remote sensing

MDPI

Article

Multi-Pooling Context Network for Image Semantic Segmentation

Qing Liu, Yongsheng Dong *, Zhiqiang Jiang, Yuanhua Pei, Boshi Zheng, Lintao Zheng and Zhumu Fu

School of Information Engineering, Henan University of Science and Technology, Luoyang 471023, China
* Correspondence: ysdong@haust.edu.cn

Abstract: With the development of image segmentation technology, image context information plays an increasingly important role in semantic segmentation. However, due to the complexity of context information in different feature maps, simple context capture operations can easily cause context information omission. Rich context information can better classify categories and improve the quality of image segmentation. On the contrary, poor context information will lead to blurred image category segmentation and an incomplete target edge. In order to capture rich context information as completely as possible, we constructed a Multi-Pooling Context Network (MPCNet), which is a multi-pool contextual network for the semantic segmentation of images. Specifically, we first proposed the Pooling Context Aggregation Module to capture the deep context information of the image by processing the information between the space, channel, and pixel of the image. At the same time, the Spatial Context Module was constructed to capture the detailed spatial context of images at different stages of the network. The whole network structure adopted the form of codec to better extract image context. Finally, we performed extensive experiments on three semantic segmentation datasets (Cityscapes, ADE20K, and PASCAL VOC2012 datasets), which fully proved that our proposed network effectively alleviated the lack of context extraction and verified the effectiveness of the network.

Keywords: semantic segmentation; context information; convolutional neural network; attention module

Citation: Liu, Q.; Dong, Y.; Jiang, Z.; Pei, Y.; Zheng, B.; Zheng, L.; Fu, Z. Multi-Pooling Context Network for Image Semantic Segmentation. *Remote Sens.* **2023**, *15*, 2800. https://doi.org/10.3390/rs15112800

Academic Editors: Jiaojiao Li, Qian Du, Wei Li, Bobo Xi, Jocelyn Chanussot, Rui Song and Yunsong Li

Received: 2 April 2023
Revised: 10 May 2023
Accepted: 12 May 2023
Published: 28 May 2023

1. Introduction

Image segmentation is an important part of computer vision, and semantic segmentation is a basic task of image segmentation. Semantic segmentation involves pixel-level semantic image processing, which is mainly utilizes the relationship between pixels and their surroundings. The development of deep learning has led to the widespread use of image semantic segmentation in real-life applications, such as medical imaging [1–3], assisted driving [4–7], and radar image processing [8–10]. Context information usually represents the relationship between its own pixels and surrounding pixels, which is crucial for visually understanding tasks. The main principle of image semantic segmentation is to give corresponding semantic expression to all pixels in the image. This expression not only pays attention to the meaning of its own pixels, but also needs to express the relationship between its own pixels and surrounding pixels. Therefore, context information is an important factor in image semantic segmentation. Contextual information is not only often used in the field of segmentation, but is also a common method of problem solving in other areas [11–13]. We divide the context information into semantic context information and spatial context information according to different image feature maps. Semantic context information is often contained in low-resolution, high-level feature maps, which is mainly used to distinguish pixel categories. The spatial context information is mainly used in a high-resolution, low-level feature map to help the pixel restore the spatial details. The combination of these two context information types greatly improves the quality of image semantic segmentation.

With the development of the convolutional neural network, more and more methods have been used to capture rich semantic context information. For example, Context-Reinforced Semantic Segmentation [14] proposes a context-enhanced semantic segmentation network to explore the advanced semantic context information in a feature graph. It embeds the learned context into the segmentation reasoning based on FCN [15] to further enhance the modern semantic segmentation method. The Co-Occurrent Features Network [16] designs a special module to learn fine-grained spatial information representation and constructs overall contextual feature information by aggregating co-occurrence feature probabilities in co-occurrence contexts. Context Encoding for Semantic Segmentation [17] is used to capture the semantic information in the scene using the encoding and decoding module to selectively filter the information with the same class of features. The Context Deconvolution Network for Semantic Segmentation [18] proposes a context deconvolution network and focuses on the semantic context association in decoding network. The Gated Path Selection Network [19] has developed a gated path selection network. In order to dynamically select the required semantic context, the gate prediction module is further introduced. Unlike previous efforts to capture semantic context information, its network can adaptively capture dense context. LightFGCNet [20] has designed a lightweight global context capture method and combines feature information from different regions during the upsampling phase to enable better global context extraction across the network. BCANet [21] has designed a boundary-guided context aggregation module to capture the correlation context between pixels in the boundary region and other pixels to facilitate the understanding of the semantic information of the overall image. DMAU-Net Network [22] presents an attention-based multiscale maximum pooling dense network, which designs an integrated maximum pool module to improve the image information feature extraction ability in the encoder section, thereby improving the network segmentation efficiency. The Multiscale Progressive Segmentation Network [23] presents a multiscale progressive segmentation network that gradually divides image targets into small, large, and other scales and cascades them into three distinct subnetworks to achieve the final image segmentation result. The Semantic Segmentation Network [24] presents a semantic segmentation network that combines multi-path structure, attention weighting, and multi-scale encoding. It captures spatial information, semantic context information, and semantic map information of images through three parallel structures. The Combining Max-Pooling Network [25] combines the traditional wavelet algorithm with a convolutional neural network pooling operation to propose a new multi-pooling scheme, and it uses this scheme to create two new stream architectures for semantically segmenting images.

There are many ways to use spatial context information. For example, the CBAM [26] aggregates spatially detailed information about pixels through pooling operations and generates different spatial context descriptors through a spatial attention module to capture spatial detail context information. The spatial context is generally found in high-resolution feature maps or in the connection of pixels to other pixels. As a result, they cannot capture spatial context information for objects that reside at different scales. The Feature Pyramid Transformer [27] uses specially designed converters to form feature pyramids in a top-down or opposite interaction to capture high-resolution spatial context. To reduce the computational effort needed to capture more spatial context, the Fast Attention Network [28] captures the same spatial context at a fraction of the computational cost by using different orders of spatial attention. The HRCNet [29] maintains spatial contextual information through a specific network structure, obtains global contextual information during the feature extraction phase, and uses a feature-enhanced feature pyramid structure to fuse contextual information at different scales. The CTNet [30] has designed a spatial context module and a channel context module to capture the semantic and spatial context between different pixel features by exploring inter-pixel correlations.

