational resources. However, these regular down-sampling methods result in the loss of texture information of the input images. GSRPnet utilizes the reverse Pixel Shuffle method to reduce image size and minimize the loss of UGS information.

2. In GSRPnet, multi-scale feature maps (as shown in Figure 7, F2-F4) are concatenated to offer multi-level UGS feature information for the MLP layer. This strategy of concatenating feature maps ensures the provision of multiple and different levels of UGS information for subsequent MLP layer decisions. The key advantage of this strategy is that GSRPnet becomes more responsive to both non-UGS and UGS boundaries. This strategy's effectiveness is also evident in other comparative models using similar approaches, such as the Unet [27] model with a skip-connection strategy, which also exhibits clear boundaries. However, the concatenation strategy in GSPRnet is more effective.
3. From the experiments on both the GS2018 dataset and WHDLD, as shown in Tables 4 and 5, GSRPnet achieved the best results in all the four precision measurement metrics. Specifically, GSRPnet achieved Pre, F1-score, IoU, and OA values of 89.328%, 89.427%, 81.344%, and 92.256% and 87.830%, 86.025%, 76.992%, and 90.373% for the GS2018 dataset and WHDLD, respectively. In addition, the number of model parameters (Params) is also a key indicator for measuring the effectiveness of a model. Generally, increasing the depth of the network or increasing the number of model parameters using other means can improve the recognition accuracy of a model. However, this improvement comes at the cost of increased computational resources. In comparison with the other four models, GSPRnet utilizes the second fewest Params (only slightly higher than Unet by 3.21 million), yet it achieves the best results. This clearly demonstrates that GSRPnet aims not only to enhance accuracy by indiscriminately increasing the number of network parameters. It substantiates the effectiveness of GSRPnet, emphasizing that its accuracy improvement is not solely reliant on blindly expanding the number of network parameters.

In general, this paper thoroughly demonstrates the potential of extracting UGS at a high spatial resolution using GSRPnet and provides a feasible scheme for UGS detection. However, GSRPnet still has room for further improvement. In future work, we will aim to delve deeper into enhancing the precision of the UGS segmentation task.

6. Conclusions

In this paper, we demonstrate that a UGS is a low-rank feature. This implies that the accurate determination of whether a location is UGS can be achieved with just one or a few pixels around it, and the segmentation accuracy is not highly dependent on the depth of the model. Excessive down-sampling, which means a large receptive field, does not contribute significantly to accuracy. In contrast, preserving more texture and spectral features aids in enhancing the model's accuracy. Therefore, we propose a novel UGS segmentation network, named GSRPnet, which, with a small number of parameters, improves the accuracy of UGS segmentation. The feature extraction backbone network used in GSRPnet, i.e., RPS-ResNet, is an enhancement of ResNet-50. Particularly, it replaces the original down-sampling convolutional layers and max-pooling layers with the reverse Pixel Shuffle method, transforming relationships between adjacent pixels into relationships between layers. This minimizes the loss of UGS feature information caused by the down-sampling process. Experimental results on GaoFen-2 remote sensing imagery show that, with a parameter count of only 17.999 M, GSRPnet outperforms U-net, PSPnet, Segnet, DeepLab V3+, and SGCN-Net in terms of precision, the F1-score, IoU, and OA. This strongly validates the correctness of our proposed approach.

Author Contributions: M.J. and H.S. conceived of and designed the experiments. Y.L. provided the original remote sensing data. M.J. and X.Z. made the dataset and wrote this paper. H.S. and Y.L. revised this paper and gave some appropriate suggestions. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key R&D Program of China (2022YFF0711600) and the Special Science and Technology Innovation Program for Carbon Peak and Carbon Neutralization of Jiangsu Province (Grant No. BE2022612).

Data Availability Statement: The GaoFen-2 satellite data and the WHDLD data used in this study can be accessed from [24,26], respectively.

Acknowledgments: The authors acknowledge the data support from Yangtze River Delta Science Data Center, National Earth System Science Data Center, and the National Science & Technology Infrastructure of China (http://geodata.nnu.edu.cn/, accessed on 30 September 2023).

