remote sensing

MDPI

Article

MECA-Net: A MultiScale Feature Encoding and Long-Range Context-Aware Network for Road Extraction from Remote Sensing Images

Yongshi Jie [1], Hongyan He [1], Kun Xing [1,*], Anzhi Yue [2], Wei Tan [1], Chunyu Yue [1], Cheng Jiang [1] and Xuan Chen [1]

[1] Beijing Institute of Space Mechanics and Electricity, China Academy of Space Technology, Beijing 100094, China
[2] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China
* Correspondence: xingkunfeixiang@163.com

Abstract: Road extraction from remote sensing images is significant for urban planning, intelligent transportation, and vehicle navigation. However, it is challenging to automatically extract roads from remote sensing images because the scale difference of roads in remote sensing images varies greatly, and slender roads are difficult to identify. Moreover, the road in the image is often blocked by the shadows of trees and buildings, which results in discontinuous and incomplete extraction results. To solve the above problems, this paper proposes a multiscale feature encoding and long-range context-aware network (MECA-Net) for road extraction. MECA-Net adopts an encoder–decoder structure and contains two core modules. One is the multiscale feature encoding module, which aggregates multiscale road features to improve the recognition ability of slender roads. The other is the long-range context-aware module, which consists of the channel attention module and the strip pooling module, and is used to obtain sufficient long-range context information from the channel dimension and spatial dimension to alleviate road occlusion. Experimental results on the open DeepGlobe road dataset and Massachusetts road dataset indicate that the proposed MECA-Net outperforms the other eight mainstream networks, which verifies the effectiveness of the proposed method.

Keywords: road extraction; convolutional neural network; multiscale feature; long-range context; remote sensing images

Citation: Jie, Y.; He, H.; Xing, K.; Yue, A.; Tan, W.; Yue, C.; Jiang, C.; Chen, X. MECA-Net: A MultiScale Feature Encoding and Long-Range Context-Aware Network for Road Extraction from Remote Sensing Images. *Remote Sens.* **2022**, *14*, 5342. https://doi.org/10.3390/rs14215342

Academic Editor: Marc Bosch

Received: 8 September 2022
Accepted: 22 October 2022
Published: 25 October 2022

1. Introduction

Road extraction is crucial for urban planning [1,2], intelligent transportation [3,4], vehicle navigation [5,6] and emergency relief [7,8]. The rapid development of high-resolution optical remote sensing technology lays a foundation for road extraction tasks, provides sufficient data support for the extraction of large-scale road networks, and effectively promotes road extraction technology. The visual interpretation method requires many interpreters to participate in the work, which takes a lot of time. Automatic road extraction can reduce labor costs and significantly improve work efficiency, so it has important research significance and practical application value. However, roads in remote sensing images vary greatly in scale and are often blocked by the shadows of trees and buildings, which poses a great challenge to road extraction.

The exploration and improvement of automatic road extraction methods have become a research hotspot, and these methods could be classified into two categories: traditional methods and deep learning methods.

Traditional road extraction methods include the pixel-based method and object-oriented method. The pixel-based method mainly uses the differences in spectral and gray features of image pixels to extract roads. For example, Miao et al. [9] realized road extraction by using a series of operations such as the rough segmentation of roads, generation of probability maps, thresholding, and kernel density estimation and measurement. Sghaier

et al. [10] first exploited Canny edge detection to extract candidate road boundaries and then used beamlet transform for multiscale inference to extract road regions. Mu et al. [11] employed the Otsu method to obtain the road binary map, then used open operation to remove the areas that do not contain roads and finally conducted edge detection to obtain the road results. Liu et al. [12] extracted the parallel lines in the image with the geometric knowledge base of rural roads, and then used the knowledge reasoning method to group and connect them to extract the complete road. The pixel-based method is suitable for extracting rural roads with a simple background.

Object-oriented methods usually use a segmentation technique to divide the image into a series of regions and then extract the road from the regions. Yu et al. [13] combined an object-oriented method and the Markov random field method for road extraction. Huang et al. [14] combined object-oriented and conditional random field methods and used three features of road color, texture, and histogram gradient to represent the objects obtained by initial segmentation. Then, they used CRF for reasoning to obtain the road results. Li et al. [15] first performed initial segmentation to obtain road regions and then used binary partition trees to extract roads hierarchically. Maboudi et al. [16] incorporated spatial, spectral, and texture features based on object-oriented methods and then combined the fuzzy logic system and ant colony algorithm to extract roads. The object-oriented method has a certain applicability and is suitable for extracting a single type of urban road in the image. However, the traditional road extraction method needs to manually design features to extract roads in specific scenes, which makes it difficult to extract different roads in a complex background, so its application scope is limited.

In recent years, deep learning has been widely used in remote sensing information extraction [17–19], and researchers have turned to deep learning methods to realize road extraction [20]. Deep learning methods mainly use convolutional neural networks (CNNs) to automatically extract road features from a large amount of sample data, and the pixel-by-pixel classification of roads and backgrounds is achieved through stacked convolution, pooling, and upsampling operations. The fully convolutional network (FCN) [21], for the first time, realizes pixel-level dense prediction using CNN. Meanwhile, the method based on FCN has been widely used in road extraction from remote sensing images. The CasNet proposed by Cheng et al. [22] consists of two cascaded subnetworks. The encoder of each subnetwork is composed of multiple stacked convolutions and pooling layers, and the decoder is composed of the corresponding convolution layers and upsampling layers. Based on the encoder and decoder, the extraction of roads and road centerlines can be realized. Buslaev et al. [23] combined residual learning and UNet [24], and they used pretrained ResNet34 [25] as an encoder for feature extraction. The DenseUNet network designed by Xin et al. [26] achieves road feature reuse through dense connection modules, and it transfers encoder features to the decoder using skip connections. Compared with the traditional methods, these FCN-based methods significantly improve the results of road extraction, but there are still some shortcomings. For example, these FCN-based methods mainly use repeated convolutional layers to extract features, which cannot deal with slender and occluded roads effectively.

Some studies focus on the integration of multiscale features to improve road extraction accuracy. Gao et al. [27] proposed the multiple feature pyramid network (MFPN), which obtains multiscale features through a feature pyramid to achieve road extraction of different scales. Zhou et al. [28] added expansive convolution with a series–parallel structure based on LinkNet34 [29] to obtain multiscale features and expand receptive fields. Meanwhile, they proposed a network called D-LinkNet, which won the championship of the DeepGlobe Road Extraction Competition in 2018. Based on the encoder–decoder, He et al. [30] integrated the Atrous Spatial Pyramid Pooling (ASPP) module to extract multiscale road features. Lu et al. [31] paralleled convolution kernels of different sizes to integrate the multiscale features of roads. Liu et al. [32] designed a multiscale dilated convolution module to extract multiscale features. The RDRCNN proposed by Gao et al. [33] uses residual learning to build the encoder, then uses cascaded dilated convolutions to

expand the receptive field of the network, and finally uses the tensor voting algorithm for postprocessing operations. The JointNet [34] is a kind of network that can extract buildings and roads, which adds densely connected dilated convolutions to expand the receptive field in the skip connection part. The MRENet designed by Shao et al. [35] uses dilated convolutions and a pyramid pooling module [36] to obtain multiscale feature information, and can simultaneously extract road surface and road center line. Tran et al. [37] added a pyramid pooling module on the basis of LinkNet34, thus constituting a PP-LinkNet network for road extraction. The RoadNet proposed by Liu et al. [38] can learn the multiscale features of roads, and can simultaneously extract the road surface, road edge and road center line. These studies have proved that the fusion of multiscale features can improve the performance of road extraction networks, but these methods have shortcomings in obtaining a long-range context.

In the last few years, the attention mechanism has attracted much attention and has been applied to road extraction. Wang et al. [39] designed a novel road extraction network called NL-LinkNet, which introduces a nonlocal module belonging to the self-attention mechanism into the encoder of LinkNet34 to obtain long-distance dependence. Zhu et al. [40] proposed GCB-NET, which utilizes a global context-aware module to model and distribute long-range road features. Xie et al. [41] proposed HsgNe for road extraction, which introduces bilinear pooling between the LinkNet's encoder and decoder to acquire higher-order spatial information and obtain long-distance spatial feature information. Based on D-LinkNet, Wu et al. [42] proposed the AD-LinkNet network, which integrates a channel attention mechanism to strengthen feature channels and obtain global features. The road extraction method proposed by Lin et al. [43] uses the idea of SENet [44] to calibrate different weights on the channels of the feature map and model the context relationship between channels. The Attention UNet proposed by Oktay et al. [45] adds an attention gate on the basis of UNet to enhance salient features and suppress irrelevant features. The CADUNet proposed by Li et al. [46] uses cascaded global attention modules and core attention modules to improve the integrity of road results. The attention mechanism helps to improve the performance of road extraction, but existing methods often use a single mechanism, from which is difficult to obtain sufficient long-range context information to alleviate road occlusion.

