remote sensing

MDPI

Article

ASPP$^+$-LANet: A Multi-Scale Context Extraction Network for Semantic Segmentation of High-Resolution Remote Sensing Images

Lei Hu *, Xun Zhou, Jiachen Ruan and Supeng Li

School of Computer and Information Engineering, Jiangxi Normal University, Nanchang 330022, China; 202141600066@jxnu.edu.cn (S.L.)
* Correspondence: hulei@jxnu.edu.cn

Abstract: Semantic segmentation of remote sensing (RS) images is a pivotal branch in the realm of RS image processing, which plays a significant role in urban planning, building extraction, vegetation extraction, etc. With the continuous advancement of remote sensing technology, the spatial resolution of remote sensing images is progressively improving. This escalation in resolution gives rise to challenges like imbalanced class distributions among ground objects in RS images, the significant variations of ground object scales, as well as the presence of redundant information and noise interference. In this paper, we propose a multi-scale context extraction network, ASPP$^+$-LANet, based on the LANet for semantic segmentation of high-resolution RS images. Firstly, we design an ASPP$^+$ module, expanding upon the ASPP module by incorporating an additional feature extraction channel, redesigning the dilation rates, and introducing the Coordinate Attention (CA) mechanism so that it can effectively improve the segmentation performance of ground object targets at different scales. Secondly, we introduce the Funnel ReLU (FReLU) activation function for enhancing the segmentation effect of slender ground object targets and refining the segmentation edges. The experimental results show that our network model demonstrates superior segmentation performance on both Potsdam and Vaihingen datasets, outperforming other state-of-the-art (SOTA) methods.

Keywords: high-resolution remote sensing images; semantic segmentation; ASPP module; local attention network model; activation function

Citation: Hu, L.; Zhou, X.; Ruan, J.; Li, S. ASPP$^+$-LANet: A Multi-Scale Context Extraction Network for Semantic Segmentation of High-Resolution Remote Sensing Images. *Remote Sens.* **2024**, *16*, 1036. https://doi.org/10.3390/rs16061036

Academic Editor: Melanie Vanderhoof

Received: 18 October 2023
Revised: 9 March 2024
Accepted: 12 March 2024
Published: 14 March 2024

1. Introduction

Remote sensing (RS) images can be used to observe natural and artificial phenomena on the Earth's surface. In the field of RS, semantic segmentation of high-resolution RS images entails a pixel-level classification task where the objective is to assign a semantic label to each pixel in the image [1]. These semantic labels mean different ground objects.

Recently, RS images have achieved spatial resolution at the centimeter scale, empowering the discernment of minute details and targets within high-resolution RS imagery. The challenge of semantic segmentation in RS images persists due to issues such as redundant information, noise interference, misclassification of tiny targets, and insufficient smoothness in the edges of ground objects. To solve this problem, this paper proposes a multi-scale context network ASPP$^+$-LANet based on LANet, which improves the segmentation performance of ground object targets at different scales and refines the edges of ground object targets.

The rapid progress of deep neural networks, especially Convolutional Neural Networks (CNNs), has greatly advanced semantic segmentation in RS images. In 2015, Long et al. [2] first proposed the concept of Fully Convolutional Networks (FCNs), an encoder-decoder structure network, which used an anti-convolutional layer instead of the fully connected layer in traditional CNNs. In the same year, the Unet network was introduced by Ronneberger et al. [3], featuring a U-shaped encoder–decoder architecture with inter-layer

skip connections. In 2017, the SegNet network was proposed by Badrinarayanan et al. [4], which implemented an encoder–decoder structure. The key innovation was situated in the decoder, where instead of using deconvolution for upsampling, pooling indices were utilized to conduct non-linear upsampling during the respective encoder's max-pooling steps. The mentioned networks all employed an encoder–decoder structure with robust feature extraction capabilities. Nevertheless, without further refinement, the direct connection between shallow texture information and deep semantic information causes underutilization of feature information, leading to insufficient discrimination between shallow information and deep information. To address these issues, a multi-scale feature extraction module was introduced into the convolutional network by researchers. In 2017, the Pyramid Scene Parsing Network (PSPNet) was introduced by Zhao et al. [5], which proposed the Pyramid Pooling Module (PPM) to aggregate diverse regional contexts. In addition, Chen et al. [6–9] successively proposed DeepLab series networks for extracting multi-scale contextual features. Among them, based on DeepLab v3 [8], a decoder structure was added to DeepLab v3+ [9], which integrated the low-level features of the encoder output with the high-level features of the Atrous Spatial Pyramid Pooling (ASPP) output. Furthermore, attention mechanisms have been extensively employed in semantic segmentation networks. In 2020, a Local Attention Network (LANet) was proposed by Ding et al. [10], introducing a patch-level-based attention mechanism for extracting contextual information. Two approaches were suggested for enhancing the feature representation: the chunked attention module enhances the embedding of contextual information, while the attention embedding module enriches the semantic information of the underlying features by embedding the local focus of the high-level features. The differences in physical information content and spatial distribution are effectively addressed, the disparities between high-level and low-level features are bridged, and significant success in the field of remote sensing image segmentation is achieved. Due to these excellent features, we chose it as our benchmark network. In 2021, Li et al. [11] proposed a Multi-Attention Network (MANet), which designed a novel linear-complexity kernel attention mechanism to alleviate the computational demands of attention. In 2023, a novel three-branch network architecture, PIDNet, was proposed by Xu et al. [12]. PIDNet comprises three branches designed to parse detailed, contextual, and boundary information. Additionally, boundary attention is employed to facilitate the fusion of detailed and contextual branches.

In recent years, the Vision Transformer (ViT) [13] has demonstrated remarkable performance in the field of RS image segmentation due to its powerful self-attention-based global context modeling capability [14–18]. Among them, in 2022, Wang et al. [18] proposed the UnetFormer network for real-time urban scene segmentation in RS images. An efficient global–local attention mechanism known as the Global–Local Transformer Block (GLTB) was implemented by the network to integrate both global and local information within the decoder. A lightweight transformer-based decoder was developed using GLTB and Feature refinement head, which aimed to enhance the network's capability to extract multi-scale contextual features and effectively improve the network's segmentation performance in semantic segmentation of RS images. In 2022, Zhang et al. [19] proposed a hybrid deep neural network, Swin-CNN, combining a transformer and a CNN. The model follows an encoder–decoder structure. A novel universal backbone dual transformer is employed in the encoder module to extract features, thus aiming to enhance long-range spatial dependency modeling. The decoder module leverages some effective blocks and successful strategies from a CNN-based remote sensing image segmentation model. In the middle of the framework, spatial pyramid pooling blocks based on depthwise separable convolutions are applied to obtain multi-scale context.

As previously noted, the incorporation of multi-scale and attention modules into the semantic segmentation network of RS images has been shown to effectively enhance the network's segmentation performance. Accordingly, we designed a new ASPP$^+$ module by augmenting an additional feature extraction channel to the ASPP module, redesigning the dilation rates, and introducing the CA mechanism [20], thereby effectively enhancing

the network's segmentation capability. The utilization of parallel dilated convolutions has been found to enhance the receptive field and capture target features of varying scales. Additionally, the incorporation of the attention module allows the model to prioritize meaningful features and acquire contextual information more effectively. Furthermore, we introduced the FReLU activation function [21] to enhance the network's generalization capability, filter out noise and low-frequency information, and retain more higher-frequency information so as to effectively improve the segmentation performance of slender ground object targets and refine the segmentation edges.

In conclusion, the main contributions of this paper include the following three aspects as follows:

(1) We propose a multi-scale context extraction network for semantic segmentation of high-resolution RS images, ASPP$^+$-LANet, by improving the LANet structure, which effectively tackles the issue of unclear segmentation in various-sized ground objects, slender ground objects, and ground object edges. By adding a new multi-scale module, the segmentation accuracy of ground objects at different scales has been improved. By introducing the activation function, the segmentation accuracy of slender ground objects and ground object edges has been improved.

(2) We designed a novel ASPP$^+$ module to effectively enhance the segmentation accuracy of ground objects at different sizes. This module adds an additional feature extraction channel to ASPP. In addition, we redesigned its dilation rates and introduced the CA mechanism. The attention mechanism can focus on more meaningful areas, improving the overall segmentation progress.

(3) We introduced the FReLU activation function. By integrating it with the LANet network, the performance of ASPP$^+$-LANet has been improved. The activation function can filter out noise and low-frequency information and retain more higher-frequency information so as to effectively enhance the segmentation accuracy of slender ground objects and ground object edges.

2. Materials and Methods

2.1. Materials

In this paper, we design a series of comparative experiments using Potsdam and Vaihingen from the ISPRS dataset [22] in order to evaluate our proposed method.

2.1.1. Potsdam Datasets

The Potsdam dataset consists of 38 images, each with a size of 6000 × 6000 pixels and a spatial resolution of approximately 5 cm [23]. In the Potsdam region, there are six land cover classes, as shown in Figure 1: impervious surfaces, buildings, low vegetation, trees, cars, and clutter/background. The clutter/background class primarily includes water bodies and objects defined as outside the designated classes, which are typically irrelevant semantic objects in urban scenes. To ensure sufficient experimental data, the dataset was preprocessed prior to the experiments, involving image cropping and data augmentation. The images were uniformly cropped into 512 × 512 pixels and subjected to horizontal and vertical flipping for data augmentation. After filtering out images with problematic labels, the dataset was divided into training, validation, and testing sets in a 6:2:2 ratio.

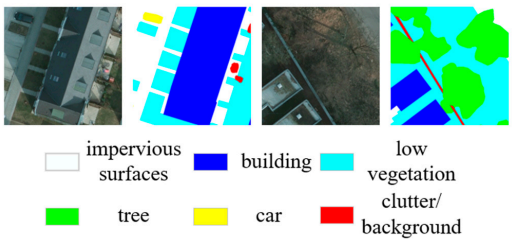


Figure 1. Partial example plot of the Potsdam dataset.

2.1.2. Vaihingen Datasets

The Vaihingen dataset consists of a total of 33 images, each of which has a varying size, with an average dimension of 2496 × 2064 pixels and a spatial resolution of approximately 9 cm [23]. The label categories and color representations are the same as those in the Potsdam dataset, as shown in Figure 2. Prior to the experiments, the images were cropped into 512 × 512 pixels and augmented by horizontal and vertical flips. After filtering out images with problematic labels, the dataset was split into a training set with 6020 image blocks, a validation set with 2006 image blocks, and a test set with 2052 image blocks.

