agriculture

MDPI

Article

Semantic Segmentation Algorithm of Rice Small Target Based on Deep Learning

Shuofeng Li [1], Bing Li [1,*], Jin Li [1], Bin Liu [1] and Xin Li [2]

[1] College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin 150001, China
[2] Beijing Aerospace Automatic Control Institute, Beijing 100854, China
* Correspondence: libing265@hrbeu.edu.cn

Abstract: At present, rice is generally in a state of dense adhesion and small granular volume during processing, resulting in no effective semantic segmentation method for rice to extract complete rice. Aiming at the above problems, this paper designs a small object semantic segmentation network model based on multi-view feature fusion. The overall structure of the network is divided into a multi-view feature extraction module, a super-resolution feature building module and a semantic segmentation module. The extraction ability of small target features is improved by super-resolution construction of small target detail features, and the learning ability of the network for small target features is enhanced and expanded through multi-view. At the same time, a dataset of quality inspection during rice processing was constructed. We train and test the model on this dataset. The results show that the average segmentation accuracy of the semantic segmentation model in this paper reaches 87.89%. Compared with the semantic segmentation models such as SegNet, CBAM, RefineNet, DeepLabv3+ and G-FRNet, it has obvious advantages in various indicators, which can provide rice quality detection and an efficient method of rice grain extraction.

Keywords: semantic segmentation; deep learning; small target; feature fusion module

check for updates

Citation: Li, S.; Li, B.; Li, J.; Liu, B.; Li, X. Semantic Segmentation Algorithm of Rice Small Target Based on Deep Learning. *Agriculture* **2022**, *12*, 1232. https://doi.org/10.3390/agriculture12081232

Academic Editors: Nen Fu Huang and Ho-Hsien Chen

Received: 17 July 2022
Accepted: 13 August 2022
Published: 16 August 2022

1. Introduction

Agriculture is regarded as the foundation of the national economy, and food is the top priority. At present, there are still obvious problems such as over-processing of rice in China. Due to over-processing, billions of kilograms of rice are lost every year, which has become an important factor affecting national food security and restricting agricultural efficiency and farmers' income. To reduce grain loss, it is necessary to carry out accurate quality inspection of rice during processing to promote the control precision of rice processing. Rice segmentation is the basis for assessing rice quality. At present, rice is generally densely adhered and has a small granular volume. The pixels of a single grain of rice are less than 3% of the overall image, resulting in no effective rice semantic segmentation method to extract complete rice.

Recently, many researchers have conducted extensive and in-depth research on how to perform effective and automatic segmentation of small targets. The image segmentation method based on graph theory uses graph division to solve the segmentation problem and achieves image segmentation by optimizing the objective function. Algorithms based on the idea of graph theory include GrabCut [1], GraphCut [2], OneCut [3], etc., but this method is too computationally intensive and is suitable for images with the same intra-class similarity. Deep learning methods include Convolutional Neural Networks (CNN) [4], Recurrent Neural Networks (RNN) [5] and Generative Adversarial Networks (GAN) [6], etc. Fully Convolutional Neural Network (FCN) was proposed in 2015 to classify images at the pixel level [7]. The Pointwise Spatial Attention Network (PSANet) method published in ECCV in 2018 utilizes a learned attention mechanism [8] to improve prediction accuracy. The Asymmetric Convolutional Network (ACNet) proposed by Xinxin Hu et al. [9] in

2019 utilizes the Residual Neural Network (ResNet) [10] to improve prediction accuracy. In 2019, JunFu et al. [11] proposed a new type of scene segmentation network (DANet) and proposed a dual attention module, which achieved SOTA effects on both CitySpace and COCO datasets. In 2021, the Hybrid Multi-Attention Network (HMANet) was proposed by Ruigang Niu et al. [12]. In 2019, Zhi Tian et al. [13] proposed a new upsampling module. Chen et al. [14] solved the problem of feature resolution reduction caused by downsampling with hole convolution. Chao Peng et al. [15] proposed an improved global convolutional network method, Aggregate context information from different regions through a pyramid pooling method [8]. Chen et al. [16] proposed an encoder–decoder structure with atrous separable convolution (Deep Lab V3+) for image semantic segmentation and further explored the Xception structure [17]. Raj et al. [18] proposed a multi-dimensional version of convolutional VGG-16. Roy et al. [19] fused coarse to fine for segmentation. Yi Lu et al. [20] proposed a Graph-FCN method. Yuhui Yuan et al. [21] proposed an object context representation method for semantic segmentation. A bidirectional segmentation network (BiSeNet) [22] and a lightweight encoder–decoder network (LEDNet) [23] were also proposed one after another. Li et al. [24] proposed a Deep Feature Aggregation Network (DFANet). Yunchao Wei et al. [25] proposed using atrous convolution to provide a convenient method for inferior supervised and half-supervised semantic segmentation. Anton et al. [26] proposed a novel loss function (gated CRF loss) for weakly supervised image semantic segmentation. Guolei Sun et al. [27] proposed a weakly supervised semantic segmentation method to mine cross-image semantics. Unsong Fan et al. [28] proposed a multi-estimation method for weakly supervised semantic segmentation. Liang-Chieh Chen et al. [29] proposed a half-supervised learning method in video sequences using urban scene segmentation.

To sum up, a lot of research has been performed on small object segmentation. However, the segmentation accuracy of small objects in the scene is still very low, especially small objects in a densely glued state. Small objects have always been a difficult problem to be solved in the field of object segmentation due to their small image size, high density and little information. Semantic segmentation in the case of dense adhesion is the premise of rice quality detection. So it is important to conduct research on semantic segmentation methods for rice. In this paper, an optimized bilateral segmentation network is introduced, and a multi-branch convolution module (MBC-module) is designed to improve the spatial branch network to extract low-level spatial features, increase the horizontal connection of the pyramid structure and use atrous convolution to spread the range of perception. A squeeze incentive network is added to improve the feature learning ability, and finally a dataset of densely bonded rice is constructed to test the algorithm. Compared with the above method, the segmentation accuracy of the rice small target has been improved by at least 1%, which verifies the effectiveness and feasibility of the method.

