applied sciences

MDPI

Article

Image Semantic Segmentation Fusion of Edge Detection and AFF Attention Mechanism

Yijie Jiao, Xiaohua Wang *, Wenjie Wang and Shuang Li

College of Electronic Information, Xi'an University of Polytechnic, Xi'an 710600, China
* Correspondence: wangxiaohua@xpu.edu.cn

Abstract: Deep learning has been widely used in various fields because of its accuracy and efficiency. At present, the improvement of image semantic segmentation accuracy has become the area of most concern. In terms of increasing accuracy, improved semantic segmentation models have attracted more attention. In this paper, a hybrid model is proposed to solve the problems of edge splitting and small objects disappearing from complex scene images. The hybrid model consists of three parts: (1) an improved HED network, (2) an improved PSP-Net, (3) an AFF attention mechanism. Continuous edges can be obtained by combining the improved HED network with an improved PSP-Net. The AFF attention mechanism can improve the segmentation effect of small target objects by enhancing its response recognition ability for specific semantic scenes. The experiments were carried out on Cityspaces, SIFT Flow, NYU-V2 and CamVid datasets, and the experimental results show that the segmentation accuracy of our method is improved by 2% for small target objects, and by 3% for scenes with complex object edges.

Keywords: semantic segmentation; edge segment; AFF; small object; complex scenes

check for updates

Citation: Jiao, Y.; Wang, X.; Wang, W.; Li, S. Image Semantic Segmentation Fusion of Edge Detection and AFF Attention Mechanism. *Appl. Sci.* **2022**, *12*, 11248. https://doi.org/10.3390/app122111248

Academic Editor: Agostino Forestiero

Received: 13 September 2022
Accepted: 4 November 2022
Published: 6 November 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Image semantic segmentation is a basic task in image understanding. It has been widely used in the fields of image scene understanding [1], visual tracking [2,3], autonomous driving [4,5], robot navigation [6,7], remote perception [8,9], industrial testing [10], video scene analysis [11], and medical image analysis [12,13]. The existing semantic segmentation models have good effect when the object is large or the object is obviously different from the background. However, when there are many objects or the detected objects are small in the scene, the segmentation results feature the disappearance of small objects, incomplete edge segmentation, edge adhesion and so on.

Researchers have conducted a lot of work to address these questions. On the basis of the convolutional neural network (CNN), a full convolutional network (FCN) [14] is proposed to segment the different resolution images. On this basis, a jump connection method was created to screen, adjust and merge features of different scales to obtain accurate semantic features. Meanwhile, an expanded convolutional layer was used in the FCN to increase the receptive field in the feature extraction network [15]. Based on the FCN, an encoder–decoder network structure was designed; the encoder is used to reduce the resolution of the feature map, and the decoder is used to improve the sampling of the feature map to restore the resolution [16]. DeepLabV1 extends its receptive field by extending convolution and uses fully connected conditional random fields at the end of the network; segmentation accuracy is improved while keeping the size of the feature map unchanged [17]. On the basis of DeepLabV1, DeeplabV2 [18] has the structure of avoiding the spatial pyramid pool (ASPP). It uses dilated convolution with different expansion rates to integrate multi-scale information and increase the multi-scale receptive field. An ICNet network processes images hierarchically, extracts features from images with different resolutions, and then uses cascading feature fusion (CFF) to improve the accuracy of semantic segmentation [19]. By introducing full-scale skip connections, Unet3+ network [20] integrates low-level details and high-level semantics in full-scale feature mapping; its long join mode helps to recover the information loss caused by downsampling. The DeeplabV3+ [21] network adopts the cross-layer fusion method, which combines shallow detail features with deep abstract features to improve the segmentation accuracy of high-resolution images. PSP-Net [22] has introduced the pyramid pooling module (PPM); PPM aggregates contextual information from different regions and obtains global information. Based on the deep Res-Net [23] network, an effective optimization strategy was developed to achieve high-precision segmentation. All the above methods improve the accuracy of semantic segmentation, but this "improvement" depends too much on the complexity of the scene. The above method is suitable for object segmentation in simple scenes, but for complex scenes, there will still be discontinuous edge segmentation, edge splitting and small objects disappearing.

Aimed at the problem of complex scene image segmentation, a model is proposed in this paper. Based on the improved PSP-Net semantic segmentation framework, the model introduces improved edge detection network holistically-nested edge detection (HED) and an attentional feature fusion (AFF) [24] attention mechanism. In this model, the improved HED network extracts the edge information of objects, and the feature fusion module in AFF fuses the edge information with improved PSP-Net semantic information. At the same time, AFF has the semantic association structure between its channels, so the response recognition ability of specific semantic events is enhanced. Thus, small objects are adequately identified. There are several important works in this paper which make a contribution to image semantic segmentation:

1. We use the MobileNetV3 [25] network instead of the original PSP-Net network encoder, and obtain an improved PSP-Net network for image semantic segmentation. MobileNetV3 can extract deep semantic information, shallow contour information and middle-layer hybrid features. The features are incorporated into the decoder by means of hops.
2. Based on the original HED model, we proposed an improved HED network. We use the output of five edges, respectively, during the convolution layer. The last full-link layer in HED is removed to obtain a complete convolutional network, and expanded convolution is introduced to increase the receptive field for feature extraction.
3. To better fuse the semantic features and edge features, we designed a fusion module by introducing AFF into the improved PSP-Net and the improved HED network. AFF fuses the edge information with the deep semantic information; the model can enhance the attention correspondence between the edge feature information and the semantic information, and the accuracy of image segmentation in complex scenes is improved.

2. Method

2.1. Hybrid Model of Improved PSP-Net Network Combined with Improved HED

In order to accurately detect and segment image edges, improved HED is integrated into the improved PSP-Net network. The network structure can obtain shallow edge information and deep semantic information; its structure is shown in Figure 1.

The network consists of three major parts: a semantic segmentation network, an edge detection network, and a semantic and edge feature fusion module.

I. Semantic segmentation module: composed of the improved PSP-Net model, the module integrates multi-scale and multi-level features to extract image semantic information.

II. Edge detection network: composed of the improved HED model, this part detects and learns edge features of images through feature extraction and fusion, so more detailed edge information can be obtained.

III. Semantic and edge feature fusion module: the edge information is mapped to semantic features to realize the fusion of semantic information and edge information.

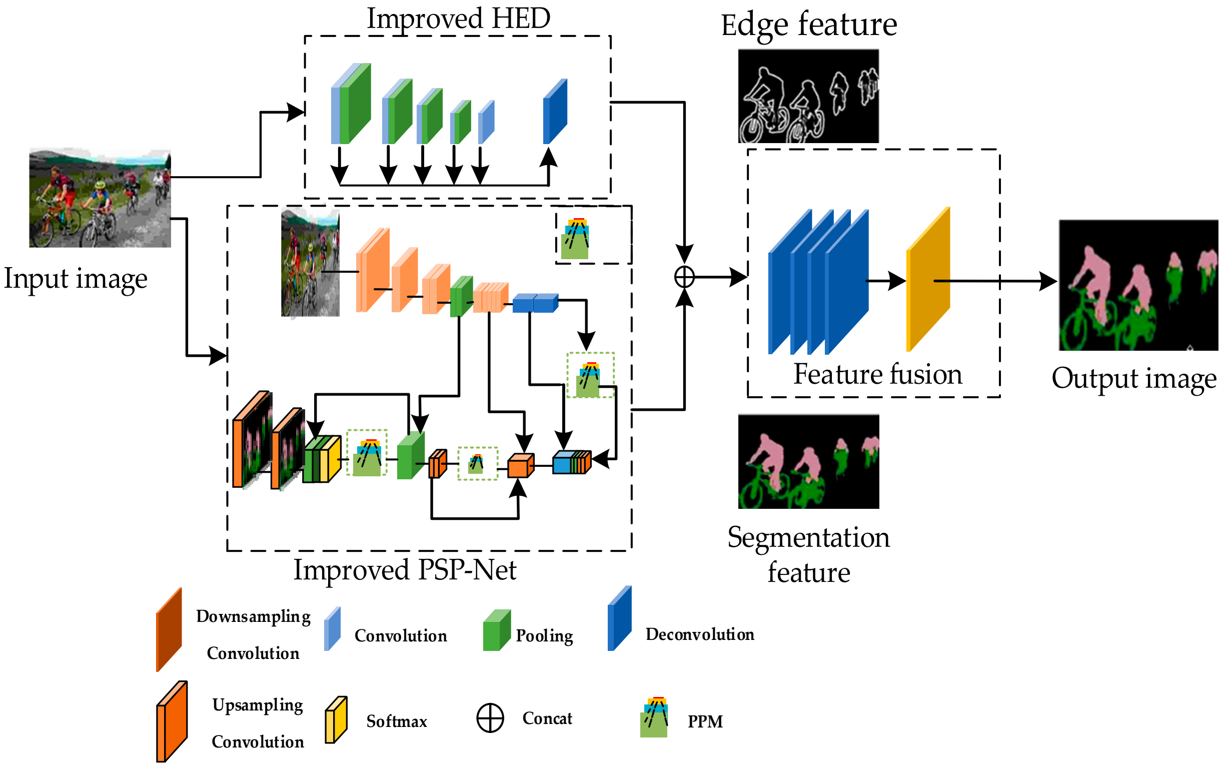

Figure 1. Improved PSP-Net combined with improved HED network.