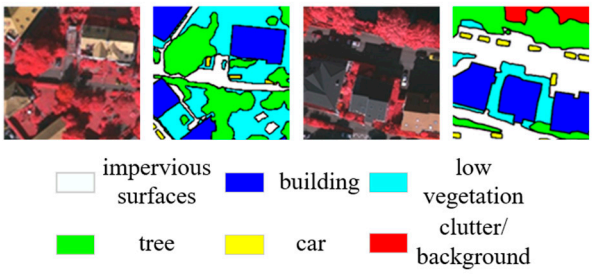


Figure 2. Partial example plot of the Vaihingen dataset.

2.2. Methods

In this section, we provide a comprehensive overview of the proposed network model, ASPP$^+$-LANet. Firstly, we present a concise summary of the network structure, highlighting the general motivation and structure. Subsequently, we explore the intricacies of two pivotal modules: the ASPP+ module and the FReLU activation function. Through the examination of these components, we aim to present a thorough understanding of the ASPP$^+$-LANet network.

2.2.1. Overall Network Structure

We propose a multi-scale context extraction network for semantic segmentation of high-resolution RS images, ASPP$^+$-LANet, as illustrated in Figure 1. Like LANet [10], our network is built upon the FCN framework [2] and employs the pre-trained ResNet50 [24] as the backbone network. It consists of two parallel branches for high-level and low-level feature extraction, incorporating multiple feature extraction and enhancement modules within these branches.

There are two motives in this paper: (1) improving the segmentation performance of ground object targets at different scales and (2) enhancing the segmentation effect of slender ground object targets and refining the segmentation edges. To achieve these goals, we added two independent modules to the LANet network: (1) the ASPP$^+$ module, which facilitates the fusion of multi-scale features; (2) the FReLU activation function [21], which enhances the network's generalization ability, and filters out noise as well as low-frequency information.

Specifically, we integrated the FReLU activation function into the activation layer at the residual module of the backbone network ResNet50 and added an ASPP$^+$ module on the high-level feature extraction branch, as indicated by the green dashed box in Figure 3. In the branch of high-level feature extraction, the high-level features generated by ResNet50 extract multi-scale contextual information through the ASPP$^+$ module and then enhance their feature representation through the Patch Attention Module (PAM) [10]. In the low-level feature extraction branch, the low-level features generated by convolution are first feature-enhanced by the PAM, and then the semantic information of the low-level features is enriched by embedding the local focus of the high-level features through the Attention Embedding Module (AEM) [10], which enables the low-level features to enhance the high-level semantic without losing spatial information. Ultimately, the features produced by the upper and lower parallel branches are merged to derive our conclusive segmentation output.

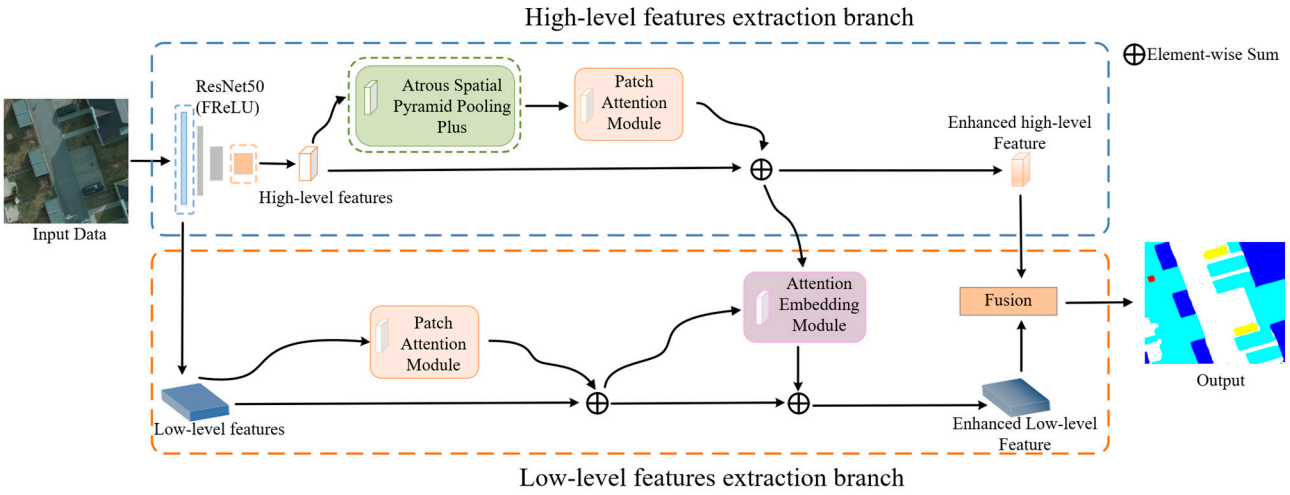


Figure 3. Overall structure of the ASPP$^+$-LANet network.

2.2.2. ASPP$^+$ Module

In RS images, challenges such as imbalanced ground object classes and significant variations in ground object scales exist. In such scenarios, it is difficult to extract target features only by a single scale. To address this issue, the paper proposes an improved multi-scale context extraction module, ASPP$^+$, with the structure shown in Figure 4. It mainly consists of two components: the first component is the parallel dilated convolution multi-scale feature extraction module, employing five parallel dilated convolution branches to capture feature information of different scales; the second component is the global feature and context extraction module, taking charge of acquiring global feature and contextual information. Ultimately, the output features from both components are concatenated to form a multi-scale feature map.

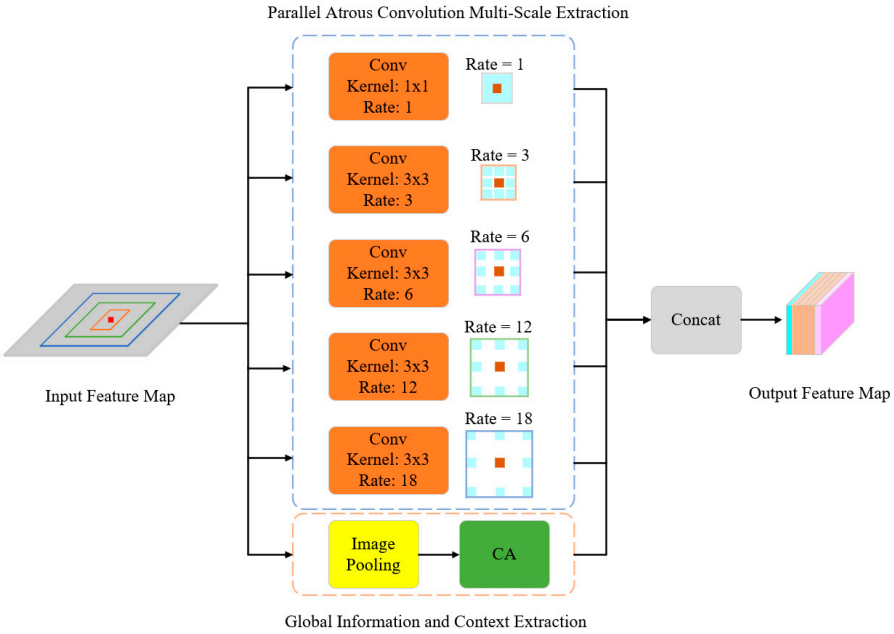


Figure 4. ASPP$^+$ module structure.

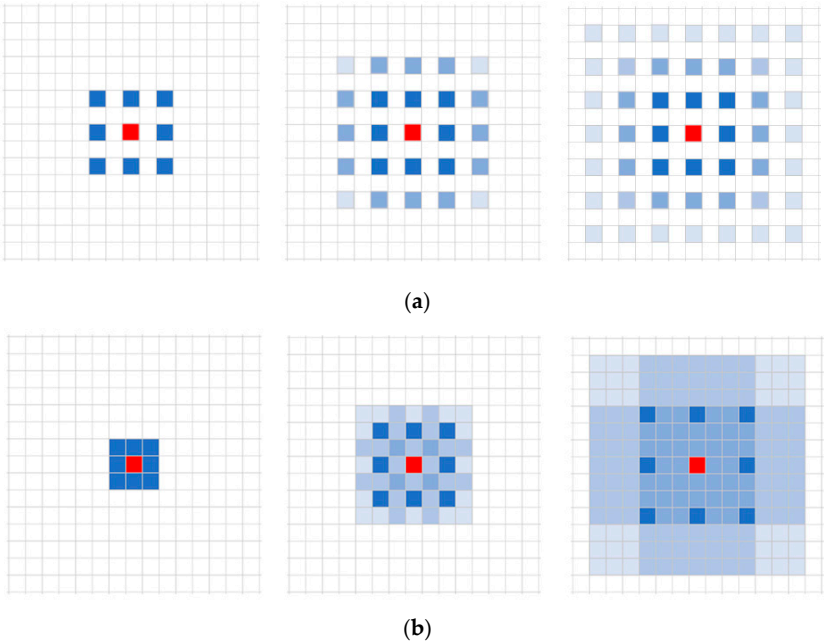
In Figure 4, the orange box represents dilated convolutions with different dilation rates, where except for the first convolution kernel with a size of 1×1, the remaining four convolution kernels are all 3×3. Additionally, they have a stride of 1 and no padding. The Image Pooling module performs global average pooling. CA [20] refers

to the Coordinate Attention module. "Concat" represents the operation of concatenating features. By concatenating the output features from these two parts, the model achieves the functionality of multi-scale context feature extraction. The dilated convolutions with different dilation rates improve the receptive field and capture target features at different scales. The attention module allows the model to focus more on meaningful features and acquire contextual information.

(1) Parallel Dilated Convolution Multi-Scale Feature Extraction

Parallel convolution can alter the receptive field of the convolutional kernel, acquiring the feature information of different scales. However, multiple parallel convolutions can increase the number of parameters and computational complexity of the network. Inspired by the ASPP module [9], using extended convolution instead of standard convolution can obtain feature information at different scales while reducing the number of parameters and computational complexity.

While the replacement of dilated convolutions has played a significant role, the setting of the dilation rate remains a challenge. The consecutive use of the same dilation rate in atrous convolutions will result in discontinuity of the convolution kernel, leading to a "grid effect", as shown in Figure 5a. On the other hand, a reasonable dilation rate, as depicted in Figure 5b, not only avoids the loss of relevant information but also captures the target context of different scales [25]. According to the literature [25], the dilation rate should follow the following principles:



(**a**)

(**b**)

Figure 5. Schematic diagram of the "grid effect" [25]. (**a**) The "grid effect" in atrous convolutions. (**b**) Reasonable combination of dilation rates in atrous convolutions.

(a) The combination of dilation rates should not have a common factor greater than 1, as it would still lead to the occurrence of the "grid effect".