2. Materials and Methods

In the present study, a semantic segmentation network of small targets using multi-view feature fusion is designed for the rice small objects with dense adhesion and similar features. This network model improves the extraction ability of small target features by super-resolution construction of small target detail features. The overall structure of the network is divided into a multi-view feature extraction module, a super-resolution feature construction module and a semantic segmentation module. The learning capability of the network for small target features is enhanced and expanded through multiple views to provide support for the subsequent construction of super-resolution features. Finally, the features constructed by super-resolution are semantically segmented. The overall network model structure is shown in Figure 1.

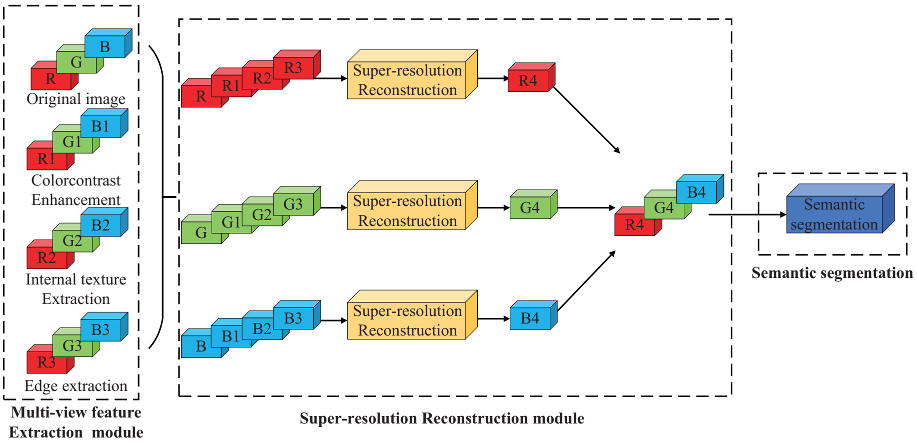
B
G
R
Original image
B1
G1
R1
Colorcontrast Enhancement
B2
G2
R2
Internal texture Extraction
B3
G3
R3
Edge extraction
Multi-view feature Extraction module
R R1 R2 R3
Super-resolution Reconstruction
R4
G G1 G2 G3
Super-resolution Reconstruction
G4
B B1 B2 B3
Super-resolution Reconstruction
B4
R4 G4 B4
Semantic segmentation
Semantic segmentation
Super-resolution Reconstruction module

Figure 1. Block diagram of the semantic segmentation module. It consists of three parts. The left side represents the multi-view feature extraction module. The middle part represents the super-resolution feature reconstruction module. The right side represents the semantic segmentation module.

The specific details of each module are described below.

2.1. Multi-View Feature Extraction Module

In this paper, three kinds of views are generated by color contrast enhancement, internal texture extraction and edge extraction for the small rice target to improve the detailed features of the tiny object from various angles. Histogram equalization is used to increase the contrast of small objects, which is actually to perform nonlinear transformation on the image and update the pixel value of each photo. This allows approximately the same number of pixel values in different grayscale spans. At the same time, the brightness of the highest point in the middle of the previous histogram is improved, while the brightness of the lowest point of the left and right ends has decreased. The histogram of the resulting photo is a relatively flat segmented graph, and the function expression for histogram equalization of the small target is:

$$S_i = T(r_i) = \sum_{i=0}^{k-1} \frac{n_i}{n} \tag{1}$$

In Equation (1), k is the number of gray levels.

In the current investigation, the LBP algorithm is used to collect the texture features of the small target. The LBP algorithm is executed on a 3 × 3 scale. The value of the pixel in the middle of the range is regarded as the comparison value, and the value of the surrounding 8 pixels is subtracted. If the pixel value of the center point is smaller than the pixel value of these points, the point is regarded as 1. Instead, the value is 0. Therefore, a 0-1 encoding with a length of 8 bits can be received, and this value can be used as the LBP value of the middle pixel in the range to represent the details of the 3 × 3 area, as shown in Figure 2.

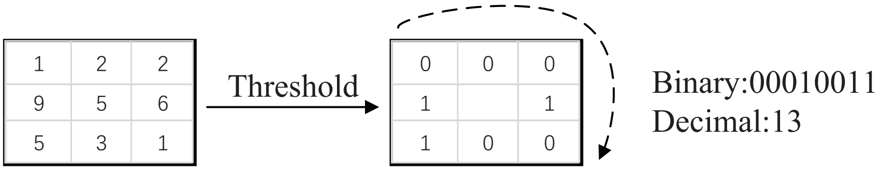

Figure 2. The flow of the LBP algorithm. The value greater than the center point is assigned as 1, and the value less than the center point is assigned as 0. Read clockwise to obtain binary coded data.

In the current study, the Sobel operator is used to extract the edge of the small target. The algorithm includes two types of 3 × 3 matrices, which are 0-degree and 90-degree directions, respectively. Two-dimensional convolution of these matrices with the image yields estimated values of luminance differences in the 0-degree and 90-degree directions, respectively. Equation (2) is as follows: G_x and G_y are the approximations of the partial derivatives of the gray values at 0 degrees and 90 degrees directions, respectively.

$$G_x = \begin{pmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{pmatrix} * A \qquad G_y = \begin{pmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{pmatrix} * A \tag{2}$$

For each point, we can obtain the gradient in two directions, and the estimated value of the gradient can be derived by Equation (3).

$$G = \sqrt{G_x^2 + G_y^2} \tag{3}$$

2.2. Super-Resolution Feature Building Blocks

Since small target features are usually small and not obvious in detail, we use the super-resolution feature construction to improve the features of tiny targets. Although the images may be upscaled by basic bilinear interpolation, the photographs become blurry and have lower characteristics when they are upscaled. Furthermore, tiny objects are still difficult to distinguish from the backdrop. In the present study, two branches are constructed; they are feature extraction and picture reconstruction, as shown in Figure 3.