These methods have excellent performance in extracting semantic context and spatial context information. For better image semantic segmentation, not only rich semantic context, but also sufficient spatial context information is required. We believe that a good

combination of these two context information types can better complete the semantic segmentation task and improve the segmentation quality. Therefore, we designed a new network structure: the Multi-Pooling Context Network (MPCNet). The MPCNet captures feature context information in different stages through encoding and decoding structures. Specifically, we designed a Pooling Context Aggregation Module (PCAM), which is composed of multiple pooling operations and dilated convolutions. The application captures rich semantic context information in low-resolution high-level feature map to improve the utilization of semantic-related context in a high-level feature map. In addition, a Spatial Context Module (SCM) was proposed, which is composed of maximum pooling and average pooling. It captures the spatial context in a low-level feature map and provides the output to the encoder in the form of a jump connection to form each decoding stage, so as to better restore the spatial details of pixels. Our MPCNet captures rich semantic context information through the encoder and combines the spatial context information from the decoder that is captured by jump connection to form the encoding and decoding structure of the whole network, which not only improves the information conversion rate of pixels, but also increases the utilization rate of the context information, thus improving the quality of semantic segmentation.

The following are our main contributions:

(1) We constructed a Multi-Pooling Context Network (MPCNet), which captures rich semantic context information through the encoder and restores the spatial context information through the decoder formed by the jump connection. The whole network realizes the effective combination of semantic context and spatial context with the encoding and decoding structure, thus completing the semantic segmentation task.
(2) We designed a Spatial Context Module (SCM), which is composed of different types of pooling layers. It transfers the spatial information in the low-level feature map at the encoding stage to each decoding stage through the jump connection, improves the information utilization of the spatial context, and, thus, increases the pixel location of the semantic category.
(3) We designed a Pooling Context Aggregation Module (PCAM) consisting of a combination of different pooling operations and dilation convolution. It cooperates with the encoder to capture different contexts in the high-level feature graph, thereby creating rich semantic contextual information for pixel classification.

2. Related Work

In this section, we introduce some relevant semantic context and spatial context information capture methods and popular semantic segmentation models.

2.1. Semantic Context Information

An image is composed of several pixels. Semantic segmentation is mainly performed to label several pixels in the image. Successfully partitioning each pixel requires rich semantic context information. Semantic context information can effectively improve the semantic classification of pixel images. In recent years, semantic context has fully verified its effectiveness in semantic segmentation methods. For example, PSPNet [31] collects the feature information of pixels by pooling the pyramids of different sizes to obtain rich semantic context for the semantic segmentation of images. ParseNet [32] uses the average feature of the layer to increase the information of each location, and then adds the semantic context to the full convolutional network to improve the image segmentation quality. DeepLabV3+ [33] is designed to expand the convolutional composition of the atrous spatial pyramid pool module to capture rich semantic contextual information, thereby improving the segmentation performance of the network. DDRNet [34] establishes two parallel depth branches and uses the two-branch structure to search the semantic context in the low-resolution feature map. The Gated Full Fusion for Semantic Segmentation [35] (GFF) uses gates to selectively fuse semantic contexts at all levels in a fully connected way, uses gate control units to control the propagation of useful semantic contexts, and

suppresses additional contextual information noise. These methods improve the semantic segmentation performance through their unique network design. They pay more attention to large-scale pixel semantic information. On the contrary, our network method aims to combine the multi-scale semantic context information in the low-resolution feature map and achieve the purpose of multi-scale semantic context information and increase the feature receptive field through different pooling combinations and dilated convolutions, so as to capture more relevant semantic context information.

2.2. Spatial Context Information

Several pixels in the image are closely connected; pixels themselves and between pixels have different meanings. In the process of semantic segmentation, it is necessary to know the different meanings between the pixel itself and other pixels, but these meanings are often contained in the spatial context information. At present, many methods are exploring how to better capture the spatial detail context information of images. One example is the SpaceMeshLab—featuring Spatial Context Memorization. Furthermore, the Meshgrid Through Convergence Consus For Semantic Segmentation [36] proposed a spatial context memo, which preserves the input dimension through the bypass branch of this spatial context and constantly communicates with the backbone network to capture its spatial context information. Context Encoding and Multi-Path Decoding [37] propose a scale selection scheme, thereby selectively fusing information from different scale features, preserving the rich spatial context information fraction in the feature scale, and improving the segmentation performance of pixel spatial details. BiSeNetV2 [38] introduces a new feature fusion module to effectively combine spatial and semantic context information, interactively explore spatial and semantic context information, and find different pixels for semantic segmentation. SGCPNet [39] devises a spatial detail-oriented context propagation strategy that uses shallow spatial detail to guide the global context and also effectively recovers lost spatial detail information. These methods have performed well in completing the capture of spatial context information, and have a good restoration and reconstruction effect on pixel spatial details, whether from the multi-scale or multi-branch. The difference is that our method compensates for the lost spatial context in the down-sampling process by combining pooling operations and transfers it to the corresponding image up-sampling stage in the down-sampling stage, which greatly compensates for the spatial details of image segmentation.

3. Methodology

In this section, we first explain the framework of our Multi-Pooling Context Network (MPCNet) and present the main principles of the two proposed modules—the Pooling Context Aggregation Module (PCAM) and the Spatial Context Module (SCM).

3.1. Overview

The structure of the Multi-Pooling Context Network for semantic segmentation (MPCNet) proposed by us is shown in Figure 1. The network uses codec as its main architecture that uses the pre-training residual network ResNet101 [40] as the encoding stage. Since down-sampling loses the spatial details of the image, we used a 3×3 convolution with a step size of 2 instead of the down-sampling operation of the backbone network. In the last resolution stage, we set the step size to 1 and used a 3×3 dilation convolution with a dilation rate of 2 instead of the convolution. In this way, the image features are retained at resolutions of $1/4$, $1/8$, $1/16$, and $1/16$, and the number of channels corresponding to each resolution is 256, 512, 1024 and 2048, respectively. These four feature resolutions also represent four different coding stages. In order to capture more semantic contexts, we applied the Pooling Context Aggregation Module (PCAM) in the last coding stage. At the same time, the Spatial Context Module (SCM) was used to capture the spatial context information of the first three coding stages, and the spatial information of the first three coding stages formed the decoding stage in the form of jump connection with the flow fusion [41]

and the output of PCAM module. In this way, the spatial details of the corresponding image encoding phase will exist in the corresponding decoding phase.

