machines

MDPI

Article

EfferDeepNet: An Efficient Semantic Segmentation Method for Outdoor Terrain

Yuhai Wei, Wu Wei * and Yangbiao Zhang

College of Automation Science and Engineering, South China University of Technology, Guangzhou 510641, China
* Correspondence: weiwu@scut.edu.cn

Abstract: The recognition of terrain and outdoor complex environments based on vision sensors is a key technology in practical robotics applications, and forms the basis of autonomous navigation and motion planning. While traditional machine learning methods can be applied to outdoor terrain recognition, their recognition accuracy is low. In order to improve the accuracy of outdoor terrain recognition, methods based on deep learning are widely used. However, the network structure of deep learning methods is very complex, and the number of parameters is large, which cannot meet the actual operating requirements of of unmanned systems. Therefore, in order to solve the problems of poor real-time performance and low accuracy of deep learning algorithms for terrain recognition, this paper proposes the efficient EfferDeepNet network for pixel level terrain recognition in order to realize global perception of outdoor environment. First, this method uses convolution kernels with different sizes in the depthwise separable convolution (DSC) stage to extract more semantic feature information. Then, an attention mechanism is introduced to weight the acquired features, focusing on the key local feature areas. Finally, in order to avoid redundancy due to a large number of features and parameters in the model, this method uses a ghost module to make the network more lightweight. In addition, to solve the problem of pixel level terrain recognition having a negative effect on image boundary segmentation, the proposed method integrates an enhanced feature extraction network. Experimental results show that the proposed EfferDeepNet network can quickly and accurately perform global recognition and semantic segmentation of terrain in complex environments.

Keywords: EfferDeepNet network; terrain recognition; semantic segmentation; outdoor environment

Citation: Wei, Y.; Wei, W.; Zhang, Y. EfferDeepNet: An Efficient Semantic Segmentation Method for Outdoor Terrain. *Machines* **2023**, *11*, 256. https://doi.org/10.3390/machines11020256

Academic Editor: Dan Zhang

Received: 3 January 2023
Revised: 26 January 2023
Accepted: 7 February 2023
Published: 9 February 2023

1. Introduction

In recent years, more and more mobile robots have been used in unmanned ground vehicles, logistics and distribution, services, and other fields [1]. The complexity of terrain is the main factor that interferes with robots in the efficient completion of tasks. Especially in outdoor environments, terrain characteristics can have great changes and uncertainties. Therefore, it is very important to ensure the stable motion of the robot in complex outdoor terrain [2]. It is necessary both to ensure the accuracy of terrain recognition and to meet the real-time task of the robot.

Pixel-level terrain recognition is the semantic segmentation of the surrounding environment and label classification of each pixel in the image to achieve the recognition of environmental objects [3]. In the process of robot work, pixel-level annotation can help the robot to identify specific objects, which is conducive to the perception of the environment. Pixel-level terrain recognition is a key technology for autonomous navigation of robots, and can be used for path planning and adaptive adjustment of the distance and speed of robots or unmanned systems [4]. In complex outdoor environments, pixel-level terrain recognition technology combined with deep learning algorithms can help robots to perceive the overall environmental information more comprehensively.

The outdoor complex terrain has obvious geometric features as well as rich texture features. For the key texture features, traditional terrain recognition methods mainly

use grayscale, color, and shape features to carry out pixel-level label prediction for the input image [5]. Traditional terrain recognition algorithms consume a lot of computational memory, and have poor migration ability and accuracy for different data [6]. With the improvement of computer hardware performance and the rapid development of deep learning algorithms, terrain recognition technology based on convolutional neural networks (CNN) has strong nonlinear modeling ability and strong real-time performance for complex task processing [7].

Therefore, this paper proposes an efficient EfferDeepNet network to solve the problems of low accuracy and poor real-time of outdoor terrain recognition. This paper combines the backbone feature network and enhanced feature network to extract and recognize the semantic information of outdoor environment terrain, which can meet the accuracy and speed requirements of recognition at the same time. Therefore, the main contributions of this paper are as follows:

- To extract rich multi-scale feature information and obtain more feature details, different sizes of convolution kernels are introduced in the feature extraction stage of depthwise separable convolution, meaning that the convolution layer has different receptive fields and the extracted feature information is more abundant.
- To solve the problem of network models wasting iterative time on irrelevant feature areas, this paper introduces an efficient channel attention (ECA) mechanism in the channel domain, which is conducive to improving the semantic segmentation accuracy of the network model and ensuring high calculation speed.
- To solve the problem of large parameters and high complexity in the backbone feature network, this paper introduces a ghost module to lighten the network model in order to improve the real-time performance of the algorithm.
- To further improve the accuracy of the semantic segmentation region and obtain clear semantic feature edges, the features of local key regions are extracted through an enhanced feature network.

The rest of this paper is structured as follows. Section 2 outlines relevant works on terrain recognition technology. Section 3 introduces the current mainstream semantic segmentation methods for terrain recognition and introduces the proposed algorithm in detail based on these methods. Section 4 shows the test performance of our algorithm on public datasets. Finally, a brief summary is provided in Section 5.

In the rest of paper, the following terms are defined: "hyperparameters" refers to the parameter set before the network model starts training in the context of machine learning; "*Swish*" is a self-generated activation function, defined as $swish(x) = x\rho(\beta x)$, where β is a learnable parameter or a fixed hyper-parameter; *ReLU* stands for rectified linear unit, and is a nonlinear activation function commonly used in artificial neural networks.

2. Related Work

Pixel-level terrain recognition technology refers to the semantic segmentation of terrain scene data [8]. Semantic segmentation refers to label prediction for each pixel in an image, which is a pixel-level classification task [9]. In other words, the task of pixel-level terrain recognition is to understand the meaning of each pixel in the image of a terrain scene at the semantic level (for example, the pixel is a car, zebra crossing, or pedestrian). Pixel-level terrain recognition technology is conducive to a robot's understanding of scene information, and is one of the core tasks in the robotics field.

Traditional terrain recognition methods based on machine learning extract visual features from image data through manual operation, then classify the features using a classifier. In 2002, Ojala et al. [10] first proposed the local binary pattern (LBP) operator. This method can extract features from the local information of the input image, and performs well as a texture descriptor in unsupervised texture segmentation [11]. In 2011, Khan and Komma et al. [12] extracted feature information from visual environment data collected by outdoor mobile robots based on the LBP method. Their method uses a random forest model as a classifier to quickly classify the feature information of complex terrain under

extreme weather changes [13]. In 2012, Khan and Masselli et al. [14] further compared the speed up robust features (SURF) descriptor with the local binary patterns (LBP) descriptor and the local ternary patterns (LTP) descriptor, then classified the feature information extracted from the descriptor using the random forest method [15]. Their results showed that the SURF descriptor performs better in high-resolution image classification, and has strong robustness against interference from environmental factors. In 2012, Filitchkin and Byl et al. [16] used a bag of visual words (BOVM) approach based on the SURF descriptor to represent image features for terrain classification tasks. In 2018, Kim et al. used a multi-resolution directional filter to obtain data statistics on the different directions of each pixel in the image in order to obtain rotation-invariant features [17].