2.1.1. Improved PSP-Net Network

The improved PSP-Net network is obtained by using the network MobileNetV3 instead of the original encoder in PSP-Net network. MobileNetV3 can extract deep semantic information, shallow contour information and mixed features in the middle layer. Features are incorporated into the decoder by means of a jump connection. The improved PSP-Net framework is shown in Figure 2.

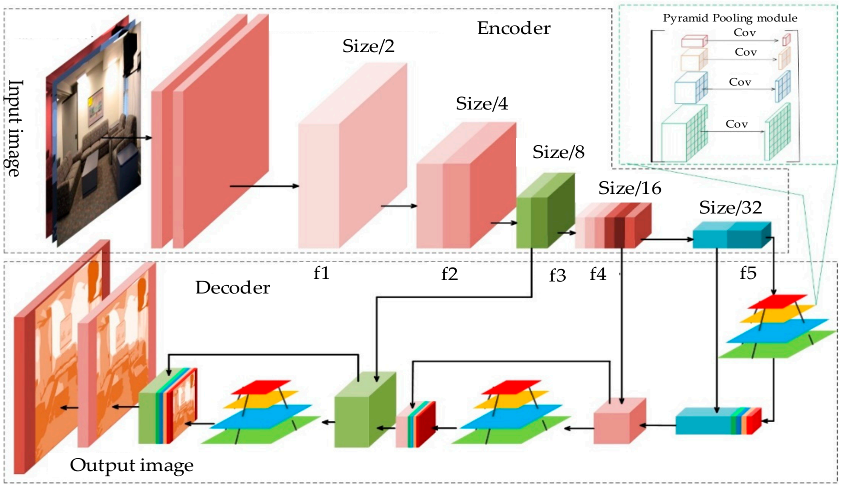

Figure 2. Improved PSP-Net network.

Encoder Network

Our improved PSP-Net network uses the encoder network MobileNetV3 to extract multi-level feature information. The structure is shown in Figure 3.

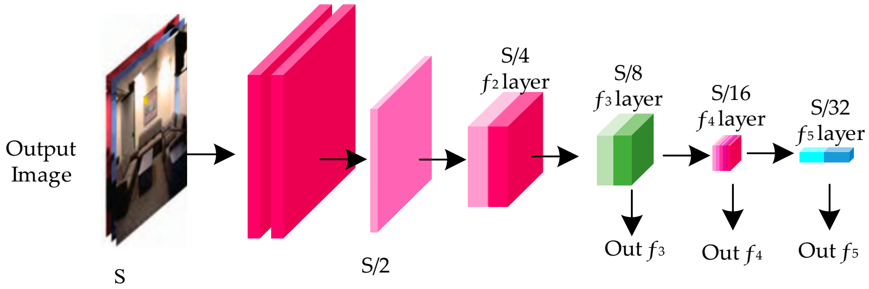

Figure 3. Encoder network.

The individual layers in Figure 3 are Benck modules of different scale sizes. The pixel size of the input image is S; f_1, f_2, f_3, f_4 and f_5 layers represent feature map layers whose size is 1/2, 1/4, 1/8, 1/16, and 1/32, respectively, of the original image. In CNN, with the increase of network layers, the features of interest points gradually shift from local features to global features [26]. Thus, the intermediate network layer is the best transition layer between local contour features and global semantic features. In order to prevent the feature map of the input graph from falling into local optima, layers f_1 and f_2 are discarded. Finally, layers f_3, f_4 and f_5 are selected as the multilevel output feature map layers of the encoder, the output of the encoder is the input of the decoder, layers f_3, f_4 and f_5 are connected to the decoding network by jumping.

Decoder Network

The main task of the decoder network is to adjust the size of feature map by upsampling operation, and then use PPM to fuse the feature map in the encoder network. The decoding network of the method in this paper is shown in Figure 4.

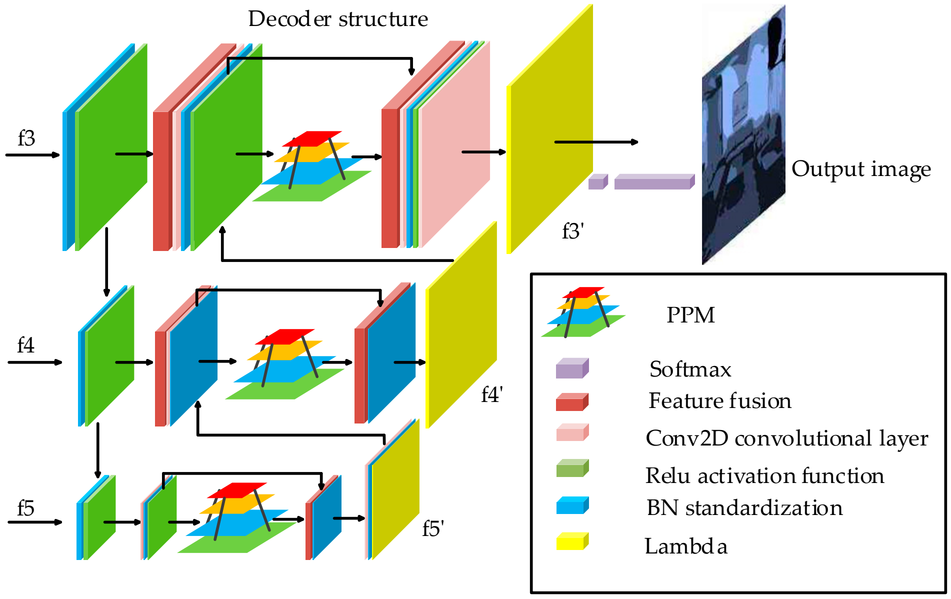

Figure 4. Decoder network structure.

In Figure 4, the feature maps input to the decoder network have different scales, so it is necessary to adjust the feature maps of different scales before performing feature fusion. The f_5 layer feature map can be input to PPM to obtain a new feature map f_5 layer feature and new features are fused through concatenating [27]. Then, the upsampling operation is used to adjust the feature map to ensure its size is consistent at different scales. Then, 1 × 1 convolution is used to align the channel numbers of the two feature maps, the fused feature map is upsampled to obtain f'_5; f'_5 is the corresponding feature layer of the f_5 feature map. Similarly, the f'_5 feature map and the f_4 feature image are channel-aligned, fused into a new feature map, and then the subsequent PPM operations are repeated to obtain the f'_4 feature map. In the same way, the f'_4 and f_3 feature maps are fused. Finally, high-precision semantic segmentation can be achieved by multi-level feature fusion.

2.1.2. Improved HED Network Model

In this paper, HED [28,29] is improved to obtain the edge detection network. The improved model can effectively extract features of different scales, and learn features of other scales while doing so. The improved model integrates deep semantic features and shallow contour edge features. The accuracy of contour information in semantic segmentation results is improved. In the model, the output of each side is upsampled by the bilinear difference algorithm, restored to the size of the original image, and then the output result is fused through the fusion layer.

The improved HED network is shown in Figure 5. The improved HED network takes the VGG-16 model as the main frame, it induces the output of 5 sides, respectively, while carrying out convolution. The last full-link layer in HED is removed and a full convolutional network is obtained. At the same time, dilated convolution is introduced to increase the receptive field of feature extraction. After upsampling at each level, loss and sigmoid layers are added to calculate the loss of each layer, respectively. After the fusion of features at each level, the loss of the sigmoid layer is calculated.

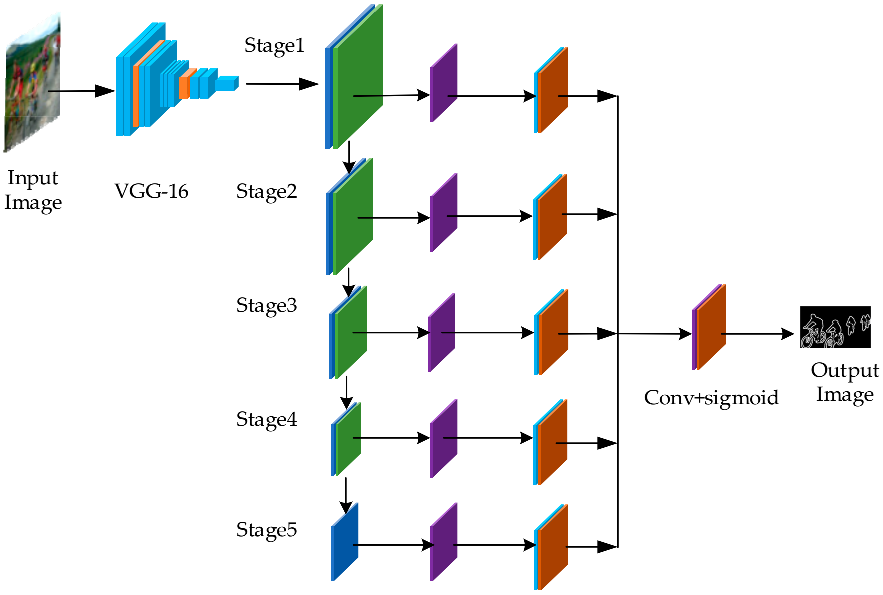

Figure 5. Improved HED network.