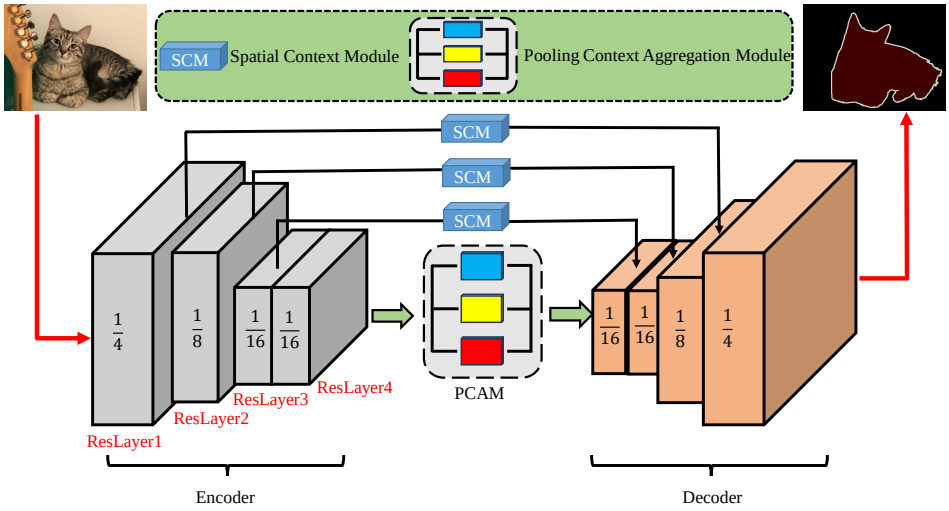

Figure 1. Overview of MPCNet. ResNet is used as the encoder backbone network, and its four different resolution layers such as ResLayer1, ResLayer2, ResLayer3, and ResLayer4 are used as the encoder stage. The PCAM obtains the semantic context of high-level features at the coding stage. The SCM sends the spatial context extracted in each encoding stage to the decoder in the form of jump connection. The whole network uses an encoding and decoding structure for semantic segmentation. (Best in color).

Note that our MPCNet aims to capture more context information for semantic segmentation. MPCNet captures three parts of the context in the encoder's high-level feature map by PCAM to form rich semantic context information, divides several categories of image pixels, and then transfers the spatial context of the image pixels to the decoder in the form of skip connection with the spatial context captured by SCM at each stage of encoding, thus restoring the spatial details of the image pixels. In order to better capture the context information, our entire model uses a codec–decode structure, extracts the context information of the image using the backbone network as the encoder to reduce the resolution, captures the semantic context information through PCAM, and combines it with the spatial detail context captured by SCM in the form of jump connection. By sampling step-by-step to form the decoder, each module structure of the whole network is clear, simple, and easy to implement.

3.2. Spatial Context Module

With the continuous down-sampling of the convolutional neural network, the low-resolution pixels of the image will lose the spatial detail information, thus resulting in blurred target boundaries. To reduce the loss of spatial detail, the spatial position of the target pixels was improved. We built the Spatial Context Module (SCM). Figure 2 shows our proposed Spatial Context Module (SCM) structure. It can be seen from Figure 2 that SCM is an integrated design of the whole module, which can be flexibly applied to any network structure. Next, let us introduce SCM in detail.

First, we used high-resolution feature map as input, but because the number of feature map channels in each stage was different, we used common convolution to unify the number of channels, then used maximum pooling and average pooling operations to collect different weight information of feature map, and then fused different weight information. The context weight obtained was calculated by sigmoid function, and then all the weight information output by sigmoid was selected by using the features of the unified channel, filtering out redundant information, and preserving relevant spatial details. To prevent the gradients from disappearing due to the increase in network depth, we initialized the

connection of spatial contextual information to ensure smooth transmission of the gradients. For spatial context module output O_{output}, the specific expression is

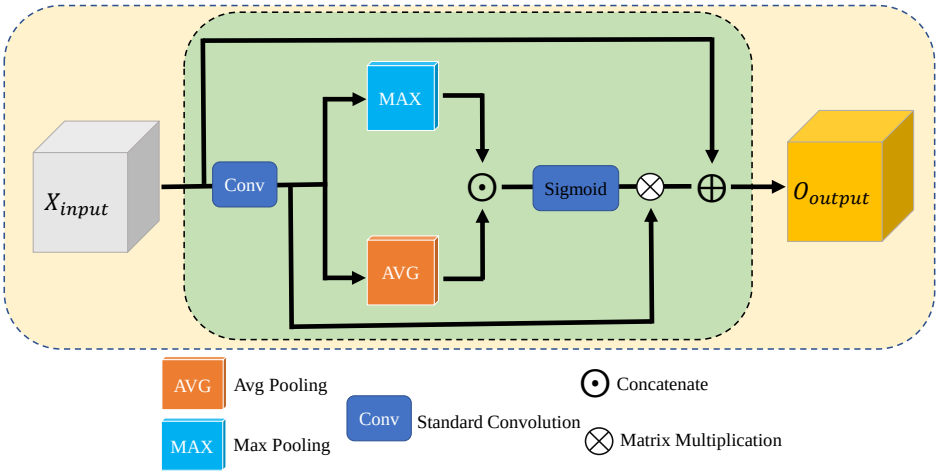

Figure 2. Overview of Spatial Context Module (SCM). It captures the spatial context information in the high-level feature graph through different pooling operations.

$$O_{output} = Sig\left[Max\left(Conv\left(X_{input}\right)\right) \odot Avg\left(Conv\left(X_{input}\right)\right)\right] \otimes Conv\left(X_{input}\right) \oplus X_{input}, \quad (1)$$

where Max represents maximum pooling, Avg represents average pooling, $Conv$ represents standard convolution, Sig represents sigmoid function, X_{input} represents high-resolution input features, $\odot$ represents concat, $\otimes$ represents matrix element multiplication, and $\oplus$ represents element summation.

Our Spatial Context Module aims to capture spatial details in high-resolution feature maps. First, we used the channel number of the convolution uniform feature map and then used the pooling operation to obtain different information weights. Because the maximum pooling can obtain more prominent pixel information weights on the image, and the average pooling can obtain additional target information, we used two parallel poolings to capture the weight information of the image, then used the probability function to effectively select it, and finally filtered out redundant information and output spatial details between image pixels. This preserved effective spatial context information in the high-resolution feature map.

3.3. *Pooling Context Aggregation Module*

Semantic context is crucial for semantic segmentation. Semantic information of dense pixels is generally reserved in low-resolution feature images, so it is necessary to reduce the resolution of the image to extract rich semantic information. However, in an image with complex background, we should not only pay attention to the semantic information of low resolution, but also pay attention to the context information between its own semantic pixels and surrounding pixels. In order to better capture the rich context information with low resolution, we designed the Pooling Context Aggregation Module (PCAM). Figure 3 shows the structure of PCAM. From Figure 3, we can see that PCAM is composed of three parts. Next, we will introduce the Pooling Context Aggregation Module in detail.