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

References

1. Alexandratos, N.; Bruinsma, J. *World Agriculture Towards 2030/2050: The 2012 Revision*; FAO: Rome, Italy, 2012. [CrossRef]
2. Zhong, J.; Li, Z.; Sun, Z.; Tian, Y.; Yang, F. The Spatial Equilibrium Analysis of Urban Green Space and Human Activity in Chengdu, China. *J. Clean. Prod.* **2020**, *259*, 120754. [CrossRef]
3. Kumar, K.S.; Bhaskar, P.U.; Padmakumari, K. Estimation of land surface temperature to study urban heat island effect using Landsat ETM+ image. *Int. J. Eng. Sci. Technol.* **2012**, *4*, 771–778.
4. Kuang, W.; Dou, Y. Investigating the Patterns and Dynamics of Urban Green Space in China's 70 Major Cities Using Satellite Remote Sensing. *Remote Sens.* **2020**, *12*, 1929. [CrossRef]
5. Chen, J.; Kinoshita, T.; Li, H.; Luo, S.; Su, D.; Yang, X.; Hu, Y. Toward Green Equity: An Extensive Study on Urban Form and Green Space Equity for Shrinking Cities. *Sustain. Cities Soc.* **2023**, *90*, 104395. [CrossRef]
6. Wolch, J.R.; Byrne, J.; Newell, J.P. Urban Green Space, Public Health, and Environmental Justice: The Challenge of Making Cities 'Just Green Enough'. Landsc. *Urban Plan.* **2014**, *125*, 234–244. [CrossRef]
7. Chen, M.; Dai, F.; Yang, B.; Zhu, S. Effects of Urban Green Space Morphological Pattern on Variation of PM2.5 Concentration in the Neighborhoods of Five Chinese Megacities. *Build. Environ.* **2019**, *158*, 1–15. [CrossRef]
8. Margaritis, E.; Kang, J. Relationship between Green Space-Related Morphology and Noise Pollution. *Ecol. Indic.* **2017**, *72*, 921–933. [CrossRef]
9. Dadvand, P.; Nieuwenhuijsen, M. Green Space and Health. In *Integrating Human Health into Urban and Transport Planning: A Framework*; Nieuwenhuijsen, M., Khreis, H., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 409–423. ISBN 978-3-319-74983-9.
10. Bertram, C.; Rehdanz, K. The Role of Urban Green Space for Human Well-Being. *Ecol. Econ.* **2015**, *120*, 139–152. [CrossRef]
11. Hills, A.P.; Farpour-Lambert, N.J.; Byrne, N.M. Precision Medicine and Healthy Living: The Importance of the Built Environment. *Prog. Cardiovasc. Dis.* **2019**, *62*, 34–38. [CrossRef]
12. Su, Y.; Huang, G.; Chen, X.; Chen, S.; Li, Z. Research progress in the eco-environmental effects of urban green spaces. *Acta Ecol. Sin.* **2011**, *31*, 7287–7300.
13. Kellison, T. An Overview of Sustainable Development Goal 11. In *The Routledge Handbook of Sport and Sustainable Development*; Routledge: London, UK, 2022; pp. 261–275.
14. Men, G.; He, G.; Wang, G. Concatenated Residual Attention UNet for Semantic Segmentation of Urban Green Space. *Forests* **2021**, *12*, 1441. [CrossRef]
15. Gandhi, G.M.; Parthiban, S.; Thummalu, N.; Christy, A. Ndvi: Vegetation Change Detection Using Remote Sensing and Gis—A Case Study of Vellore District. *Procedia Comput. Sci.* **2015**, *57*, 1199–1210. [CrossRef]
16. Zhou, X.; Li, L.; Chen, L.; Liu, Y.; Cui, Y.; Zhang, Y.; Zhang, T. Discriminating Urban Forest Types from Sentinel-2A Image Data through Linear Spectral Mixture Analysis: A Case Study of Xuzhou, East China. *Forests* **2019**, *10*, 478. [CrossRef]
17. Zhang, X.; Liu, L.; Chen, X.; Gao, Y.; Xie, S.; Mi, J. GLC_FCS30: Global Land-Cover Product with Fine Classification System at 30 m Using Time-Series Landsat Imagery. Earth Syst. *Sci. Data* **2021**, *13*, 2753–2776. [CrossRef]
18. Ardila, J.P.; Bijker, W.; Tolpekin, V.A.; Stein, A. Context-Sensitive Extraction of Tree Crown Objects in Urban Areas Using VHR Satellite Images. *Int. J. Appl. Earth Obs. Geoinf.* **2012**, *15*, 57–69. [CrossRef]
19. Chu, X.; Zheng, A.; Zhang, X.; Sun, J. Detection in Crowded Scenes: One Proposal, Multiple Predictions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
20. Qin, Z.; Wang, H.; Li, X. Ultra Fast Structure-Aware Deep Lane Detection. In *Proceedings of the Computer Vision—ECCV 2020*; Glasgow, UK, 23–28 August 2020, Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 276–291.
21. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 7464–7475.
22. Wang, Z.; Ma, Y.; Zhang, Y. Review of Pixel-Level Remote Sensing Image Fusion Based on Deep Learning. *Inf. Fusion* **2023**, *90*, 36–58. [CrossRef]