To solve the above problems, this paper proposes a multiscale feature encoding and long-range context-aware network (MECA-Net). In MECA-Net, the multiscale feature encoding module is responsible for extracting multiscale features, and the long-range context-aware module is used to obtain sufficient long-range context information from both channel and spatial dimensions.

The main contributions of this paper are as follows:

1. A multiscale feature encoding module (MFEM) is designed to extract multiscale features and improve the network's ability to extract roads of different scales.
2. A long-range context-aware module (LCAM) is proposed, which uses the channel attention module (CAM) and strip pooling module (SPM) to obtain sufficient long-range context information from the channel and spatial dimensions and improve the continuity of road extraction results.
3. A road extraction network called MECA-Net is proposed to extract slender roads and alleviate the occlusion of roads. The effectiveness of MECA-Net is verified on the public DeepGlobe dataset and the Massachusetts dataset.

The rest of this paper is organized as follows: Section 2 introduces the composition of MECA-Net and the basic principle of each component module in detail. Section 3 introduces the experimental details, including the experimental dataset, evaluation method, experimental setup, and analysis of experimental results. Sections 4 and 5 present the discussion and conclusion of this paper, respectively.

2. Methodology

This section describes the proposed method in detail. Section 2.1 describes the overall structure of our designed network; Section 2.2 describes the multiscale feature encoding module; Section 2.3 introduces the long-range context-aware module; Section 2.4 describes the loss function used for network training.

2.1. Overview of the Network Structure

The overall structure of MECA-Net is shown in Figure 1, which includes three parts: encoder, decoder, and skip connection. MECA-Net is based on LinkNet34, and its encoder is composed of a pretrained ResNet34. The multiscale feature encoding module (MFEM) is added in the skip connection part to extract multiscale features in different stages and pass the multiscale features to the decoder. The decoder is mainly composed of the decoder module for upsampling and the long-range context-aware module (LCAM). The input of LCAM is the fusion result of skip connection features and the corresponding output features of the decoder module. The input of MECA-Net is an RGB image with a resolution of $512 \times 512 \times 3$, and the output is a binary map with a dimension of $512 \times 512 \times 1$, where white pixels and black pixels represent the road and background, respectively.

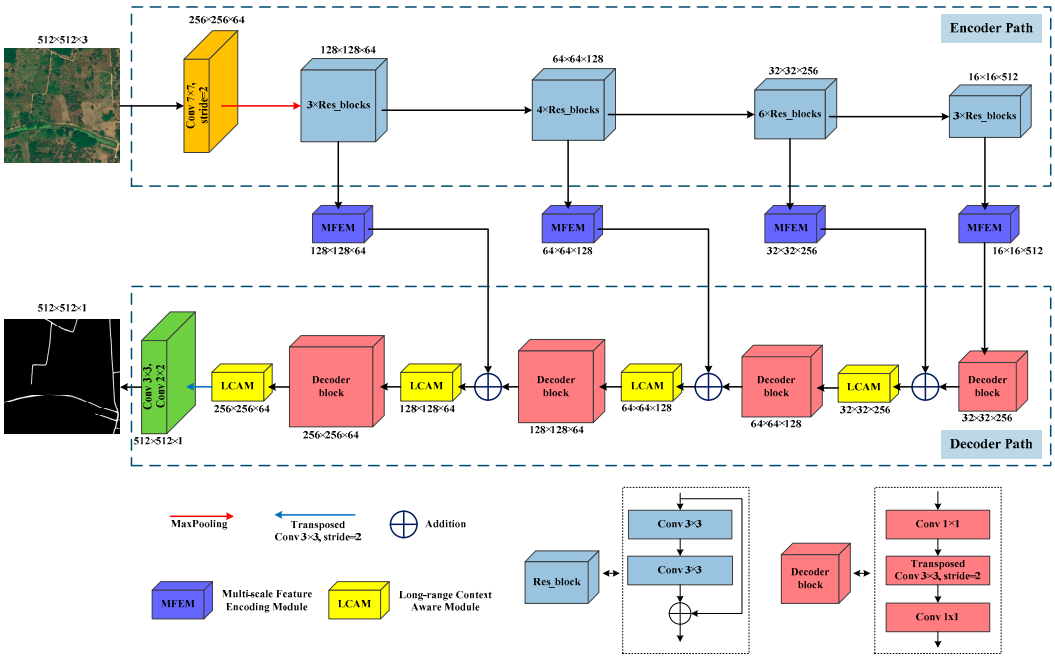


Figure 1. The overall structure of the proposed MECA-Net.

2.2. Multiscale Feature Encoding Module

One challenge in road extraction is the identification of slender roads. To overcome this challenge, inspired by the research [47] and [48], this paper proposes the MFEM, and its core idea is to aggregate multiscale features using convolution kernels of three different sizes: 1×1, 3×3, and 5×5. The convolution layers with different kernel sizes can extract features of different scales, which can provide multiscale feature representations for the next stage after aggregation [47]. The MFEM is added to the skip connection part to extract and aggregate multiscale features at different stages of the backbone network and provide feature information of different scales for the decoding process, thus enhancing the ability to extract roads of different scales. The structure of MFEM is illustrated in Figure 2.

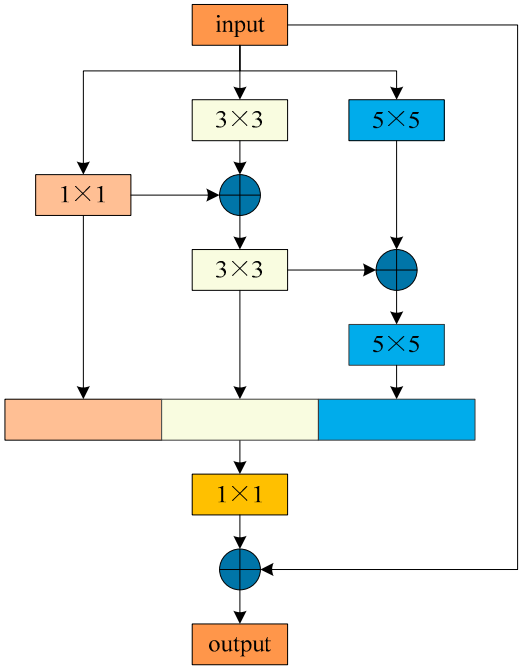


Figure 2. The structure of MFEM.

The input feature of MFEM is denoted as $\boldsymbol{x}$ and the output feature as $\boldsymbol{y}$. The MFEM has three branches, and the input feature $\boldsymbol{x}$ is inputted to the three branches in the form of $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, and $\boldsymbol{x}_3$, respectively. In the three branches, the input features are firstly extracted by convolutional layers of 1×1, 3×3, and 5×5, respectively. These three convolution operations are denoted as F_1, F_2, and F_3, and the BN [49] layer and ReLU [50] activation function are added after each convolution layer. To achieve the effective aggregation of features of different scales, the output features of the previous branch are fused with those of the current branch, and the fused features are optimized by a convolution operation. The output features of the three branches are represented as $\boldsymbol{y}_1$, $\boldsymbol{y}_2$, and $\boldsymbol{y}_3$, respectively. The output features of each branch are concatenated, and the dimensions are reduced through the 1×1 convolution layer, and then the output features of module $\boldsymbol{y}$ are obtained after adding them with the input features $\boldsymbol{x}$. The MFEM can be representred as:

$$\boldsymbol{x}_i = \boldsymbol{x}, i = 1, 2, 3 \tag{1}$$

$$\boldsymbol{y}_i = \begin{cases} F_i(\boldsymbol{x}_i), \ i = 1 \\ F_i\left(F_i(\boldsymbol{x}_i) + \boldsymbol{y}_{i-1}\right), i = 2, 3 \end{cases} \tag{2}$$

$$\boldsymbol{y} = W_{1\times1}(\mathrm{CONCAT}(\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{y}_3)) + \boldsymbol{x} \tag{3}$$

where $W_{1\times1}$ represents 1×1 convolution, and CONCAT represents a concatenate operation.

2.3. Long-Range Context-Aware Module

In remote sensing images, roads are often blocked by the shadows of roadside trees and buildings, resulting in the discontinuity of road extraction results. To maintain the continuity and integrity of the road topology, it is necessary to obtain long-range context information and use it to improve the continuity of the results. Therefore, this paper proposes the LCAM, which consists of the channel Attention Module (CAM) and the strip pooling module (SPM) in parallel, as shown in Figure 3, to capture long-range context from the channel dimension and spatial dimension.

