sensors

MDPI

Article

Dual Crisscross Attention Module for Road Extraction from Remote Sensing Images

Chuan Chen [1], Huilin Zhao [2], Wei Cui [2,*] and Xin He [2]

[1] TUM Department of Aerospace and Geodesy, Technical University of Munich, 80333 Munich, Germany; chuan.chen@tum.de
[2] School of Resources and Environmental Engineering, Wuhan University of Technology, Wuhan 430070, China; zhaohl2016@whut.edu.cn (H.Z.); 2962575697@whut.edu.cn (X.H.)
* Correspondence: cuiwei@whut.edu.cn; Tel.: +86-136-2860-8563

Abstract: Traditional pixel-based semantic segmentation methods for road extraction take each pixel as the recognition unit. Therefore, they are constrained by the restricted receptive field, in which pixels do not receive global road information. These phenomena greatly affect the accuracy of road extraction. To improve the limited receptive field, a non-local neural network is generated to let each pixel receive global information. However, its spatial complexity is enormous, and this method will lead to considerable information redundancy in road extraction. To optimize the spatial complexity, the Crisscross Network (CCNet), with a crisscross shaped attention area, is applied. The key aspect of CCNet is the Crisscross Attention (CCA) module. Compared with non-local neural networks, CCNet can let each pixel only perceive the correlation information from horizontal and vertical directions. However, when using CCNet in road extraction of remote sensing (RS) images, the directionality of its attention area is insufficient, which is restricted to the horizontal and vertical direction. Due to the recurrent mechanism, the similarity of some pixel pairs in oblique directions cannot be calculated correctly and will be intensely dilated. To address the above problems, we propose a special attention module called the Dual Crisscross Attention (DCCA) module for road extraction, which consists of the CCA module, Rotated Crisscross Attention (RCCA) module and Self-adaptive Attention Fusion (SAF) module. The DCCA module is embedded into the Dual Crisscross Network (DCNet). In the CCA module and RCCA module, the similarities of pixel pairs are represented by an energy map. In order to remove the influence from the heterogeneous part, a heterogeneous filter function (HFF) is used to filter the energy map. Then the SAF module can distribute the weights of the CCA module and RCCA module according to the actual road shape. The DCCA module output is the fusion of the CCA module and RCCA module with the help of the SAF module, which can let pixels perceive local information and eight-direction non-local information. The geometric information of roads improves the accuracy of road extraction. The experimental results show that DCNet with the DCCA module improves the road IOU by 4.66% compared to CCNet with a single CCA module and 3.47% compared to CCNet with a single RCCA module.

Keywords: remote sensing; semantic segmentation; road extraction; attention mechanism; geometric information; directionality

Citation: Chen, C.; Zhao, H.; Cui, W.; He, X. Dual Crisscross Attention Module for Road Extraction from Remote Sensing Images. *Sensors* **2021**, *21*, 6873. https://doi.org/10.3390/s21206873

Academic Editors: Chiman Kwan and Gwanggil Jeon

Received: 17 August 2021
Accepted: 14 October 2021
Published: 16 October 2021

1. Introduction

Road extraction has become a popular topic as a branch subject of semantic segmentation, and many complicated deep learning methods have been developed [1–3] and improved continuously to pursue a higher accuracy.

Some researchers have focused on the loss function. He et al. [3] proposed an encoder-decoder network model with a special loss function, which was optimized by road structure constraints. This method improved the detection of road coherence to a certain extent. In addition, many deep learning models based on the fully convolution network (FCN) have

been proposed [1,4]. Some researchers have analysed the effective area of the receptive field in traditional convolutional neural networks (CNNs). The active receptive field presents a Gaussian distribution, which will lead to difficulties for each pixel in obtaining contextual information [5]. These road extraction methods were generally at the pixel level, which would lead to local receptive area limits; therefore, each pixel will not receive enough global information [6]. However, the roads in remote sensing images have multidirectional characteristics. Due to the lack of perception of global information, the geometric information of the road has not been fully utilized.

To solve the poor receptive field problem, many constructions have been proposed to expand the receptive field range. Wang et al. [2] proposed the non-local neural network, which calculates the feature similarity between each pixel and other pixels on the feature map. In this way, the global information can be integrated into the centre pixel. However, this method needs to calculate the relationship of each pixel pair on the feature map, which will lead to a high computational complexity. In specific semantic segmentation problems, it will generate information redundancy. Some researchers have focused on the convolution process and produced dilated convolutions to enlarge the receptive field [7]. However, the dilated convolution has some flaws. When using dilated convolutions, the effective receptive field of the pixels at the edge of the feature map is quite different from that of the pixels at the centre. The uncertainty of the effective receptive field will make it untargeted to solve semantic segmentation problems.

Following the idea of reducing computational cost and information redundancy, Huang et al. [8] generated the CCNet by using a crisscross shape attention module, which can expand the receptive field in the vertical and horizontal directions at the same time. Using a recurrent mechanism, CCNet can let each pixel receive global information with a relatively low computational cost. However, this mechanism will also lead to a recurrent dilemma, which means that the similarity calculation of two pixels will be influenced by the intermediate pixel. In extreme cases, the heterogeneity of the intermediate nodes will reduce the similarity of homogeneous pixels in the oblique direction.