Although the traditional image segmentation methods mainly use grayscale, color, texture, and shape features to perform pixel-level label prediction for the input image in order to maximize the difference between regions and the similarity within regions [18], traditional image segmentation methods cannot accurately segment complex scene features, and their edge processing capability does not achieve the desired results [19]. With the development of deep learning, semantic segmentation technology based on deep learning has achieved new breakthroughs in segmentation accuracy and speed. Semantic segmentation based on deep learning mainly uses the fully supervised learning method, and uses the image data manually labeled with pixel level labels to train and predict the network. In 2015, Shelhamer et al. [20] proposed a full convolutional network (FCN) which uses deconvolution layers to directly upsample feature layers after feature extraction in convolution layers, thereby obtaining semantic segmentation results with the same size of output and input images. An FCN network realizes pixel-to-pixel and end-to-end network training and pixel-level classification through a full convolution structure, which shows excellent performance in fully supervised learning in the field of semantic segmentation. However, FCN networks have problems in that the semantic information location is easy to lose and the global context information is easy to ignore, which affects the accuracy of prediction [21].

Based on the idea of FCN networks and the aforementioned problems, researchers have proposed a series of representative semantic segmentation networks. In 2015, Ronneberger et al. [22] proposed the U-Net network based on an encoder–decoder structure. In the encoder part, the network extracts features through a convolution operation and a downsampling operation. In the decoder part, the deep features and shallow features are fused by splicing, then the deconvolution operation is used to restore the feature map to the original resolution. The design of the decoder saves more context information and enhances the precision of semantic segmentation [23]. In 2015, Visin et al. [24] proposed the ReNet network, which builds a correlation model between pixels by cascading the memory characteristics of multiple RNN networks, thereby solving the problem of insufficient use of global context information in FCN networks. A semantic segmentation network based on RNN can record historical information to obtain the sequence features of pixels and effectively use context information to refine image segmentation. In 2017, Zhao et al. [25] proposed the PSPNet network. By introducing the pooling module of the pyramid, the network splices the feature information of four different scales and then obtains a pyramid feature layer containing global information and multi-scale information. Finally, the network performs a convolution operation to generate the final semantic segmentation map. The PSPNet network provides effective global context information for pixel-level scene resolution [26]. Most researchers have achieved good results in terms of precision, which is essential in real-time operation. Efficient real-time capability can enable unmanned systems to achieve long-term autonomous operation.

Therefore, to solve the problems of poor image boundary segmentation and low real-time performance encountered in pixel-level terrain recognition based on complex scenes, the present paper proposes a lightweight EfferDeepNet network for feature extraction and segmentation of pixel level semantic information of outdoor complex terrain. Based on the idea of the EfficientNet network, this method proposes a backbone feature extraction

network with better performance. At the same time, based on the idea of the DeepLabV3+ network, an enhanced feature extraction network is introduced to further capture more precise object boundaries. By combining the advantages of the above two network models, our proposed approach effectively reduces the parameters and complexity of the model, and has high accuracy and real-time performance in outdoor scene recognition.

3. Basic Materials and Methods

3.1. EfficientNet Network Model

The EfficientNet network is based on neural architecture search (NAS) technology [27]. Its performance is excellent, and the test results on similar tasks are outstanding [28]. The principle of NAS is that a given search space can include different structures of submodules in the network. According to the search strategy, the structure and parameters can be combined in the search space to form different neural network structures. Then, the network model is trained on the training set and its performance is evaluated on the verification set. The performance evaluation strategy is the core element of neural structure search. Through the evaluation of search results and iterative search, the best network structure can be obtained. The specific process of NAS is shown in Figure 1.

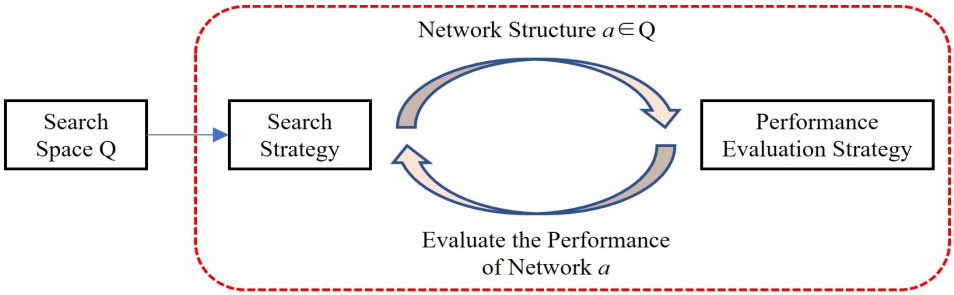

Figure 1. Model structure of NAS. With a search space Q in the search strategy, the performance of network a is continuously evaluated through the search strategy to find the optimal neural network structure.

Through the neural structure search of NAS, Google designed and searched EfficientNet B0–B7, a total of eight networks with different sizes. EfficientNet B1–B7 is a series of networks formed on the basis of the EfficientNet-B0 network structure by reducing and enlarging the hyperparameters. The network structure of EfficientNet-B0 is shown in Table 1.

Table 1. Network structure of EfficientNet-B0.

Stage	Operator	Resolution	Channels	Layers	Strides
1	Conv 3 × 3	224 × 224	32	1	2
2	MBConv 1, 3 × 3	112 × 112	16	1	1
3	MBConv 6, 3 × 3	112 × 112	24	2	2
4	MBConv 6, 5 × 5	56 × 56	40	2	2
5	MBConv 6, 3 × 3	28 × 28	80	3	2
6	MBConv 6, 5 × 5	14 × 14	112	3	1
7	MBConv 6, 5 × 5	14 × 14	192	4	2
8	MBConv 6, 3 × 3	7 × 7	320	1	1
9	Conv 1 × 1 & Pooling & FC	7 × 7	1280	1	1

In the table, *Operator* represents the input operation module, *Resolution* represents the length and width of the input image or intermediate feature channel, *Channels* represents the number of output feature channels in the current stage, *Layers* represents the stacking time of the *Operator* module, and *Strides* represents the steps of the current module.

It can be seen that the EfficientNet-B0 network is composed of simple stacks of MBConv modules. The overall structure of the network is relatively simple, and the most important element is the MBConv module. The structure of the MBConv module is shown in Figure 2.

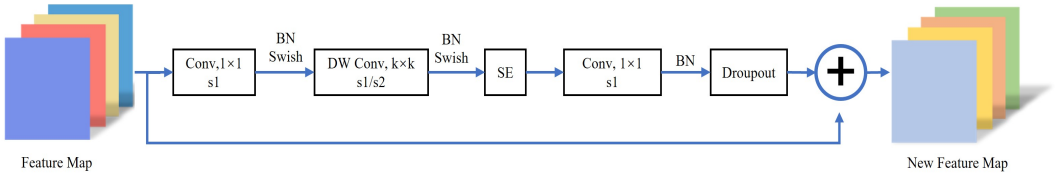

Figure 2. Module structure of MBConv. The *Droupout* layer performs random inactivation, which causes the activation value of the neurons to stop working with a certain probability.

The MBConv module performs a dimension-increasing operation on the input feature layer, then convolves the output feature information through the batch normalization (BN) layer and *Swish* activation functions, respectively [29]. Next, the squeeze-and-excitation (SE) attention module is connected in order to learn the correlations between channels. Finally, the reduced-dimension feature information is passed through the *Droupout* layer and combined with the input feature layer.

The main contribution of the EfficientNet network model is to explore the number of convolution cores, the depth of the network, and the resolution of the input image. Through the optimal configuration strategy, these three hyperparameters greatly improve the network performance. At the same time, increasing the number of feature layers in each layer of the network means that more feature information can be extracted, thereby improving the recognition accuracy of the network. However, for the problem of semantic information recognition and segmentation of outdoor large-scale scenes, the real-time performance needs to be further improved.