(b) Assuming that dilation rates corresponding to N convolutional kernel sizes $k \times k$ of atrous convolutions are $[r_1, \ldots, r_i, \ldots, r_n]$, it is required that Equation (1) satisfies $M_2 \leq k$.

$$M_i = max[M_{i+1} - 2r_i, M_{i+1} - 2(M_{i+1} - r_i), r_i] \quad (1)$$

where r_i represents the dilation rate of the i-th atrous convolution and M_i represents the maximum dilation rate for the i-th layer of atrous convolution, with a default value of $M_n = r_n$.

Therefore, this paper follows the aforementioned design principles and obtains a set of most appropriate dilation rates (1, 2, 4, 8, 12) through several comparative experiments (detailed experimental procedures described in Section 3.3.4), which significantly enhances the segmentation performance of ground object targets at different sizes.

Additionally, to enhance the model's generalization ability, we incorporate batch normalization and ReLU activation functions [26] after each convolutional layer. Finally, we connect the five parallel dilated convolution branches to form the parallel dilated convolution multi-scale feature extraction module, as depicted by the blue dashed box in Figure 4. The expression is represented as:

$$\left\langle C^1_{1\times1}(X)\cdot C^d_{3\times3}(X)\right\rangle, d = 2,4,8,12 \tag{2}$$

where $\langle\cdot\rangle$ represents feature concatenation, which refers to each feature being spliced along the channel dimension. $C^1_{1\times1}$ denotes a 1×1 convolution with a dilation rate of 1. $C^d_{3\times3}$ represents a 3×3 convolution with a dilation rate of d. X denotes the input feature.

(2) Global features and contextual information extraction

Global feature extraction refers to the generalization and integration of features from the entire feature map to obtain global contextual information. The global feature and context extraction module, as illustrated by the orange dashed box in Figure 4, begins by performing global average pooling on the input feature. It then utilizes a CA module to emphasize meaningful features, thereby capturing global contextual information. The expression can be represented as:

$$CA(GAP(X)) \tag{3}$$

where $CA(\cdot)$ represents Coordinate Attention. $GAP(\cdot)$ denotes Global Average Pooling. X represents the input features.

In conclusion, based on the aforementioned information, we can obtain an improved multi-scale context extraction module, referred to as the ASPP$^+$ module. Its overall representation is illustrated by Equation (4).

$$\left\langle C^1_{1\times1}(X)\cdot C^d_{3\times3}(X)\cdot CA(GAP(X))\right\rangle,\ d = 3,6,12,18 \tag{4}$$

where $\langle\cdot\rangle$ represents feature concatenation, which refers to the concatenation of each feature along the channel dimension. $C^1_{1\times1}$ denotes a 1×1 convolution with a dilation rate of 1. $C^d_{3\times3}$ represents a 3×3 convolution with a dilation rate of d. X denotes the input feature. $CA(\cdot)$ represents Coordinate Attention. $GAP(\cdot)$ denotes Global Average Pooling.

2.2.3. FReLU

In RS images, there always exists interference from noise and low-frequency information, which makes it challenging for existing image semantic segmentation networks to achieve satisfactory results for slender and limbic ground object targets. Activation functions, on the other hand, play a crucial role in enhancing network generalization, filtering out noise and low-frequency information, and preserving high-frequency information, which can help resolve this issue. Therefore, this paper conducted comparative experiments with different activation functions on the LANet network (detailed in Section 3.3.5), and the results indicate that incorporating FReLU into the LANet network yields the best performance.

This paper focuses on improving the bottleneck residual module within the ResNet50 backbone network, as illustrated in Figure 6. The activation functions in each convolutional layer of the bottleneck module are replaced with FReLU. Similar to ReLU [27] and PReLU [28], FReLU utilizes the max() function as a simple non-linear function. Whereas ReLU is defined as $y = \max(x, 0)$ and PReLU as $y = \max(x, px)$, FReLU adds a negligible spatial condition overhead and extends the conditional part to a two-dimensional condition

that depends on the spatial context of each pixel, as illustrated in Figure 7. It can be represented as $y = \max(x, T(x))$, where T(·) denotes the two-dimensional spatial representation. The function definition of FReLU is as follows:

$$f(x_{c,\,i,\,j}) = max(x_{c,\,i,\,j},\ T(x_{c,\,i,\,j})) \tag{5}$$

$$T(x_{c,\,i,\,j}) = x^{w}_{c,\,i,\,j} \cdot\ p^{w}_{c} \tag{6}$$

where (i, j) represents the pixel position in two-dimensional space; c denotes the c-th channel; $T(x_{c,i,j})$ represents the two-dimensional condition; $x^{w}_{c,i,j}$ denotes the parameterized pool window centered on the input pixel of the nonlinear activation function on the c-th channel at position (i, j) in two-dimensional space; and p^{w}_{c} represents coefficients that are shared by this window in the same channel.

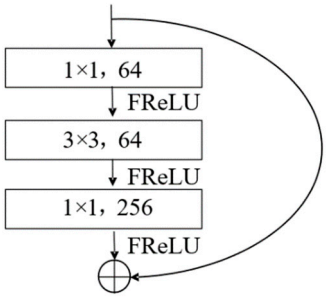


Figure 6. BottleNeck structure of ResNet50.

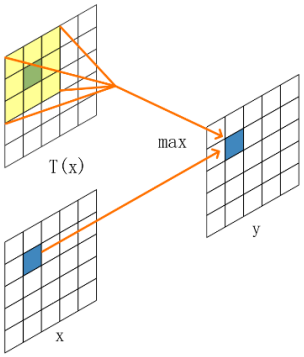


Figure 7. Schematic diagram of FReLU.

High-resolution RS images often exhibit complex backgrounds, leading to challenges in achieving accurate semantic segmentation, especially for slender and limbic ground object targets. FReLU, by incorporating spatial context information as a non-linear function condition, possesses superior contextual capturing capabilities. It effectively filters out noise and low-frequency information while preserving high-frequency details. The results show that FReLU can significantly improve the segmentation effect of slender ground objects and refine the segmentation edges.

3. Experiments and Results

In this section, we conducted a series of comparative experiments and ablation studies to validate the effectiveness of our proposed method. Initially, we delineated three evaluation metrics utilized for quantitative analysis. Following that, we furnished comprehensive details regarding the network's parameter configurations and experimental setups. Subsequently, we performed comparative experiments with other SOTA methods to assess and compare the performance of our proposed network. Additionally, we conducted ablation studies to evaluate the performance of our network under various configuration settings. We analyzed the experimental results in terms of segmentation accuracy, visual effects, and

ablation studies. Ultimately, to bolster the credibility of our experiments, we conducted an investigation into the optimal dilation rate for the ASPP+ module. Furthermore, we undertook experiments to evaluate the performance differences of various activation functions on the baseline network, LANet.

3.1. Evaluation Criteria

To quantitatively evaluate the efficacy of our proposed method, this paper utilizes three evaluation metrics for comprehensive comparison and analysis: Pixel Accuracy (PA), F1 Score (F1), and Mean Intersection over Union (MIoU). The formulas for these metrics are as follows:

PA refers to the proportion of correctly predicted pixels of a certain category to the total number of pixels.

$$PA = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

F1 takes into account both the precision and recall of a classification model and enables it to be seen as the harmonic mean of precision and recall.

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$F1 = \frac{2 * \text{Precision} * \text{Recall}}{Precision + Recall} \tag{10}$$

MIoU refers to the average Intersection over Union (IoU) of each class in the dataset.

$$IoU = \frac{TP}{TP + FP + FN} \tag{11}$$

$$MIoU = \frac{1}{n}\sum_{i=1}^{n} IoU \tag{12}$$

where TP stands for True Positive, indicating the number of pixels in the predicted results that belong to a certain class and are indeed of that class; FP stands for False Positive, signifying the number of pixels in the predicted results that belong to other classes but are mistakenly classified as that class; TN stands for True Negative, depicting the number of pixels in the predicted results that belong to other classes and are indeed of other classes; FN stands for False Negative, referring to the number of pixels in the predicted results that belong to a certain class but are mistakenly classified as other classes. n represents the number of classes. i represents the i-th class.

3.2. Implementation Details

In this article, our network and other comparative networks are implemented in the PyTorch deep learning framework, and experiments are conducted on a 64-bit Windows 10 system server. The server is equipped with an Intel Core i9-12900k CPU (3.20 GHz), 128 GB of memory, and an NVIDIA GeForce RTX 4090 graphics card.

During the training process, referring to some model [18] and synthesizing our hardware and our experimental results, the experimental parameters were set as follows: the batch size was set to 6, the learning rate was set to 0.025, the total epochs was set to 400, the momentum was set to 0.9, adaptive moment estimation optimizer (Adam) [29] was used to optimize our model, and stochastic gradient descent (SGD) was employed for the

optimization training. Additionally, a "poly" learning rate decay strategy was utilized to dynamically adjust the learning rate using the following expression:

$$l = l_{ini}\left(1 - \frac{e}{e_{max}}\right)^{0.9} \tag{13}$$

where l represents the current learning rate, l_{ini} stands for the initial learning rate, e denotes the current training epoch, and e_{max} refers to the maximum number of training epochs.

3.3. Experiment Results

3.3.1. Segmentation Precision Analysis

To validate the efficacy of our proposed method, we conducted comparisons with several classical network models, including UNet [3], SegNet [4], DeepLab V3+ [9], LANet [10], MANet [11], UnetFormer [18], and Swin-CNN [19] on the Potsdam and Vaihingen datasets. The evaluation metrics for each method are presented in Tables 1 and 2. The tables clearly demonstrate that LANet's experimental results outperform classical semantic segmentation networks such as UNet, SegNet, and so on. Nonetheless, the utilization of a single-scale feature extraction approach in LANet results in diminished segmentation performance when confronted with ground object targets of varied sizes. Consequently, we implemented enhancements to LANet by integrating the ASPP$^+$ module and the FReLU activation function. This integration effectively enhances the segmentation performance for ground object targets at different scales, as well as slender ground objects and ground objects' edges.

Table 1. Segmentation accuracy of different methods on the Potsdam dataset.

Method	Parameters(M)	PA/%	F1/%	MIoU/%	Kappa
UNet	17.27	92.66	78.08	71.35	0.9492
SegNet	29.45	92.61	77.61	70.84	0.9491
DeepLab V3+	21.94	90.00	72.24	64.17	0.8913
LANet	23.81	93.29	78.77	72.29	0.9496
MANet	35.86	92.06	76.89	69.86	0.9256
UNetFormer	11.28	91.23	75.01	67.51	0.9138
Swin-CNN	66	94.56	81.68	76.62	0.9521
ASPP$^+$-LANet	27.46	95.53	82.57	77.81	0.9552

Table 2. Segmentation accuracy of different methods on the Vaihingen dataset.