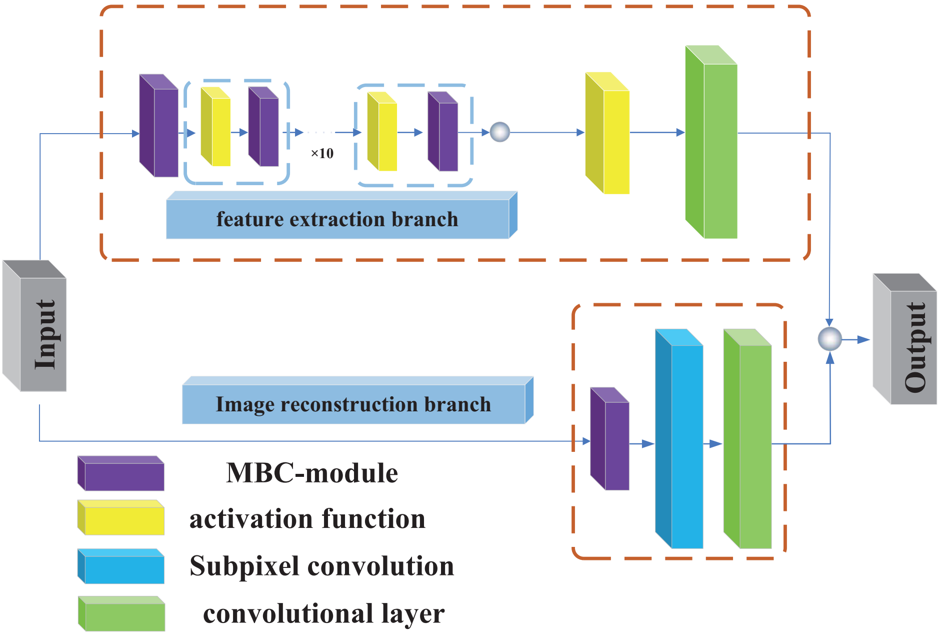

Figure 3. Structure diagram of small target feature super-resolution reconstruction. The module consists of two parts including a feature extraction branch and a feature reconstruction branch.

Reconstructing leftover photos and creating better quality feature maps are both performed using the feature extraction. To create a high-quality image, the reconstruction branch adds the feature map from the previous phase to the upsampled image pixel by pixel. The loss function formula is shown in Equation (4).

$$L(\hat{y}, y; \theta) = \frac{1}{N} \sum_{i=1}^{N} \sum_{s=1}^{L} \rho(\hat{y}_s^{(i)} - y_s^{(i)}) = \frac{1}{N} \sum_{i=1}^{N} \sum_{s=1}^{L} \rho((\hat{y}_s^{(i)} - x_s^{(i)}) - r_s^{(i)}) \tag{4}$$

In Equation (4), s and r_s are residual images, x_s is an enlarged blurred image, $\hat{y}_s$ is a clear image generated by analysis, and y_s is the original clear image (ground truth). Moreover, y_s=x_s+r and $\rho(\cdot)$=$\sqrt{x^2+\varepsilon^2}$ are the Charbonnier penalty functions. In addition to this, θ is a large number of parameters of the module to be corrected. N is the number of images processed each time. *L* is the height of the pyramid. All convolutional layers contain 64 filters, and their size is 3 × 3. PReLUs are activation functions. Small images can be obtained by cropping; they have lower resolution. Through the saliency path, these images can be reconstructed. We also adjusted the super-resolution upsampling rate to 2× to prevent the reconstructed tiny pictures from looking out of proportion to the inspection speed of the back end inspection path.

2.3. Semantic Segmentation Model

In this paper, an improved bilateral segmentation network is used for semantic segmentation for small targets of rice, which optimizes the two-way branches of the original network's spatial path and context path and improves the feature merge module and attention exquisiteness module, thereby further improving the performance of semantic segmentation. The overall network block diagram is shown in Figure 4. The specific details of each part of the optimization are described in detail below.

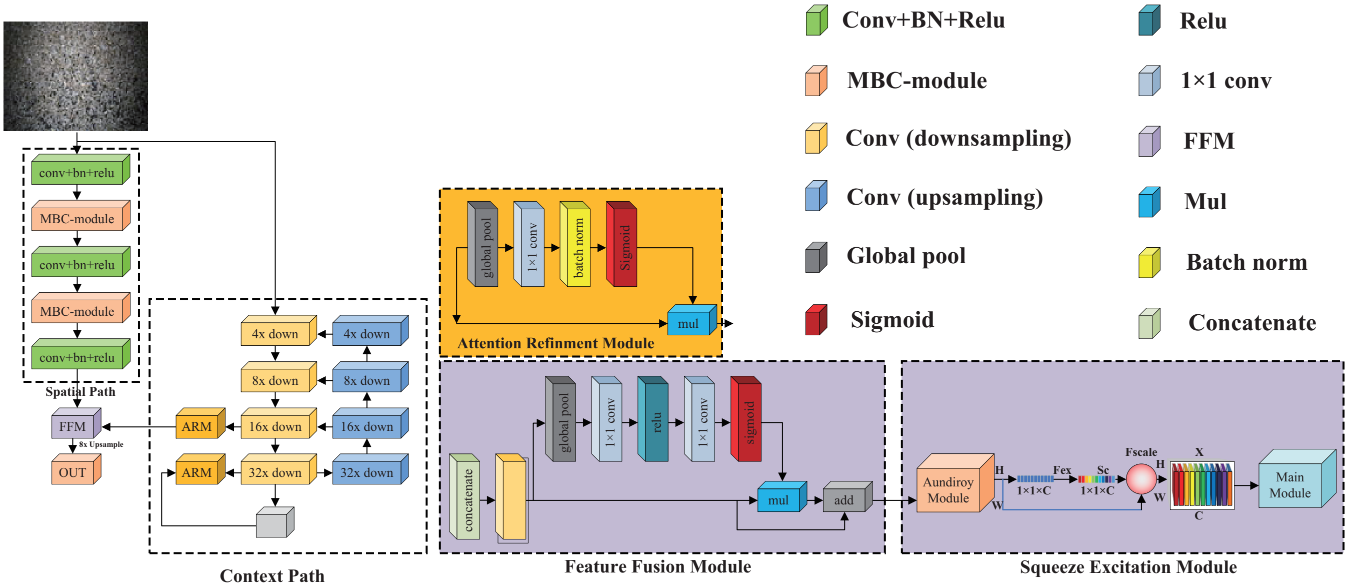

Figure 4. Block diagram of semantic segmentation network. This module mainly includes four parts: context path, attention refinement module, feature fusion module and squeeze excitation module.