Edge detection using the improved HED network can be divided into five different stages. Each stage consists of a convolution layer, pooling layer and dilated convolution at different scales. That is, five different side branches are distinguished, and features are obtained by means of convolution, pooling and void convolution in each stage. The cavity rate of the cavity convolution is 2, and the convolution kernel is 3 × 3.

Each stage parameters of the improved HED network are shown in Table 1.

Table 1. Parameters of each stage in the improved HED network.

Network Layer	Convolution Kernel Size
Stage1(Pooling1 + Dilated Conv + Sigmoid)	[3, 3, 64]
Stage2(pooling2 + Dilated Conv + Sigmoid)	[3, 3, 128]
Stage3(pooling3 + Dilated Conv + Sigmoid)	[3, 3, 256]
Stage4(pooling4 + Dilated Conv + Sigmoid)	[3, 3, 512]
Stage5(pooling5 + Dilated Conv + Sigmoid)	[3, 3, 512]

In the improved HED network, under the same computational conditions, dilated convolution [30,31] provides a larger receptive field without loss of resolution, which can better capture multi-scale context information and the size of the output feature image will not be changed. Figure 6 is the schematic of dilated convolution.

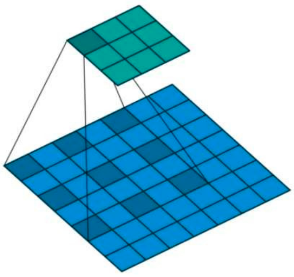

Figure 6. Dilated convolution schematic.

The larger the receptive field value of the neuron in the dilated convolution is, the more pixel range the convolution kernel can contact. This means that more global semantic level features are extracted when extracting features. On the other hand, if the receptive field value is smaller, the features contained in it tend to be local and detailed. Therefore, the value of the receptive field can be used to roughly judge the abstraction level of each layer.

2.1.3. Semantic and Edge Feature Fusion Network

The improved HED and the improved PSP acquire edge information and semantic information, respectively, and these two kinds of feature information need to be effectively fused. In this paper, canonical correlation analysis (CCA) [32] is used for feature fusion structure, as shown in Figure 7.

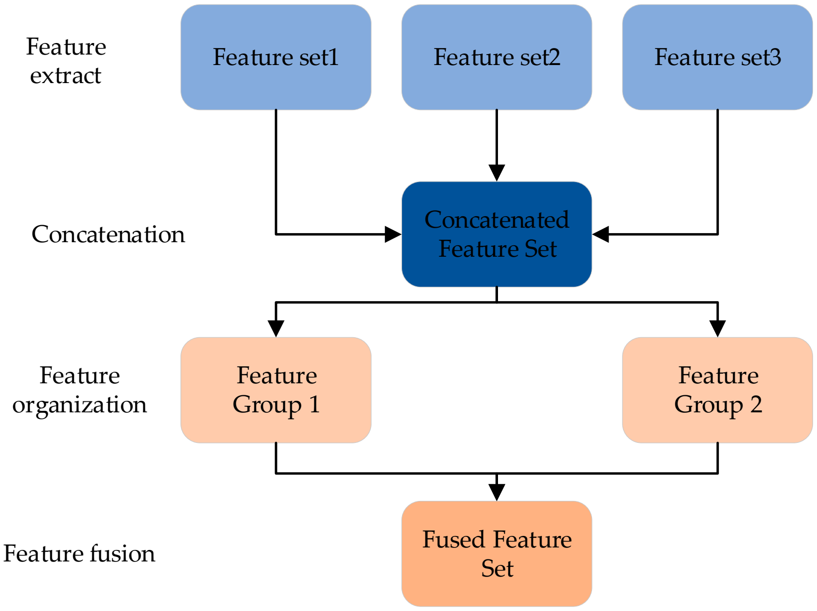

Figure 7. CCA feature fusion network structure.

As shown in Figure 7, the features extracted from semantic segmentation and edge detection are pre-integrated and recombined in a concatenated manner, and the features of the same nature are initially classified. The main purpose of this part is to make the classification feature information of the target object in the image correspond to the edge feature information of the object, and finally achieve feature fusion. In this way, CCA feature fusion structure enriches the expression of feature information and avoids the influence of noise caused by the fusion of two different features.

2.1.4. Experiment of Hybrid Model

Dataset Selection

In this experiment, the NYU-V2 indoor environment and CamVid outdoor environment datasets are used. The NYU-V2 dataset is a famous public indoor image segmentation and depth image dataset. This dataset includes 1449 fine-scale indoor images of 464 scenes in 3 cities and 407,024 images without fine labels. The CamVid dataset is a street view dataset from the perspective of driving a car. The dataset contains 701 images, the training set includes 367 images, the validation set contains 101 images, the test set contains 233 photos, and the image resolution is 960 × 720. The dataset provides ground truth tags that associate each pixel with one of the 32 semantic classes.

Training

Distributed training is used to train our improved PSP-Net network combined with improved HED, which is showed in Figure 1. First, the improved HED network is trained independently. Set batch_size to 5, 50 rounds, every 10 rounds of a cycle, learning rate set to 0.001, momentum coefficient set to 0.9. After the training, the model is transplanted to the semantic segmentation network through transfer learning, and the final training of the whole model is completed. Experiments were conducted on NYU-V2 dataset and CamVid dataset, respectively. Model training LOSS values are shown in Figures 8 and 9, respectively. As can be seen from the Figures 8 and 9, with the progress of training, the loss value steadily decreases to the stable value, which indicates the convergence of the model.

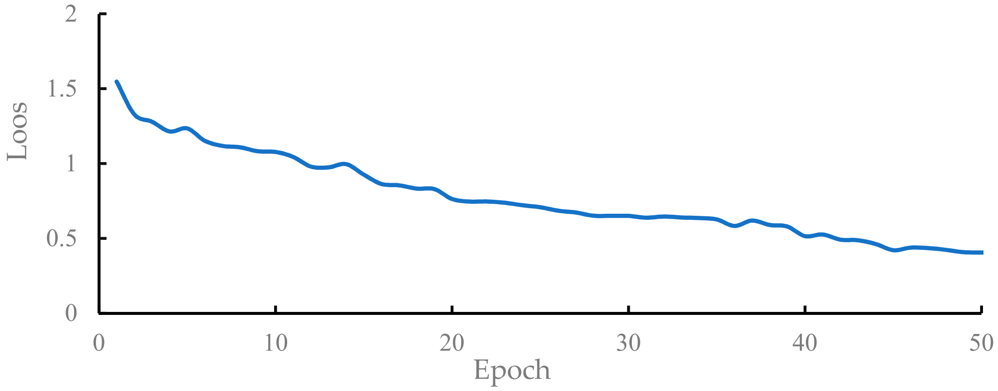

Figure 8. The LOSS changes based on the NYU-V2 dataset.

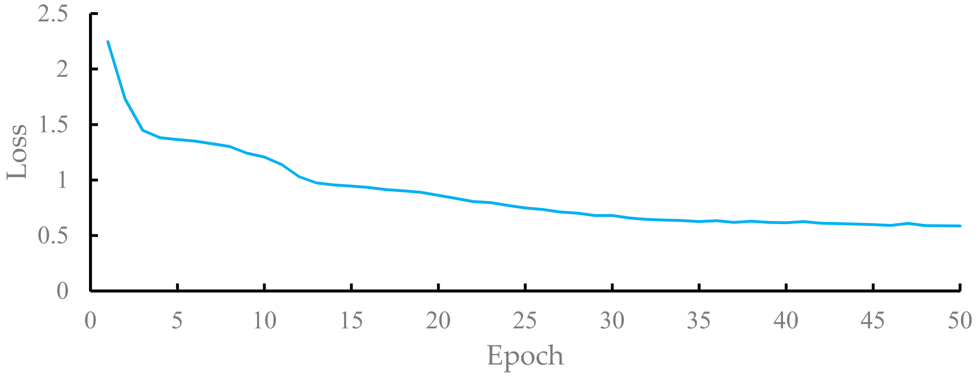

Figure 9. The LOSS changes based on the CamVid dataset.

Experiments Based on the NYU-V2 Dataset

Our semantic and edge feature fusion network, Seg-Net [33], PSP-Net, and Deeplabv3+, used the NYU-V2 dataset for object segmentation experiment. The experimental results are shown in Table 2.