3.2. DeepLabV3+ Network Model

DeepLabV3+ is a semantic segmentation method based on encoder–decoder structure [30]. This method introduces a spatial pyramid pooling module or encoder–decoder structure into the deep neural network when performing semantic segmentation tasks. At the same time, by constantly exploring the backbone feature extraction network, depthwise separable convolution is applied to the atrous spatial pyramid pooling (ASPP) and decoder module in order to construct a faster and stronger encoder–decoder structure.

As shown in Figure 3, the Deeplabv3+ network is divided into an encoder and a decoder. The encoder is mainly composed of a backbone feature extraction network and an ASPP module. To solve the problem of slow semantic segmentation caused by the huge number of parameters of the DeeplabV3+ network, a lightweight MobileNetV2 network is used as the backbone feature extraction network. This network is a lightweight model based on depthwise separable convolution, and can extract shallow semantic feature information and deep semantic feature information from MobileNetV2 network. After the deep semantic feature information is enhanced by the ASPP module, the upper sampling operation of bilinear interpolation is used to connect the deep and shallow semantic feature information. Finally, the decoder module decodes both the shallow semantic feature information and the deep semantic feature information and outputs the semantic segmentation results.

3.3. Proposed Method: EfferDeepNet Network Structure

The EfferDeepNet network is mainly composed of a backbone feature extraction network and enhanced feature extraction network. The backbone feature extraction network uses a modified version of the EfficientNet network with better performance, whith the aim of reducing the size of the network and enhancing its image feature extraction ability. On the basis of the EfficientNet network, pyramid convolution and an attention mechanism are introduced, enabling the network to extract terrain with obvious texture features more

effectively. At the same time, the lightweight design of the network ensures the real-time performance of network reasoning while improving the portability and real-time performance of the network on the hardware equipment of unmanned systems. The enhanced feature network is based on the DeepLabV3+ semantic segmentation algorithm, further improving the extraction and recognition of terrain semantic information.

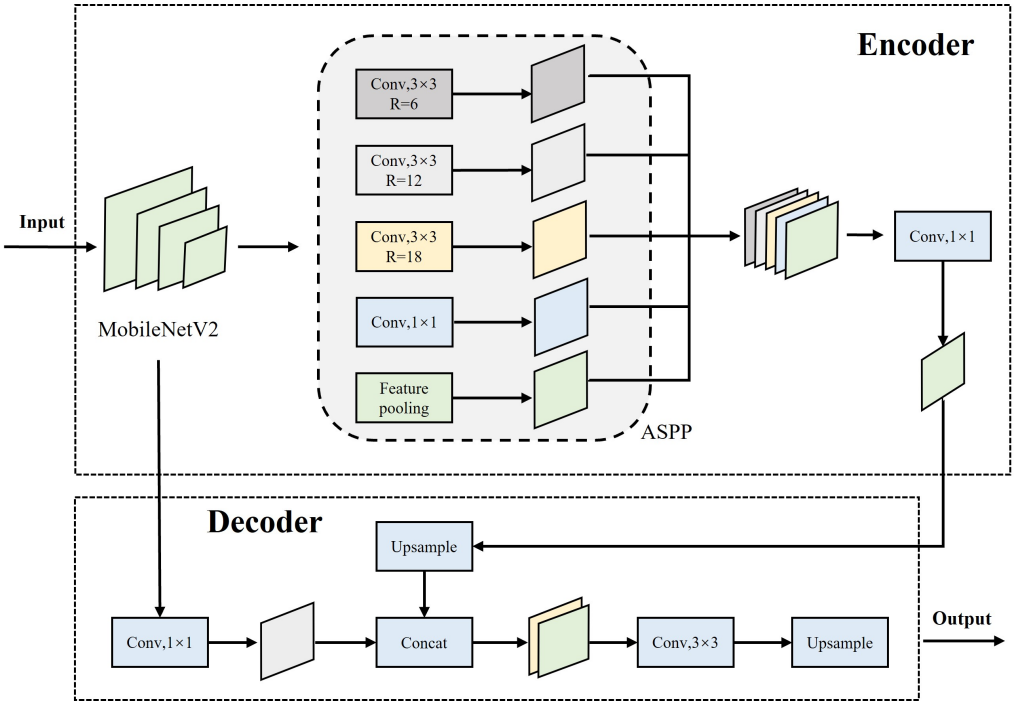

Figure 3. The network model of Deeplabv3+.

3.3.1. Backbone Feature Network: EfferNet

- Feature Extraction by Depthwise Separable Convolution

For feature extraction, the EfficientNet network combines traditional ordinary convolution and depthwise separable convolution to extract features from the input image. Basic features such as points, lines, and angles can be extracted from the shallow layer of the network, while abstract features can be extracted from the deep layer of the network. In EfficientNet networks, only single-size convolution kernels are used in a single stage. For example, only 3×3 convolution kernels are used in a single stage of the network. The common convolution used in the EfficientNet network is shown in Figure 4. In the EfferNet network, we use convolution kernels of different sizes in the shallow stage of the network rather than using single-size convolution kernels. By using convolution kernels of different sizes in the different stages of the EfferNet network, the receptive fields of the convolution layer of the network model at each stage are different, meaning that the extracted features are more abundant, in turn allowing for more effective identification of image information. The structure of the pyramid convolution module used in this paper is shown in Figure 5.

In the PConv module, the input feature information passes through convolution kernels of different sizes, making the output feature information more suitable for semantic classification. The advantage of this structure is that the network can capture multi-scale feature information at all stages. Diversified convolution kernels can provide receptive fields of different sizes, with small convolution kernels able to extract more details and large convolution kernels able to extract contextual information and larger objects. Therefore, pyramid convolution can reduce the layer dependency of the network and extract rich multi-scale information while reducing the network depth. The PConv module fuses and complements this information, which is conducive to improving the network's identifica-

tion performance while maintaining a similar number of parameters and computation cost to the original model.

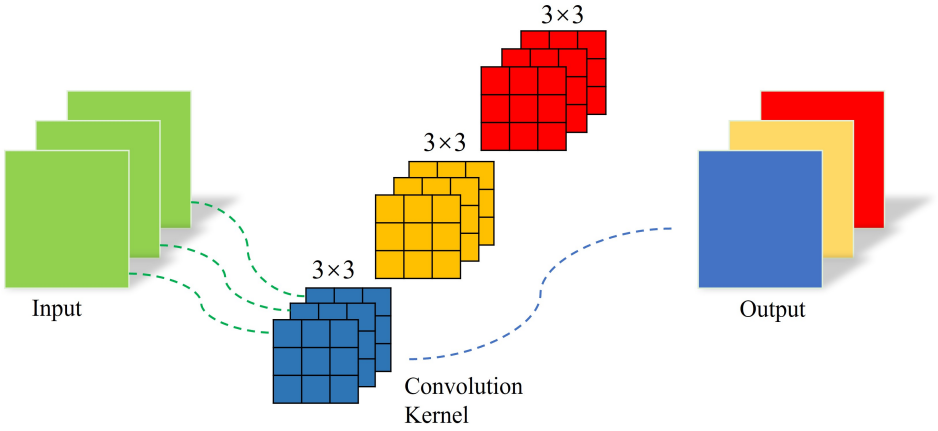

Figure 4. Convolution operation diagram of the EfficientNet network. The convolution kernel sizes of the network model are all the same.