(1) Spatial path framework building via multi-branch convolution.
In semantic segmentation tasks, the fully convolutional structure is usually adopted by existing segmentation networks, consisting of an up- and down-sampling process. However, the down-sampling process for small objects will lose important spatial features, resulting in the inability to obtain accurate segmentation results during upsampling. Therefore, we design a multi-branch convolution module (MBC-module) to optimize the spatial branch network to extract low-level spatial features. Its internal structure is divided into a multi-branch convolutional layer and a connected dilated convolutional layer through which multiple branched convolution kernels are connected to obtain receptive fields of different scales, as shown in Figure 5. Bypass pruning is added to the multi-branch convolution part to reduce the large number of convolution kernel channels. First, a 1 × 1 convolution kernel is used to realize the interaction and information integration of the channels and to reduce the dimensionality of the number of convolution kernel channels. The two 5 × 5 convolution kernels were subsequently replaced with two 3 × 3 convolution kernels to reduce the amount of parameters and at the same time enhance the nonlinear ability of the model. In addition, 1 × *n* and *n* × 1 convolution kernels are further used to replace the original convolution kernels to enhance the width and height features. At the

same time, bypass pruning is set to reduce the number of channels of the convolution kernel and reduce the amount of parameter calculation. The improved spatial path is shown in Figure 6.

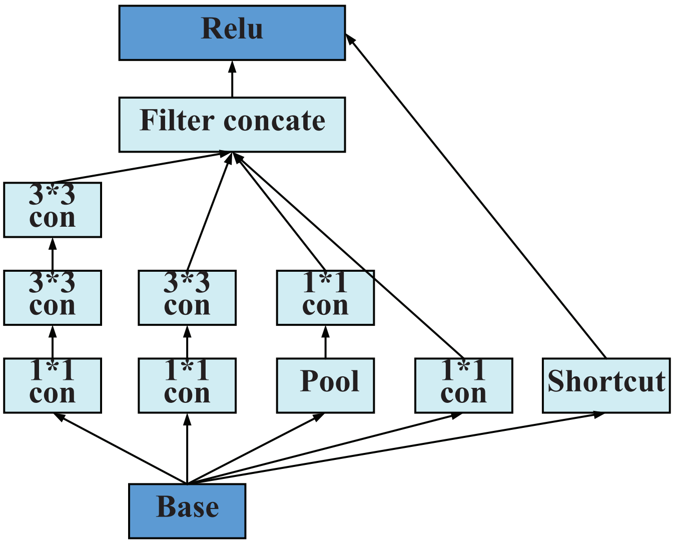

Figure 5. Multi-branch convolution module (MBC-module). By replacing the large convolution kernel with a smaller size convolution kernel, not only can the parameters be reduced, but also the nonlinear ability of the model can be improved.

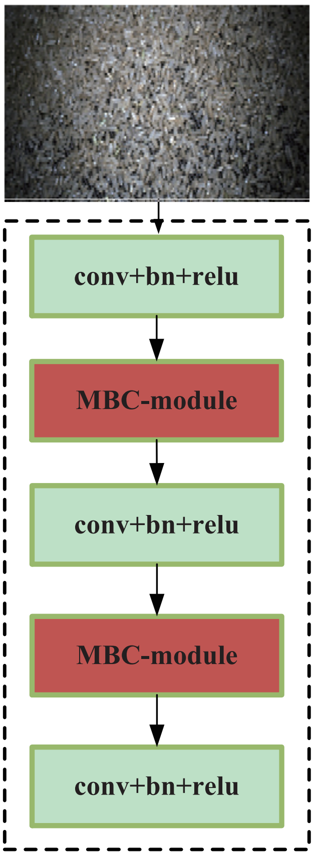

Figure 6. Improved spatial path. It includes three layers of convolution layers and two layers of MBC-modules.

(2) The context path structure based on feature pyramid is established.
Unlike spatial branching networks which need to preserve rich underlying spatial features, contextual branching networks are designed to provide a larger receptive field. In the current investigation, we improve a network branching contextual branching network that utilizes residual networks, global average pooling to obtain large perceptual fields by considering the increase of perceptual fields and the requirement of computational power. To merge the surface feature maps with high precision and the internal feature maps with sufficient semantic information in the context path, the lateral connection of the pyramid structure is added, as shown in Figure 7. Therefore, it is possible to rapidly compose a feature pyramid with rich semantic messages at all sizes from a unitary image of a unitary size, and a residual network is used in the context path to quickly down-sample the feature map to obtain a large perceived range. With these fast down-sampled feature maps, a rich semantic context message is encoded. Furthermore, an overall average pooling is added to the last part of the residual module to receive the global receptive field. The improved context path is shown in Figure 8.

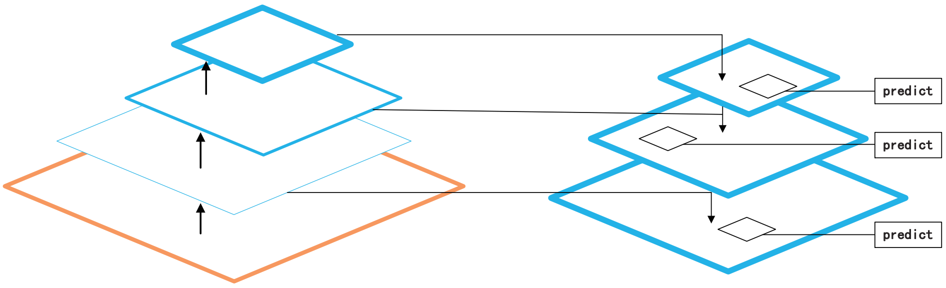

Figure 7. Feature pyramid fusion structure diagram. The horizontal connection of the feature pyramid is added to fuse the information of the shallow feature map and the deep feature map.