Method	Parameters(M)	PA/%	F1/%	MIoU/%	Kappa
UNet	17.27	98.03	81.83	79.53	0.9637
SegNet	29.45	96.82	80.21	76.77	0.9433
DeepLab V3+	21.94	92.77	73.31	67.33	0.8721
LANet	23.81	97.55	80.82	77.77	0.9465
MANet	35.86	98.08	81.81	79.55	0.9677
UNetFormer	11.28	96.73	80.08	76.52	0.9429
Swin-CNN	66	97.98	81.66	78.86	0.9625
ASPP$^+$-LANet	27.46	98.24	81.99	79.83	0.9689

As shown in Table 1, our proposed method, ASPP$^+$-LANet, achieves the following performance metrics on the Potsdam dataset: PA reaches 95.53%, F1 reaches 82.57%, and MIoU reaches 77.81%, which is improved by 2.24%, 3.80%, and 5.52%, respectively, compared to the baseline LANet network. Furthermore, our method demonstrates superior performance compared to existing semantic segmentation networks. This notable performance can be attributed to two key factors. Primarily, our proposed ASPP$^+$ module enhances the network's ability to extract multi-scale features by setting appropriate dilation rates, thereby effectively improving the segmentation accuracy for ground object targets of different sizes. Moreover, the introduction of the FReLU activation function filters out

noise and low-frequency information while preserving high-frequency information, thereby improving segmentation performance for slender and limbic ground object targets.

As shown in Table 2, our proposed method, ASPP$^+$-LANet, achieves the following performance metrics on the Vaihingen dataset: PA reaches 98.24%, F1 reaches 81.99%, and MIoU reaches 79.83%, which is improved by 0.69%, 1.17%, and 2.06%, respectively, compared to the baseline LANet network. Furthermore, our method demonstrates superior performance compared to existing semantic segmentation networks.

3.3.2. Renderings Analysis

To better highlight the feasibility of the proposed method in this paper, we selected six representative test targets for analysis on the Potsdam and Vaihingen datasets. Additionally, we conducted a subjective visual comparison analysis among the classical semantic segmentation methods, as illustrated in the figure below.

By comparing the visualization results on the Potsdam dataset, as shown in Figures 8 and 9, it can be observed that our proposed method achieves superior segmentation accuracy on ground object targets of different scales compared to other comparative methods. Additionally, Figures 10 and 11 demonstrate that our proposed method achieves superior segmentation accuracy on slender ground objects and ground object edges compared to other comparative methods. Moreover, Figures 12 and 13 reveal that our proposed method outperforms other comparative methods in terms of missing detections and false detections. The above experimental results validate the efficacy of our proposed ASPP$^+$-LANet model. After integrating the ASPP$^+$ module and the FReLU activation function, there was indeed a noticeable improvement in the segmentation performance of ground object targets at varying scales in the Potsdam dataset. Moreover, it also enhances the segmentation effect for slender ground object targets, refining the segmentation edges. These results demonstrate the effectiveness of our approach.

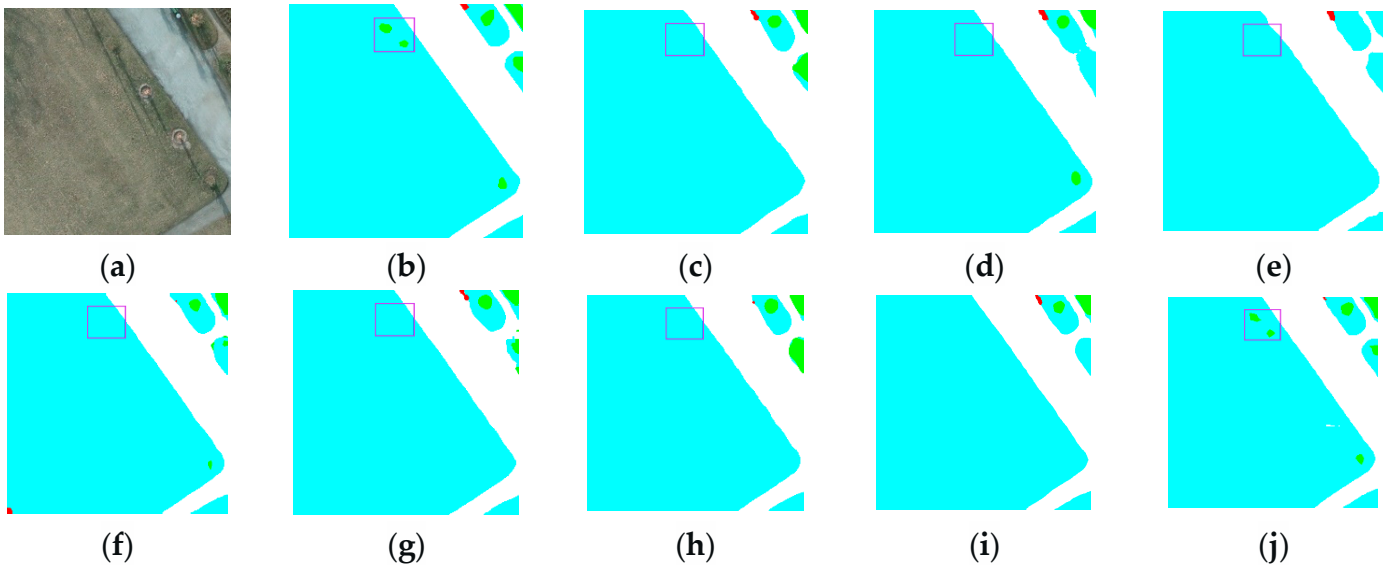


Figure 8. Visual comparison of semantic segmentation for small object features on the Potsdam dataset. (**a**) Image, (**b**) Ground truth, (**c**) UNet, (**d**) SegNet, (**e**) DeepLabv3+, (**f**) LANet, (**g**) MANet, (**h**) UnetFormer, (**i**) Swin-CNN, (**j**) ASPP$^+$-LANet. The colors represent the same types of ground object as shown in Figure 1, and the same applies to other similar images.

By analyzing the visualization results on the Vaihingen dataset, as shown in Figures 14 and 15, it can be observed that our proposed method achieves superior segmentation accuracy on ground object targets of different scales compared to other comparative methods. Additionally, Figures 16 and 17 demonstrate that our proposed method achieves better segmentation accuracy on slender ground objects and ground object edges compared

to other comparative methods. Moreover, Figures 18 and 19 reveal that our proposed method outperforms other comparative methods in terms of missing detections and false detections. The above experimental results validate the effectiveness of our proposed ASPP$^+$-LANet model. After integrating the ASPP$^+$ module and the FReLU activation function, there was indeed a noticeable improvement in the segmentation performance of ground object targets at varying scales in the Vaihingen dataset. Moreover, it also enhances the segmentation effect for slender ground object targets, refining the segmentation edges. These results further demonstrate the effectiveness of our approach.

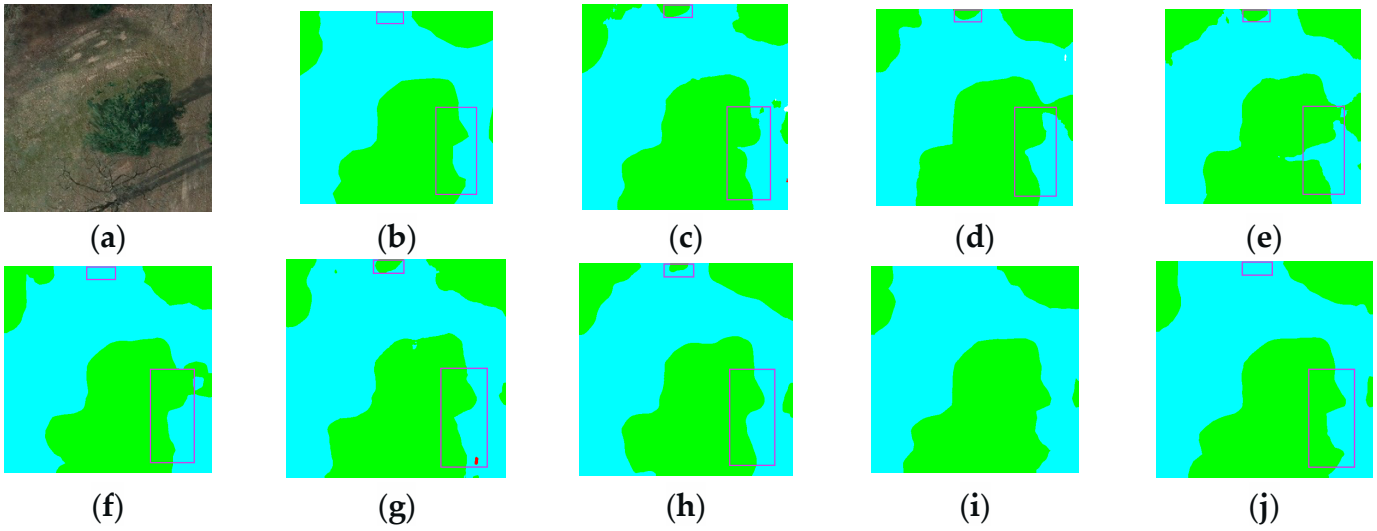


Figure 9. Visual comparison of semantic segmentation for large object features on the Potsdam dataset. (**a**) Image, (**b**) Ground truth, (**c**) UNet, (**d**) SegNet, (**e**) DeepLabv3+, (**f**) LANet, (**g**) MANet, (**h**) UnetFormer, (**i**) Swin-CNN, (**j**) ASPP$^+$-LANet.