Table 2. Accuracy and speed of image semantic segmentation network on NYU-V2 dataset.

Method	Input	BaseNet	MIOU(Val)	MIOU(Test)	FPS
Seg-Net	640 × 480	VGG-16	50.7	58.1	14
PSP-Net	640 × 480	ResNet50	65.8	66.5	80.5
Deeplabv3+	640 × 480	Xception	70.1	72.5	79.2
Our semantic and edge	640 × 480	MobileNet3	70.1	73.3	52.65

Compared with the Seg-Net network, our semantic and edge feature fusion network improves the segmentation accuracy by 15.2%, and the processing speed is also greatly improved. Compared with PSP-Net, the segmentation accuracy is improved by 6.8% because our semantic and edge feature fusion network uses improved HED so that the model can better obtain edge features. Compared with the Deeplabv3+ network, our semantic and edge feature fusion network slightly improves the segmentation accuracy. Table 3 shows the segmentation accuracy and MIOU for each subclass in the NYU-V2 dataset for the four networks mentioned above.

Table 3. MIOU comparison of different image segmentation networks.

Name	Seg-Net	PSP-Net	DeepLabV3+	Our Semantic and Edge
Wall	67.2	59.4	62.4	65.2
Floor	73.8	66.2	70.3	70.1
Closet	49.6	39.1	43.9	43.8
Bed	51.4	28	40.1	37.8
Chair	41.2	26.6	36	31.2
Sofa	46.2	19.8	33.6	36.2
Table	31.3	20.5	24.2	25.6
Door	22.6	13.9	18.1	18.2
Window	32.8	23.4	26.4	25.2
Bookshelf	32.1	26.9	28	35.2
Painting	49.6	41.2	43.8	50.3
Counter	42.9	21.1	29.5	41.5
Shutters	50	40.1	26.1	51.2
Table	11.1	5.6	6.8	11.2
Shelf	5.9	2.3	4	6.2
Curtain	34	14.9	19.1	37.2
Comb table	25.3	7.6	14.9	25.4
Pillow	27.9	12.8	18.7	28.3
Mirror	10.8	0.8	1.3	19.5
Carpet	16.7	10.6	11	19.4
Clothing	11.4	2	4.6	13.2
Ceiling	54.6	40.9	46.2	58.2
Book	28	21	25.8	31.2
Refrigerator	16.7	4.9	7.5	20.4
TV set	28.5	12.7	25.3	30.1
Paper	20.5	10.7	15	21.3
Box	7.2	2.5	3.6	9.7
Whiteboard	42.9	11	40.6	45.3
People	24.5	1.2	11	30.1
Bedside table	34.9	4.8	14.5	32.2
Toilet	59.4	25	46.8	60.6
Sink	41.7	12.8	35.1	43.4
Lamp	3.05	12.1	22.8	39.1
Tub	30.6	14	8.5	24.7
Handbag	4.3	1.6	3.4	5.2
Person	14.3	4.7	9.6	18.8
Bath towel	30.6	10.4	13.5	32.6
Towel	23.3	6.6	10.2	24.2
Other	21.2	20.6	4.8	10.2

It can be seen from Table 3 that our semantic and edge feature fusion network is better than others. As shown in Figure 10.

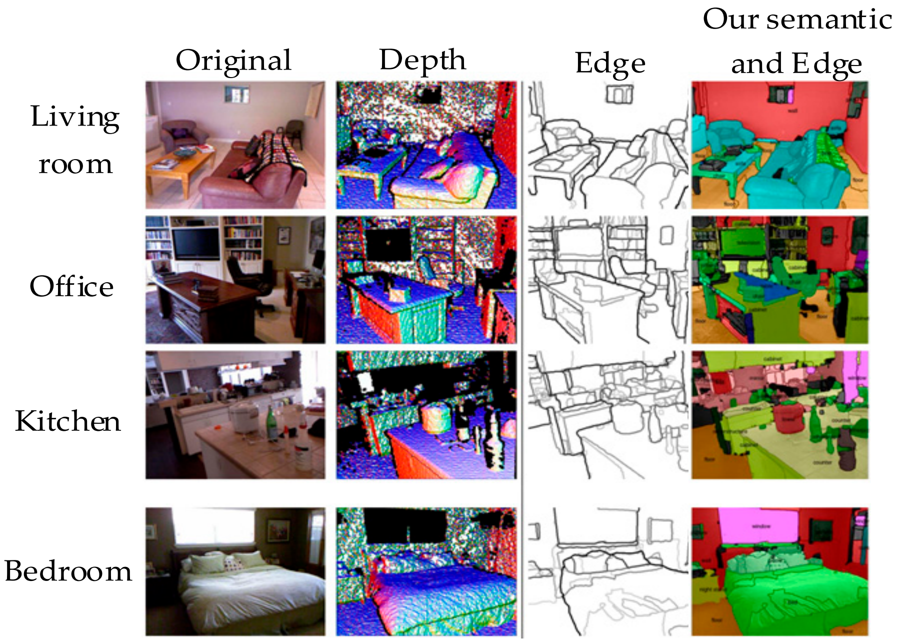

Figure 10. Segmentation results on the NYU-V2 dataset figure.

The results in Figure 10 show that our semantic and edge feature fusion network is more effective, and that small objects are clear. In the second line, the "Office" scene, the objects are many and small, which poses a significant challenge to image semantic segmentation. However, the improved model is used to segment the small objects, and it can be seen that the practical effect is better. In the first and fourth rows of the scene, we can see that the edges of large objects are continuous, and the edges of the "sofa" and "bed" in the figure are more continuous.

Experiments Based on CamVid Dataset

Our semantic and edge feature fusion network, together with FCN-8s, Seg-Net, PSP-Net, and Deeplabv3+, were used on the CamVid dataset for the object segmentation experiment. The experimental results are shown in Table 4.

Table 4. Accuracy and speed of image semantic segmentation network on CamVid dataset.

Method	**Input**	**BaseNet**	**MIOU(Val)**	**MIOU(Test)**	**FPS**
Seg-Net	2048 × 1024	VGG-16	67.5	66.2	30.1
PSP-Net	2048 × 1024	ResNet50	69.0	68.3	72.3
Deeplabv3+	2048 × 1024	Xception	75.6	73.5	74.8
FCN-8S	2048 × 1024	VGG-16	62.2	61.8	8
Our semantic and edge	2048 × 1024	MobileNet3	76.1	74.9	51.2

Compared with FCN-8s, the segmentation accuracy of our semantic and edge feature fusion network is improved by 13.1%, and the processing speed is also greatly improved. Compared with Seg-Net, PSP-Net, and DeeplabV3+, the segmentation accuracy is improved by 8.7%, 6.6%, and 1.4%, respectively.

The segmentation results based on the CamVid dataset are shown in Figure 11.

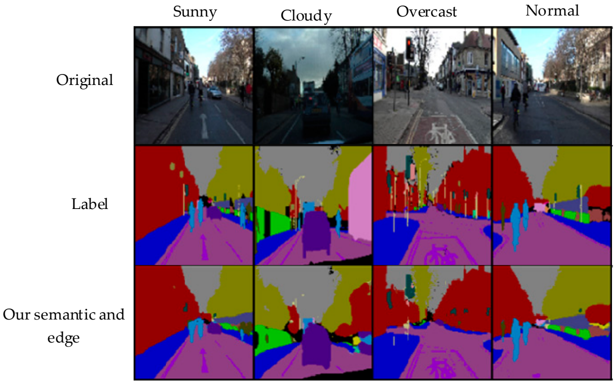

Figure 11. Segmentation results on the CamVid dataset.

The results in Figure 11 show that our semantic and edge feature fusion network can segment objects such as "street" and "building" better, and the edge is more continuous. This is because the edge features are fully utilized in the improved HED network. As can be seen from the semantic segmentation results of different weather conditions, our method can accurately segment "car", "building" and other objects. However, the segmentation effect on small objects needs to be improved. Therefore, on the base of our semantic and edge feature fusion network, the AFF attention mechanism is introduced to solve the problem of imprecision of small objects.

2.2. AFF Attention Mechanism

2.2.1. Architecture of Attention Mechanism

In the deep convolutional neural network (DCN) [34] based on the attention mechanism, the function of the attention mechanism is to filter information and effectively allocate resources for the neural network.

The use of attention mechanisms in deep convolutional neural networks can be divided into six types. Figure 12 shows the architectural diagram of the six different attention mechanism networks.