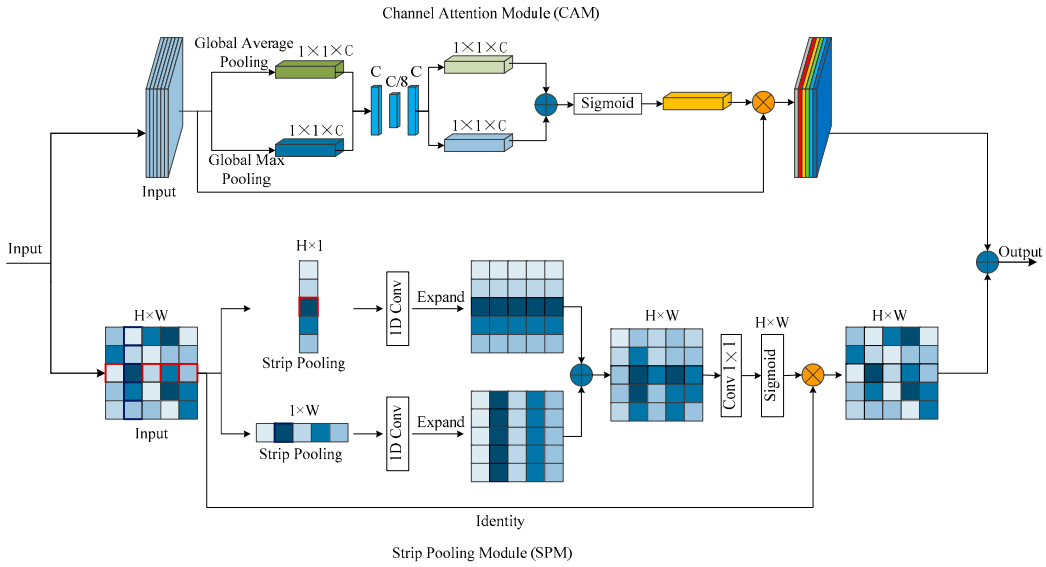


Figure 3. The structure of LCAM.

2.3.1. Channel Attention Module

The intrinsic relationship between feature channels plays an important role in the road extraction task, but the existing extraction methods often ignore this relationship. Therefore, CAM [51] is introduced to model the long-range dependence relationship between feature channels and obtain the long-range context information of channel dimensions. The structure of CAM is presented in Figure 3. The input features of CAM are the input features of LCAM, which are represented as $\boldsymbol{x} \in R^{H\times W\times C}$, and the output features are represented as $\boldsymbol{y}^{cam} \in R^{H\times W\times C}$. The input features are passed through the global average pooling (GAP) layer and the global max pooling (GMP) layer, respectively, to obtain the features with a dimension of $1 \times 1 \times$ C, and then they are inputted to two cascaded 1×1 convolutional layers. Then, the result of adding the two output features of the 1×1 convolutional layers is the input to the Sigmoid activation function and multiplied with the input feature $\boldsymbol{x}$ to obtain the output feature of the CAM, i.e., $\boldsymbol{y}^{cam}$. The process of CAM can be expressed as:

$$\boldsymbol{y}_{gap} = W^2_{1\times1}\left(W^1_{1\times1}(\mathrm{GAP}(\boldsymbol{x}))\right) \tag{4}$$

$$\boldsymbol{y}_{gap} = W^2_{1\times1}\left(W^1_{1\times1}(\mathrm{GAP}(\boldsymbol{x}))\right) \tag{5}$$

$$\boldsymbol{y}^{cam} = \sigma\left(\boldsymbol{y}_{gap} + \boldsymbol{y}_{gmp}\right) \otimes \boldsymbol{x} \tag{6}$$

where $W^1_{1\times1}$ represents the first 1×1 convolutional operation, which is used to reduce the number of channels C to C/8; $W^2_{1\times1}$ represents the second convolutional layer, which is used to restore the number of feature channels to C; σ represents the Sigmoid function, and $\otimes$ represents element-wise multiplication.

2.3.2. Strip Pooling Module

This paper uses SPM [52] to obtain long-range context information from the spatial dimension. The structure of SPM is shown in Figure 3. This module uses horizontal stripe pooling and vertical stripe pooling to obtain long-range context information from different directions. The input feature of this module is denoted as $\boldsymbol{x} \in R^{H\times W}$, and the channel dimension is omitted here for the convenience of representation. The feature $\boldsymbol{x}$ is fed into two parallel branches on which horizontal and vertical strip pooling operations

are performed. The output of horizontal strip pooling is $\boldsymbol{y}^h \in R^{H\times 1}$, and the calculation process is as follows:

$$y_i^h = \frac{1}{W}\sum_{j=0}^{W-1} x_{ij},\ \ i = 0, 1, \ldots, H-1 \tag{7}$$

The output of vertical strip pooling is denoted as $\boldsymbol{y}^v \in R^{1\times W}$, and the calculation process is as follows:

$$y_j^v = \frac{1}{H}\sum_{i=0}^{H-1} x_{ij},\ \ j = 0, 1, \ldots, W-1 \tag{8}$$

Horizontal strip pooling and vertical strip pooling operations are followed by one-dimensional convolution with a kernel size of 3. The expand operation is adopted to expand the output features of the one-dimensional convolution into $H \times W$, which are denoted as $\boldsymbol{y}_E^h \in R^{H\times W}$ and $\boldsymbol{y}_E^v \in R^{H\times W}$, respectively. Then, the expanded features of the two branches are added and input to the 1×1 convolutional layer and Sigmoid function, and then they are multiplied with the input features to obtain the output feature of SPM, i.e., $\boldsymbol{y}^{spm} \in R^{H\times W}$.

$$\boldsymbol{y}^{spm} = \sigma\left(W_{1\times 1}\left(\boldsymbol{y}_E^h + \boldsymbol{y}_E^v\right)\right) \bigotimes \boldsymbol{x} \tag{9}$$

where σ represents the Sigmoid function, $W_{1\times 1}$ represents 1×1 convolution, and $\bigotimes$ represents element-wise multiplication.

The result of adding the output features of CAM and SPM is taken as the output feature of LCAM, i.e., $\boldsymbol{y}^{out}$.

$$\boldsymbol{y}^{out} = \boldsymbol{y}^{cam} + \boldsymbol{y}^{spm} \tag{10}$$

2.4. Loss Function

Binary cross-entropy loss (BCE) is commonly used in binary semantic segmentation tasks. Its calculation formula is as follows:

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}(g_i \times log(p_i) + (1 - g_i) \times log(1 - p_i)) \tag{11}$$

where N is the number of image pixels, g_i represents the value of the ith pixel label, and p_i represents the prediction probability of the corresponding pixel.

In the road extraction task, there are only two categories: road and background. The activation function of the output layer of our proposed road extraction network is the Sigmoid function, and the output is the probability map with a pixel value between 0 and 1. The threshold of 0.5 is used to distinguish the road and background. When the probability is greater than 0.5, the corresponding pixel is predicated as the road; otherwise, it is predicated as the background.

The calculation formula of the BCE loss function shows that this loss function will calculate the loss of each pixel and then calculate the average by treating all pixels with the same weight. This is not suitable for the road extraction task because the road occupies a small proportion of the image, which will result in the imbalance of positive and negative samples and weaken the loss of pixels belonging to the road category. To solve this problem, this paper adds the Dice loss function [53] to the BCE loss function. The Dice loss function measures the similarity between labels and prediction results, which can address the imbalance between positive and negative samples. The Dice loss function can be calculated as:

$$L_{Dice} = 1 - \frac{2\sum_{i=1}^{N}(g_i \times p_i)}{\sum_{i=1}^{N} g_i^2 + \sum_{i=1}^{N} p_i^2} \tag{12}$$

The overall loss function of network training is $L_{BCE+Dice}$:

$$L_{BCE+Dice} = L_{BCE} + L_{Dice} \tag{13}$$

3. Experiments

This section describes the contents relating to the experiments in detail. Section 3.1 introduces the datasets used for the experiments; Section 3.2 presents the evaluation metrics for network performance; Section 3.3 describes the experimental settings; Section 3.4 analyzes the experimental results.

3.1. Dataset

The public DeepGlobe dataset [54] and Massachusetts dataset [55] were adopted to carry out the experiments.

(1) The DeepGlobe dataset images were collected from three regions: Thailand, Indonesia, and India, covering urban and suburban scenes. The spatial resolution of each image was 0.5m, and the size was 1024 × 1024 pixels. Referring to the research [56,57], this paper used 6226 labeled images in the dataset for experiments and split them into a training set with 4696 images and a test set with 1530 images. Then, the 4696 images were divided into 3756 training images and 940 validation images at the ratio of 8:2. To train the model with limited memory, the training and validation images were cropped into 512 × 512 image tiles. Finally, the dataset contained 15024 training image tiles, 3760 validation image tiles, and 1530 testing images with their original size. The sample images and the corresponding labels of the DeepGlobe dataset are shown in Figure 4.