23. Xu, Z.; Zhou, Y.; Wang, S.; Wang, L.; Li, F.; Wang, S.; Wang, Z. A Novel Intelligent Classification Method for Urban Green Space Based on High-Resolution Remote Sensing Images. *Remote Sens.* **2020**, *12*, 3845. [CrossRef]
24. Shi, Q.; Liu, M.; Marinoni, A.; Liu, X. UGS-1m: Fine-Grained Urban Green Space Mapping of 31 Major Cities in China Based on the Deep Learning Framework. Earth Syst. *Sci. Data* **2023**, *15*, 555–577. [CrossRef]
25. Liu, W.; Yue, A.; Shi, W.; Ji, J.; Deng, R. An Automatic Extraction Architecture of Urban Green Space Based on DeepLabv3plus Semantic Segmentation Model. In Proceedings of the 2019 IEEE 4th International Conference on Image, Vision and Computing (ICIVC), Xiamen, China, 5–7 July 2019; pp. 311–315.
26. Shao, Z.; Zhou, W.; Deng, X.; Zhang, M.; Cheng, Q. Multilabel Remote Sensing Image Retrieval Based on Fully Convolutional Network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 318–328. [CrossRef]
27. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Munich, Germany, 5–9 October 2015, Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241.
28. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
29. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
30. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
31. Zhou, G.; Chen, W.; Gui, Q.; Li, X.; Wang, L. Split Depth-Wise Separable Graph-Convolution Network for Road Extraction in Complex Environments from High-Resolution Remote-Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]
32. Li, X.; Gong, P.; Zhou, Y.; Wang, J.; Bai, Y.; Chen, B.; Hu, T.; Xiao, Y.; Xu, B.; Yang, J.; et al. Mapping Global Urban Boundaries from the Global Artificial Impervious Area (GAIA) Data. *Environ. Res. Lett.* **2020**, *15*, 094044. [CrossRef]
33. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
34. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems*; Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2015; Volume 28.
35. Ho, J.; Jain, A.; Abbeel, P. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*; Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 6840–6851.
36. Chen, L.-C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**, arXiv:1706.05587.
37. Shi, W.; Caballero, J.; Huszar, F.; Totz, J.; Aitken, A.P.; Bishop, R.; Rueckert, D.; Wang, Z. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
38. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
39. 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] [PubMed]
40. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2015**, arXiv:1412.6980.
41. Guo, R.; Niu, D.; Qu, L.; Li, Z. SOTR: Segmenting Objects with Transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 7157–7166.
42. Ziwen, C.; Patnaik, K.; Zhai, S.; Wan, A.; Ren, Z.; Schwing, A.G.; Colburn, A.; Fuxin, L. AutoFocusFormer: Image Segmentation off the Grid. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 18227–18236.
43. Dong, J.; Zhang, D.; Cong, Y.; Cong, W.; Ding, H.; Dai, D. Federated Incremental Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 3934–3943.