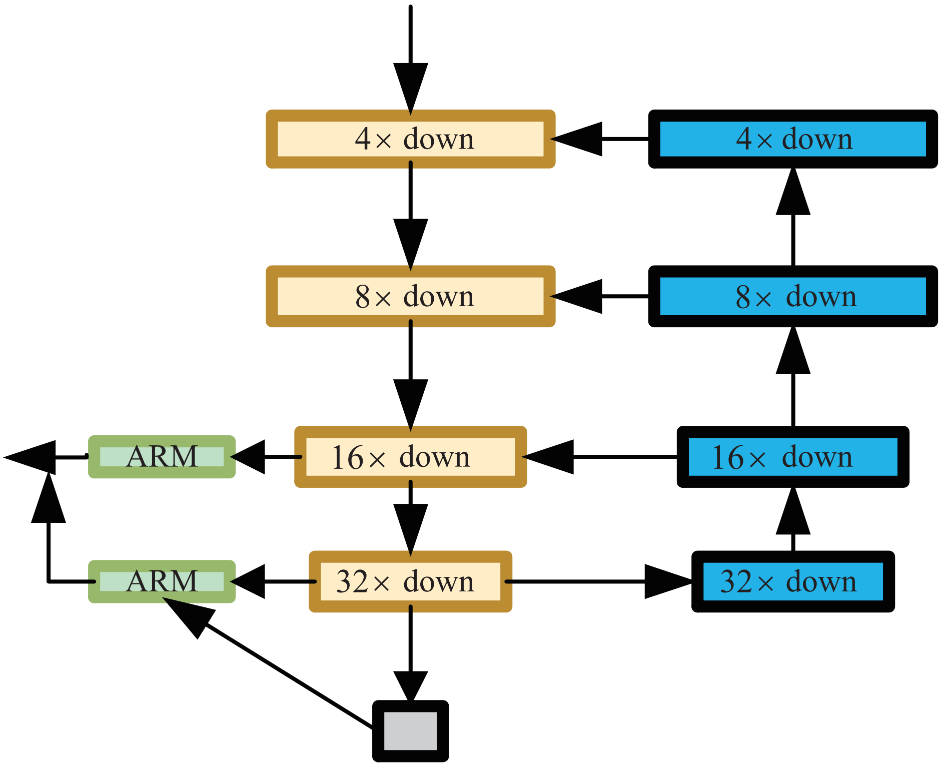

Figure 8. Improved contextual path. Expand the receptive field by using residual networks and global average pooling.

(3) Design of feature fusion module and attention refinement module.
Since the feature representation levels of the two paths are different, the method cannot simply summarize these features. Most of the spatial information captured by the spatial path encodes rich detailed information. Furthermore, the contextual information is mainly encoded by the output features of the contextual path. In summary, the resulting features of spatial paths are poor, while the resulting features of contextual paths are excellent. Then, a customized feature merge module is needed to merge these features, as shown in Figure 9.
In this paper, we consider that feature fusion in which both feature maps are fused with a larger receptive field can better utilize the spatial information of the low-level features and the semantic information of the high-level features. Therefore, in the current study, to not increase the computational effort, we use atrous convolution to expand the receptive field instead of conventional convolution. Atrous convolution improves the perception range by joining atrous into the standard convolution kernel. Contrasted with the ordinary convolution operation, atrous convolution has one more parameter to be adjusted, which is called the atrous rate. It represents the number of spaces between each pixel in the filter, as shown in Figure 10.

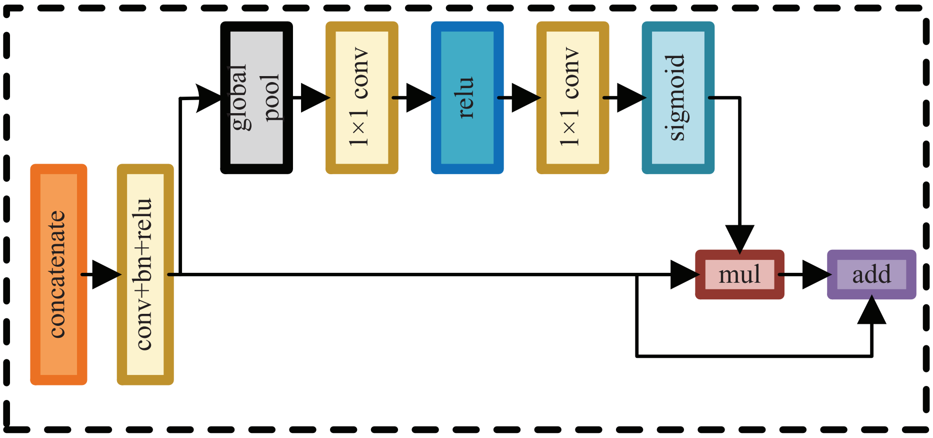

Figure 9. Feature fusion module. It is used to fuse the low-level features output by the spatial path with the high-level features output by the context path.

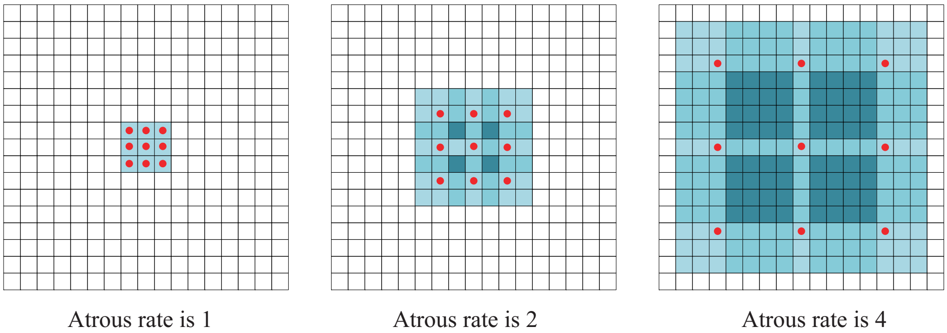

Figure 10. Atrous convolution. Increase the receptive field by injecting atrous.