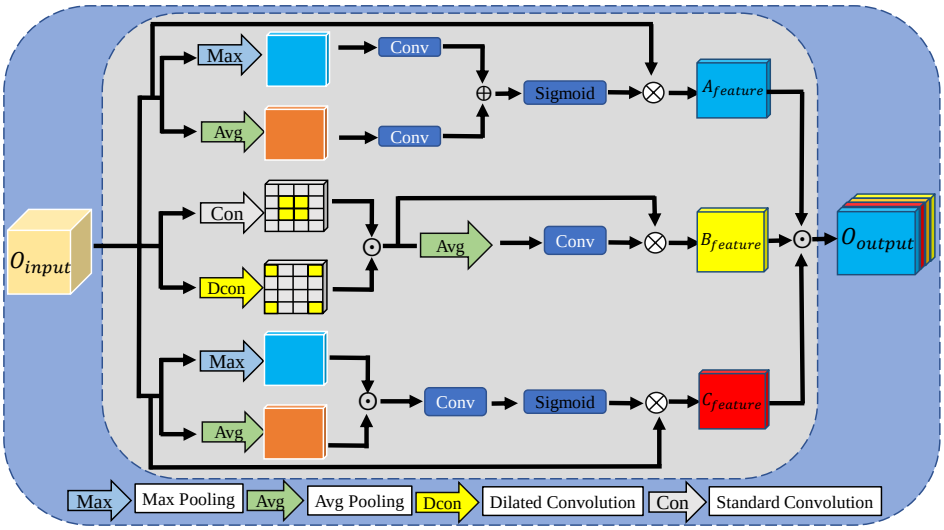

Figure 3. Overview of Pooling Context Aggregation Module(PCAM). It is mainly composed of three parts of context information by capturing the semantic context in the low-resolution feature map.

The Pooling Context Aggregation Module (PCAM) is composed of three different parts, and the corresponding capture $A_{feature}$, B_{feture}, and C_{feture} has three parts of context information. First, the input low-resolution feature O_{input} performs maximum pooling and average pooling operations, and it then uses 1×1 convolution to capture the context information between its channels after each pooling module. The maximum pooling channel information and average pooling channel information are fused to form a complete channel context weight. The weight probability is expressed using the sigmoid function, and then the channel weight is selected with the initial input characteristics to remove redundant channel information, as well as preserve complete and rich channel context information $A_{feature}$. Next, in the second part, we use ordinary and dilation convolution to expand the receptive field of the input features, as well as fuse and retain contextual information between pixels. Then, average pooling and convolution are used to select weights for feature links, remove redundant information, retain useful information between pixels, increase connectivity between pixels, and capture contextual information between pixels B_{feture}. The last part is the spatial context information that is captured by the spatial context module C_{feture}. The captured three-part context information is fused to form a low-resolution semantic context O_{output}. The formal description of output is as follows:

$$O_{output} = A_{feature} \odot B_{feature} \odot C_{feature}, \tag{2}$$

where $A_{feature}$, B_{feture}, and C_{feture} represent channel context information, context information between pixels, and spatial context information, respectively. They are specifically expressed as follows:

$$A_{feature} = Sig\left[Conv\left(Max\left(O_{input}\right)\right) \oplus Conv\left(Avg\left(O_{input}\right)\right)\right] \otimes O_{input}, \tag{3}$$

$$B_{feature} = Conv\left(Avg\left(Conv\left(O_{input}\right) \odot Dconv\left(O_{input}\right)\right)\right) \otimes \left(Conv\left(O_{input}\right) \odot Dconv\left(O_{input}\right)\right), \tag{4}$$

$$C_{feature} = Sig\left[Conv\left(\left(Max\left(O_{input}\right) \odot Avg\left(O_{input}\right)\right)\right)\right] \otimes O_{input}, \tag{5}$$

where Max represents maximum pooling, Avg represents average pooling, $Conv$ represents standard convolution, $Dconv$ represents 3×3 dilated convolution. Sig represents sigmoid function, O_{input} represents low-resolution input features, $\odot$ represents concat, $\otimes$ represents matrix element multiplication, and $\oplus$ element summation.

Our proposed Pooling Context Aggregation Module aims to capture rich semantic context information of low-resolution feature maps through different pooling and convolu-

tion operations. The channel weight is expressed by probability through maximum pooling and average pooling, and the context information between its channels is obtained; in order to preserve the connection between pixels, we use dilated convolution to capturing the context information between pixels; because the low-resolution feature map also contains spatial details, we use the spatial context module to capture its spatial context. Unlike the high-resolution spatial context module, we remove the unified channel convolution and initialization connection. The whole low-resolution semantic context is composed of these three parts of context information. It not only divides the semantic categories of each pixel, but also distinguishes itself and surrounding pixels by certain pixel categories. It ensures the semantic correctness of different pixels.

4. Experimental Results

In this section, we compare numerical and segmentation results with ten image semantic segmentation methods from recent years on the PASCAL VOC2012 dataset [42], the Cityscape dataset [43], and the ADE20K MIT dataset [44].

4.1. Datasets and Experimental Settings

In this subsection, we first introduce the three semantic segmentation datasets used for network training, and then detail the specific parameter details of the experiments.

4.1.1. PASCAL VOC2012

PASCAL VOC 2012 is a computer vision competition dataset. It is divided into three sections according to the data training requirements: training, evaluation, and test sets. Each set has roughly 1400 images. The categories of these images include not only humans and animals, but also driving tools, indoor scenes, etc. There are 21 categories covering many objects in our lives.

4.1.2. Cityscapes

Cityscapes is a vehicle driving dataset. It has a total of 19 street view category labels, and the dataset is divided into three parts, including a training, evaluation, and test dataset. The corresponding images are 2979, 500, and 1525, respectively, and each image has a high resolution of 2048 $\times$ 1024.

4.1.3. ADE20K MIT

The ADE20K dataset is MIT's open scene understanding dataset. It contains over 20 K images of over 3000 object classes. Because of the complexity of the classes, the samples in the dataset have different resolutions of up to 2400 $\times$ 1800 pixels.

4.1.4. Experimental Settings

We implemented our network on a single GPU using the Python language, which used ResNet101 with a dilated convolution strategy as the backbone of the network. Specifically, we replaced the pooling module with dilated convolution and resolved the size of Resnet's final output feature map to 1/16, thus avoiding 1/8, which would use too much GPU memory, and ensuring sufficient contextual information.

Our experiments generally refer to most previous work [33,45,46] using pixel accuracy (PA), intersection over union (IoU), and mean intersection of union (mIoU) as evaluation metrics [47]. A combination of random gradient descent (SGD) [48] and cross-entropy loss with a small batchsize dataset setup was used to train the network weights. For all datasets, we used a horizontal random flip and random scaling. For the Cityscapes dataset, we used a learning rate of 0.01, set the batchsize to 8, and set the training iterations to 160 K. For the ADE20K and PASCAL VOC2012 datasets, we set the learning rate to 0.007, set the batchsize to 12, and set training iterations to 100 K.