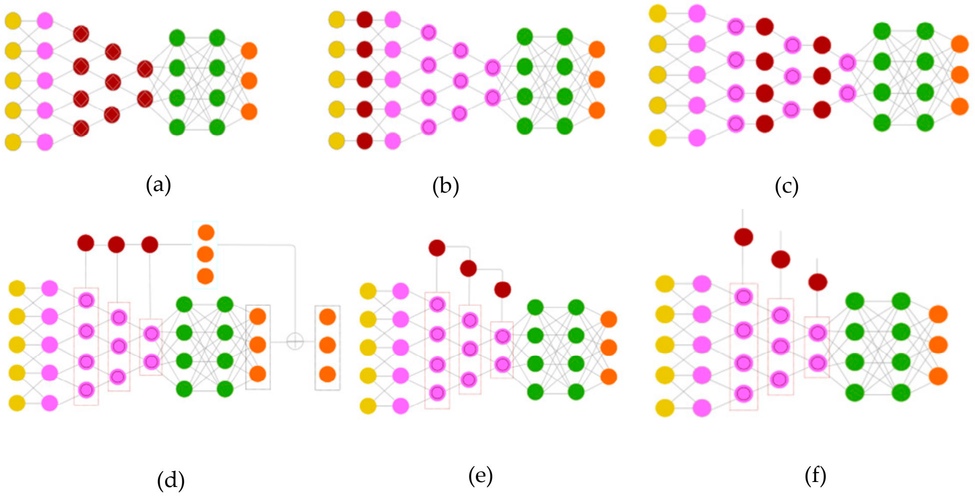

Figure 12. The architecture of the attention mechanism network. In (**a**) DCN attention pooling, the attention mechanism replaces the classic CNN [35] pooling mechanism. In (**b**) DCN attention input, the attention module is the filtering mask of the input data. In (**c**) DCN attention layer, the attention mechanism is applied between the convolutional layers. In (**d**) DCN attention prediction, the attention mechanism assists the model in the prediction process. In (**e**) DCN residual attention mechanism, the mechanism extracts information from the feature map and provides the remaining input connections at the next layer. In (**f**) DCN attention output, the attention mechanism captures significant activations of features of other architectures, or other instances of the same architecture.

2.2.2. AFF Attention Mechanism

The core of the AFF attention mechanism is to propose a multi-scale channel attention module (MS-CAM) and the application method of the AFF attention mechanism. We will improve our model based on AFF. The main framework of AFF attention mechanism is shown in Figure 13. The MS-CAM structure is shown in Figure 14.

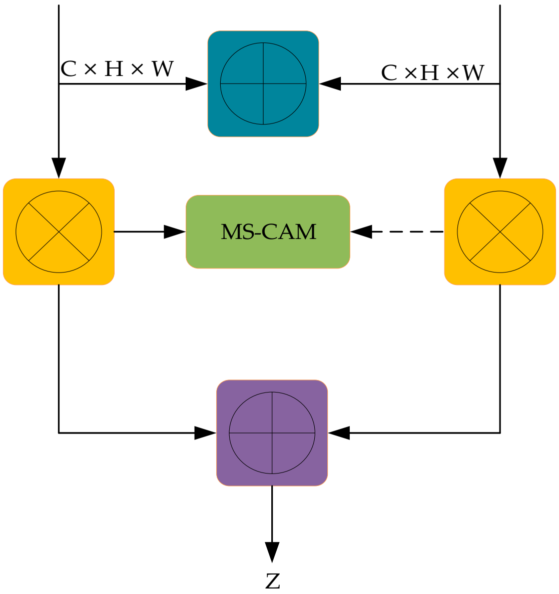

Figure 13. Structure of the AFF attention mechanism.

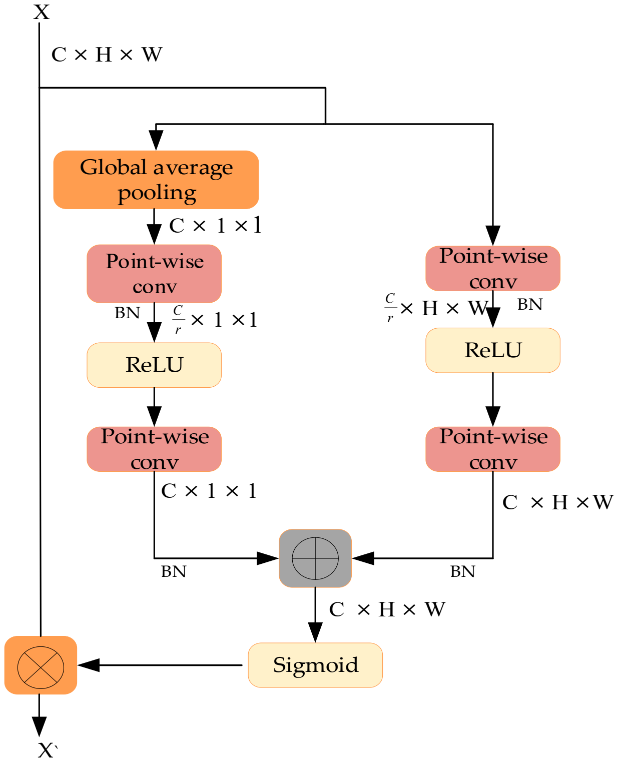

Figure 14. Structure of the MS-CAM.

As can be seen from Figure 14, the function of the AFF attention mechanism is to fuse features on the base of MS-CAM. The calculation method of the AFF attention mechanism is as follows:

I. Set the output as $X \in \mathrm{R}^{\mathrm{C} \times \mathrm{H} \times \mathrm{W}}$ in the multi-scale channel of the attention structure. The three different pathways in the MS-CAM structure acquire different channel characteristic information. The first path is the acquisition of global features, and the global feature $G(X)$ is obtained using a global average pooling method and two point-wise convolutions (point-wise conv) methods. The second way is the acquisition of local features. The local feature $L(X)$ is obtained by convolution point-by-point with the original features of the third way. The formula is as follows:

$$X' = X \otimes M(X) = X \otimes \partial(L(X) \oplus g(X)) \tag{1}$$

In Equation (1), $M(X) \in \mathrm{R}^{\mathrm{C} \times \mathrm{H} \times \mathrm{W}}$ is the weight of the multi-scale channel. $\oplus$ means broadcast addition. $\otimes$ means element-wise multiplication.

II. In the AFF attention mechanism structure, the outputs are set as $X \in \mathrm{R}^{\mathrm{C} \times \mathrm{H} \times \mathrm{W}}$ and $Y \in \mathrm{R}^{\mathrm{C} \times \mathrm{H} \times \mathrm{W}}$. They are different layers of the feature image. If the structure of MS-CAM is M, the final AFF attention mechanism is:

$$Z = M(X \cup Y) \otimes X + (1 - M(X \cup Y)) \otimes Y \tag{2}$$

In Equation (2), $Z \in \mathrm{R}^{\mathrm{C} \times \mathrm{H} \times \mathrm{W}}$ is the fused feature. $\cup$ is the initial feature integral. $M(X \cup Y)$ is between 0 and 1, it enables the network to perform soft selection or weighted between X and Y.

3. Proposed Model

3.1. Model Architecture

In this paper, the improved HED is introduced into the improved PSP-Net network, and we obtain a hybrid network. Moreover, the AFF attention mechanism is introduced in four positions of the hybrid network, and the final model of this paper is obtained. The final model is shown in Figure 15.

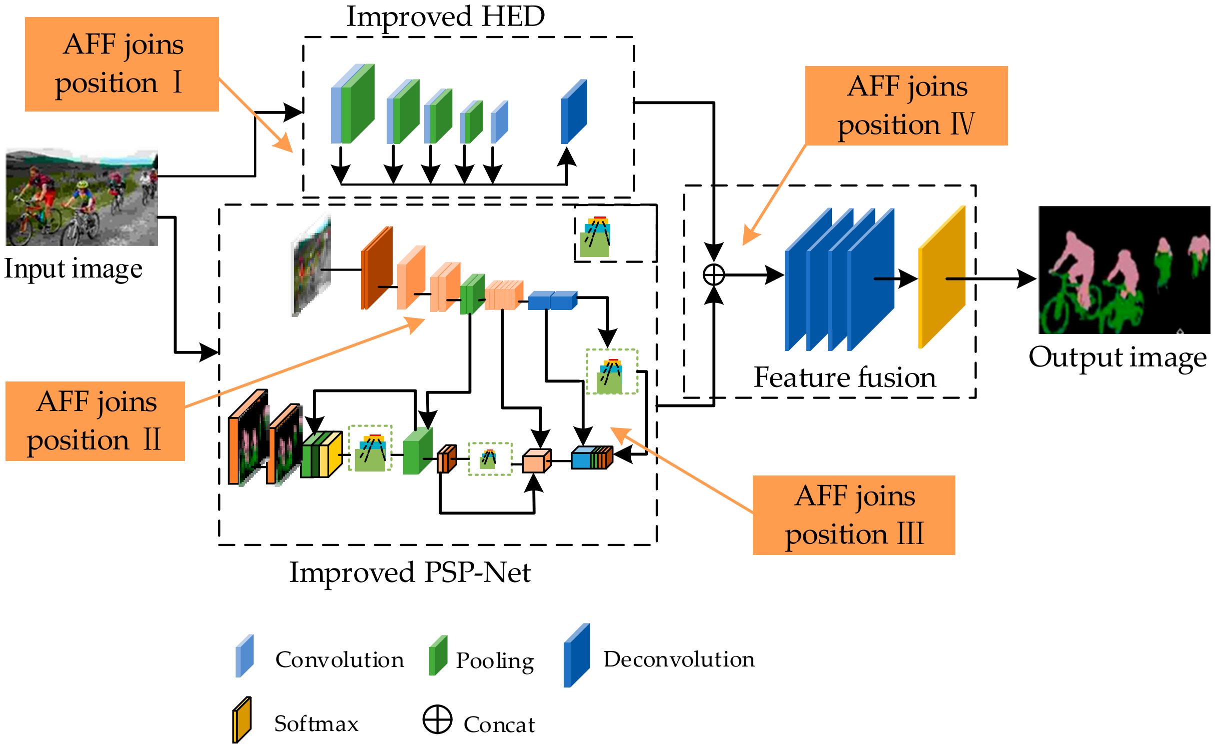

Figure 15. A semantic segmentation network architecture combining edge detection and AFF attention mechanism.