In the present study, a squeeze excitation module is added to inhibit the incorrect information channel of the attention refinement module and increase the work speed of the module, as shown in Figure 11. The secondary module is connected to the backbone network by two different methods. One of the branches is the result of the assist module, which is integrated through a 1 × 1 convolution before entering the main branch. Another branch complements an attention mechanism structure to the deep auxiliary network between the connections of the two networks. The purpose is to fix the output properties of the assist module. The work process can be classified as squeezing and activation. First, the feature map is condensed, and the two-dimensional feature is transformed into one-dimensional through average pooling. In addition, these feature maps are transformed from two-dimensional feature maps to one-dimensional feature maps through 1 × 1 pooling, which can more perfectly show the arrangement of feature values of all channels and further improve the effect of feature learning.

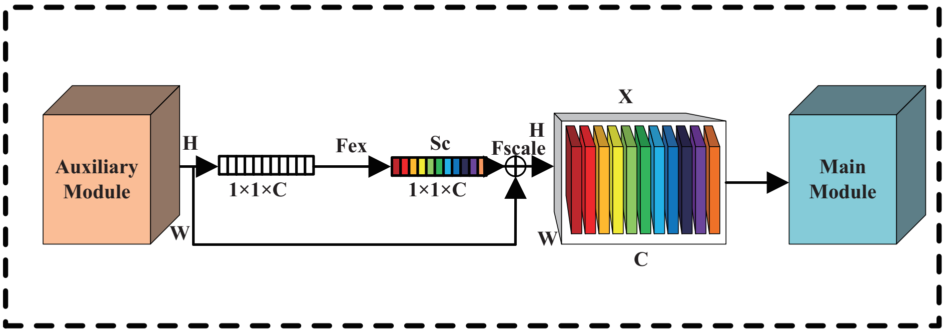

Figure 11. The squeeze excitation module improves the ability to increase learning by squeezing and motivating.

3. Results and Discussion

In the current investigation, we construct a rice segmentation dataset of 9000 images, including 3000 images of long-grain rice, 3000 images of short-grain rice and 3000 images of round-grain rice. We use NVIDIA 1080Ti GPU to conduct all experiments on the Tensorflow platform, and the effect diagram obtained through the trained model is shown in Figure 12.

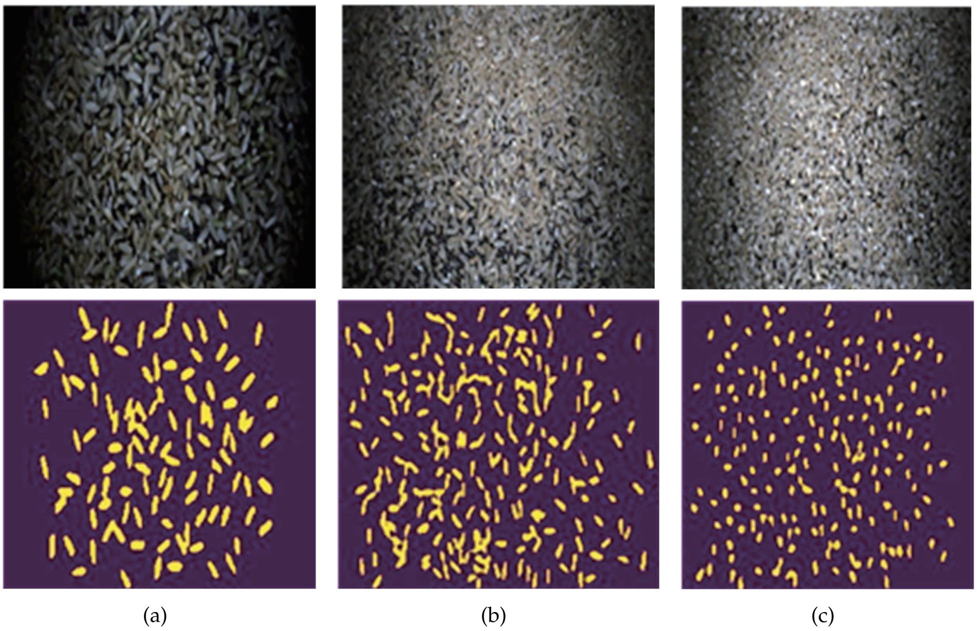

Figure 12. Effect diagram of rice grain semantic segmentation: (**a**) long rice; (**b**) short rice; (**c**) round rice.

Tables 1–3 are the segmentation results on the long rice, the short rice, and the round rice datasets, respectively, and Table 2 represents the semantic segmentation effect on the ISPRS Potsdam dataset. From the experimental results, the proposed semantic segmentation network outperforms the other compared networks. Compared with the G-FRNet model in the long-grain rice dataset, MIoU/ F1-score/Accuracy increased by 1.40%, 1.89% and 1.55%, respectively. Compared with the G-FRNet model in the short-grain rice dataset, MIoU/ F1-score/Accuracy increased by 1.31%, 2.09% and 1.95%, respectively. Compared with the

G-FRNet model in the round-grain rice dataset, MIoU/F1-score/Accuracy increased by 1.23%, 2.07% and 1.17%, respectively.

Table 1. Semantic segmentation results of long-grain rice dataset. The index results of MIOU, F1-score and accuracy of each algorithm.

Model	MIOU (%)	F1-Score (%)	Accuracy (%)
FCN-32s	46.12	55.23	58.76
FCN-8s	53.86	68.71	70.54
U-net [30]	56.77	69.36	71.87
SegNet [31]	59.63	78.25	80.35
CBAM [32]	61.32	80.74	82.71
RefineNet [33]	63.94	82.52	83.65
DeepLabv3+	63.45	83.97	85.93
G-FRNet [34]	64.91	85.19	86.42
Network (ours)	66.31	87.28	87.97

Table 2. Semantic segmentation results of short-grain rice dataset. The index results of MIOU, F1-score and accuracy of each algorithm.