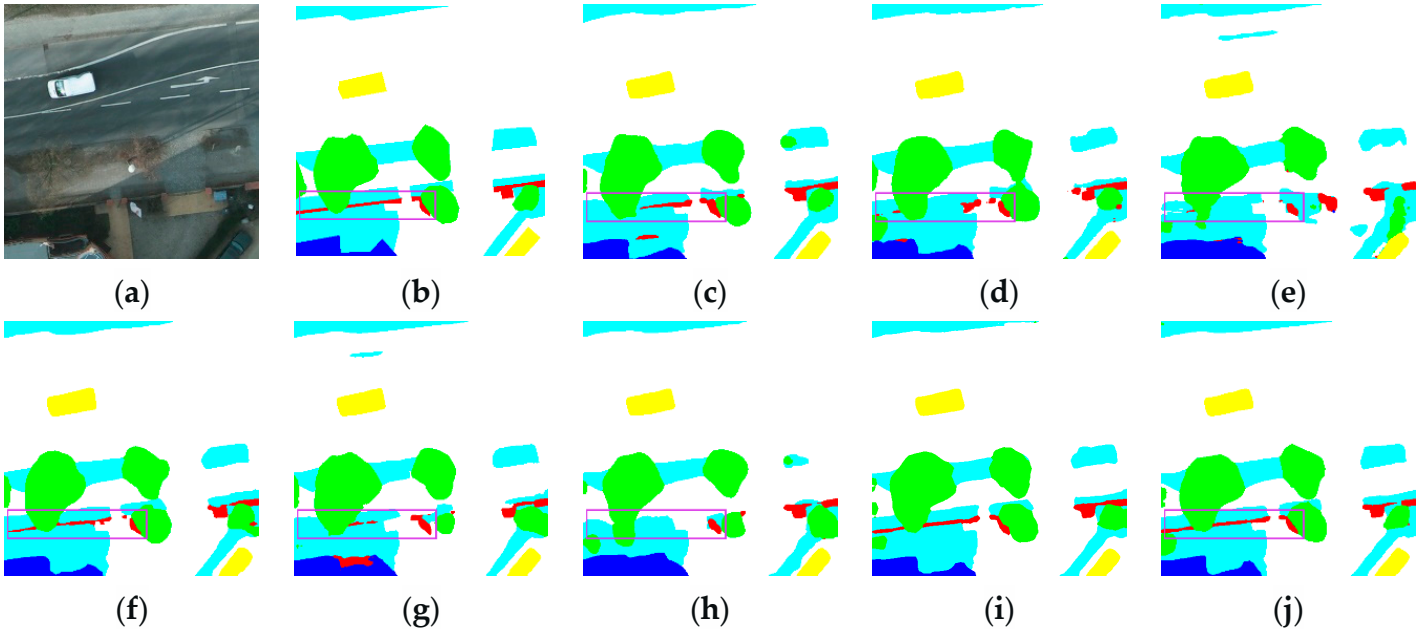


Figure 10. Visual comparison of semantic segmentation for slender ground objects features on the Potsdam dataset. (**a**) Image, (**b**) Ground truth, (**c**) UNet, (**d**) SegNet, (**e**) DeepLabv3+, (**f**) LANet, (**g**) MANet, (**h**) UnetFormer, (**i**) Swin-CNN, (**j**) ASPP$^+$-LANet.

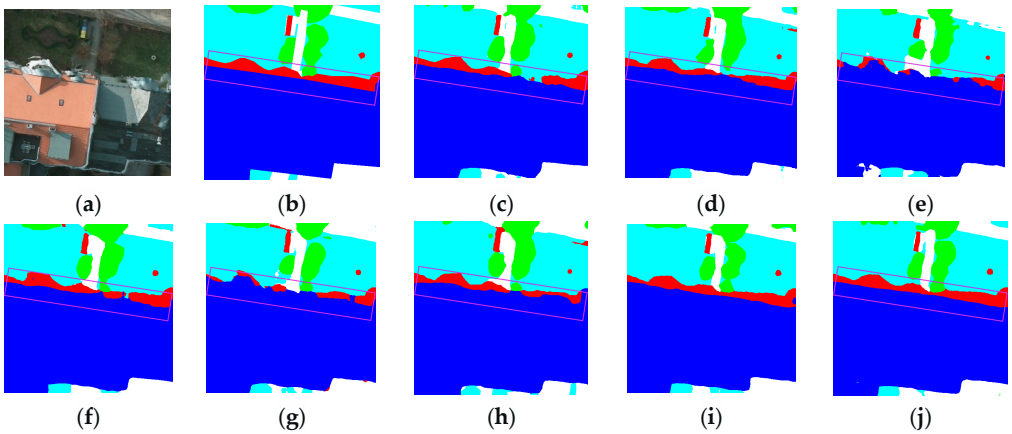


Figure 11. Visual comparison of semantic segmentation for limbic ground objects features on the Potsdam dataset. (**a**) Image, (**b**) Ground truth, (**c**) UNet, (**d**) SegNet, (**e**) DeepLabv3+, (**f**) LANet, (**g**) MANet, (**h**) UnetFormer, (**i**) Swin-CNN, (**j**) ASPP^{+}-LANet.

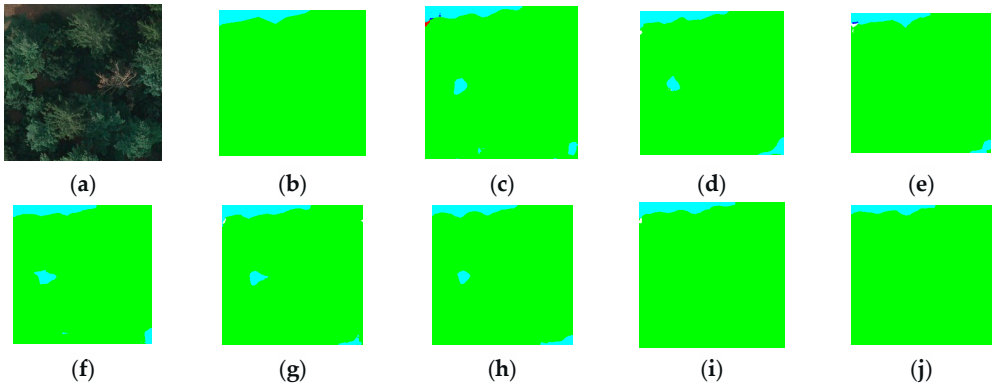


Figure 12. Visual comparison of semantic segmentation for the missing detection of object features in the Potsdam dataset. (**a**) Image, (**b**) Ground truth, (**c**) UNet, (**d**) SegNet, (**e**) DeepLabv3+, (**f**) LANet, (**g**) MANet, (**h**) UnetFormer, (**i**) Swin-CNN, (**j**) ASPP^{+}-LANet.

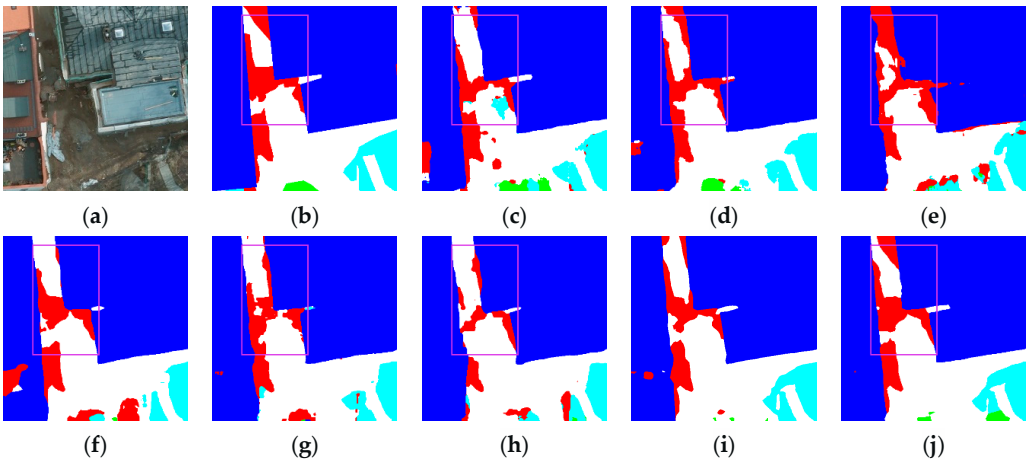


Figure 13. Visual comparison of semantic segmentation for the false detection of object features in the Potsdam dataset. (**a**) Image, (**b**) Ground truth, (**c**) UNet, (**d**) SegNet, (**e**) DeepLabv3+, (**f**) LANet, (**g**) MANet, (**h**) UnetFormer, (**i**) Swin-CNN, (**j**) ASPP^{+}-LANet.

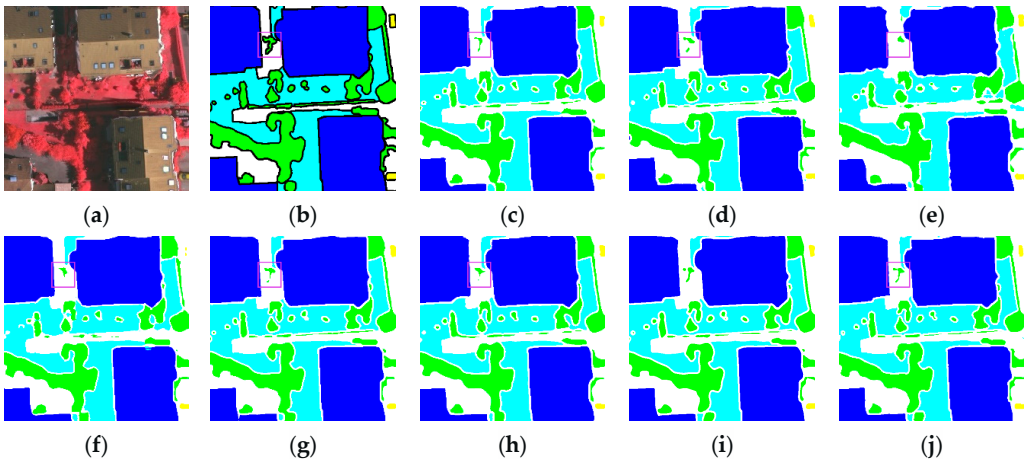


Figure 14. Visual comparison of semantic segmentation for small object features on the Vaihingen dataset. (**a**) Image, (**b**) Ground truth, (**c**) UNet, (**d**) SegNet, (**e**) DeepLabv3+, (**f**) LANet, (**g**) MANet, (**h**) UnetFormer, (**i**) Swin-CNN, (**j**) ASPP$^+$-LANet.

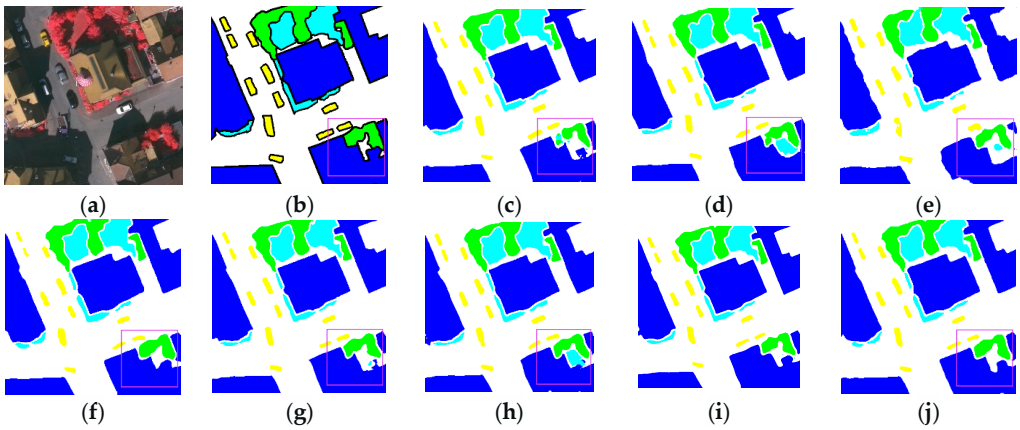


Figure 15. Visual comparison of semantic segmentation for large object features on the Vaihingen dataset. (**a**) Image, (**b**) Ground truth, (**c**) UNet, (**d**) SegNet, (**e**) DeepLabv3+, (**f**) LANet, (**g**) MANet, (**h**) UnetFormer, (**i**) Swin-CNN, (**j**) ASPP$^+$-LANet.