4.2. Ablation Experiments with MPCNet

In this section, we designed ablation experiments on two modules of the MPCNet (Pooling Context Aggregation Module (PCAM) and Spatial Context Module (SCM)) for the Cityscape dataset. In the ablation experiments that follow, we set the training iterations to 100K for the convenience of the experiments.

4.2.1. Ablation Experiment for PCAM

To demonstrate the effectiveness of our proposed PCAM in MPCNet, we performed ablation experiments on its components. Table 1 shows our proposed PCAM ablation experiments on the ResNet101 backbone network for the Cityscapes dataset. We divided PCAM into two parts for ablation experiments—one containing only channel context $A_{feature}$ and spatial context $C_{feature}$ and one containing the context information between pixels $B_{feature}$. From Table 1, we can see that, regardless of whether it contained only channel context $A_{feature}$ and spatial context $C_{feature}$ or context information between pixels $B_{feature}$, the PA and mIOU of the segmentation pixels were greatly reduced, and the results were not as good as those of the three merges.

Table 1. PA and mIoU of our PCAM module for the Cityscapes dataset ($A_{feature}$, $B_{feature}$, and $C_{feature}$ denote the channel context, context information between pixels, and spatial context of our proposed PCA module, respectively). (Note that the bold indicates the best value for that column).

Method	$A_{feature}$	$B_{feature}$	$C_{feature}$	PA (%)	mIOU (%)
ResNet101				90.77	71.25
ResNet101	✓		✓	94.76	77.88
ResNet101		✓		94.27	77.56
ResNet101	✓	✓	✓	**95.81**	**78.05**

To further evaluate the advancement of our PCAM, we compared the results with PCAM using several classic context extraction modules: PPM [31], ASPP [33], and MMP [49]. To increase the fairness of the comparison data, we set consistent training parameters in the comparison experiments. Table 2 shows the results of the module comparison. From Table 2, we can see that our PCAM achieved 97.92% PA and 78.24% mIOU based on the same parameter settings, which outperformed those with the PPM, ASPP and MMP. The main reason is that our proposed PCAM aggregates the channel context, spatial context, and inter-pixel context of the low-resolution feature map, thereby making maximum use of the pixel information of the low-resolution feature map to capture more semantic context information.

Table 2. Comparison of PA and mIoU of our PCAM module with the other three modules (PPM, ASPP, MPM) for the Cityscapes dataset. (Note that the bold indicates the best value for that column).

Method	BaseNet	PPM	ASPP	MPM	PCAM	PA (%)	mIOU (%)
MPCNet	ResNet101	✓				94.56	76.43
MPCNet	ResNet101		✓			95.15	77.68
MPCNet	ResNet101			✓		95.01	77.21
MPCNet	ResNet101				✓	**97.92**	**78.24**

4.2.2. Ablation Experiment for SCM

In order to verify the validity of the SCM module, we conducted an experimental comparative analysis on the backbone network ResNet101 using SPM [49] with the same ability to capture spatial context information. Table 3 shows the experimental analysis for the Cityscapes dataset. From Table 3, we can see that the SCM module was superior to the SPM in the ResNet101 baseline network, and its performance reached 76.74% mIOU. The main reason is that our proposed SCM filters spatial information through different

pooling, saves spatial location information in different stages, and transfers spatial details to decoders through skip connections, thereby greatly restoring the pixel location to maintain the consistency of semantic and spatial details.

Table 3. Comparison of PA and mIoU of proposed PCAM module with the SPM modules for the Cityscapes dataset. (Note that the bold indicates the best value for that column).

Method	SPM	SCM	PA (%)	mIOU (%)
ResNet101			90.77	71.25
ResNet101	✓		94.76	75.58
ResNet101		✓	**96.21**	**76.74**

4.3. Segmentation Performances and Comparisons

In this subsection, to demonstrate the segmentation performance of our proposed MPCNet, numerical and visualization results were compared with ten segmentation methods for three image semantic segmentation datasets.

4.3.1. PASCAL VOC2012

To validate the effectiveness of our proposed MPCNet, we conducted a numerical experimental comparison with excellent semantic segmentation algorithms of recent years on the VOC2012 dataset. Table 4 shows comparison of the PA and mIOU for the PASCAL VOC2012 dataset with ten other methods. Since some of the methods did not run on this dataset, we ran the pixel precision (PA) of the FCN [15], PSPNet [31], DeepLab [50], Denseaspp [51], OCNet [52], and DeepLabV3+ [33] on the same device. The results of the OCRNet [53], OCNet [52], and ANN [54] were derived from SA-FFNet [55]. From Table 4, we can see that our method obtained 94.83% PA and 77.48% mIOU. Under ResNet101, our PA was 0.99% to 6.1% higher than other methods. Our MPCNet could achieve an mIOU of 77.48%, which was 1.06% higher than the SA-FFNet [55]. A comparison of different experimental values reveals that our MPCNet maintains good pixel accuracy.

Table 4. Comparison of our proposed MPCNet's PA and mIOU for the PASCAL VOC2012 dataset with ten other methods.

Method	BaseNet	PA (%)	mIOU (%)
FCN [15]	ResNet101	88.73	62.20
DeepLab [50]	ResNet101	92.84	78.51
PSPNet [31]	ResNet101	93.11	82.60
DeepLabv3+ [33]	ResNet101	93.78	80.57
Denseaspp [51]	ResNet101	93.68	75.27
ANN [54]	ResNet101	93.20	72.79
DANet [56]	ResNet101	93.38	80.40
OCRNet [53]	ResNet101	93.47	74.69
OCNet [52]	ResNet101	93.80	75.55
SA-FFNet [55]	ResNet101	93.84	76.42
MPCNet (ours)	ResNet101	94.83	77.48

4.3.2. Cityscapes

In this section, we conducted a comparative experiment for the Cityscapes dataset. Table 5 shows comparison of the PA and mIOU for the Cityscapes dataset. Considering the rigor of the experiment, we also retested the pixel accuracy for DeepLab [50], FCN [15], DeepLabV3+ [33], PSPNet [31], OCNet [52], Denseaspp [51], DANet [56], ANN [54], and OCRNet [53], as well as the mIOU of the FCN [15] for the Cityscapes dataset. From Table 5, we can see that our PA was 97.92%, which was 1.67% higher than other methods. The mIOU was 78.24%, which was 5.11% higher than other methods. Therefore, our MPCNet still has advantages regarding the PA and mIOU.