Figure 4. Sample images and labels of the DeepGlobe dataset.

(2) The Massachusetts dataset covers urban, suburban, and rural scenarios and contains 1108 training images, 14 validation images, and 49 testing images. The size of each image is 1500 × 1500 and the resolution is 1m. The original training set and validation set images were seamlessly cropped into 512 × 512 image tiles. Finally, the training set contained 7972 image tiles and the validation set contained 126 image tiles. The testing set used 49 images with their original size. The sample images and the corresponding labels of the Massachusetts dataset are shown in Figure 5.

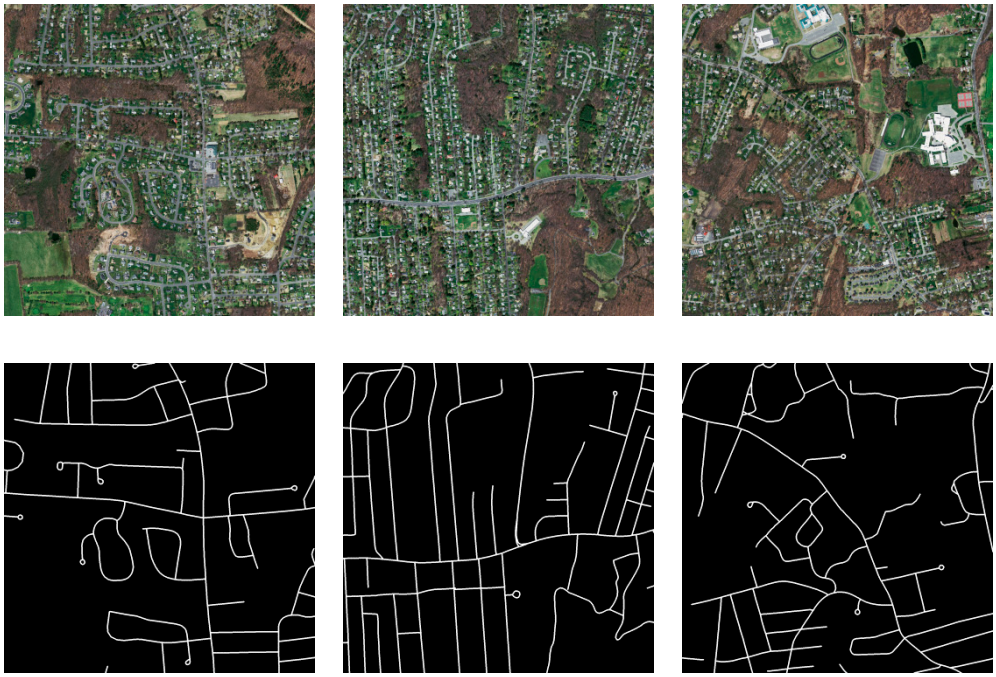


Figure 5. Sample images and labels of the Massachusetts dataset.

3.2. Evaluation Metrics

Four evaluation metrics were used to measure the accuracy of road extraction results obtained by each network, including *IoU*, *Precision*, *Recall*, and *F1*. *IoU* represents the ratio between the intersection and union of prediction results and labels. Precision represents the proportion of correctly predicted pixels among the pixels predicted as roads. *Recall* represents the proportion of pixels correctly predicted as roads among all road pixels. *F1* is the harmonic mean of *Precision* and *Recall*, which is a comprehensive evaluation metric. The calculation formulas of the above four evaluation metrics are shown as follows:

$$IoU = \frac{TP}{TP + FP + FN} \tag{14}$$

$$Precision = \frac{TP}{TP + FP} \tag{15}$$

$$Recall = \frac{TP}{TP + FN} \tag{16}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{17}$$

where *TP* (True Positive) represents the number of road pixels that are correctly predicted as road, *FP* (False Positive) represents the number of background pixels that are incorrectly predicted as road, and *FN* (False Negative) represents the number of road pixels that are incorrectly predicted as background.

3.3. Experimental Settings

To enhance the generalization ability, random rotation, random horizontal flip, random vertical flip, and random Gaussian blur were used to augment the training data. All the experiments in this paper were conducted on an NVIDIA GeForce GTX 1080 Ti (11 GB memory), and the PyTorch [58] deep learning library was used to construct, train, and test the model. The optimizer used in this paper was SGD, and the momentum and weight decay were set to 0.9 and 5e-4, respectively, and the batch size was set to 4. The number of epochs was set to 150. Besides, the "poly" learning rate strategy was adopted, and the

learning rate was multiplied by $\left(1 - \frac{iter}{max\_iter}\right)^{power}$. Among them, the initial learning rate was set to 0.01, and the power value was set to 0.9.

3.4. Experimental Results

To evaluate the performance of MECA-Net, it was compared with eight mainstream semantic segmentation networks for road extraction, including UNet [24], SegNet [59], LinkNet [29], DeepLabv3+ [60], D-LinkNet [28], RoadNet [38], PP-LinkNet [37], and NL-LinkNet [39]. UNet is a classical encoder–decoder structure network, which is widely used in road extraction tasks. SegNet is also a commonly used network for road extraction. LinkNet takes ResNet as the feature extraction network and combines the skip connection features with the decoder features by adding. It is one of the most commonly used baseline networks for road extraction at present. DeepLabv3+ uses multiple dilation convolutions with different dilation rates to obtain multiscale information. D-LinkNet uses LinkNet34 as the baseline and adds dilated convolutions of series–parallel structures to obtain multiscale features. RoadNet can learn multiscale and multilayered road features. PP-LinkNet uses the pyramid pooling module to extract multiscale features. NL-LinkNet introduces a nonlocal computing module based on LinkNet34 to obtain long-range context. To make a fair comparison, the same experimental settings were used in the training process of all networks. Additionally, LinkNet, DeepLabv3+, D-LinkNet, PP-LinkNet, NL-LinkNet, and MECA-Net all used ResNet34 as the backbone network.

Experimental results on the DeepGlobe dataset. As illustrated in Figure 6, the first row shows that each method can correctly extract wide roads, but the extraction results of the comparison methods for slender roads between farmland were not ideal. RoadNet, SegNet, NL-LinkNet, UNet, and PP-LinkNet rarely extracted the slender road, and LinkNet extracted a portion of the roads. DeepLabv3+ and D-LinkNet extracted relatively more roads attributed to the aggregation of multiscale features. Compared with these methods, the results extracted by MECA-Net were the most complete. The second row shows the image under the background of woodland. The road in the image is narrow and slender, which easily leads to missed detection. The results of other methods were missing in different degrees, but MECA-NET obtained the best extraction result for narrow and slender roads. The road in the image in the third row is seriously blocked by roadside trees and their shadows. The results of all methods were discontinuous to a certain extent, but the continuity of the results obtained by MECA-Net was the best. The fourth row shows a residential scene, where the shadows of buildings block the road. Other methods failed to identify the roads obscured by building shadows, and their results were all discontinuous and incomplete. In contrast, MECA-Net shows obvious advantages in solving this problem, and it maintained the continuity and integrity of the road topology. The roads in the images in the fifth and sixth rows are obscured by dense trees. In this case, MECA-Net dealt with the problem of result discontinuity and achieved better results by obtaining sufficient long-range context information.

The quantitative results of each network on the DeepGlobe dataset are presented in Table 1. The results show that MECA-Net improved the road extraction ability under the effect of the multiscale feature encoding module and the long-range context-aware module, and the results were better than those of other methods. Among all the comparison methods, MECA-Net had the highest results in IoU, Recall, and F1 (65.15%, 79.41%, and 78.90%, respectively), which are 0.69%, 1.88%, and 0.51% higher than those of the baseline network LinkNet34, respectively. Compared with other classical encoder–decoder networks such as UNet, the IoU, Recall, and F1 of MECA-Net improved by 0.93%, 3.25%, and 0.69%, respectively. Compared with NL-LinkNet, which only uses nonlocal operations to obtain long-range context in the spatial dimension, the IoU, Precision, Recall, and F1 of MECA-Net improved by 1.31%, 0.39%, 1.55%, and 0.97%, respectively. Compared with D-LinkNet, MECA-Net not only aggregated multiscale features but also obtained long-range context information from the channel and spatial dimensions, and its IoU, Recall, and F1 improved by 0.47%, 1.04%, and 0.35%, respectively. The IoU, Precision, Recall and F1 of MECA-Net

were 2.36%, 1.02%, 2.49% and 1.76% higher than those of RoadNet, and were 0.91%, 1.3%, 0.01% and 0.67% higher than those of PP-LinkNet, respectively.