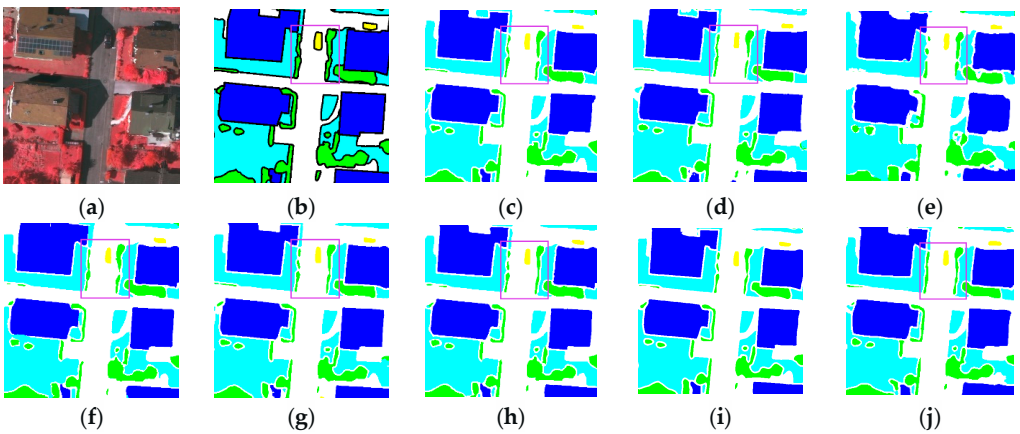


Figure 16. Visual comparison of semantic segmentation for slender ground objects features on the Vaihingen dataset. (**a**) Image, (**b**) Ground truth, (**c**) UNet, (**d**) SegNet, (**e**) DeepLabv3+, (**f**) LANet, (**g**) MANet, (**h**) UNetFormer, (**i**) Swin-CNN, (**j**) ASPP$^+$-LANet.

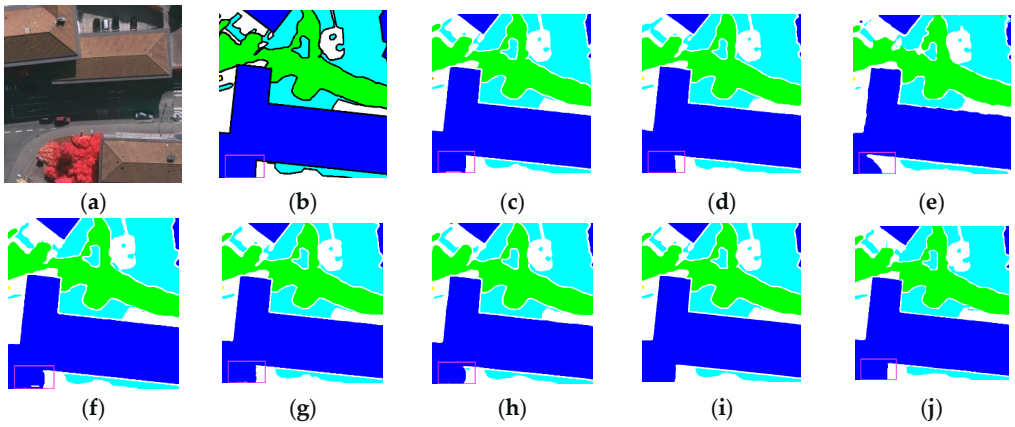


Figure 17. Visual comparison of semantic segmentation for limbic ground objects features on the Vaihingen dataset. (**a**) Image, (**b**) Ground truth, (**c**) UNet, (**d**) SegNet, (**e**) DeepLabv3+, (**f**) LANet, (**g**) MANet, (**h**) UnetFormer, (**i**) Swin-CNN, (**j**) ASPP^{+}-LANet.

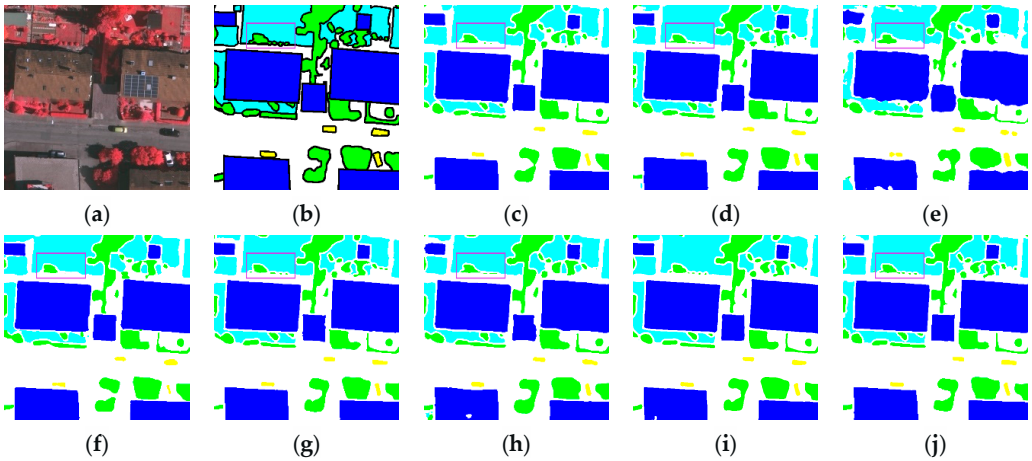


Figure 18. Visual comparison of semantic segmentation for the missing detection of object features in the Vaihingen dataset. (**a**) Image, (**b**) Ground truth, (**c**) UNet, (**d**) SegNet, (**e**) DeepLabv3+, (**f**) LANet, (**g**) MANet, (**h**) UnetFormer, (**i**) Swin-CNN, (**j**) ASPP^{+}-LANet.

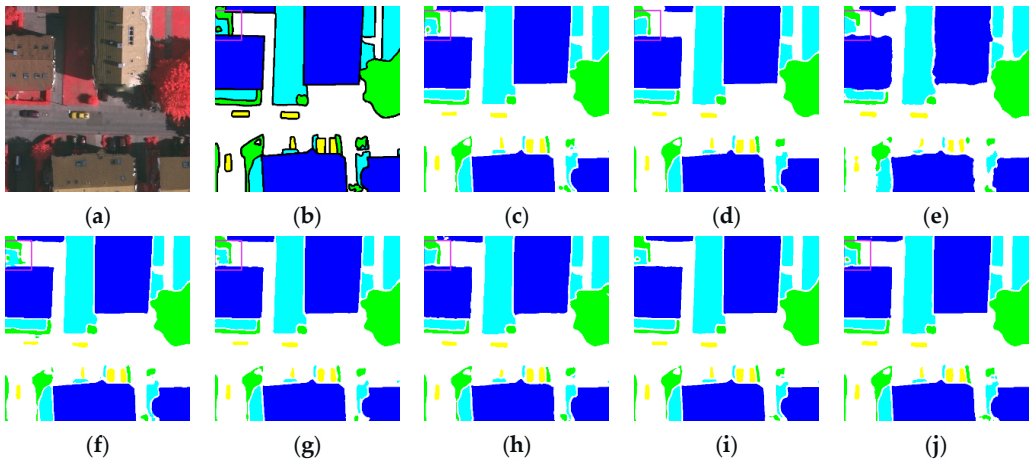


Figure 19. Visual comparison of semantic segmentation for the false detection of object features in the Vaihingen dataset. (**a**) Image, (**b**) Ground truth, (**c**) UNet, (**d**) SegNet, (**e**) DeepLabv3+, (**f**) LANet, (**g**) MANet, (**h**) UnetFormer, (**i**) Swin-CNN, (**j**) ASPP^{+}-LANet.

3.3.3. Ablation Experiments Analysis

To effectively capture detailed features from high-resolution RS images and overcome the technical challenges in accurately segmenting ground object targets at various scales, we propose the ASPP$^+$ module. Building upon the ASPP module, the ASPP$^+$ module adds a feature extraction channel, redefines the dilation rates, and introduces CA mechanisms, thereby effectively improving the segmentation performance of ground object targets at different scales. Moreover, in order to enhance the segmentation performance of slender ground object targets and refine the segmentation edges, we replaced the activation function on the backbone network (ResNet50) with FReLU. This alteration assists in filtering out noise and low-frequency information while preserving more high-frequency information, thereby further improving the segmentation accuracy of RS images. We conducted corresponding ablation experiments to individually verify the effectiveness of the ASPP module, ASPP$^+$ module, and FReLU activation function. The results of the ablation experiments on the Potsdam and Vaihingen datasets are presented in Tables 3 and 4.

Table 3. Results of ablation experiments on the Potsdam dataset.

Method	PA/%	F1/%	MIoU/%
LANet	93.29	78.77	72.29
LANet + ASPP	93.71	79.46	73.29
LANet + ASPP$^+$	93.86	79.80	73.75
LANet + FReLU	95.22	82.05	77.06
ASPP$^+$-LANet	95.53	82.57	77.81

Table 4. Results of ablation experiments on the Vaihingen dataset.

Method	PA/%	F1/%	MIoU/%
LANet	97.55	80.82	77.77
LANet + ASPP	97.77	81.31	78.65
LANet + ASPP$^+$	97.80	81.42	78.79
LANet + FReLU	97.76	81.30	78.59
ASPP$^+$-LANet	98.24	81.99	79.83

According to Table 3, it can be observed that the inclusion of the ASPP module leads to improvements in all performance metrics compared to the baseline network, LANet. Furthermore, by further refining the ASPP module, we were able to achieve even more significant enhancements in the performance metrics compared to the initial inclusion of the ASPP module. By incorporating the FReLU activation function, significant improvements can be observed in all performance metrics compared to the baseline network, LANet. Finally, by integrating the ASPP$^+$ module and the FReLU activation function into the LANet network, we further improved the overall performance metrics. The metrics such as PA, F1, and MioU reached 95.53%, 82.57%, and 77.81% respectively. Compared to the baseline network, LANet, there were increases of 2.24%, 3.80%, and 5.52% in PA, F1, and MioU, respectively.