Table 5. Comparison of our proposed MPCNet's PA and mIOU for the Cityscapes dataset with ten other methods.

Method	BaseNet	PA (%)	mIOU (%)
FCN [15]	ResNet101	94.85	66.61
DeepLab[50]	ResNet101	95.78	79.30
PSPNet [31]	ResNet101	96.49	78.40
DeepLabv3+ [33]	ResNet101	96.66	79.55
Denseaspp [51]	ResNet101	95.85	80.60
ANN [54]	ResNet101	95.16	81.30
DANet [56]	ResNet101	95.45	81.50
OCRNet [53]	ResNet101	95.29	81.80
OCNet [52]	ResNet101	96.53	81.40
SA-FFNet [55]	ResNet101	96.25	73.13
MPCNet (ours)	ResNet101	97.92	78.24

4.3.3. ADE20K

To further validate our proposed MPCNet, we performed experiments on a larger ADE20K dataset. Table 6 shows the mIOU and PA of the MPCNet and ten other methods. It can be seen From Table 6 that the pixel accuracy of the MPCNet was 82.55%, and the mIOU was 38.04%. The results of these two methods still have certain advantages over the other ten methods. The ADE20K dataset has a large number of images and complex pixel types. Our proposed MPCNet extracts different pixel semantic contexts through the PCAM, uses the SCM to compensate for the missing spatial details, and uses codec mode to increase the capture of complex information. Our proposed MPCNet achieved different segmentation performance, so it is effective.

Table 6. Comparison of our proposed MPCNet's PA and mIOU for the ADE20K dataset with ten other methods.

Method	BaseNet	PA (%)	mIOU (%)
FCN [15]	ResNet101	76.32	29.47
SegNet [57]	ResNet101	68.59	21.63
DeepLab[50]	ResNet101	80.26	33.87
PSPNet [31]	ResNet101	81.56	41.68
DeepLabv3+ [33]	ResNet101	82.31	36.42
Denseaspp [51]	ResNet101	81.75	34.55
ANN [54]	ResNet101	81.37	45.24
DANet [56]	ResNet101	82.27	36.33
OCRNet [53]	ResNet101	81.88	45.28
OCNet [52]	ResNet101	82.10	45.04
MPCNet (ours)	ResNet101	82.55	38.04

4.4. Visual Comparison

To demonstrate the proposed visual advantage of the MPCNet, we compared three methods for the Cityscapes dataset in Figure 4, namely, PSPNet, OCNet, and DeepLabv3+. From Figure 4, it can be seen that small targets in a complex background, such as traffic lights, people in the distance, bicycles, etc., were all pixel categories that were difficult to segment. In contrast, our method had a better segmentation result than the other methods and could be successfully segmented. In addition, the loss of spatial detail information for pixels was successfully alleviated, such as the division and positioning of "human contour" and "overlapping vehicle" in line 5. From the perspective of the segmentation effect, our proposed MPCNet can provide the context information needed for segmentation and can accurately segment the image.

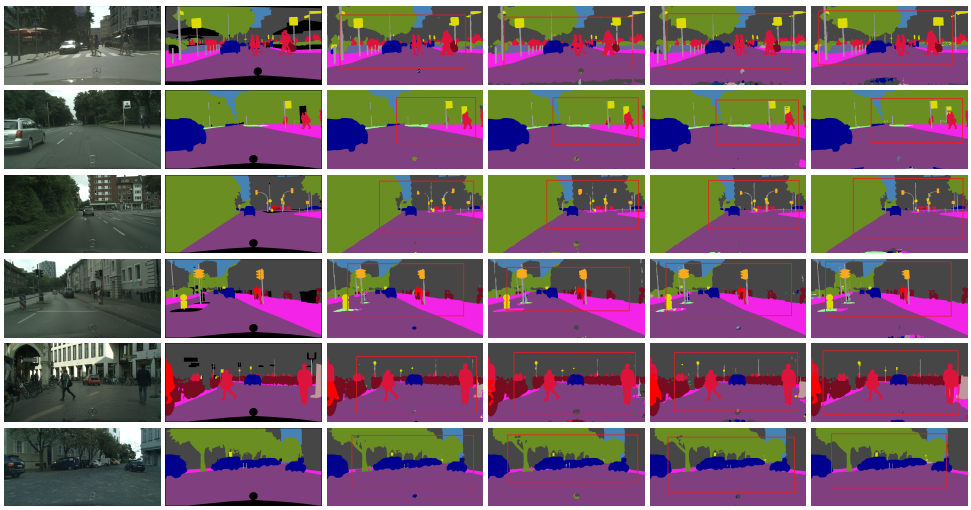

Figure 4. Comparison of the visual segmentation results of our proposed MPCNet with the other three methods for the Cityscapes dataset: (**a**) Image. (**b**) Ground Truth. (**c**) PSPNet [31]. (**d**) OCNet [52]. (**e**) DeepLabv3+ [33]. (**f**) Ours.

To further verify the validity of our method, we compared our proposed MPCNet with three methods for the VOC2012 dataset in Figure 5. From Figure 5, it can be seen that both vehicle and animal MPCNets could result in the semantics being classified correctly and the outline being clear. We propose that PCAM constructs a semantic context by capturing different contextual information and semantically dividing pixels. The SCM improves the spatial positioning ability of each semantic category and ensures that the outline of the category is clear. Therefore, from the perspective of visual analysis, our proposed MPCNet is effective in the application of semantics segmentation.