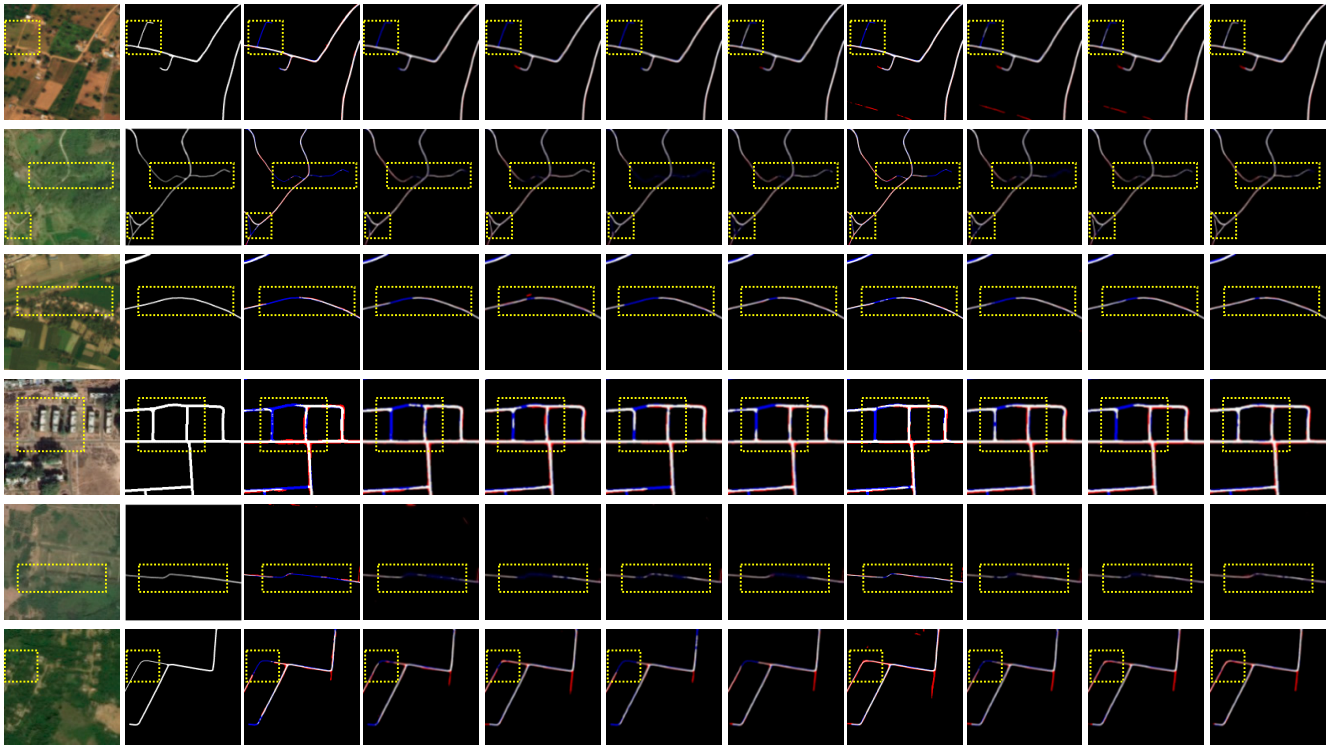


Figure 6. Comparison of visualization results between MECA-Net and other road extraction methods on the DeepGlobe dataset. (**a**) Image; (**b**) Label; (**c**) RoadNet results; (**d**) SegNet results; (**e**) NL-LinkNet results; (**f**) UNet results; (**g**) DeepLabv3+ results; (**h**) PP-LinkNet results; (**i**) LinkNet results; (**j**) D-LinkNet results; (**k**) MECA-Net results. White: true positive. Black: true negative. Blue: false negative. Red: false positive.

Experimental results on the Massachusetts dataset. As shown in Figure 7, the first row of images is a slender road in the suburban scene, and the MECA-Net proposed in this paper had better extraction results. The roads in the image of the second row are slender, annular, and not obvious, resulting in poor recognition results for each method. In contrast, MECA-Net had better extraction results. The road in the third-row image is seriously occluded, and the loss of road features is serious. However, the proposed method identified as many occluded roads as possible. Similarly, the trees in the images in the fourth, fifth and sixth rows significantly block the road. It was found that the extraction results of MECA-Net were more complete and continuous than those of other methods.

Table 1. Quantitative evaluation results of different networks on the DeepGlobe dataset.

Networks	IoU (%)	Precision (%)	Recall (%)	F1 (%)
RoadNet [38]	62.79	77.37	76.92	77.14
SegNet [59]	63.73	79.52	76.25	77.85
NL-LinkNet [39]	63.84	78.00	77.86	77.93
UNet [24]	64.22	**80.37**	76.16	78.21
DeepLabv3+ [60]	64.23	78.00	78.44	78.22
PP-LinkNet [37]	64.24	77.09	79.40	78.23
LinkNet [29]	64.46	79.27	77.53	78.39
D-LinkNet [28]	64.68	78.73	78.37	78.55
MECA-Net (ours)	**65.15**	78.39	**79.41**	**78.90**

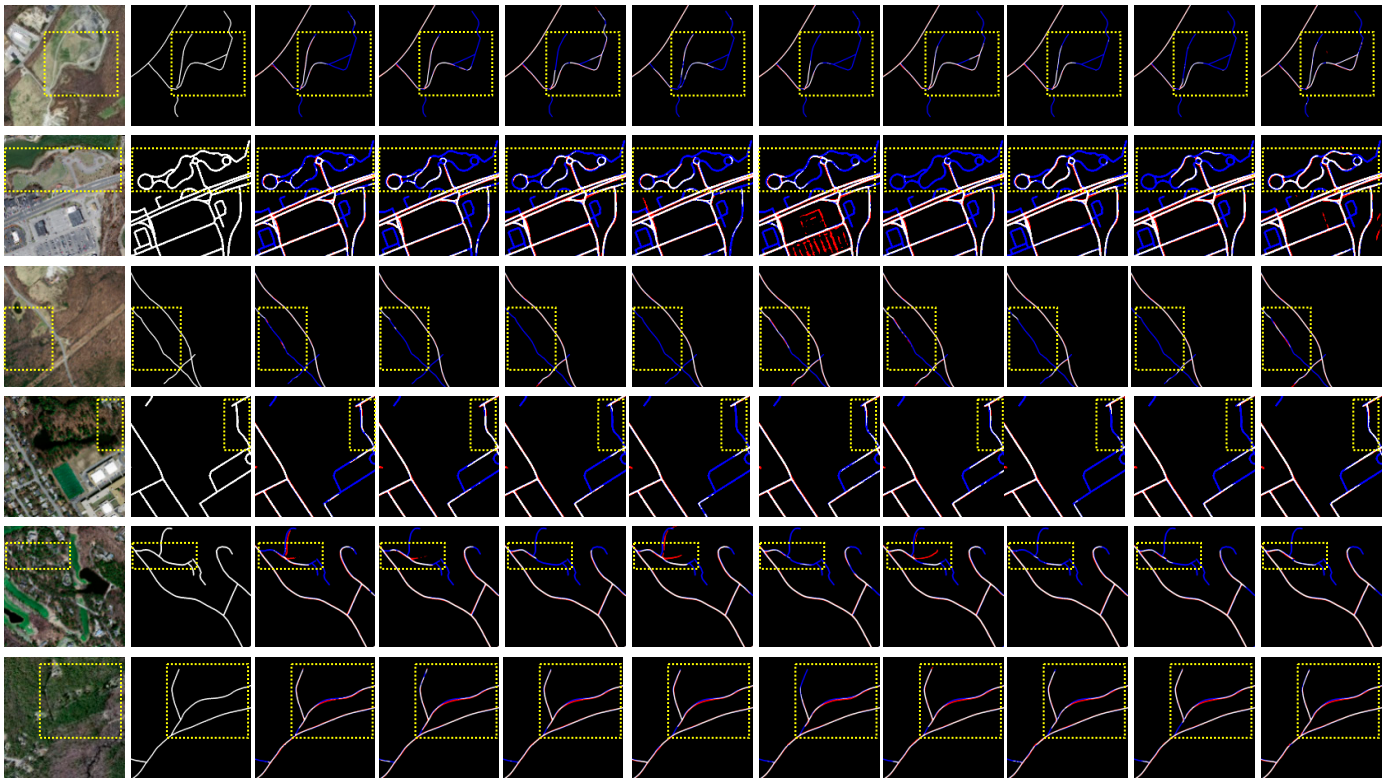


Figure 7. Comparison of visualization results between MECA-Net and other road extraction methods based on the Massachusetts dataset. (**a**) Image; (**b**) Label; (**c**) RoadNet results; (**d**) SegNet results; (**e**) NL-LinkNet results; (**f**) UNet results; (**g**) DeepLabv3+ results; (**h**) PP-LinkNet results; (**i**) LinkNet results; (**j**) D-LinkNet results; (**k**) MECA-Net results. White: true positive. Black: true negative. Blue: false negative. Red: false positive.