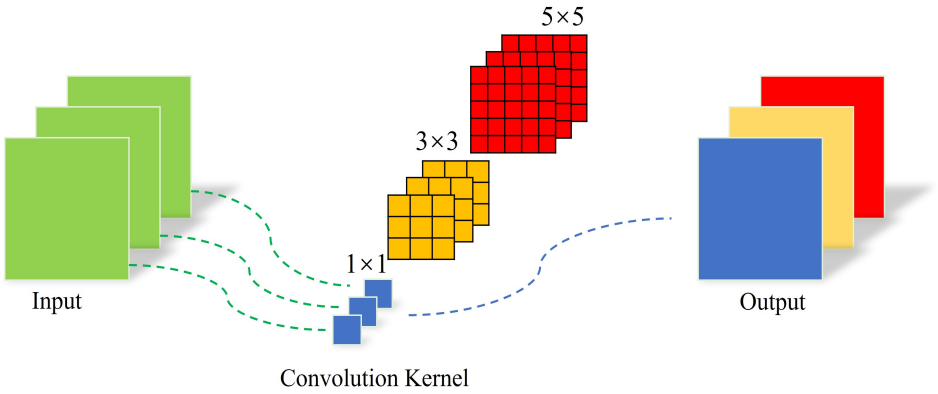

Figure 5. Convolution operation diagram of the PConv module in the EfferNet network. The proposed network model uses convolution kernels of different sizes.

- Attention Mechanism

In the field of image recognition, the use of attention mechanisms generally includes three categories, namely, the spatial domain, channel domain, and hybrid domain. An attention mechanism enables a model to focus on key local information while ignoring irrelevant areas by assigning different weights to input features in order to make more accurate decisions [31]. The MBConv module in the EfficientNet network adopts the squeeze-and-excitation (SE) channel domain attention mechanism. In this algorithm, there are two problems when using the SE module: first, the fully connected layer reduces the dimension in order to reduce the computational load, and reducing the dimensionality of the $1 \times 1 \times C$ feature layer through the first fully connected layer is unfavorable for the prediction of channel domain attention; second, it is inefficient and unnecessary to use the fully connected layer to capture the correlations between all channels [32].

To avoid these problems involving the SE module, we replace the SE module with efficient channel attention (ECA) based on the channel domain in the MBConv module in order to improve the overall performance of the network. As shown in Figure 6. For the input features, the global average pooling layer is first used for conversion to a $1 \times 1 \times C$ feature layer. Then, one-dimensional convolution with a convolution kernel of size k is used to capture the interaction information between each channel and its k adjacent channels. Finally, the weight of each channel and the recalibrated feature layer are obtained

through the activation function. This approach can ensure that the whole module has fewer parameters and faster calculation speed without reducing the dimension of the module. Our experimental results show that the ECA module effectively improves the semantic segmentation accuracy of the network model.

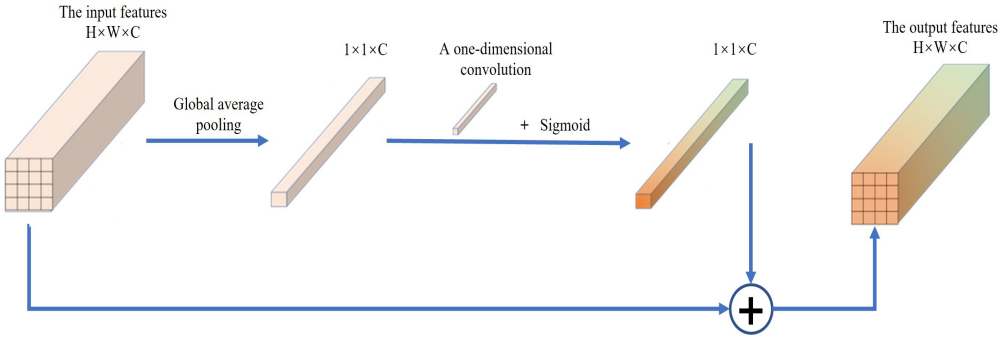

Figure 6. Structure of the ECA module; "+" means that the output results of the previous stage are added. *Sigmoid* is used as the activation function of the neural network for the output of hidden layer neurons, and the value range is (0, 1).

- Network Lightweight

A lightweight network means that, while maintaining the accuracy of the network model, its parameters and computation cost are reduced in order to solve performance problems involving the low memory and computing power available in the hardware devices of unmanned systems [33]. Methods of obtaining lightweight networks are mainly divided into two types, namely, structure design and model compression.

The parameters and complexity of a network model can be reduced by lightweight design without loss of accuracy. Therefore, the ghost module is used in the EfferNet network to achieve a lightweight design, as shown in Figure 7. By visualizing the middle feature layer of the trained deep neural network, it can be seen that there are many middle feature layers in the network dealing with similar situations. These huge and redundant feature maps are very important for network information recognition. We can perform convolution operation and linear transformation on the input features to generate similar feature maps while greatly reducing the amount of parameters and computation required.

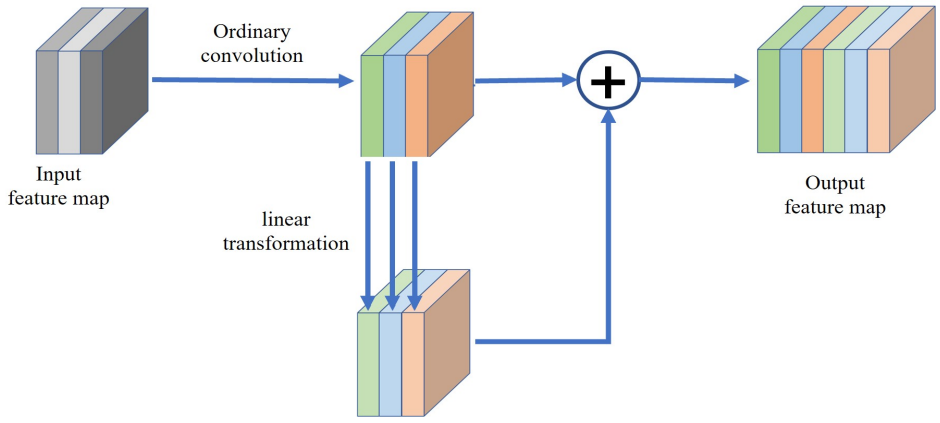

Figure 7. Ghost module structure; "+" means that the output results of the previous stage are added.

3.3.2. Enhanced Feature Extraction Network

To achieve more accurate of feature recognition of environmental terrain, our proposed method further incorporates an enhanced feature extraction network and makes the semantic segmentation area more accurate by introducing an encoder–decoder structure. The encoder is used to effectively extract the feature layer, while the decoder obtains clear object boundaries through gradual recovery of spatial information.

The encoder uses the ASPP module to capture the features output from the backbone feature extraction network. Based on ordinary convolution, the expansion rate parameter is introduced into the atrous convolution, resulting in the convolution layer with the same size of convolution kernel obtaining a larger receptive field. The structure of atrous convolution is shown in Figure 8.