For example, there are two pixels P1 and P2 in the Figure 1:

Figure 1. Recurrent dilemma.

P1 and P2 are two points on the road, while S1 and S2 are two points on the non-road part. S1 and S2 are in the vertical and horizontal directions of P1 and P2. Using a recurrent mechanism [8], the relation between P1 and P2 needs to contain the transitions in S1 and S2, which can be described as P1→S1→P2 and P1→S2→P2. In the similarity calculation, by processing the feature vector, the similarity of P1 and S1 is α_1. The similarities of S1 and P2, P1 and S2, S1 and P2 are α_2, α_3, and α_4, respectively. The similarity of P1 and P2, which is directly calculated by the feature vectors, is α_5. Due to the heterogeneity of S1 and P1, α_1 is relatively smaller than α_5. Moreover, α_2, α_3, and α_4 are also relatively smaller than α_5 for the heterogeneity of corresponding points. In the recurrent mechanism, the similarity of P1 and P2 will be $\alpha_1 \times \alpha_2 + \alpha_3 \times \alpha_4$, which can be regarded as the second-order small amount compared to α_5, which is why the recurrent mechanism will lead to the distortion of similarity. In the methodology section, we will illustrate this in detail by the formula.

Consequently, the key shortcoming of CCNet is that the crisscross shape attention module has poor directionality when extracting multidirectional objects, especially roads. Considering the multidirectionality of roads, we add two extra attention areas, with one line in the 45° direction and another line in the 135° direction, to let each pixel receive more directional information. These two lines are called the RCCA module. The CCA module and RCCA module can let each pixel receive contextual information from eight directions. The computational cost of RCCA module is two-thirds of the computational cost of CCA module.

In the attention process, the generation of weights is an important step, where the activation function is often used. Some researchers choose the sigmoid function to output the final weights in the spatial attention module [9]. The activation function can be regarded as a selection of the feature. According to the requirements of semantic segmentation work, we proposed an HFF, which can largely reduce the influence of the heterogeneous part in the attention area.

Then we proposed a SAF module that can distribute the weight of the CCA module and RCCA module according to the specific direction of roads in images. The combination of the SAF module, CCA module and RCCA module is called the DCCA module. The computational cost of DCCA module is five-thirds of the computational cost of CCA module.

The innovations are based on the following aspects.

In the road extraction area, we proposed the RCCA module, which expands the attention area to pixels in the oblique direction for the first time. Consequently, it can solve the recurrent dilemma in DCCA module was designed by distributing the weight of the CCA module and RCCA module according to the specific road shape. The DCCA module can let each pixel receive contextual information from eight directions and not only local pixel information from convolutions. Therefore, it can be regarded as the combination of eight directional nonlocal attention mechanisms and the local convolution mechanism.

A heterogeneous filter function is created to suppress the influence of heterogeneous regions in the attention process. Processed by the heterogeneous filter function, each pixel can largely receive contextual information from homogeneous areas, which can promote the extraction accuracy.

In the following sections, we will introduce the related work, methodology, experiment, and conclusion. In the related work section, we will give an overview of the research content related to our work. Then, in the methodology section, we will give a detailed description of the implementation method of the DCCA module and the grafted network. In the experimental section, we will show the advantages of the DCCA module on the attention of road directionality through experimental analysis.

2. Related Work

2.1. Semantic Segmentation Based on Deep Learning

With the advent of LeNet in 1998, convolutional neural networks (CNNs) began to be widely used in image information processing [10]. The main elements of CNNs were

also determined at this time, including convolutional layer, pooling layer, fully connected layer, etc. Researchers designed AlexNet and let people see the potential of CNN in image processing for the first time in the ImageNet competition in 2012 [11]. Since then, the image processing capabilities have been transferred to semantic segmentation, and various networks, including different architectures, exist in the semantic segmentation stage.

The deep learning algorithm of semantic segmentation affixes class labels to each pixel. That is, the general workflow can be regarded as the interpretation of the pixels. After years of development, there are many prominent semantic segmentation neural networks, such as FCN [4], U-Net [12], SegNet [13], and DeepLab [14]. However, FCNs and other networks based on the CNN structure limit the range of the receptive field and can only obtain short-range context information. Although the traditional deep convolutional neural network obtains global context information by superimposing multiple convolutions, related studies have shown that the actual perception range of this method is smaller than the theoretical expected value [6]. Some researchers have found that this kind of receptive field has an irregular Gaussian distribution around the central pixel [5]. Consequently, the long-distance dependency relation inside the sample cannot be processed properly. To address this problem, Chen et al. proposed the Atrous Spatial Pyramid Pooling (ASPP) module with multiscale dilated convolution to integrate context information [14–16]. On this basis, Zhao et al. further proposed Pyramid Scene Parsing Network (PSPNet) with a pyramid pooling module to capture contextual information [17]. These kinds of methods based on dilated convolution still have the deflection that they obtain information from a small number of surrounding points and cannot form a dense context information structure. At the same time, methods based on pooling lose too much spatial information and thus cannot effectively meet the pixel-by-pixel classification requirements of semantic segmentation. To effectively obtain the global context information of the pixel, PSPNet learns to summarize the contextual information of each pixel by the predicting attention map [18]. A non-local network uses a self-attention mechanism to enable each pixel to perceive the features of pixels at all other locations, which can produce more powerful pixel-level characterization capabilities [2].