In Figure 15, The I AFF is added to the improved HED to accelerate the feature fusion of five bypass structures with different output scales. It also controls the unity of the channels. The II AFF joins to the encoder part of the semantic segmentation network; this causes the pixels associated with the feature and their channels to converge more closely to the location of attention. The III AFF is introduced in the semantic segmentation network decoder section; AFF attention mechanism is parallel with PPM, which can accelerate the learning speed and propagation depth of features. The IV AFF is introduced in feature fusion part of edge feature and the semantic information; it can enhance the attentional correspondence between edge feature information and semantic information and the segmentation accuracy of the model is guaranteed effectively.

3.2. AFF Attention Mechanism

The way the attention mechanism is used in the model is not simply incorporated or concatenated into the system. In this paper, we need to verify experimentally where the AFF attention mechanism is introduced to produce the best results. We need to do experiments to determine the location of AFF attention mechanism. Appropriate location of AFF can achieve the best feature fusion and make the model segmentation the most efficient. In this paper, the AFF attention mechanism is used in the following four ways shown in Figure 16.

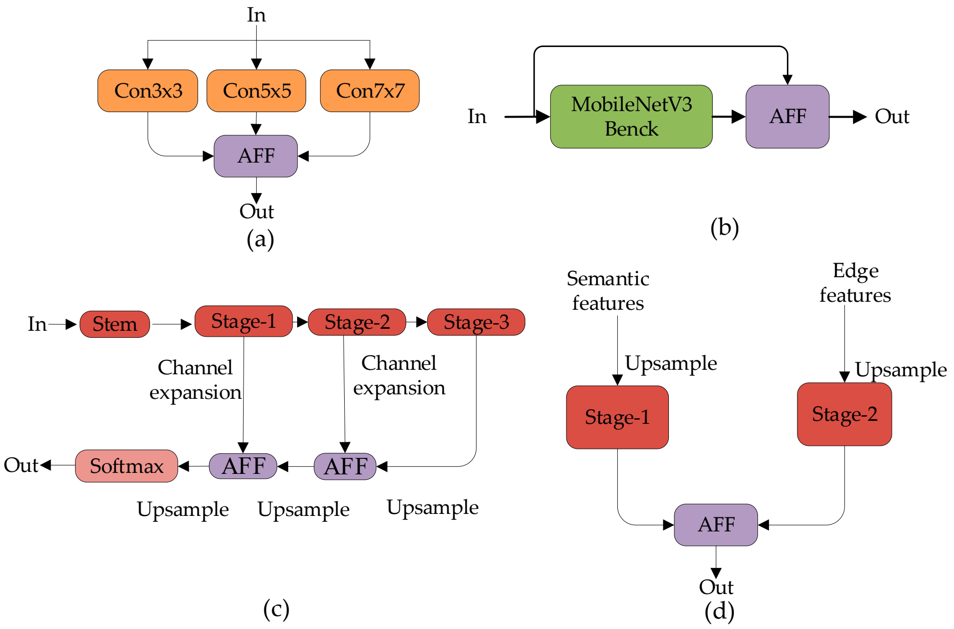

Figure 16. Four different joining positions. (**a**) corresponds to position I in Figure 15, adding AFF to the improved HED network; (**b**) corresponds to position II in Figure 15, adding AFF to the encoder MobileNetV3 network; (**c**) corresponds to position III in Figure 15, adding AFF to the PSP-Net decoder; (**d**) corresponds to position IV in Figure 15, adding AFF to the fusion of semantic features and edge features.

3.3. Loss Function

We design the loss function according to the positions of the AFF attention mechanism. The loss function of the model is the sum of the loss weights of the improved HED network and the improved PSP-Net network. That is, $L_{\mathrm{sfe}} = L_{\mathrm{sf}} + L_{\mathrm{se}}$. L_{sfe} is the loss function of the improved HED network and the improved PSP-Net network, L_{sf} is the loss function of the improved HED network, L_{se} is the loss function of the improved PSP-Net network.

$$L_{\mathrm{sf}} = \mathrm{argmin}(L_s(W,w) + L_f(W,w,h)) \tag{3}$$

$$L_{\mathrm{se}} = (1-\gamma)(L_s + L_f) \tag{4}$$

The loss function of the I AFF is:

$$L_{\mathrm{sf}-A} = \theta_{1c}[\mathrm{argmin}(L_s(W,w) + L_f(W,w,h))] + \theta_{2c}L_A(m,m^*) \tag{5}$$

In Equation (5), L_{sf-A} is the loss function when the AFF is introduced into the improved HED network, L_s is the loss function of the five side output networks in the improved HED network. L_f is the loss function of edge information fusion. W is the set of parameters in the network, w is the set of parameters of the output layer, and h is expressed as the weight of the ensemble. $\theta_{1\mathrm{c}}$ and $\theta_{2\mathrm{c}}$ are the parameters of the equilibrium loss function $L_{\mathrm{sf}-A}$, which are both set to 0.5 in this paper.

The loss function of the II AFF is:

$$\begin{gathered} L_{E1-A} = -\theta 1a\{\sum_{j=1}^{l}(1-y^j)\ln[(1-p_8^j)(1-p_{16}^j)(1-p_{32}^j)] \\ +\sum_{j=1}^{l}y^j\ln(p_8^jp_{16}^jp_{32}^j) + \beta\sum||\omega||_2\} - \theta_{2a}L_A(m,m^*) \end{gathered} \tag{6}$$

In Equation (6), L_{E1-A} is the loss function when the AFF is added to the encoder side. It is the sum of the loss function of encoder and the loss function of AFF. The encoder side uses the MobileNetV3 network to extract multi-level feature information. The feature maps of 1/8, 1/16 and 1/32 sizes of the coding layer are selected for training, so the probability feature maps of class j of the 1/8, 1/16 and 1/32 layers are defined as p_8^j, p_{16}^j and p_{32}^j, respectively, which are the expected output of class *j*. $L_A(m, m^*)$ is the loss function of the AFF attention mechanism, a per-pixel cross-entropy loss function based on the mask. *m*, *m** are the masks generated by the attention mechanism and their corresponding mask labels. θ_{1b} and θ_{2b} are balanced loss functions with the value 1.

The loss function of the III AFF is

$$L_{E2-A} = \theta_{1b}\{-\sum_{j=1}^{l}[y^j\ln\overset{\cap}{k^j} + (1-y^j)\ln(1-\overset{\cap}{k^j}) + \beta\sum||\omega||^2\} + \theta_{2b}L_A(m,m^*) \tag{7}$$

In Equation (7), L_{E2-A} is the loss function when AFF is added to the PSP-Net decoder side. It is the sum of the loss function on the decoder side and the loss function of AFF. $\overset{\cap}{k^j}$ is the jth probabilistic feature layer of the network output layer. θ_{1b} and θ_{2b} are the parameters of the equilibrium loss function L_{E2-A}. In this paper, both are set to 1.

The loss function of the IV AFF is

$$L_{sfe-A} = \theta_{1d}[\gamma(L_{E1} + L_{E2} + L_{Dice}) + (1-\gamma)(L_s + L_f)] + \theta_{2d}L_A(m,m^*) \tag{8}$$

In Equation (8), $L_{\mathrm{sfe}-A}$ is the loss function after adding AFF to the semantic and edge feature fusion part. It is the sum of the loss function of the edge detection network, the loss function of the semantic segmentation network, and the loss function of AFF. L_{E1}, L_{E2}, and L_{Dice} are the loss functions of encoder networks, decoders, and semantic segmentation networks, respectively. $\theta_{1\mathrm{d}}$ and $\theta_{2\mathrm{d}}$ are the parameters of the balance loss function $L_{\mathrm{sfe}-A}$, which are both set to 0.5 in this paper.

4. Experiment and Results

4.1. Implementations Details

In order to verify adding the AFF attention mechanism to the position which could achieve the best semantic segmentation effect, we conducted experiments on the four locations in Figure 15 on the Cityscapes and SIFT Flow datasets. The operating platform consisted of a Windows 10 operating system, NVIDIA RTX2080Ti GPU (Nvidia, Santa Clara, CA, USA), and TensorFlow-GPU framework (Google, Mountain View, CA, USA).