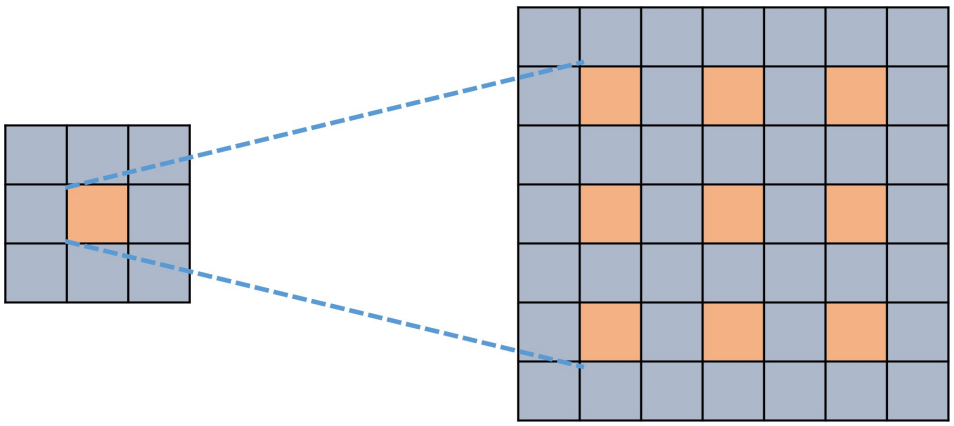

Figure 8. Atrous convolution structure.

The output features of the four convolution layers of the ASPP module pass through the batch normalization (BN) layer and *ReLU* activation function. Then, the output of the pooled layer is connected to a convolution kernel with a size of 1×1, and the resolution is restored through the bilinear oversampling layer for fusion with the features of the other convolution layers. Finally, the dimensions of multi-scale features are reduced by the common convolution layer, BN layer, and activation function, and the output enhanced feature layer is used to recover the decoder spatial information. The structure of the decoder is shown in Figure 9.

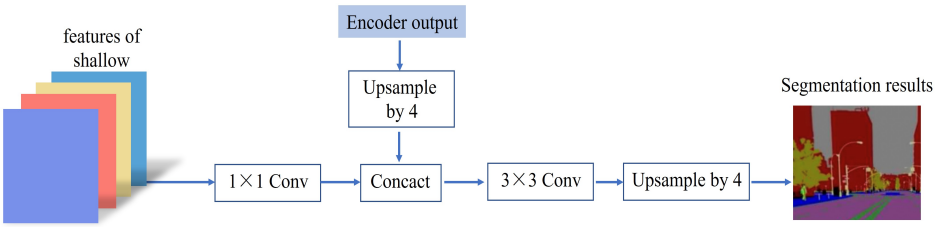

Figure 9. The structure of the decoder.

The decoder takes the shallow features extracted from the backbone feature extraction network as input and uses the ordinary convolution layer with a 1×1 convolution kernel, BN layer, and activation function to reduce the dimension. Then, the results of the enhanced feature layer extracted by the encoder after four upsampling operations are stacked and a 1×1 convolution layer is used to adjust the number of channels to match the number of label categories in the full network. Finally, the width and height of the feature layer are restored by upsampling to the width and height of the input terrain recognition image in order to carry out terrain recognition and label prediction of pixels. In this way, the spatial boundary information of the picture can be recovered through the decoder module.

To sum up, the EfferDeepNet network uses the EfferNet network as the backbone feature extraction network. Then, the semantic features from the backbone feature extraction network are used as the encoder input of the enhanced feature extraction network. The overall structure of the network is shown in Figure 10. In the backbone feature extraction network part, the Fused-MGPConv module combines pyramid convolution with convolution using kernel sizes of 3×3 and 5×5. At the same time, the ECA module is connected to perform global average pooling of the input feature map, after which the weight of each channel is obtained through the sigmoid activation function. In the deep stage of the

network structure, a 1 × 1 convolution kernel is used and the MGPConv module is built using the ghost module. The enhanced feature extraction network is mainly composed of the optimized encoder–decoder structure.

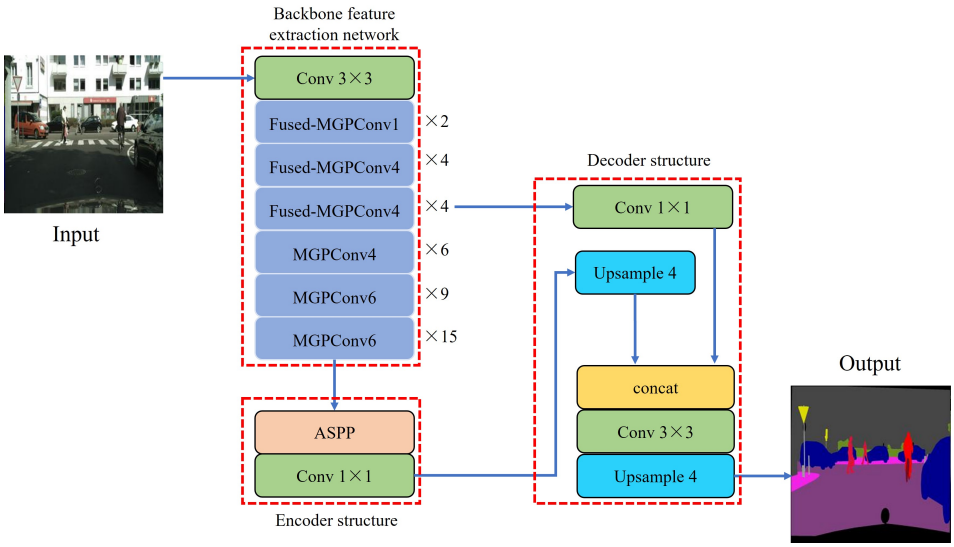

Figure 10. Overall structure of the EfferDeepNet network.

4. Experimental Results

The EfferDeepNet network proposed in this paper is implemented based on the *Pytorch* framework. *Pytorch* is an open source neural network framework from *Facebook* that can support powerful GPU acceleration functions and has a rich ecosystem. In addition, the system is simple and flexible to use and has a fast operation speed. The hardware and software specifications used when training the EfferDeepNet network are shown in Tables 2 and 3, respectively.

Table 2. Hardware used for network training.

CPU Version	RAM Version	GPU Version
Intel Core i7-7800X 3.50 GHz	64 GB	Nvidia GeForce GTX 1080Ti

Table 3. Software used for network training.

Operating System	Python Version	Pytorch Version
Ubuntu 16.04	3.8.11	1.90

In addition, we used the *Cityscapes* open dataset for experimental testing. *Cityscapes* is a semantic segmentation dataset of urban streetscapes published by Mercedes-Benz in 2015, and represents a new large-scale dataset in the field of autonomous navigation. The *Cityscapes* dataset consists of street scenes from more than fifty different cities; captured frames from videos are saved as original pictures and labeled for pixel-level terrain recognition. The *Cityscapes* dataset provides 5000 fine annotated images and 20,000 rough annotated images, and includes nineteen different categories of sidewalks, buildings, and walls.

4.1. Comparison of Semantic Segmentation Accuracy

In the training process of the EfferDeepNet network, we set the number of network iterations as $Epoch = 100$ and the learning rate as 0.001. *Adam* was selected as the network optimizer, and cross entropy was used as the loss function. The original image resolution of the *Cityscapes* dataset is 1024 × 2048. In the experimental test, using the large resolution image as the input of the training network seriously affects the performance of the algorithm. Therefore, we preprocessed the original images and tags to improve the semantic

recognition of pixel-level terrain in the EfferDeepNet network. In the image preprocessing stage, we used regional interpolation to reduce the image resolution to 512 × 1024. The use of regional interpolation can avoid jagged edges in the images and provide a better overall effect. In addition, further normalization of the image can accelerate both the convergence of the network and the generalization ability of the model. The pixel-level prediction results on the test set are shown in Figure 11. From the results of semantic segmentation, it can be seen that EfferDeepNet can achieve relatively complete segmentation of the main body of the target in all nineteen categories with a high accuracy rate.