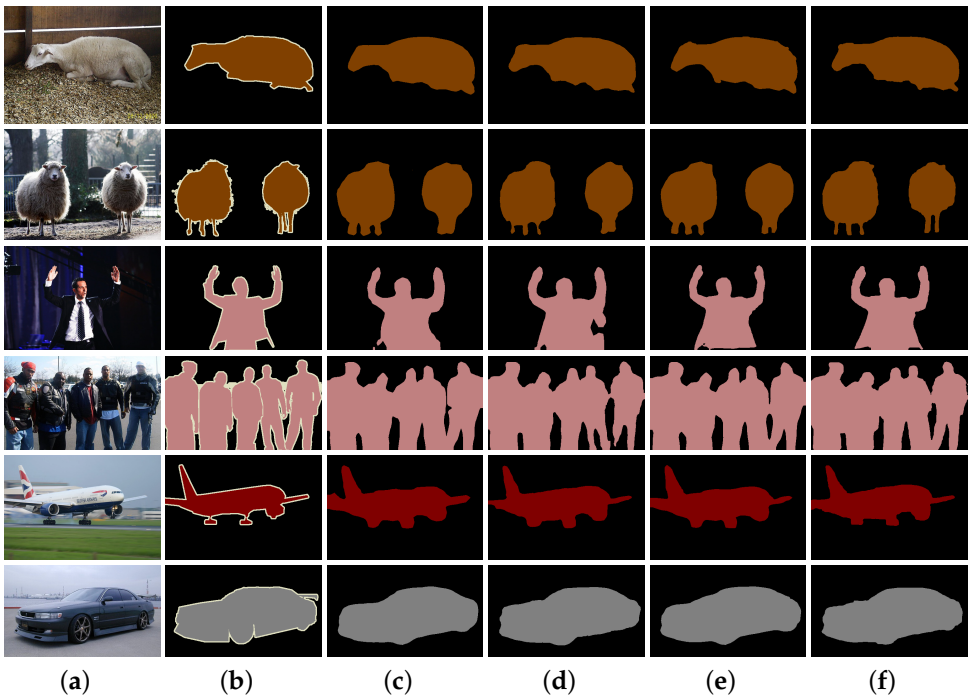

Figure 5. Comparison of the visual segmentation results of our proposed MPCNet with the other three methods for the PASCAL VOC dataset: (**a**) Image. (**b**) Ground Truth. (**c**) PSPNet [31]. (**d**) OCNet [52]. (**e**) DeepLabv3+ [33]. (**f**) Ours.

5. Conclusions

In this paper, we proposed a Multi-Pooling Context Network (MPCNet) for semantic segmentation. Specifically, our proposed PCAM aggregates the semantic context information in the high-level feature graph through three parts of feature information, increases the semantic exploitation of pixels in the low-resolution feature graph, and classifies different pixels in the image into semantic categories. Our proposed SCM captures the spatial contextual information of high-resolution features and passes it to the decoder in the form of a jump connection to enhance the spatial localization of semantic categories. The stable structure of the network using coding and decoding ensures that the contextual information is fully utilized, thus better improving the segmentation results. Experimental results show that our proposed MPCNet is effective.

Our method has initially alleviated the problem of insufficient context information capture in simple images, but the segmentation effect for complex backgrounds and multi-category pixel images still needs to be improved. For different complex background image processing, not only sufficient context information is needed, but also more attention should be paid to the relationships between pixels. For example, overlapping target objects, small target objects, and multi-shape target objects constitute the difficulties of semantic segmentation of complex images, and are also the focus of our research work in the future.

Author Contributions: Q.L.: Methodology, Writing—Original Draft Preparation, Experiments. Y.D.: Conceptualization, Methodology, Writing–Reviewing and Editing. Z.J.: Methodology, Investigation, Experiments. Y.P.: Methodology, Investigation, Experiments. B.Z.: Methodology, Investigation, Experiments. L.Z.: Methodology, Investigation, Experiments. Z.F.: Methodology, Investigation, Writing—Reviewing and Editing. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Natural Science Foundation of Henan under Grant 232300421023.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data generated and analyzed during this study are available from the corresponding author by request.

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

References

1. Li, P.; Liu, Y.; Cui, Z.; Yang, F.; Zhao, Y.; Lian, C.; Gao, C. Semantic graph attention with explicit anatomical association modeling for tooth segmentation from CBCT images. *IEEE Trans. Med. Imaging* **2022**, *41*, 3116–3127. [CrossRef] [PubMed]
2. Song, J.; Chen, X.; Zhu, Q.; Shi, F.; Xiang, D.; Chen, Z.; Fan, Y.; Pan, L.; Zhu, W. Global and local feature reconstruction for medical image segmentation. *IEEE Trans. Med. Imaging* **2022**, *41*, 2273–2284. [CrossRef] [PubMed]
3. Wang, Q.; Du, Y.; Fan, H.; Ma, C. Towards collaborative appearance and semantic adaptation for medical image segmentation. *Neurocomputing* **2022**, *491*, 633–643. [CrossRef]
4. Dai, Y.; Zheng, T.; Xue, C.; Zhou, L. SegMarsViT: Lightweight mars terrain segmentation network for autonomous driving in planetary exploration. *Remote. Sens.* **2022**, *14*, 6297. [CrossRef]
5. Li, X.; Zhao, Z.; Wang, Q. ABSSNet: Attention-based spatial segmentation network for traffic scene understanding. *IEEE Trans. Cybern.* **2021**, *52*, 9352–9362. [CrossRef]
6. Liu, Q.; Dong, Y.; Li, X. Multi-stage context refinement network for semantic segmentation. *Neurocomputing* **2023**, *535*, 53–63. [CrossRef]
7. Wang, H.; Chen, Y.; Cai, Y.; Chen, L.; Li, Y.; Sotelo, M.A.; Li, Z. SFNet-N: An improved SFNet algorithm for semantic segmentation of low-light autonomous driving road scenes. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 21405–21417. [CrossRef]
8. Liu, B.; Hu, J.; Bi, X.; Li, W.; Gao, X. PGNet: Positioning guidance network for semantic segmentation of very-high-resolution remote sensing images. *Remote. Sens.* **2022**, *14*, 4219. [CrossRef]
9. Wang, H.; Chen, X.; Zhang, T.; Xu, Z.; Li, J. CCTNet: Coupled CNN and transformer network for crop segmentation of remote sensing images. *Remote Sens.* **2022**, *14*, 1956. [CrossRef]
10. Nie, J.; Zheng, C.; Wang, C.; Zuo, Z.; Lv, X.; Yu, S.; Wei, Z. Scale–Relation joint decoupling network for remote sensing image semantic segmentation. *IEEE Trans. Geosci. Remote. Sens.* **2022**, *60*, 1–12. [CrossRef]