Through experiments, we determined where AFF should be introduced into the hybrid module. From Section 3.2 of the paper, we know that the AFF attention mechanism can be introduced in the encoder, decoder, edge detection network, and the fusion of semantic and edge features. We determined the optimal location of the attention mechanism through experimentation. The evaluation criteria are MIOU. From Section 3.1, we know there are four different scenarios:

I. AFF is added to the improved HED network.
II. AFF is added to the MobileNetV3 encoder.
III. AFF is added to the PSP-Net decoder.
IV. AFF is added to the feature fusion part of the improved HED network and the improved PSP-Net.

4.2. Dataset

The Cityscapes dataset is used in autonomous driving. It has street view images of more than 50 different cities; there are 8 large categories and 30 small categories of images in the dataset. The user can optimize the dataset according to the needs of the design, and 19 subcategories are used in this paper. The dataset contains the original image and the corresponding classification label image. These label images can be used in combination with the original image to better complete the tasks of image instance segmentation, semantic segmentation, and object detection. In this experiment, image enhancement is used to improve the generalization computing ability of the model and avoid data over fitting. The image enhancement of the Cityscapes dataset is shown in Figure 17.

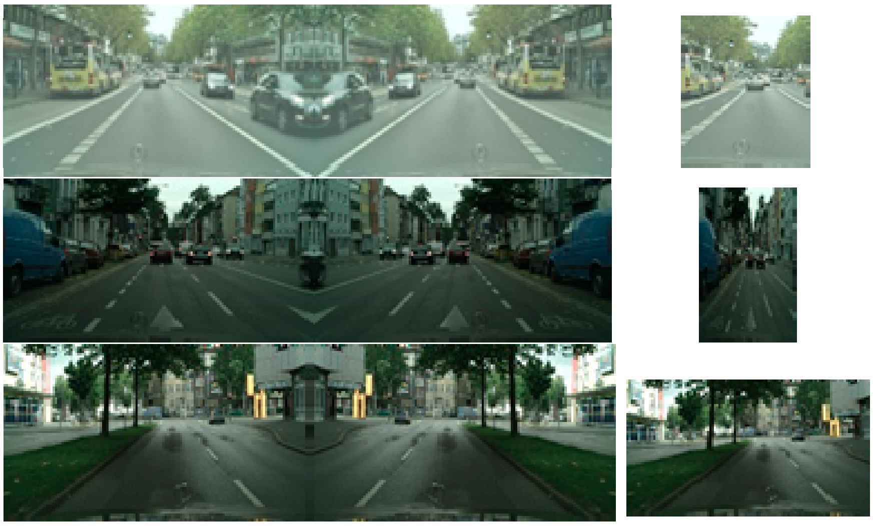

Figure 17. Image enhancement of the Cityscapes dataset.

4.3. Experiment Results and Analysis

The model was trained and verified according to the experimental scheme designed in Section 4.1. The experimental results are shown in Tables 5 and 6.

Table 5. Feature fusion MIOU comparison of different semantic segmentation networks on Cityscapes dataset.

Experiment	I	II	III	IV
MIOU	82.56	80.37	81.89	83.38

Table 6. Feature fusion MIOU comparison of different semantic segmentation networks on SIFT Flow dataset.

Experiment	I	II	III	IV
MIOU	82.71	79.94	80.63	83.68

As shown in Tables 5 and 6, the semantic segmentation of AFF is added to the feature fusion part of the improved HED network and the improved PSP-Net is better than others. The main reasons are as follows: The PPM in PSP-Net architecture can extract features very well, and AFF can accelerate the speed of feature learning and the depth of feature propagation. When AFF is introduced into the encoder of MobileNetV3, the AFF attention mechanism will make the structure of the model complicated. Unnecessary hyperparameters will be generated, which makes model training slow and wastes resources. When AFF is introduced into improved HED network, the feature fusion of five bypass structures with different output scales is accelerated, and channels are unified.

In Figure 18, Experiment I and Experiment IV were analyzed by image visualization on the Cityscapes dataset. Experiment IV is the proposed model in this paper.

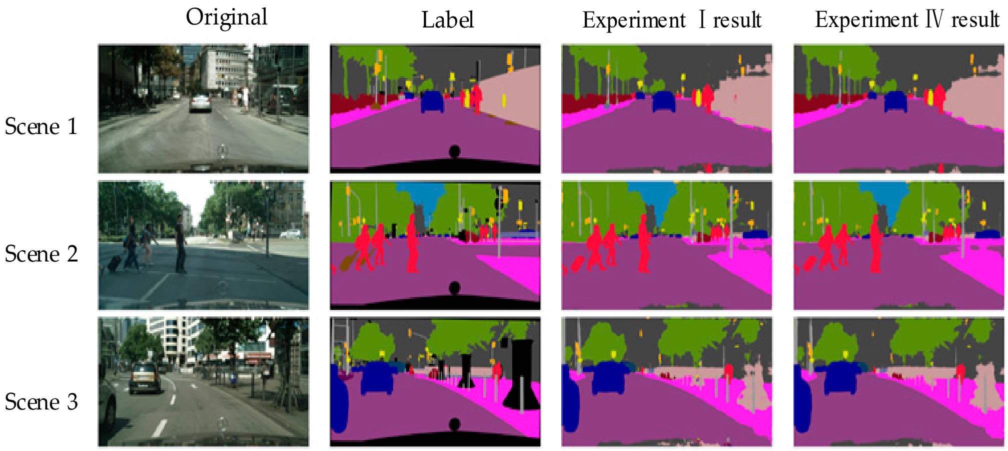

Figure 18. Images of Experiment I and Experiment IV on the Cityscapes dataset.

In Figure 19, Experiment I and Experiment IV were analyzed by image visualization on the SIFT Flow dataset.

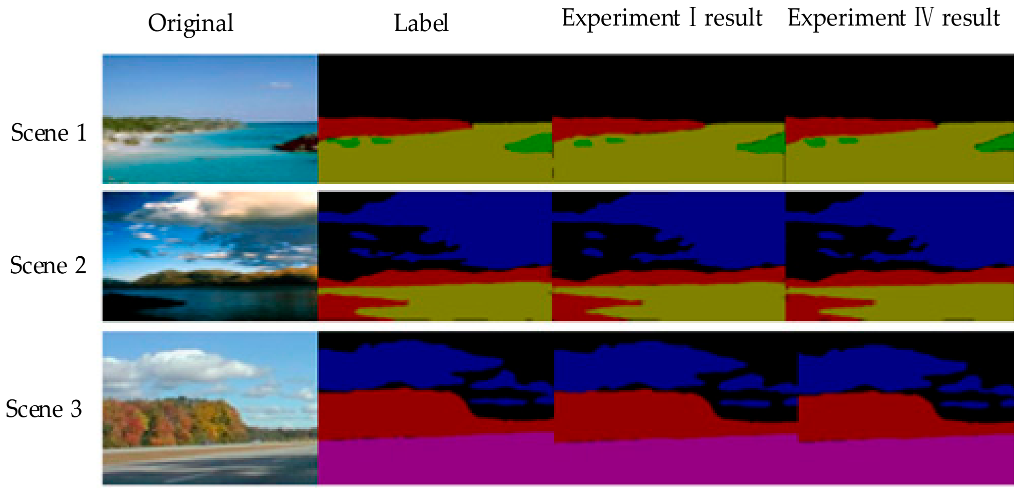

Figure 19. Images of Experiment I and Experiment IV on the SIFT Flow dataset.

In Figures 18 and 19, Experiment IV, the segmentation edges of "car", "vegetation", "road", "light pole" and "land", "ocean", "cloud" are continuous, and the small objects "light pole" and "cloud" are finely segmented. In Experiment I, the edge is well segmented, but there are still some problems in the segmentation of some small objects.

From the experimental results shown in Tables 5 and 6, it can be seen that in Experiment IV, adding the AFF attention mechanism to the feature fusion part of the improved HED network and the improved PSP-Net has better effect than the other three positions.

So, the proposed model in this paper is shown in Figure 15 with the AFF joining position IV, the loss function is Equation (8) in Section 3.3.

4.4. Performance Comparison

Finally, the proposed model is compared with the other semantic segmentation algorithms; the MIOU of each class of objects is calculated. Table 7 lists the results. From the experimental data shown in Table 7, it can be seen that the segmentation accuracy of our proposed model is better than that of other methods.

Table 7. MIOU.