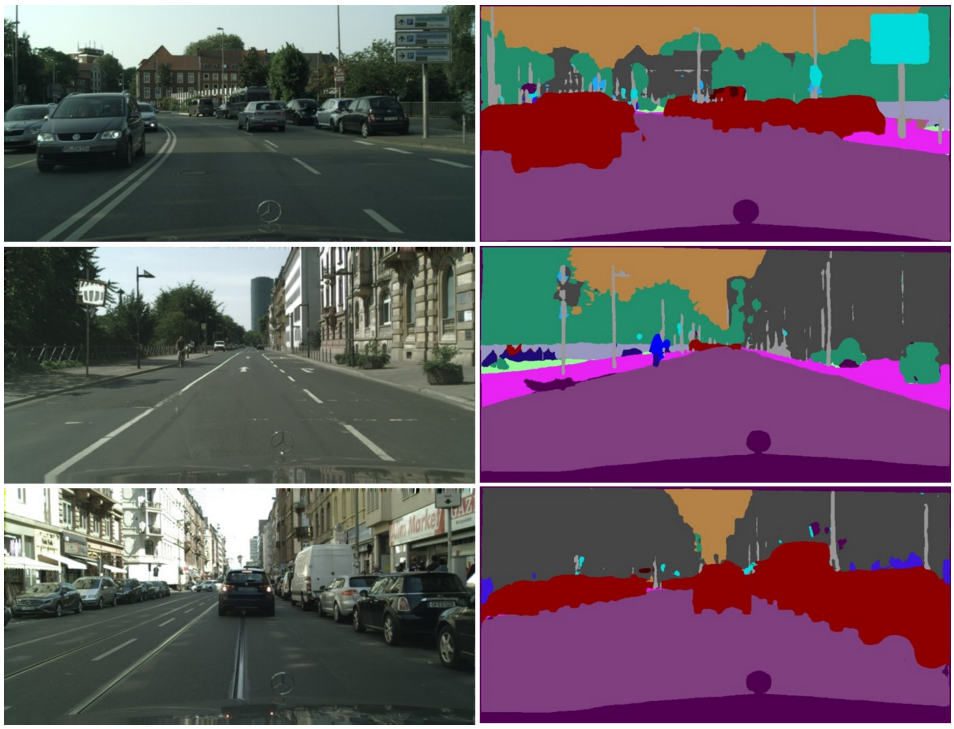

Figure 11. Prediction results of the EfferDeepNet network.

Table 4 shows a comparison of the precision of the EfferDeepNet network and other mainstream segmentation networks on the *Cityscapes* dataset. Here, mIOU is the mean intersection over union and mPA is the mean pixel accuracy. The experimental results show that our proposed EfferDeepNet network achieves 83.2% on mIOU, which is considered a high semantic segmentation accuracy. The accuracy of the EfficientNet network in terms of mIOU is 82.5%, which is lower than that of our proposed EfferDeepNet network. In addition, the accuracy of EfferDeepNet in terms of mPA reaches 89.3%, which is higher than the compared network models.

Table 4. Precision comparison between the proposed EfferDeepNet network and similar networks on the Cityscapes dataset.

	ResNet38	**PSPNet**	**DeepLabV3+**	**EfficientNet**	**EfferDeepNet**
mIOU	80.6%	81.2%	82.1%	82.5%	83.2%
mPA	82.7%	85.4%	87.1%	88.6%	89.3%

To better compare the accuracy of each segmentation network method, the same training parameters were used. In the process of network training, we recorded and visualized the experimental results in real time, as shown in Figure 12. It can be seen from the results that the training accuracy of each network reaches convergence. The semantic segmentation accuracy of our proposed EfferDeepNet method is the best, and the required number of training iterations is lower. Compared with similar segmentation

networks, EfferDeepNet achieves convergence and higher accuracy with the least number of iterations, showing that the proposed method has advantages in semantic recognition of complex terrain.

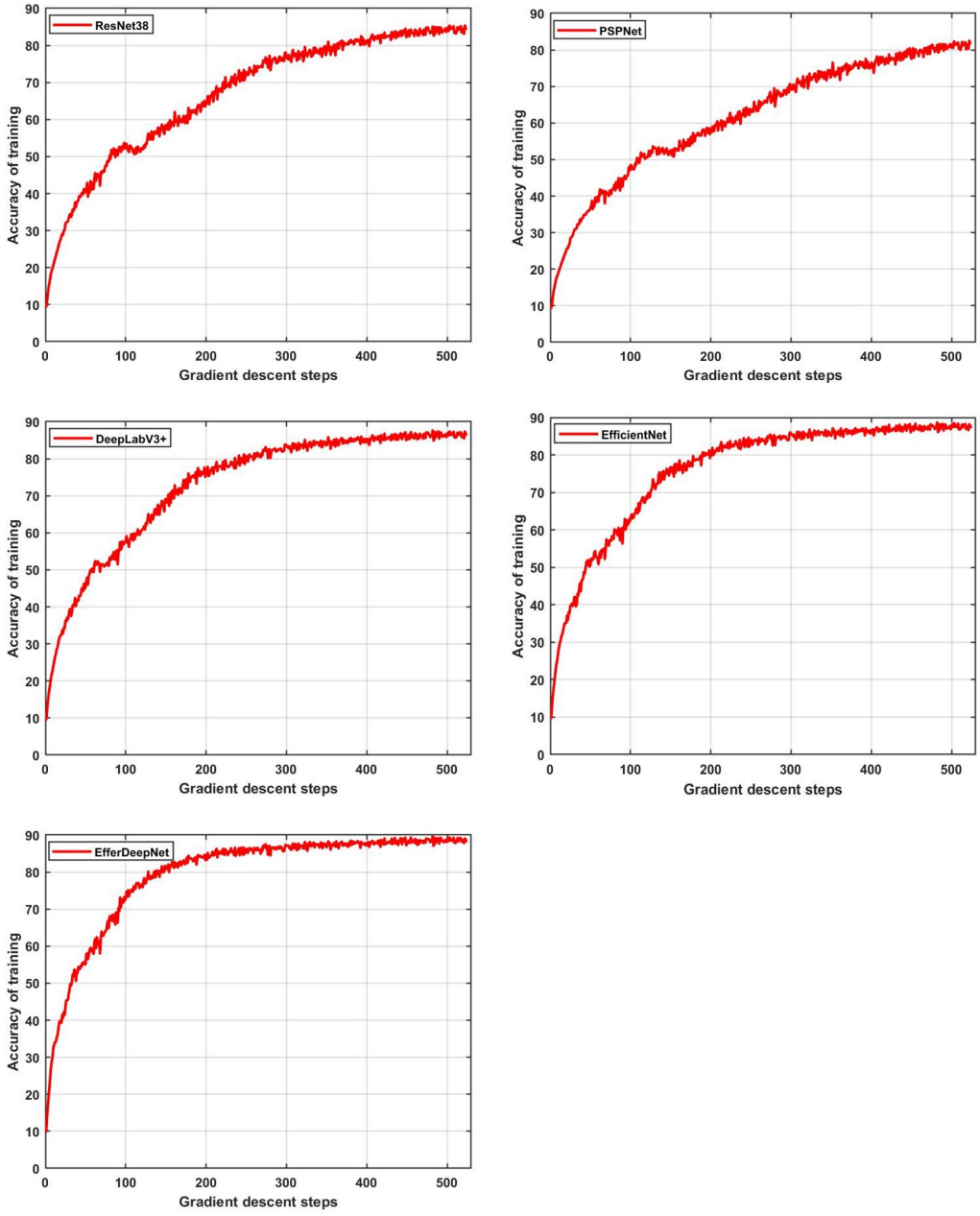

Figure 12. Accuracy comparison of different methods.