11. Dong, Y.; Jiang, Z.; Tao, F.; Fu, Z. Multiple spatial residual network for object detection. *Complex Intell. Syst.* **2022**, *9*, 1–16. [CrossRef]
12. Dong, Y.; Tan, W.; Tao, D.; Zheng, L.; Li, X. CartoonLossGAN: Learning surface and coloring of images for cartoonization. *IEEE Trans. Image Process.* **2021**, *31*, 485–498. [CrossRef] [PubMed]
13. Dong, Y.; Yang, H.; Pei, Y.; Shen, L.; Zheng, L.; Li, P. Compact interactive dual-branch network for real-time semantic segmentation. *Complex Intell. Syst.* **2023**, *2023*, 1–11. [CrossRef]
14. Zhou, Y.; Sun, X.; Zha, Z.J.; Zeng, W. Context-reinforced semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 4046–4055.
15. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
16. Zhang, H.; Zhang, H.; Wang, C.; Xie, J. Co-occurrent features in semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 548–557.
17. Zhang, H.; Dana, K.; Shi, J.; Zhang, Z.; Wang, X.; Tyagi, A.; Agrawal, A. Context encoding for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7151–7160.
18. Fu, J.; Liu, J.; Li, Y.; Bao, Y.; Yan, W.; Fang, Z.; Lu, H. Contextual deconvolution network for semantic segmentation. *Pattern Recognit.* **2020**, *101*, 107152. [CrossRef]
19. Geng, Q.; Zhang, H.; Qi, X.; Huang, G.; Yang, R.; Zhou, Z. Gated path selection network for semantic segmentation. *IEEE Trans. Image Process.* **2021**, *30*, 2436–2449. [CrossRef]
20. Chen, Y.; Jiang, W.; Wang, M.; Kang, M.; Weise, T.; Wang, X.; Tan, M.; Xu, L.; Li, X.; Zhang, C. LightFGCNet: A aightweight and focusing on global context information semantic segmentation network for remote sensing imagery. *Remote Sens.* **2022**, *14*, 6193. [CrossRef]
21. Ma, H.; Yang, H.; Huang, D. Boundary guided context aggregation for semantic segmentation. *arXiv* **2021**, arXiv:2110.14587.
22. Yang, Y.; Dong, J.; Wang, Y.; Yu, B.; Yang, Z. DMAU-Net: An Attention-Based Multiscale Max-Pooling Dense Network for the Semantic Segmentation in VHR Remote-Sensing Images. *Remote Sens.* **2023**, *15*, 1328. [CrossRef]
23. Hang, R.; Yang, P.; Zhou, F.; Liu, Q. Multiscale progressive segmentation network for high-resolution remote sensing imagery. *IEEE Trans. Geosci. Remote. Sens.* **2022**, *60*, 1–12. [CrossRef]
24. Lin, Z.; Sun, W.; Tang, B.; Li, J.; Yao, X.; Li, Y. Semantic segmentation network with multi-path structure, attention reweighting and multi-scale encoding. *Vis. Comput.* **2023**, *39*, 597–608. [CrossRef]
25. De Souza Brito, A. Combining max-pooling and wavelet pooling strategies for semantic image segmentation. *Expert Syst. Appl.* **2021**, *183*, 115403. [CrossRef]
26. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
27. Zhang, D.; Zhang, H.; Tang, J.; Wang, M.; Hua, X.; Sun, Q. Feature pyramid transformer. In *Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020*; Proceedings, Part XXVIII 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 323–339.
28. Hu, P.; Perazzi, F.; Heilbron, F.C.; Wang, O.; Lin, Z.; Saenko, K.; Sclaroff, S. Real-time semantic segmentation with fast attention. *IEEE Robot. Autom. Lett.* **2020**, *6*, 263–270. [CrossRef]
29. Xu, Z.; Zhang, W.; Zhang, T.; Li, J. HRCNet: High-resolution context extraction network for semantic segmentation of remote sensing images. *Remote Sens.* **2020**, *13*, 71. [CrossRef]
30. Li, Z.; Sun, Y.; Zhang, L.; Tang, J. CTNet: Context-based tandem network for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 9904–9917. [CrossRef] [PubMed]
31. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
32. Liu, W.; Rabinovich, A.; Berg, A.C. Parsenet: Looking wider to see better. *arXiv* **2015**, arXiv:1506.04579.
33. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 March 2018; pp. 801–818.
34. Hong, Y.; Pan, H.; Sun, W.; Jia, Y. Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes. *arXiv* **2021**, arXiv:2101.06085.
35. Li, X.; Zhao, H.; Han, L.; Tong, Y.; Yang, K. Gff: Gated fully fusion for semantic segmentation. *arXiv* **2019**, arXiv:1904.01803.
36. Kim, T.; Kim, J.; Kim, D. SpaceMeshLab: Spatial context memoization and meshgrid atrous convolution consensus for semantic segmentation. In Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AL, USA, 19–22 September 2021; pp. 2259–2263.
37. Ding, H.; Jiang, X.; Shuai, B.; Liu, A.Q.; Wang, G. Semantic segmentation with context encoding and multi-path decoding. *IEEE Trans. Image Process.* **2020**, *29*, 3520–3533. [CrossRef]
38. Yu, C.; Gao, C.; Wang, J.; Yu, G.; Shen, C.; Sang, N. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *Int. J. Comput. Vis.* **2021**, *129*, 3051–3068. [CrossRef]

39. Hao, S.; Zhou, Y.; Guo, Y.; Hong, R.; Cheng, J.; Wang, M. Real-Time semantic segmentation via spatial-detail guided context propagation. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *2022*, 1–12. [CrossRef]
40. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
41. Li, X.; You, A.; Zhu, Z.; Zhao, H.; Yang, M.; Yang, K.; Tan, S.; Tong, Y. Semantic flow for fast and accurate scene parsing. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 775–793.
42. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
43. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 3213–3223.
44. Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; Torralba, A. Scene parsing through ade20k dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 633–641.
45. Minaee, S.; Boykov, Y.; Porikli, F.; Plaza, A.; Kehtarnavaz, N.; Terzopoulos, D. Image segmentation using deep learning: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 3523–3542. [CrossRef]
46. Dong, Y.; Shen, L.; Pei, Y.; Yang, H.; Li, X. Field-matching attention network for object detection. *Neurocomputing* **2023**, *535*, 123–133. [CrossRef]
47. Garcia-Garcia, A.; Orts-Escolano, S.; Oprea, S.; Villena-Martinez, V.; Garcia-Rodriguez, J. A review on deep learning techniques applied to semantic segmentation. *arXiv* **2017**, arXiv:1704.06857.
48. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
49. Hou, Q.; Zhang, L.; Cheng, M.M.; Feng, J. Strip Pooling: Rethinking spatial pooling for scene parsing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 18–20 June 2020; pp. 4003–4012.
50. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
51. Yang, M.; Yu, K.; Zhang, C.; Li, Z.; Yang, K. Denseaspp for semantic segmentation in street scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 3684–3692.
52. Yuan, Y.; Huang, L.; Guo, J.; Zhang, C.; Chen, X.; Wang, J. Ocnet: Object context network for scene parsing. *arXiv* **2018**, arXiv:1809.00916.
53. Yuan, Y.; Chen, X.; Wang, J. Object-contextual representations for semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 173–190.
54. Zhu, Z.; Xu, M.; Bai, S.; Huang, T.; Bai, X. Asymmetric non-local neural networks for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 593–602.
55. Zhou, Z.; Zhou, Y.; Wang, D.; Mu, J.; Zhou, H. Self-attention feature fusion network for semantic segmentation. *Neurocomputing* **2021**, *453*, 50–59. [CrossRef]
56. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 3146–3154.
57. 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]