Name	DeepLabV3+	PSP-Net	Our Proposed Model
Average	80.15	81.23	83.38
Road	97.52	94.45	97.21
Sidewalk	82.73	81.69	86.32
Building	92.25	91.12	80.56
Wall	51.21	50.61	82.45
Fence	53.64	50.83	80.26
Pole	58.27	57.52	81.38
Traffic light	61.19	60.93	73.92
Traffic sign	70.46	68.34	92.64
Vegetation	90.52	90.15	63.11
Terrain	62.21	61.87	94.17
Sky	95.73	93.67	75.67
ocean	94.62	89.79	76.82
Person	72.95	70.11	80.76
Rider	60.34	58.29	93.82
Car	94.58	93.12	61.93
Truck	63.72	61.24	75.98
Bus	78.16	73.35	54.86
Train	53.64	53.57	55.24
Motorcycle	52.72	50.61	52.35
Bicycle	71.91	69.48	83.21

Compared to other semantic segmentation algorithms, no matter the comprehensive performance or each small classification, our method has a good segmentation effect. For small target objects, it can be found that the segmentation accuracy is improved by about 2%. For the scene with more complex edges, the segmentation accuracy is improved by about 3%.

5. Conclusions

Deep learning has been widely used in the field of image understanding. With the development of deep learning, the accuracy of image semantic segmentation has also been greatly improved. In this paper, we proposed a hybrid semantic segmentation model, which combines the improved HED network with the improved PSP-Net, and adds the AFF attention mechanism. In the improved PSP-Net network, we use the MobileNetv3 network to instead be the encoder of the original PSP-Net network. In the improved HED network, the last full link in the HED network is removed to obtain a complete convolutional network, and extended convolution is introduced to increase the acceptance domain of feature extraction. The improved HED network can effectively extract features at different scales, and learn features at other scales while doing so. The AFF attention mechanism can improve the response recognition ability for specific scenes and enhance the segmentation accuracy of small target objects. There are four places where the AFF attention mechanism can be inserted. In order to achieve the best segmentation effect, we carried out experiments on four positions, respectively. The experimental results show that adding AFF to the feature fusion part of the improved PSP-Net and the improved HED network can obtain the best effect. This hybrid model has experimented on the Cityscapes dataset and the SIFT Flow dataset. The experimental results prove that the model can not only improve the segmentation accuracy of complex scene images and small objects, but also improve the convergence speed and fitting degree of the model when training images in complex scenes.

Author Contributions: Conceptualization, X.W. and W.W.; formal analysis, X.W.; investigation, W.W.; resources, X.W.; data curation, Y.J.; writing—original draft preparation, Y.J.; writing—review and editing, S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by [Natural Science Foundation of China] grant number [51905405]; [Key Research and Development plan of Shaanxi province China] grant number [2019ZDLGY01-08]; [Innovation Capability Support Program of Shaanxi] grant number [2021TD-29]; [Key Research and Development program of Shaanxi province] grant number [2022GY-276] And The APC was funded by the above four projects.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Experiment data can obtain at https://github.com/jiaoyijie/data-for-semantic-segmentation.

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

References

1. Borji, A.; Sihite, D.N.; Itti, L. Salient Object Detection: A Benchmark. *IEEE Trans. Image Process.* **2015**, *24*, 5706–5722. [CrossRef] [PubMed]
2. Xu, Y.; Osep, A.; Ban, Y.; Horaud, R.; Leal-Taixé, L.; Alameda-Pineda, X. How to Train Your Deep Multi-Object Tracker. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 6786–6795.
3. Weinland, D.; Ronfard, R.; Boyer, E. A survey of vision-based methods for action representation, segmentation and recognition. *Comput. Vis. Image Underst.* **2011**, *115*, 224. [CrossRef]
4. Li, B.; Liu, S.; Xu, W.; Qiu, W. Real-time object detection and semantic segmentation for autonomous driving. In Proceedings of the Volume 10608, MIPPR 2017: Automatic Target Recognition and Navigation, Xiangyang, China, 19 February 2018.
5. Tseng, Y.; Jan, S. Combination of computer vision detection and segmentation for autonomous driving. In Proceedings of the 2018 IEEE/ION Position, Location and Navigation Symposium (PLANS), Monterey, CA, USA, 23–26 April 2018; pp. 1047–1052.
6. Crespo, J.; Castillo, J.C.; Mozos, Ó.M.; Barber, R. Semantic Information for Robot Navigation: A Survey. *Appl. Sci.* **2020**, *10*, 497. [CrossRef]
7. Zhang, Y.; Chen, H.; He, Y.; Ye, M.; Cai, X.; Zhang, D. Road segmentation for all-day outdoor robot navigation. *Neurocomputing* **2018**, *314*, 316–325. [CrossRef]
8. Kemker, R.; Salvaggio, C.; Kanan, C. Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 60–77. [CrossRef]
9. Chen, G.; Li, C.; Wei, W.; Jing, W.; Woźniak, M.; Blažauskas, T.; Damaševičius, R. Fully Convolutional Neural Network with Augmented Atrous Spatial Pyramid Pool and Fully Connected Fusion Path for High Resolution Remote Sensing Image Segmentation. *Appl. Sci.* **2019**, *9*, 1816. [CrossRef]
10. Xu, H.; Bai, M.; Wan, T.; Xue, T.; Tang, W. Image semantic analysis and retrieval recommendation for clothing based on deep learning. *Fangzhi Gaoxiao Jichukexue Xuebao* **2020**, *33*, 64–72.
11. Abbas, Q.; Ibrahim, M.E.; Jaffar, M.A. Video scene analysis: An overview and challenges on deep learning algorithms. *Multimed. Tools Appl.* **2017**, *77*, 20415–20453. [CrossRef]
12. Benjdira, B.; Ouni, K.; Al Rahhal, M.M.; Albakr, A.; Al-habib, A.; Mahrous, E. Spinal Cord Segmentation in Ultrasound Medical Imagery. *Appl. Sci.* **2020**, *10*, 1370. [CrossRef]
13. Jiang, F.; Grigorev, A.; Rho, S.; Tian, Z.; Fu, Y.; Sori, W.J.; Khan, A.; Liu, S. Medical image semantic segmentation based on deep learning. *Neural Comput. Appl.* **2017**, *29*, 1257–1265. [CrossRef]
14. Shelhamer, E.; Long, J.; 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 June 2015; pp. 3431–3440.

15. Chen, L.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**, arXiv:1706.05587.
16. Xing, Y.; Zhong, L.; Zhong, X. An Encoder-Decoder Network Based FCN Architecture for Semantic Segmentation. *Wirel. Commun. Mob. Comput.* **2020**, *2020*, 8861886:1–8861886:9. [CrossRef]
17. Chen, L.; Papandreou, G.; Kokkinos, I.; Murphy, K.P.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef] [PubMed]
18. Chen, L.; Papandreou, G.; Kokkinos, I.; Murphy, K.P.; Yuille, A.L. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. *arXiv* **2014**, arXiv:1412.7062.
19. Zhao, H.; Qi, X.; Shen, X.; Shi, J.; Jia, J. ICNet for Real-Time Semantic Segmentation on High-Resolution Images. In *Computer Vision–ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2018; Volume 11207.
20. Huang, H.; Lin, L.; Tong, R.; Hu, H.; Zhang, Q.; Iwamoto, Y.; Han, X.; Chen, Y.; Wu, J. UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 1055–1059.
21. Chen, L.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
22. 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.
23. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
24. Dai, Y.; Gieseke, F.; Oehmcke, S.; Wu, Y.; Barnard, K. Attentional Feature Fusion. In Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2021; pp. 3559–3568.
25. Howard, A.G.; Sandler, M.; Chu, G.; Chen, L.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for MobileNetV3. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 1314–1324.
26. Feng, X.; Zhou, S.; Zhu, Z.; Wang, L.; Hua, G. Local to Global Feature Learning for Salient Object Detection. *Pattern Recognit. Lett.* **2022**, *162*, 81–88. [CrossRef]
27. Cao, G.; Xie, X.; Yang, W.; Liao, Q.; Shi, G.; Wu, J. Feature-fused SSD: Fast detection for small objects. In Proceedings of the International Conference on Graphic and Image Processing, Hong Kong, 24–26 February 2018.
28. Xie, S.; Tu, Z. Holistically-Nested Edge Detection. *Int. J. Comput. Vis.* **2015**, *125*, 3–18. [CrossRef]
29. Liu, Y.; Cheng, M.; Hu, X.; Wang, K.; Bai, X. Richer Convolutional Features for Edge Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5872–5881.
30. Liu, G.; Reda, F.A.; Shih, K.J.; Wang, T.; Tao, A.; Catanzaro, B. Image Inpainting for Irregular Holes Using Partial Convolutions. In Proceedings of the European conference on computer vision (ECCV) 2018 ECCV, Munich, Germany, 8–14 September 2018.
31. Yu, F.; Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. *arXiv* **2015**, arXiv:1511.07122.
32. Guo, C.; Wu, D. Canonical Correlation Analysis (CCA) Based Multi-View Learning: An Overview. *arXiv* **2019**, arXiv:1907.01693.
33. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef] [PubMed]
34. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable Convolutional Networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 764–773.
35. Geng, Q.; Zhou, Z.; Cao, X. Survey of recent progress in semantic image segmentation with CNNs. *Sci. China Inf. Sci.* **2017**, *61*, 051101. [CrossRef]