Model	MIOU (%)	F1-Score (%)	Accuracy (%)
FCN-32s	47.43	56.54	61.07
FCN-8s	55.27	70.12	72.25
U-net	58.08	71.67	73.18
SegNet	61.74	79.76	81.66
CBAM	62.63	82.05	85.12
RefineNet	64.35	84.13	84.96
DeepLabv3+	64.76	85.28	87.24
G-FRNet	66.32	86.62	87.23
Network (ours)	67.63	88.51	89.18

Table 3. Semantic segmentation results of the round grain rice dataset. The index results of MIOU, F1-score and accuracy of each algorithm.

Model	MIOU (%)	F1-Score (%)	Accuracy (%)
FCN-32s	44.97	54.08	57.61
FCN-8s	52.52	67.46	69.39
U-net	55.62	68.23	73.73
SegNet	59.08	77.19	79.27
CBAM	61.27	79.39	81.56
RefineNet	62.79	81.77	82.52
DeepLabv3+	62.33	82.82	84.78
G-FRNet	63.86	84.24	85.37
Network (ours)	65.09	86.31	86.54

For small targets with densely adhered rice, spatial information and receptive field are the keys to achieve high precision. Semantic segmentation algorithms such as FCN-32s, FCN-8s, U-net, SegNet, CBAM, RefineNet, DeepLabv3+ and G-FRNet try to maintain the resolution of the input image, encode sufficient spatial information through deconvolution, and pass the pyramid pooling modules or large convolution kernels to capture enough receptive fields. Therefore, the pixel-to-pixel relationship is not fully considered. Iit lacks spatial consistency, and it is not sensitive enough to the details in the image. In particular, the resolution of the target image is much higher than the ordinary resolution, and in the case of real-time semantic segmentation, the existing methods generally use small input images or lightweight models to accelerate, thus losing most of the spatial information of the original image and destroying the spatial information, which ultimately leads to poor overall semantic segmentation results.

The algorithm in this paper first refines the contour details of densely adhered rice through the super-resolution reconstruction method of multi-feature fusion and improves the semantic segmentation accuracy of rice grains, which are small in size and have too many similar features between different targets. A spatial path is improved to preserve the spatial size of the original input image and encode the spatial information of the local detail features of small objects. At the same time, considering that each target requires a large receptive field and fast calculation speed, the context path is improved. A specific attention refinement module is proposed to refine the features of each stage, and attention vectors are calculated to guide feature learning. A lightweight model and global average pooling are utilized to provide a larger receptive field to obtain a larger receptive field that encodes high-level semantic context information. Finally, the up-sampled output features of global pooling are combined with the features of lightweight models.

Compared with FCN-32s, the algorithm in the current study has obvious advantages in the ability to obtain spatial information and semantic information of rice. Therefore, the average accuracy rate is 28.74% higher. The reason why the semantic segmentation effect of FCN-8s is better than that of FCN-32s is that the size of the last convolutional layer is enlarged by four times, so the loss of spatial information is small, but the feature extraction ability for sticky rice grains is still poor. Therefore, the average accuracy of the algorithm in this paper is 17.16% higher than that of FCN-8s. Compared with U-net, the algorithm in the present study is more prominent in the way of multi-scale feature fusion and enhances the detailed semantic features in the way of feature pyramid, so the average accuracy rate is 14.96% higher. Compared with FCN, SegNet has very similar ideas. Although the encoder and decoder networks have better segmentation performance than FCN, the network does not have enough receptive fields for the detailed features of densely bonded rice grain. Therefore, the average accuracy of our algorithm is 7.46% higher. Compared with CBAM, the algorithm in the current investigation has a better attention mechanism, and the difference between the contour features and content features of densely bonded rice is enlarged by the extrusion excitation module, so the average accuracy rate is 4.76% higher. Compared with the U-net network, RefineNet is essentially the same. Although it captures a wide range of background context information through identity mapping and chain residual network, it has poor ability to extract densely glued edge contour feature information. Therefore, the average accuracy of the algorithm in the current study is 4.18% higher than that of RefineNet. Compared with DeepLabv3+, the algorithm in the present study has almost no difference in the range of receptive field, but it is better than it in terms of rice contour feature extraction ability, so the average accuracy rate is 1.92% higher. G-FRNet uses deep features to assist shallow features to filter fuzzy and ambiguous features, effectively integrating low-level, high-level and global feature information. However, compared with the algorithm in the current investigation, the effect of using the spatial path and the context path to determine the semantic contour of the target is slightly insufficient, so the average accuracy is 1.56% higher. In the segmentation effect of long, short and round rice grains, because the feature information of short-grain rice is small, the edge contour features of rice are relatively stable and the invalid interference features are small, the semantic segmentation of short-grain rice with each network is the best effect.

4. Conclusions

Rice is mostly granular and includes small targets, most of which have dense adhesion during processing, which leads to poor detection accuracy and low accuracy of rice quality inspection. In the present study, an optimized semantic segmentation network model is introduced to perform semantic segmentation in the case of dense sticky rice to extract the rice in its complete morphology and establish three kinds of rice datasets of long-grain, short-grain, and round-grain types. On the rice test set, the average segmentation accuracy of the semantic segmentation network in this paper reaches 87.89%, which has obvious advantages in all indexes compared with other semantic segmentation models. Through the improvement of the original segmentation network in the current investigation, the feature extraction effect of rice contour is improved, and the accurate segmentation required for preprocessing during rice quality inspection is realized, thereby improving the quality of rice quality inspection. In the follow-up work, we will continue to expand the research object to segmentation of other cereal grains to achieve more accurate semantic segmentation in different grain quality inspection situations. At the same time, the division efficiency is improved, and the index analysis of the divided grains is carried out to further make greater contributions to the grain processing industry.