According to Table 4, it is evident that the addition of the ASPP module leads to improvements in all metrics compared to the baseline LANet. Subsequent modifications made to the ASPP module result in slight enhancements in the metrics compared to the initial implementation. Furthermore, the inclusion of the FReLU activation function leads to improvements in all metrics compared to the LANet baseline. However, it is worth noting that the improvement achieved by incorporating FReLU is not as significant as that observed in the Potsdam dataset. This discrepancy could be attributed to the presence of a higher number of RS images related to narrow streets in the Potsdam dataset, a characteristic absent in the Vaihingen dataset. Finally, by integrating the ASPP$^+$ module and FReLU activation function into the LANet network, we further enhanced the overall performance metrics. The metrics, including PA, F1, and MioU, reached 98.24%, 81.99%,

and 79.83%, respectively. Compared to the baseline LANet, there were improvements of 0.69%, 1.17%, and 2.06% in PA, F1, and MioU metrics, respectively.

In addition, to further visually represent the impact of each module in the ablation experiments on the results of semantic segmentation, we present the visualization of the core component ablation experiments of our method on the Potsdam and Vaihingen datasets, as shown in Figures 20 and 21. Among them, the first and second rows are used to verify the efficacy of large object detection and small object detection, respectively. From the figures, it can be observed that incorporating the ASPP$^+$ module into LANet improves the detection performance for ground object targets of different scales, surpassing both LANet alone and the results of incorporating the FReLU activation function in LANet. The third and fourth rows are used to evaluate the detection performance of slender and limbic ground object targets. The figures demonstrate that integrating the FReLU activation function into LANet enhances the detection of slender and limbic ground object targets, outperforming both LANet alone and the results of incorporating the ASPP$^+$ module in LANet. Thus, we can conclude that the efficacy of our integration of the ASPP$^+$ module and FReLU activation function in LANet has been validated.

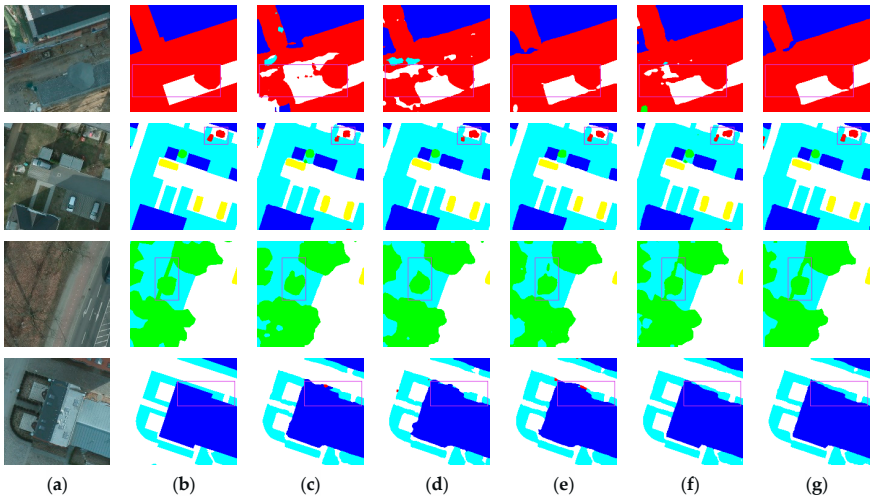


Figure 20. Visual comparisons of the ablation experiments conducted on the Potsdam dataset: (**a**) Image, (**b**) Grond Truth, (**c**) LANet, (**d**) LANet + ASPP, (**e**) LANet + ASPP$^+$, (**f**) LANet + FReLU, (**g**) ASPP$^+$-LANet.

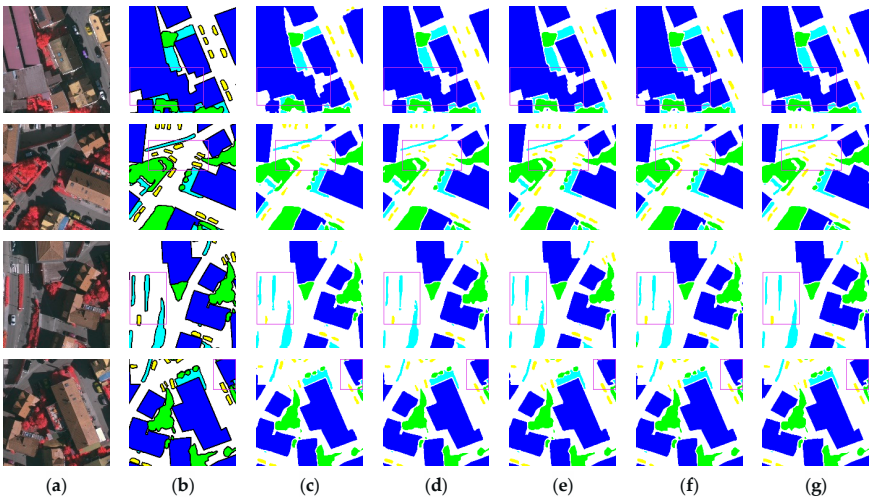


Figure 21. Visual comparisons of the ablation experiments conducted on the Vaihingen dataset: (**a**) Image, (**b**) Grond Truth, (**c**) LANet, (**d**) LANet + ASPP, (**e**) LANet + ASPP$^+$, (**f**) LANet + FReLU, (**g**) ASPP$^+$-LANet.

3.3.4. Dilation Rates Analysis of ASPP$^+$ Module

The ASPP$^+$ module is a fusion of the enhanced ASPP [9] module and the CA [20] module. This fusion facilitates the efficient extraction of multi-scale semantic features in RS images. Due to the addition of an extra feature extraction channel in ASPP, as the backbone network performs feature extraction, the resolution of the feature maps gradually decreases. The combination of (1, 6, 12, 18) is not optimal for effectively extracting multi-resolution feature maps. Insufficient utilization of smaller dilation rates hinders the segmentation capability of small targets, resulting in weaker segmentation ability of the network for ground object targets at different scales. Therefore, it is necessary to readjust the dilation rates of the atrous convolution. Considering our two distinct datasets, to avoid redundant experiments, we exclusively conducted the experimentation on the Potsdam dataset for readjusting the dilation rates. In order to effectively extract multi-scale contextual features and enhance the segmentation performance for ground object targets of varying scales, this paper follows the guidelines outlined in Section 2.2 to determine rational dilation rates. To this end, we devised five groups of experiments with different dilation rates for comparison within the ASPP$^+$-LANet network, which comprise of (1, 2, 4, 6, 8), (1, 2, 4, 8, 12), (1, 3, 6, 12, 18), (1, 3, 8, 16, 18), and (1, 3, 8, 18, 24). The experimental results are presented in Table 5. According to the evaluation metrics obtained from different combinations of dilation rates, the experiment achieved optimal results when the dilation rates were (1, 2, 4, 8, 12). This is because such dilation rate settings are well-suited for feature extraction of ground object targets at different scales in the Potsdam dataset. When the dilation rate is too large or too small, it adversely affects the effectiveness of feature extraction.

Table 5. Comparative experiments with different dilation rates of ASPP$^+$ on the ASPP$^+$-LANet.

Dilation Rate	PA/%	F1/%	MIoU/%
(1, 2, 4, 6, 8)	95.50	82.51	77.74
(1, 2, 4, 8, 12)	95.53	82.57	77.81
(1, 3, 6, 12, 18)	95.47	82.39	77.60
(1, 3, 8, 16, 18)	95.46	82.51	77.74
(1, 3, 8, 18, 24)	95.51	82.52	77.73

3.3.5. Comparative Analysis of Activation Functions

In order to validate the effectiveness of the FReLU activation function, this paper conducted experimental comparisons of different activation functions on the benchmark network, LANet. Considering the availability of two datasets, to avoid redundant experiments, we exclusively performed activation function comparisons on the Potsdam dataset. The results are summarized in the following table.

As depicted in Table 6, among the numerous activation functions examined, the incorporation of the FReLU activation function into the baseline LANet network yielded the most favorable segmentation results on the Potsdam dataset. The evaluation metrics, including PA, F1, and MIoU, exhibited remarkable values of 95.22%, 82.05%, and 77.06%, respectively. These findings highlight the superiority of the FReLU activation function in enhancing the segmentation performance, specifically for RS tasks.

Table 6. Experimental Comparisons of Different Activation Functions on the LANet Network.

Activation Function	PA/%	F1/%	MIoU/%
LANet + LeakyReLU [26]	93.34	78.73	72.31
LANet + PReLU [28]	94.37	80.65	74.94
LANet + ELU [30]	90.23	72.58	64.83
LANet + Mish [31]	89.99	73.50	65.74
LANet + DY-ReLU [32]	94.10	80.26	74.40
LANet + FReLU	95.22	82.05	77.06

4. Discussion

According to the ablation experiments, the improved model effectively improves the accuracy of building extraction, as indicated in Tables 3 and 4. Moreover, from the first and second plots of Figures 20 and 21, we can see that our proposed network performs outstandingly well in segmenting large ground object targets as well as small ground object targets. In addition, the segmentation effect of the effect map with the ASPP module alone is much better than that of the effect map with FReLU alone, which indicates that the ASPP module can indeed effectively improve the segmentation effect of ground object targets at different scales. This is due to the fact that ASPP is a multi-scale module, which can effectively enhance the network's ability to extract multi-scale contexts. From the third and fourth plots of Figures 20 and 21, we can see that our proposed network performs outstandingly well in segmenting slender ground object targets and ground object edges. However, the segmentation effect of the effect map with the ASPP module alone is much lower than that of the effect map with FReLU alone, which indicates that the FReLU module can indeed effectively improve the segmentation effect of the slender ground object targets and ground object edges. This is because FReLU is able to filter noise and low-frequency information and retain more high-frequency information, while slender ground object targets, as well as ground object edges, mostly belong to high-frequency information.

Regarding the ASPP$^+$ module, we conducted detailed experiments on its dilation rate settings, as shown in Section 3.3.4. We found that the setting of the dilation rate is not the larger or smaller as being better for different sizes of feature targets; larger segmentation targets can be segmented by convolutional kernels with larger dilation rates; on the contrary, smaller targets can be segmented by convolutional kernels with smaller dilation rates. Therefore, the dilation rate should be set reasonably and appropriately in order to make the segmentation targets of different sizes achieve effective feature extraction.

Regarding the selection of the activation function, we also conducted detailed experiments on it, as shown in Section 3.3.5. The activation function can enhance the generalization ability of the network, filter noise and low-frequency information, and retain more high-frequency information, which can effectively improve the performance of the network. However, different activation functions do not improve the performance of the network in the same way; therefore, in this paper, the activation functions proposed in recent years are compared and tested, and the most suitable activation function for the network in this paper is derived.