The quantitative results of each network on the Massachusetts dataset are presented in Table 2. The IoU and F1 of MECA-Net on the Massachusetts dataset were 65.82% and 79.39%, respectively, which are better than those of the other eight comparison methods. On the Massachusetts dataset, the IoU, Recall and F1 of MECA-Net were 0.39%, 0.68% and 0.29% higher than the baseline network LinkNet, respectively. MECA-Net outperformed RoadNet, NL-LinkNet, DeepLabv3+ and D-LinkNet in four metrics. The IoU, Recall and F1 of MECA-Net were 0.31%, 0.92% and 0.23 higher than SegNet, and 0.19%, 0.62% and 0.14% higher than UNet, respectively. The IoU, Precision and F1 of MECA-Net were 0.1%, 0.24% and 0.07% higher than PP-LinkNet, respectively.

Table 2. Quantitative evaluation results of different networks on the Massachusetts dataset.

Networks	IoU (%)	Precision (%)	Recall (%)	F1 (%)
RoadNet [38]	65.08	80.33	77.42	78.85
SegNet [59]	65.51	81.14	77.27	79.16
NL-LinkNet [39]	65.45	80.62	77.66	79.11
UNet [24]	65.63	81.00	77.57	79.25
DeepLabv3+ [60]	64.93	80.09	77.42	78.74
PP-LinkNet [37]	65.72	80.39	78.28	79.32
LinkNet [29]	65.43	80.77	77.51	79.10
D-LinkNet [28]	65.51	80.48	77.88	79.16
MECA-Net (ours)	65.82	80.63	78.19	79.39

4. Discussion

Section 4.1 discusses the ablation experiments of MECA-Net. Section 4.2 compares the number of parameters for different networks.

4.1. Ablation Study

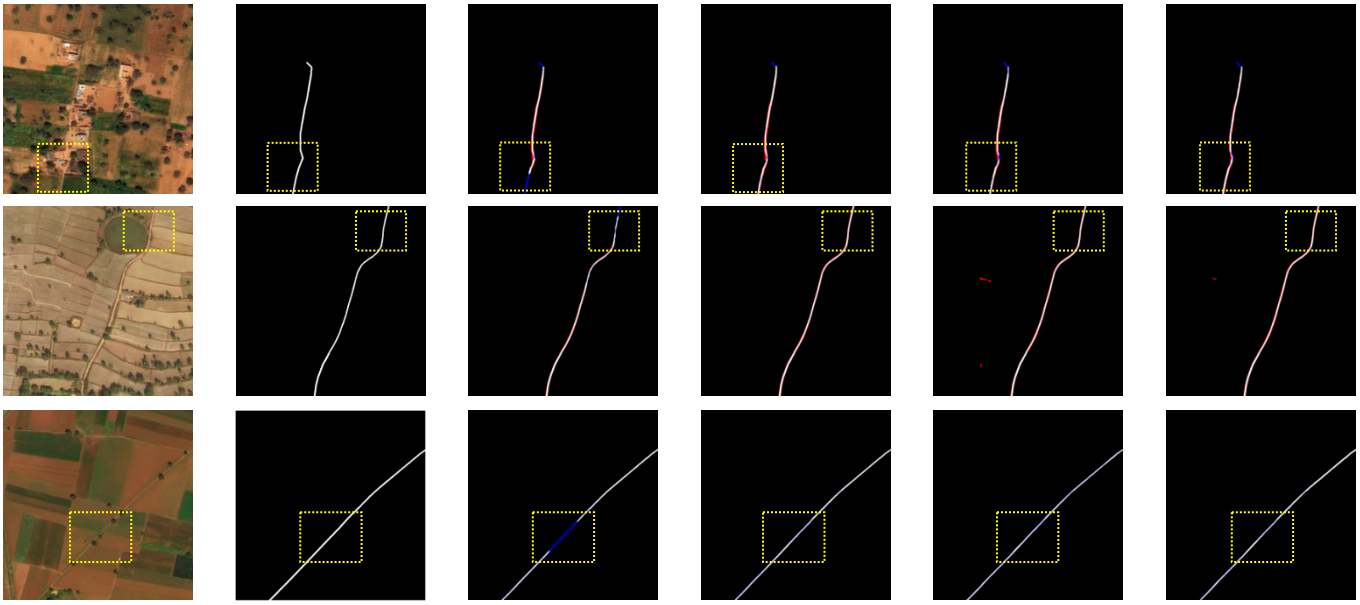
Ablation experiments were conducted for the designed network on the DeepGlobe dataset. Here, LinkNet34 was taken as the baseline, and each improvement module was gradually added on this basis to verify the effectiveness of each module in this method. The results of the ablation experiment are shown in Table 3

Table 3. Ablation experiments for the proposed method on the DeepGlobe dataset.

Networks	IoU (%)	Precision (%)	Recall (%)	F1 (%)
Baseline	64.46	79.27	77.53	78.39
Baseline + MFEM	64.91	79.21	78.23	78.72
Baseline + MFEM + CAM	65.00	78.44	79.13	78.79
Baseline + MFEM + CAM + SPM	65.15	78.39	79.41	78.90

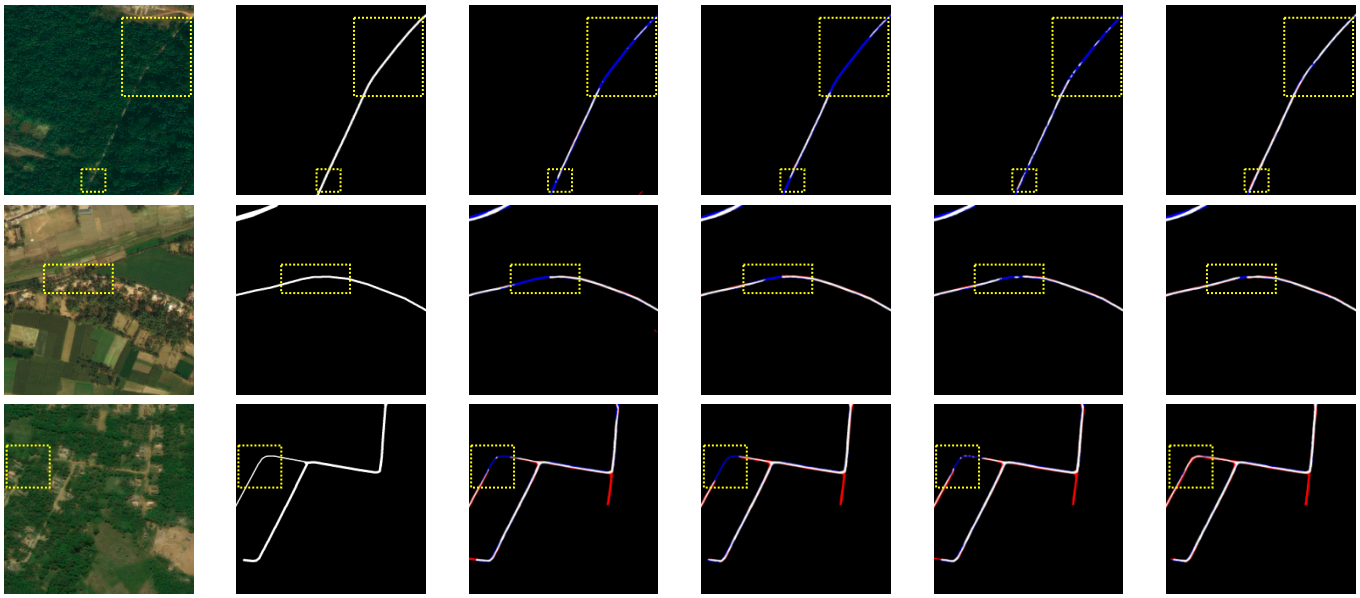
After adding the MFEM to the baseline, the IoU, Recall, and F1 were improved by 0.45%, 0.70%, and 0.33%, respectively, indicating that the MFEM improves the road extraction ability by aggregating multiscale features. The comparison of visualization results is presented in Figure 8. The results from the first row to the third row show that the baseline network recognized the slender road in the image after adding the MFEM, thus verifying the effectiveness of the MFEM.

After adding CAM on the basis of Baseline + MFEM, the IoU, Recall, and F1 were improved by 0.09%, 0.9%, and 0.07%, respectively, indicating the effectiveness of obtaining long-range context in the channel dimension to improve road extraction performance. Finally, adding SPM on the basis of Baseline + MFEM + CAM is equivalent to Baseline + MFEM + LCAM, which constitutes the MECA-NET. After adding SPM, the IoU, Recall, and F1 were improved by 0.15%, 0.28%, and 0.11%, respectively, indicating that the accuracy of road extraction can be improved by using SPM to obtain long-range context in the spatial dimension. As shown in Figure 9, the roads in the three rows of images all have the problem of occlusion. By comparison, the addition of CAM on the basis of Baseline + MFEM alleviated the occlusion problem, and the addition of SPM further improved the continuity of the extraction results of the occluded roads, which verifies the effectiveness of LCAM.