4.2. Real-Time Performance Comparison

The number of parameters affects the processing speed of network models used for semantic segmentation. Therefore, we tested and analyzed the parameters of each network model; the results are shown in Table 5. Here, MS is the model size and FPS is the frames per second. It can be seen from Table 5 that the size of the backbone feature extraction network parameters in EfferDeepNet is 16 MB, and the image processing efficiency reaches 41.7%. Compared with other methods, its image processing efficiency is better. The size of the parameters in the backbone feature extraction network of EfficientNet is 19.7 MB, further showing that our method is more lightweight and efficient.

Table 5. Real-time performance comparison between the EfferDeepNet network and other networks on the Cityscapes dataset.

	ResNet38	PSPNet	DeepLabV3+	EfficientNet	EfferDeepNet
MS	18.4 MB	20.1 MB	22.9 MB	19.7 MB	16 MB
FPS	22.6%	28.9%	12.3%	33.4%	41.7%

In addition, we conducted five repeated experiments on each network model; the real-time experimental results are shown in Figure 13. The results of this experiment visualize the average semantic segmentation times of each frame image in order to compare the efficiency of each method. The comparison results show that the proposed EfferDeepNet method has the fastest time for semantic segmentation of environmental terrain and high real-time performance. Compared with EfficientNet method, our method is much faster in semantic segmentation. Compared with other semantic segmentation methods, EfferDeepNet is able to efficiently perform semantic segmentation of terrain in complex environments.

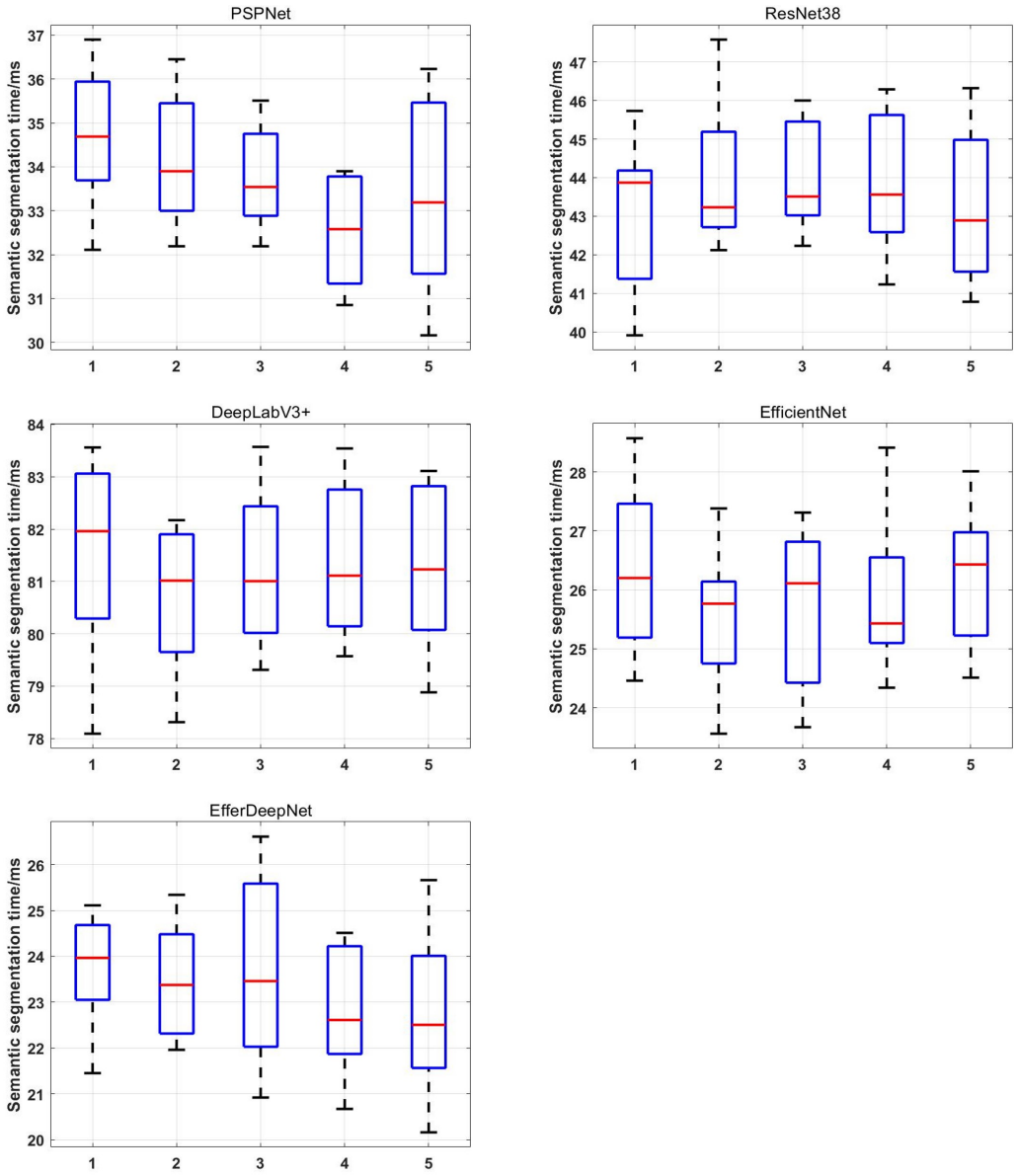

Figure 13. Average semantic segmentation time of different methods.

5. Conclusions

This paper proposes an efficient and lightweight EfferDeepNet network model for pixel-level semantic segmentation of terrain in complex environments, which is a basic necessity for robotic system to achieve autonomous work. EfferDeepNet combines the powerful texture feature extraction performance of the backbone feature network with the enhanced feature network's multi-scale feature fusion and spatial information recovery performance. This method realizes the end-to-end training task of the network, and has remarkable effects on pixel-level semantic recognition and segmentation of complex terrain. Our experimental results show that the proposed method achieves 83.2% on the mIOU index and 89.3% on the mPA index in terms of semantic segmentation accuracy. In terms of real-time performance, the FPS index of our method reaches 41.7%. In summation, the method proposed in this paper is able to efficiently perform semantic segmentation of complex terrain.