In order to improve the robustness of the model, in this paper, we use the data enhancement method to perform operations such as random flipping on the Potsdam and Vaihingen datasets. We also discuss the impact of the data enhancement method on the semantic segmentation results and, based on the analysis in Figures 22 and 23, it can be seen that the use of the data enhancement method improves the combined performance metrics over the non-use of the data enhancement method on both semantic segmentation datasets, provided that all other conditions remain consistent. This further indicates that data enhancement is one of the factors that improve the semantic segmentation results of the method proposed in this paper.

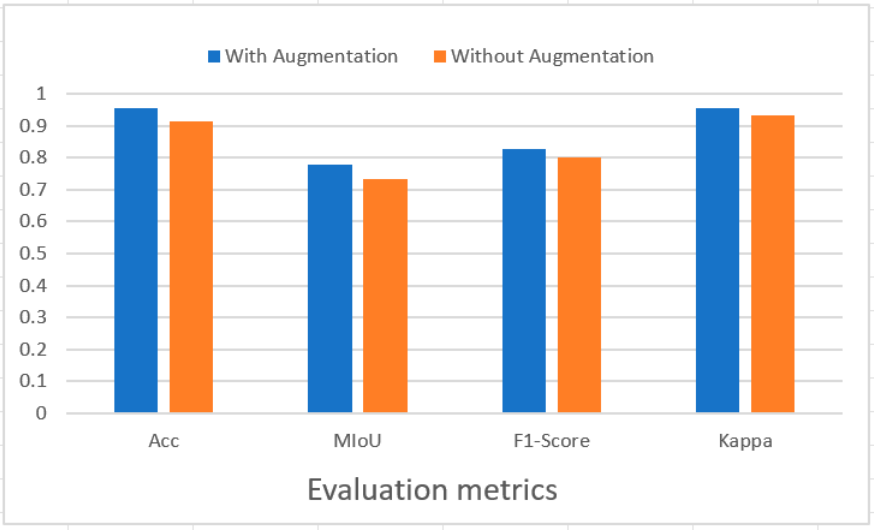


Figure 22. Effect of data augmentation on semantic segmentation results for the Potsdam dataset.

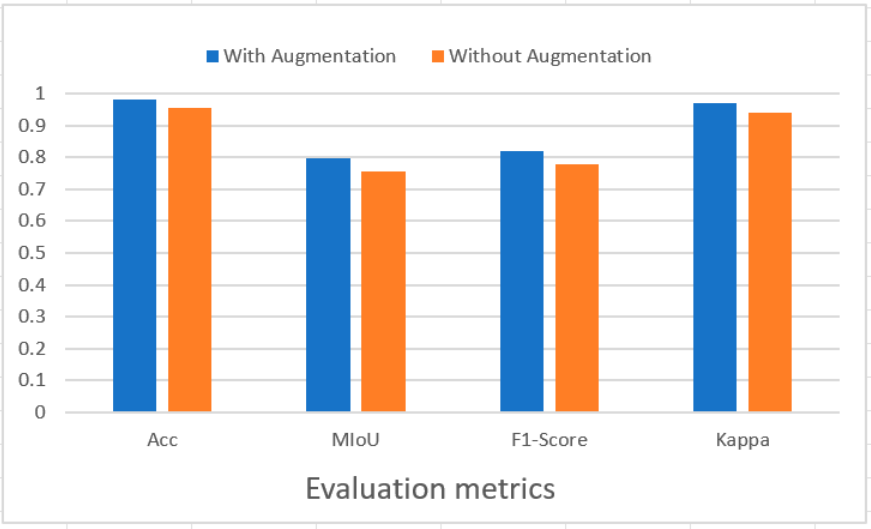


Figure 23. Effect of data augmentation on semantic segmentation results for the Vaihingen dataset.

5. Conclusions

In this paper, we propose a multi-scale context extraction network for semantic segmentation of high-resolution RS images, ASPP$^+$-LANet, aiming to fully capture the rich characteristics of ground object features. Firstly, we design a new ASPP$^+$ module, expanding upon the ASPP module by incorporating an additional feature extraction channel and redesigning the dilation rate, which effectively improves the segmentation effect of ground object features at different scales by controlling the size of the dilation rate. Furthermore, the CA mechanism has been introduced to extract meaningful features and acquire contextual information. The FReLU activation function has been incorporated to enhance the segmentation effect of slender ground object targets and refine the segmentation edges. Therefore, on the Potsdam and Vaihingen datasets, ASPP$^+$-LANet achieves superior segmentation performance for ground object targets at different scales, as well as slender and limbic ground object targets.

Nevertheless, certain limitations of our current approach must be acknowledged, especially concerning the influence of shadows on the segmentation accuracy of buildings, vegetation, and other objects, as well as the segmentation boundaries of non-smooth objects. Changes in the color of objects like buildings and vegetation can be induced by shadows. To address this issue, a more precise color division is required to distinguish between shadows

and actual objects, aiming to enhance accuracy levels. Furthermore, in the detection of non-smooth object edges, there is a need to enhance the network's capability to identify small target objects. This is crucial as object edges with jagged features can be perceived as tiny targets.

In the future, we will explore better methods to achieve higher accuracy and efficiency in RS image segmentation tasks. Firstly, we will be more specific in dividing the colors to distinguish the shadows from the actual objects; secondly, we will use the lightweight module to better optimize the network model and improve the network model segmentation efficiency and segmentation accuracy to solve the non-smooth ground objects edges problem.

Author Contributions: Conceptualization, L.H. and X.Z.; funding acquisition, L.H.; investigation, X.Z. and L.H.; methodology, X.Z. and L.H.; project administration, L.H. and S.L.; software, X.Z. and J.R.; writing–original draft, X.Z., L.H., S.L. and J.R.; writing–review and editing, L.H., X.Z. and J.R. supervision, L.H.; All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 61662033.

Data Availability Statement: This data can be found here: https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx and https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-vaihingen.aspx, all accessed on 15 October 2022.

Acknowledgments: The authors would like to thank the editor and the anonymous reviewers who provided insightful comments on improving this article.

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

References

1. Xu, S.; Pan, X.; Li, E.; Wu, B.; Bu, S.; Dong, W.; Xiang, S.; Zhang, X. Automatic Building Rooftop Extraction from Aerial Images via Hierarchical RGB-D Priors. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 7369–7387. [CrossRef]
2. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
3. 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: 18th International Conference, Proceedings, Part III 18, Munich, Germany, 5–9 October 2015; Springer: Cham, Switzerland, 2015; pp. 234–241.
4. 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] [PubMed]
5. 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, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
6. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv* **2014**, arXiv:1412.7062.
7. 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.* **2017**, *40*, 834–848. [CrossRef] [PubMed]
8. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**, arXiv:1706.05587.
9. 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 Computer Vision—ECCV 2018, Munich, Germany, 8–14 September 2018; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer: Cham, Switzerland, 2018; pp. 833–851.
10. Ding, L.; Tang, H.; Bruzzone, L. LANet: Local Attention Embedding to Improve the Semantic Segmentation of Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 426–435. [CrossRef]
11. Li, R.; Zheng, S.; Zhang, C.; Duan, C.; Su, J.; Wang, L.; Atkinson, P.M. Multiattention Network for Semantic Segmentation of Fine-Resolution Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5607713. [CrossRef]
12. Xu, J.; Xiong, Z.; Bhattacharyya, S.P. PIDNet: A Real-Time Semantic Segmentation Network Inspired by PID Controllers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 19529–19539.
13. Xu, M.; Wang, W.; Wang, K.; Dong, S.; Sun, P.; Sun, J.; Luo, G. Vision Transformers (ViT) Pretraining on 3D ABUS Image and Dual-CapsViT: Enhancing ViT Decoding via Dual-Channel Dynamic Routing. In Proceeding of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Istanbul, Turkiye, 5–8 December 2023; pp. 1596–1603.

14. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H.; et al. Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 6881–6890.
15. Strudel, R.; Garcia, R.; Laptev, I.; Schmid, C. Segmenter: Transformer for Semantic Segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 7242–7252.
16. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.
17. He, X.; Zhou, Y.; Zhao, J.; Zhang, D.; Yao, R.; Xue, Y. Swin Transformer Embedding UNet for Remote Sensing Image Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4408715. [CrossRef]
18. Wang, L.; Li, R.; Zhang, C.; Fang, S.; Duan, C.; Meng, X.; Atkinson, P.M. UNetFormer: A UNet-Like Transformer for Efficient Semantic Segmentation of Remote Sensing Urban Scene Imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *190*, 196–214. [CrossRef]
19. Zhang, C.; Jiang, W.; Zhang, Y.; Wang, W.; Zhao, Q.; Wang, C. Transformer and CNN Hybrid Deep Neural Network for Semantic Segmentation of Very-High-Resolution Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4408820. [CrossRef]
20. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 13708–13717.
21. Ma, N.; Zhang, X.; Sun, J. Funnel Activation for Visual Recognition. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 351–368.
22. Rottensteiner, F.; Sohn, G.; Gerke, M.; Wegner, J.D.; Breitkopf, U.; Jung, J. Results of the ISPRS Benchmark on Urban Object Detection and 3D Building Reconstruction. *ISPRS J. Photogramm. Remote Sens.* **2014**, *93*, 256–271. [CrossRef]
23. Lyu, Y.; Vosselman, G.; Xia, G.S.; Yilmaz, A.; Yang, M.Y. UAVid: A Semantic Segmentation Dataset for UAV Imagery. *ISPRS J. Photogramm. Remote Sens.* **2020**, *165*, 108–119. [CrossRef]
24. 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, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
25. Wang, P.; Chen, P.; Yuan, Y.; Liu, D.; Huang, Z.; Hou, X.; Cottrell, G. Understanding Convolution for Semantic Segmentation. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 1451–1460.
26. Maas, A.L.; Hannun, A.Y.; Ng, A.Y. Rectifier Nonlinearities Improve Neural Network Acoustic Models. *Proc. ICML* **2013**, *30*, 3.
27. Glorot, X.; Bordes, A.; Bengio, Y. Deep Sparse Rectifier Neural Networks. *J. Mach. Learn Res.* **2011**, *15*, 315–323.
28. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on Imagenet Classification. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1026–1034.
29. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.
30. Clevert, D.; Unterthiner, T.; Hochreiter, S. Fast and Accurate Deep Network Learning by Exponential Linear Units (elus). *arXiv* **2015**, arXiv:1511.07289.
31. Misra, D. Mish: A Self Regularized Non-monotonic Activation Function. *arXiv* **2019**, arXiv:1908.08681.
32. Chen, Y.; Dai, X.; Liu, M.; Chen, D.; Yuan, L.; Liu, Z. Dynamic Relu. In Proceedings of the European Conference on Computer Vision, Online, 23–28 August 2020; pp. 351–367.