(a) (b) (c) (d) (e) (f)

Figure 8. Comparison of visualization results before and after adding MFEM. (**a**) Image; (**b**) Label; (**c**) Baseline results; (**d**) Baseline + MFEM results; (**e**) Baseline + MFEM + CAM results; (**f**) Baseline + MFEM + CAM + SPM results. White: true positive. Black: true negative. Blue: false negative. Red: false positive.



(a) (b) (c) (d) (e) (f)

Figure 9. Comparison of visualization results before and after adding LCAM. (**a**) Image; (**b**) Label; (**c**) Baseline results; (**d**) Baseline + MFEM results; (**e**) Baseline + MFEM + CAM results; (**f**) Baseline + MFEM + CAM + SPM results. White: true positive. Black: true negative. Blue: false negative. Red: false positive.

4.2. Number of Parameters for Each Network

The number of parameters is also an important indicator of network performance. The more parameters a network has, the larger the computation requirements. The F1 and parameter number of different networks on the DeepGlobe dataset are shown in Figure 10.

It can be seen that RoadNet had the least number of parameters (14.72 million). D-linkNet and MECA-Net had more parameters, i.e., 31.10 million and 47.41 million, respectively. The number of parameters of MECA-Net was 25.77 million more than that of LinkNet, and this is because four multiscale feature encoding modules and four long-range context-aware modules were added on the basis of LinkNet. The MECA-NET achieved the highest F1 score, but it also had the largest number of parameters and needed to occupy more memory. The follow-up study will investigate how to improve the accuracy without significantly increasing the number of parameters.

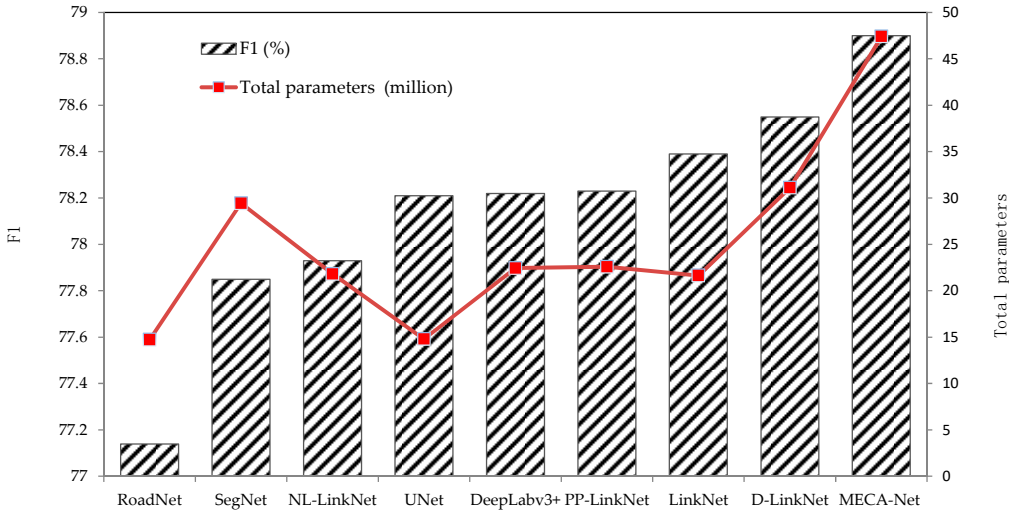


Figure 10. Comparison of the F1 value and parameter number of different networks on the DeepGlobe dataset. The bar chart and line chart represent the F1 value and the number of parameters of each network, respectively.

5. Conclusions

In this paper, a new road extraction network called MECA-Net is proposed to solve the problems of the slender road being difficult to identify and the road being obscured by the shadow of trees and buildings. MECA-Net uses the multiscale feature encoding module to extract multiscale road features, which improves the recognition ability of the model for slender roads. Meanwhile, MECA-Net uses the long-range context-aware module to obtain sufficient long-range context information from the channel and spatial dimensions, which alleviates the problem of road occlusion and improves the continuity and integrity of road extraction results. The experimental results on the open DeepGlobe dataset and Massachusetts dataset show that MECA-Net is superior to the other eight mainstream road extraction methods. The limitation of MECA-Net is that the number of parameters is relatively large, which will affect the processing efficiency. In future work, depthwise separable convolution [61] will be introduced to reduce the number of network parameters and improve the computation speed. In addition, future work will also investigate the application of the proposed method to remote sensing images from other data sources, using transfer learning-based techniques.

Author Contributions: Y.J. and K.X. designed the network and wrote the paper; Y.J. and W.T. carried out the experiments; Y.J. and C.Y. analyzed the experimental results; C.J. and X.C. processed the dataset; H.H. and A.Y. revised the paper and provided valuable suggestions. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Natural Science Foundation of China (42050202) and the CAST Young Elite Foundation.

Data Availability Statement: Not applicable.

Acknowledgments: We would like to thank the anonymous reviewers for their valuable comments and suggestions.

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

References

1. Bong, D.B.L.; Lai, K.C.; Joseph, A. Automatic road network recognition and extraction for urban planning. *Int. J. Appl. Sci. Eng. Technol.* **2009**, *5*, 209–215.
2. Hinz, S.; Baumgartner, A.; Ebner, H. Modeling contextual knowledge for controlling road extraction in urban areas. In Proceedings of the IEEE/ISPRS Joint Workshop on Remote Sensing and Data Fusion over Urban Areas, Rome, Italy, 8–9 November 2001; pp. 40–44. [CrossRef]
3. Xu, Y.; Xie, Z.; Feng, Y.; Chen, Z. Road extraction from high-resolution remote sensing imagery using deep learning. *Remote Sens.* **2018**, *10*, 1461. [CrossRef]
4. Li, Y.; Guo, L.; Rao, J.; Xu, L.; Jin, S. Road segmentation based on hybrid convolutional network for high-resolution visible remote sensing image. *IEEE Geosci. Remote Sens. Lett.* **2018**, *16*, 613–617. [CrossRef]
5. Hormese, J.; Saravanan, C. Automated road extraction from high resolution satellite images. *Procedia Technol.* **2016**, *24*, 1460–1467. [CrossRef]
6. Zhang, Z.; Liu, Q.; Wang, Y. Road extraction by deep residual u-net. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 749–753. [CrossRef]
7. Ma, H.; Lu, N.; Ge, L.; Li, Q.; You, X.; Li, X. Automatic road damage detection using high-resolution satellite images and road maps. In Proceedings of the 2013 IEEE International Geoscience and Remote Sensing Symposium, Melbourne, VIC, Australia, 21–26 July 2013; pp. 3718–3721. [CrossRef]
8. Li, Q.; Zhang, J.; Wang, N. Damaged road extraction from post-seismic remote sensing images based on gis and object-oriented method. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium, Beijing, China, 10–15 July 2016; pp. 4247–4250. [CrossRef]
9. Miao, Z.; Wang, B.; Shi, W.; Zhang, H. A semi-automatic method for road centerline extraction from VHR images. *IEEE Geosci. Remote Sens. Lett.* **2014**, *11*, 1856–1860. [CrossRef]
10. Sghaier, M.O.; Lepage, R. Road extraction from very high resolution remote sensing optical images based on texture analysis and beamlet transform. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *9*, 1946–1958. [CrossRef]
11. Mu, H.; Zhang, Y.; Li, H.; Guo, Y.; Zhuang, Y. Road extraction base on Zernike algorithm on SAR image. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; pp. 1274–1277. [CrossRef]
12. Liu, J.; Qin, Q.; Li, J.; Li, Y. Rural road extraction from high-resolution remote sensing images based on geometric feature inference. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 314. [CrossRef]
13. Yu, C.; Yi, Y. Object-based road extraction in remote sensing image using Markov random field. *Geomat. Inf. Sci. Wuhan Univ.* **2011**, *36*, 544–547. (In Chinese)
14. Huang, Z.; Xu, F.; Lu, L.; Nie, H. Object-based conditional random fields for road extraction from remote sensing image. *IOP Conf. Ser. Earth Environ. Sci.* **2014**, *17*, 012276. [CrossRef]
15. Li, M.; Stein, A.; Bijker, W.; Zhan, Q. Region-based urban road extraction from VHR satellite images using binary partition tree. *Int. J. Appl. Earth Obs. Geoinf.* **2016**, *44*, 217–225. [CrossRef]
16. Maboudi, M.; Amini, J.; Malihi, S.; Hahn, M. Integrating fuzzy object based image analysis and ant colony optimization for road extraction from remotely sensed images. *ISPRS J. Photogramm. Remote Sens.* **2018**, *138*, 151–163. [CrossRef]
17. Xue, W.; Zhao, L.; Wang, P. Satellite-earth Coordinated On-orbit Intelligent Target Recognition of Optical Remote Sensing Images. *Spacecr. Recovery Remote Sens.* **2021**, *42*, 128–137. [CrossRef]
18. Liu, W.; Nie, Y.; Chen, X.; Li, J.; Zhao, L.; Zheng, F.; Han, Y.; Liu, S. Deep Learning Method in Complex Scenes Luminous Ship Target Detection. *Spacecr. Recovery Remote Sens.* **2022**, *43*, 124–137. [CrossRef]
19. Zhang, Y.; Han, X.; Zhang, S.; Gao, W. Rapid Detection of Airport Targets Based on Visual Saliency and Convolutional Neural Network. *Spacecr. Recovery Remote Sens.* **2021**, *42*, 117–127. [CrossRef]
20. Chen, Z.; Deng, L.; Luo, Y.; Li, D.; Junior, J.M.; Gonçalves, W.N.; Awal Md Nurunnabi, A.; Li, J.; Wang, C.; Li, D. Road extraction in remote sensing data: A survey. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *112*, 102833. [CrossRef]
21. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2015, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440. [CrossRef]
22. Cheng, G.; Wang, Y.; Xu, S.; Wang, H.; Xiang, S.; Pan, C. Automatic road detection and centerline extraction via cascaded end-to-end convolutional neural network. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3322–3337. [CrossRef]
23. Buslaev, A.; Seferbekov, S.; Iglovikov, V.; Shvets, A. Fully convolutional network for automatic road extraction from satellite imagery. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops 2018, Salt Lake City, UT, USA, 18–22 June 2018; pp. 207–210. [CrossRef]
24. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention 2015, Munich, Germany, 5–9 October 2015; pp. 234–241. [CrossRef]

25. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
26. Xin, J.; Zhang, X.; Zhang, Z.; Fang, W. Road extraction of high-resolution remote sensing images derived from DenseUNet. *Remote Sens.* **2019**, *11*, 2499. [CrossRef]
27. Gao, X.; Sun, X.; Zhang, Y.; Yan, M.; Xu, G.; Sun, H.; Jiao, J.; Fu, K. An end-to-end neural network for road extraction from remote sensing imagery by multiple feature pyramid network. *IEEE Access* **2018**, *6*, 39401–39414. [CrossRef]
28. Zhou, L.; Zhang, C.; Wu, M. D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops 2018, Salt Lake City, UT, USA, 18–22 June 2018; pp. 182–186. [CrossRef]
29. Chaurasia, A.; Culurciello, E. Linknet: Exploiting encoder representations for efficient semantic segmentation. In Proceedings of the 2017 IEEE Visual Communications and Image Processing (VCIP), St. Petersburg, FL, USA, 10–13 December 2017; pp. 1–4. [CrossRef]
30. He, H.; Yang, D.; Wang, S.; Wang, S.; Li, Y. Road extraction by using atrous spatial pyramid pooling integrated encoder-decoder network and structural similarity loss. *Remote Sens.* **2019**, *11*, 1015. [CrossRef]
31. Lu, X.; Zhong, Y.; Zheng, Z.; Liu, Y.; Zhao, J.; Ma, A.; Yang, J. Multi-scale and multi-task deep learning framework for automatic road extraction. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 9362–9377. [CrossRef]
32. Liu, Z.; Wang, M.; Wang, F.; Ji, X. A residual attention and local context-aware network for road extraction from high-resolution remote sensing imagery. *Remote Sens.* **2021**, *13*, 4958. [CrossRef]
33. Gao, L.; Song, W.; Dai, J.; Chen, Y. Road extraction from high-resolution remote sensing imagery using refined deep residual convolutional neural network. *Remote Sens.* **2019**, *11*, 552. [CrossRef]
34. Zhang, Z.; Wang, Y. JointNet: A common neural network for road and building extraction. *Remote Sens.* **2019**, *11*, 696. [CrossRef]
35. Zhao, Z.; Zhou, Z.; Huang, X.; Yang, Z. MRENet: Simultaneous extraction of road surface and road centerline in complex urban scenes from very high-resolution images. *Remote Sens.* **2021**, *13*, 239. [CrossRef]
36. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890. [CrossRef]
37. Tran, A.; Zonoozi, A.; Varadarajan, J.; Kruppa, H. Pp-linknet: Improving semantic segmentation of high resolution satellite imagery with multi-stage training. In Proceedings of the 2nd Workshop on Structuring and Understanding of Multimedia heritAge Contents, Seattle, WA, USA, 12 October 2020; pp. 57–64. [CrossRef]
38. Liu, Y.; Yao, J.; Lu, X.; Xia, M.; Wang, X.; Liu, Y. RoadNet: Learning to comprehensively analyze road networks in complex urban scenes from high-resolution remotely sensed images. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 2043–2056. [CrossRef]
39. Wang, Y.; Seo, J.; Jeon, T. NL-LinkNet: Toward lighter but more accurate road extraction with nonlocal operations. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]
40. Zhu, Q.; Zhang, Y.; Wang, L.; Zhong, Y.; Guan, Q.; Lu, X.; Zhang, L.; Li, D. A global context-aware and batch-independent network for road extraction from VHR satellite imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *175*, 353–365. [CrossRef]
41. Xie, Y.; Miao, F.; Zhou, K.; Peng, J. HsgNet: A road extraction network based on global perception of high-order spatial information. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 571. [CrossRef]
42. Wu, M.; Zhang, C.; Liu, J.; Zhou, L.; Li, X. Towards accurate high resolution satellite image semantic segmentation. *IEEE Access* **2019**, *7*, 55609–55619. [CrossRef]
43. Lin, Y.; Xu, D.; Wang, N.; Shi, Z.; Chen, Q. Road extraction from very-high-resolution remote sensing images via a nested SE-Deeplab model. *Remote Sens.* **2020**, *12*, 2985. [CrossRef]
44. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141. [CrossRef]
45. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; et al. Attention u-net: Learning where to look for the pancreas. *arXiv* **2018**, arXiv:1804.03999.
46. Li, J.; Liu, Y.; Zhang, Y.; Zhang, Y. Cascaded attention DenseUNet (CADUNet) for road extraction from very-high-resolution images. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 329. [CrossRef]
47. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2015, Boston, MA, USA, 7–12 June 2015; pp. 1–9. [CrossRef]
48. Gao, S.; Cheng, M.; Zhao, K.; Zhang, X.; Yang, M.; Torr, P. Res2net: A new multi-scale backbone architecture. *IEEE Trans. Pattern Anal. Mach. Intel.* **2019**, *43*, 652–662. [CrossRef]
49. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the 32nd International Conference on Machine Learning ICML 2015, Lille, France, 6–11 July 2015; pp. 448–456.
50. Glorot, X.; Bordes, A.; Bengio, Y. Deep sparse rectifier neural networks. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA, 11–13 April 2011; pp. 315–323.
51. Woo, S.; Park, J.; Lee, J.; Kweon, I. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision 2018, Munich, Germany, 8–14 September 2018; pp. 3–19.
52. Hou, Q.; Zhang, L.; Cheng, M.; Feng, J. Strip pooling: Rethinking spatial pooling for scene parsing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2020, Seattle, WA, USA, 13–19 June 2020; pp. 4003–4012. [CrossRef]

53. Milletari, F.; Navab, N.; Ahmadi, S.A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision, Stanford, CA, USA, 25–28 October 2016; pp. 565–571. [CrossRef]
54. Demir, I.; Koperski, K.; Lindenbaum, D.; Pang, G.; Huang, J.; Basu, S.; Hughes, F.; Tuia, D.; Raskar, R. Deepglobe 2018: A challenge to parse the earth through satellite images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops 2018, Salt Lake City, UT, USA, 18–22 June 2018; pp. 172–181. [CrossRef]
55. Mnih, V.; Hinton, G.E. Learning to detect roads in high-resolution aerial images. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 210–223.
56. Singh, S.; Batra, A.; Pang, G.; Torresani, L.; Basu, S.; Paluri, M.; Jawahar, C. Self-Supervised feature learning for semantic segmentation of overhead imagery. In Proceedings of the 2018 BMVC British Machine Vision Conference, Newcastle, UK, 3–6 September 2018; Volume 1, p. 4.
57. Mei, J.; Li, R.; Gao, W.; Cheng, M. CoANet: Connectivity attention network for road extraction from satellite imagery. *IEEE Trans. Image Process.* **2021**, *30*, 8540–8552. [CrossRef]
58. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. In Proceedings of the Annual Conference on Neural Information Processing Systems 2019, Vancouver, BC, Canada, 8–14 December 2019.
59. 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]
60. 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 2018, Munich, Germany, 8–14 September 2018; pp. 801–818.
61. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520. [CrossRef]