Author Contributions: Methodology, S.L., B.L. (Bing Li), J.L., B.L. (Bin Liu) and X.L.; validation, S.L. and B.L. (Bin Liu); writing—original draft preparation, S.L. and B.L. (Bin Liu); writing—review and editing, S.L., B.L. (Bing Li), J.L. and X.L., supervision, S.L., B.L. (Bing Li), J.L. and X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Foundation of Heilongjiang Province grant number KY10400210217, Fundamental Strengthening Program Technical Field Fund grant number 2021-JCJQ-JJ-0026.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Rother, C.; Kolmogorov, V.; Blake, A. "GrabCut": Interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph. (TOG)* **2004**, *23*, 309–314. [CrossRef]
2. Boykov, Y.; Jolly, M.P. Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images. In Proceedings of the Eighth IEEE International Conference on Computer Vision, ICCV 2001, Vancouver, BC, Canada, 7–14 July 2001; Volume 1, pp. 105–112 . [CrossRef]
3. Tang, M.; Gorelick, L.; Veksler, O.; Boykov, Y. GrabCut in One Cut. In Proceedings of the 2013 IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 1769–1776. [CrossRef]
4. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
5. Pearlmutter, B. Gradient calculations for dynamic recurrent neural networks: A survey. *IEEE Trans. Neural Netw.* **1995**, *6*, 1212–1228. [CrossRef]
6. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]
7. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June2015; pp. 3431–3440. [CrossRef]
8. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6230–6239. [CrossRef]
9. Ding, X.; Guo, Y.; Ding, G.; Han, J. ACNet: Strengthening the Kernel Skeletons for Powerful CNN via Asymmetric Convolution Blocks. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 1911–1920. [CrossRef]
10. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), LasVegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
11. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual Attention Network for Scene Segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 3141–3149. [CrossRef]
12. Niu, R.; Sun, X.; Tian, Y.; Diao, W.; Chen, K.; Fu, K. Hybrid Multiple Attention Network for Semantic Segmentation in Aerial Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5603018. [CrossRef]
13. Tian, Z.; He, T.; Shen, C.; Yan, Y. Decoders Matter for Semantic Segmentation: Data-Dependent Decoding Enables Flexible Feature Aggregation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 3121–3130. [CrossRef]

14. 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]
15. Peng, C.; Zhang, X.; Yu, G.; Luo, G.; Sun, J. Large Kernel Matters—Improve Semantic Segmentation by Global Convolutional Network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1743–1751. [CrossRef]
16. 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.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2018; pp. 833–851. [CrossRef]
17. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1800–1807. [CrossRef]
18. Eigen, D.; Fergus, R. Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-scale Convolutional Architecture. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 2650–2658. [CrossRef]
19. Roy, A.; Todorovic, S. A Multi-scale CNN for Affordance Segmentation in RGB Images. In Proceedings of the Computer Vision—ECCV 2016, Amsterdam, The Netherlands, 11–14 October 2016; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2016; pp. 186–201. [CrossRef]
20. Lu, Y.; Yaran, C.; Zhao, D.; Chen, J. *Graph-FCN for Image Semantic Segmentation*; Springer: Cham, Switzerland, 2020.
21. Yuan, Y.; Chen, X.; Wang, J. *Object-Contextual Representations for Semantic Segmentation*; Springer: Cham, Switzerland, 2019.
22. Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; Sang, N. BiSeNet: Bilateral Segmentation Network for Real-Time Semantic Segmentation. In Proceedings of the Computer Vision—ECCV 2018, Munich, Germany, 8–14 September 2018; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2018; pp. 334–349. [CrossRef]
23. Wang, Y.; Zhou, Q.; Liu, J.; Xiong, J.; Gao, G.; Wu, X.; Latecki, L.J. Lednet: A Lightweight Encoder-Decoder Network for Real-Time Semantic Segmentation. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 1860–1864. [CrossRef]
24. Li, H.; Xiong, P.; Fan, H.; Sun, J. DFANet: Deep Feature Aggregation for Real-Time Semantic Segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 9514–9523. [CrossRef]
25. Wei, Y.; Xiao, H.; Shi, H.; Jie, Z.; Feng, J.; Huang, T.S. Revisiting Dilated Convolution: A Simple Approach for Weakly- and Semi-Supervised Semantic Segmentation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7268–7277. [CrossRef]
26. Lee, J.; Kim, E.; Lee, S.; Lee, J.; Yoon, S. FickleNet: Weakly and Semi-Supervised Semantic Image Segmentation Using Stochastic Inference. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 5262–5271. [CrossRef]
27. Sun, G.; Wang, W.; Dai, J.; Van Gool, L. Mining Cross-Image Semantics for Weakly Supervised Semantic Segmentation. In Proceedings of the Computer Vision—ECCV 2020, Glasgow, UK, 23–28 August 2020; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2020; pp. 347–365. [CrossRef]
28. Fan, J.; Zhang, Z.; Tan, T. Employing Multi-estimations for Weakly-Supervised Semantic Segmentation. In Proceedings of the Computer Vision—ECCV 2020, Glasgow, UK, 23–28 August 2020; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2020; pp. 332–348. [CrossRef]
29. Chen, L.C.; Lopes, R.G.; Cheng, B.; Collins, M.D.; Cubuk, E.D.; Zoph, B.; Adam, H.; Shlens, J. Naive-Student: Leveraging Semi-Supervised Learning in Video Sequences for Urban Scene Segmentation. In Proceedings of the Computer Vision—ECCV 2020, Glasgow, UK, 23–28 August 2020; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2020; pp. 695–714. [CrossRef]
30. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. Technical Report. *arXiv* **2015**, arXiv:1505.04597.
31. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. Technical Report. *arXiv* **2016**, arXiv:1511.00561.
32. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the Computer Vision–ECCV, Munich, Germany, 8–14 September 2018; Springer: Cham, Switzerland, 2018; pp. 3–19. [CrossRef]
33. Lin, G.; Milan, A.; Shen, C.; Reid, I. RefineNet: Multi-Path Refinement Networks for High-Resolution Semantic Segmentation. Technical Report. *arXiv* **2016**, arXiv:1611.06612.
34. Islam, M.A.; Rochan, M.; Bruce, N.D.B.; Wang, Y. Gated Feedback Refinement Network for Dense Image Labeling. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4877–4885. [CrossRef]