Author Contributions: Conceptualization, Y.W. and W.W.; methodology, Y.W. and Y.Z.; software, Y.W. and Y.Z.; validation, Y.W., W.W. and Y.Z.; formal analysis, Y.W. and W.W.; investigation, Y.W., W.W. and Y.Z.; resources, W.W.; data curation, Y.W. and Y.Z.; writing—original draft preparation, Y.W. and Y.Z.; writing—review and editing, Y.W.; visualization, Y.W.; supervision, Y.Z.; project administration, W.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant numbers 61573148.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Lajoie, P.Y.; Ramtoula, B.; Chang, Y.; Carlone, L.; Beltrame, G. DOOR-SLAM: Distributed, Online, and Outlier Resilient SLAM for Robotic Teams. *IEEE Robot. Autom. Lett.* **2020**, *5*, 1656–1663. [CrossRef]
2. Liao, Z.; Hu, Y.; Zhang, J.; Qi, X.; Zhang, X.; Wang, W. SO-SLAM: Semantic Object SLAM with Scale Proportional and Symmetrical Texture Constraints. *IEEE Robot. Autom. Lett.* **2022**, *7*, 4008–4015. [CrossRef]
3. Li, R.; Wang, Y.; Wang, L.; Lu, H.; Wei, X.; Zhang, Q. From Pixels to Semantics: Self-Supervised Video Object Segmentation with Multiperspective Feature Mining. *IEEE Trans. Image Process.* **2022**, *31*, 5801–5812. [CrossRef] [PubMed]
4. Yin, C.; Tang, J.; Yuan, T.; Xu, Z.; Wang, Y. Bridging the Gap Between Semantic Segmentation and Instance Segmentation. *IEEE Trans. Multimed.* **2022**, *24*, 4183–4196. [CrossRef]
5. Seyedhosseini, M.; Tasdizen, T. Semantic Image Segmentation with Contextual Hierarchical Models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 951–964. [CrossRef]
6. Liu, K.; Ye, Z.; Guo, H.; Cao, D.; Chen, L.; Wang, F.-Y. FISS GAN: A Generative Adversarial Network for Foggy Image Semantic Segmentation. *IEEE-CAA J. Autom. Sin.* **2021**, *8*, 1428–1439. [CrossRef]
7. Jin, Z.; Iqbal, M.Z.; Bobkov, D.; Zou, W.; Li, X.; Steinbach, E. A Flexible Deep CNN Framework for Image Restoration. *IEEE Trans. Multimed.* **2020**, *22*, 1055–1068. [CrossRef]
8. Jing, L.; Chen, Y.; Tian, Y. Coarse-to-Fine Semantic Segmentation From Image-Level Labels. *IEEE Trans. Image Process.* **2020**, *29*, 225–236. [CrossRef]
9. Zhang, Y.; Chen, X.; Li, J.; Wang, C.; Xia, C.; Li, J. Semantic Object Segmentation in Tagged Videos via Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 1741–1754. [CrossRef]
10. Ojala, T.; Pietikainen, M.; Maenpaa, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 971–987. [CrossRef]
11. Pan, L.; Dai, Y.; Liu, M.; Porikli, F.; Pan, Q. Joint Stereo Video Deblurring, Scene Flow Estimation and Moving Object Segmentation. *IEEE Trans. Image Process.* **2020**, *29*, 1748–1761. [CrossRef]
12. Khan, Y.N.; Komma, P.; Bohlmann, K.; Zell, A. Grid-based visual terrain classification for outdoor robots using local features. In Proceedings of the 2011 IEEE Symposium on Computational Intelligence in Vehicles and Transportation Systems, Paris, France, 11–15 April 2011; pp. 16–22.
13. 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]

14. Khan, Y.N.; Masselli, A.; Zell, A. Visual terrain classification by flying robots. In Proceedings of the IEEE International Conference on Robotics and Automation, Saint Paul, MN, USA, 14–18 May 2012; pp. 498–503.
15. Saha, S.; Mou, L.; Qiu, C.; Zhu, X.X.; Bovolo, F.; Bruzzone, L. Unsupervised Deep Joint Segmentation of Multitemporal High-Resolution Images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 8780–8792. [CrossRef]
16. Filitchkin, P.; Byl, K. Feature-based terrain classification for LittleDog. In Proceedings of the 25th IEEE/RSJ International Conference on Intelligent Robots and Systems, Vilamoura, Portugal, 7–12 October 2012; pp. 1387–1392.
17. Mirsadeghi, S.E.; Royat, A.; Rezatofighi, H. Unsupervised Image Segmentation by Mutual Information Maximization and Adversarial Regularization. *IEEE Robot. Autom. Lett.* **2021**, *6*, 6931–6938. [CrossRef]
18. Zhang, J.; Ma, C.; Yang, K.; Roitberg, A.; Peng, K.; Stiefelhagen, R. Transfer Beyond the Field of View: Dense Panoramic Semantic Segmentation via Unsupervised Domain Adaptation. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 9478–9491. [CrossRef]
19. Gu, Z.; Niu, L.; Zhao, H.; Zhang, L. Hard Pixel Mining for Depth Privileged Semantic Segmentation. *IEEE Robot. Autom. Lett.* **2021**, *23*, 3738–3751. [CrossRef]
20. Shelhamer, E.; Long, J.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651. [CrossRef]
21. Yin, R.; Cheng, Y.; Wu, H.; Song, Y.; Yu, B.; Niu, R. FusionLane: Multi-Sensor Fusion for Lane Marking Semantic Segmentation Using Deep Neural Networks. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 1543–1553. [CrossRef]
22. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv* **2015**, arXiv:1505.04597.
23. Ren, M.; Dey, N.; Fishbaugh, J.; Gerig, G. Segmentation-Renormalized Deep Feature Modulation for Unpaired Image Harmonization. *IEEE Trans. Med. Imaging* **2021**, *40*, 1519–1530. [CrossRef]
24. Visin, F.; Kastner, K.; Cho, K.; Matteucci, M.; Courville, A.; Bengio, Y. ReNet: A Recurrent Neural Network Based Alternative to Convolutional Networks. *arXiv* **2015**, arXiv:1505.00393.
25. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. In Proceedings of the 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6230–6239.
26. Yang, K.; Hu, X.; Bergasa, L.M.; Romera, E.; Wang, K. PASS: Panoramic Annular Semantic Segmentation. *IEEE Trans. Intell. Transp. Syst.* **2020**, *21*, 4171–4185. [CrossRef]
27. Tan, M.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 10–15 June 2019.
28. de Carvalho, O.L.F.; Júnior, O.A.d.C.; Albuquerque, A.O.d.; Santana, N.C.; Borges, D.L. Rethinking Panoptic Segmentation in Remote Sensing: A Hybrid Approach Using Semantic Segmentation and Non-Learning Methods. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 3512105. [CrossRef]
29. Samani, E.U.; Yang, X.; Banerjee, A.G. Visual Object Recognition in Indoor Environments Using Topologically Persistent Features. *IEEE Robot. Autom. Lett.* **2021**, *6*, 7509–7516. [CrossRef]
30. Chen, L.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Proceedings of the 15th European Conference on Computer Vision, Munich, Germany, 8–14 September 2018.
31. Luo, Z.; Li, J.; Zhu, Y. A Deep Feature Fusion Network Based on Multiple Attention Mechanisms for Joint Iris-Periocular Biometric Recognition. *IEEE Signal Process. Lett.* **2021**, *28*, 1060–1064. [CrossRef]
32. Chen, J.; Wu, Y.; Yang, Y.; Wen, S.; Shi, K.; Bermak, A.; Huang, T. An Efficient Memristor-Based Circuit Implementation of Squeeze-and-Excitation Fully Convolutional Neural Networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *33*, 1779–1790. [CrossRef]
33. Wang, Z.; Li, L.; Xue, Y.; Jiang, C.; Wang, J.; Sun, K.; Ma, H. FeNet: Feature Enhancement Network for Lightweight Remote-Sensing Image Super-Resolution. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5622112. [CrossRef]