With the advancement of the attention mechanism in the application of semantic segmentation, crisscross attention [8], a very prominent attention method, was proposed. Crisscross attention proposed measures in view of the large amount of calculation and low efficiency of non-local networks by using a crisscross shape attention module, which can expand the receptive field in the vertical and horizontal directions at the same time. However, it also has a certain problem that the attention directionality is limited. When solving the semantic segmentation problem, an attention area, which can help centre pixels obtain contextual information from eight directions with automatic weights, is more appropriate to interpret objects with complex directionality.

2.2. Attention Mechanism and Its Implementation in CNN

The attention mechanism has been widely used in natural language processing and computer vision [19,20]. The attention mechanism in computer vision simulates the human recognition process of an image, which means that the perception system does not process the entire scene at once but puts attention to certain specific parts to obtain the information with high priority. The priority of these parts is selected by preset preferences in the human brain, such as for colour, shape, and characteristics.

Some researchers have implemented attention mechanisms in image captioning tasks. They proposed soft attention and hard attention architectures with a visualization method of the attention area. Researchers at Google [21] first proposed the transformer structure based on self-attention and the multihead self-attention structure. The key, query, and value of self-attention are output by sequence-to-sequence type, which becomes the basis of subsequent attention research. Wang et al. [2] proposed a nonlocal neural network, which can remove the local receptive field limitation of convolution and capture global information effectively. Hu et al. proposed a channel attention mechanism, using global

average pooling and full connection to focus on the attention weight of channel feature extraction [22]. Some researchers have proposed the convolutional block attention module (CBAM), which can carry out spatial attention and channel attention to the feature map at the same time [9]. The embed type of the CBAM is series connection. Based on this, Fu et al. used parallel connections to embed spatial attention and channel attention in neural networks [23].

These attention methods can be summarized as self-attention families. Generally, the self-attention mechanism can capture the spatial dependence of any two positions in the feature map and obtain global context information, thereby greatly improving the performance of the semantic segmentation network [24,25].

2.3. Attention Mechanism and Its Implementation in CNN

In recent years, a variety of methods have been proposed to extract roads from remote sensing images. These methods can be generally divided into two categories: road area extraction and road centerline extraction. Road area extraction [26–31] can generate pixel-level markers of roads, and the purpose of road centerline extraction [32–35] is to detect the skeleton of the road.

Zhang et al. first applied a support vector machine (SVM) to the road extraction of remote sensing images based on edge detection [36]. Song et al. proposed a method using shape index features and support vector machines (SVMs), which put geometric features into consideration for the first time [37]. Based on this, researchers use salient features to design a multilevel framework, which can extract roads from high-resolution multispectral images [38].

With the development of deep learning, road extraction methods based on deep learning have shown better performance than non-deep learning methods. Researchers have proposed a method to detect road areas from high-resolution aerial images using restricted Boltzmann machines, which first implemented deep learning tools [27]. Compared to this method, researchers have used CNNs to extract roads and buildings and obtain better results [30]. Alvarez et al. [39] proposed an automatic road extraction method based on U-Net. Zhong et al. [40] proposed a semantic segmentation neural network that combines the advantages of residual learning and U-Net for road extraction, which simplified training and achieved better results with fewer parameters.

Zhang et al. [41] used D-Link-Net and DenseNet for high-resolution satellite image road extraction. Based on this, Peng et al. [42] proposed a multiscale enhanced road detection framework (Dense-U-Net) based on densely connected convolutional networks (Dense-Net) and U-Net, which can effectively perform feature learning and retain finer spatial details.

Generally, the method of road extraction based on deep learning focuses more on the use of road features. These methods lack attention to geometric information, such as directionality. The DCCA module fills this vacancy.

3. Methodology

In the methodology section, we will first provide a description of the framework. Then, we will introduce the implementation details of the CCA module and the grafted network. The design of the RCCA module will be described in the RCCA part. In heterogeneous filter function and output part, the HFF and SAF module will be introduced.

3.1. Framework

The neural network is based on the DCCA module, which is called DCNet. DCNet consists of three main parts: backbone part, attention part and output part. The backbone part extracts the feature by traditional convolutions, which supports obtaining the correlation between pixels in different directions. The attention part consists of the CCA module and RCCA module, which can constitute an eight-direction nonlocal attention mechanism. HFF plays an important role in the energy process inside the attention part. In

the output part, the SAF module can distribute the weights of these two modules according to the energy distribution in the sampling area and fuse the correlation information from eight directions. Considering the use of local 3×3 convolutions, the output part can realize the combined acquisition of 3×3 local information and eight directions of nonlocal information. The network structure is shown in the Figure 2.

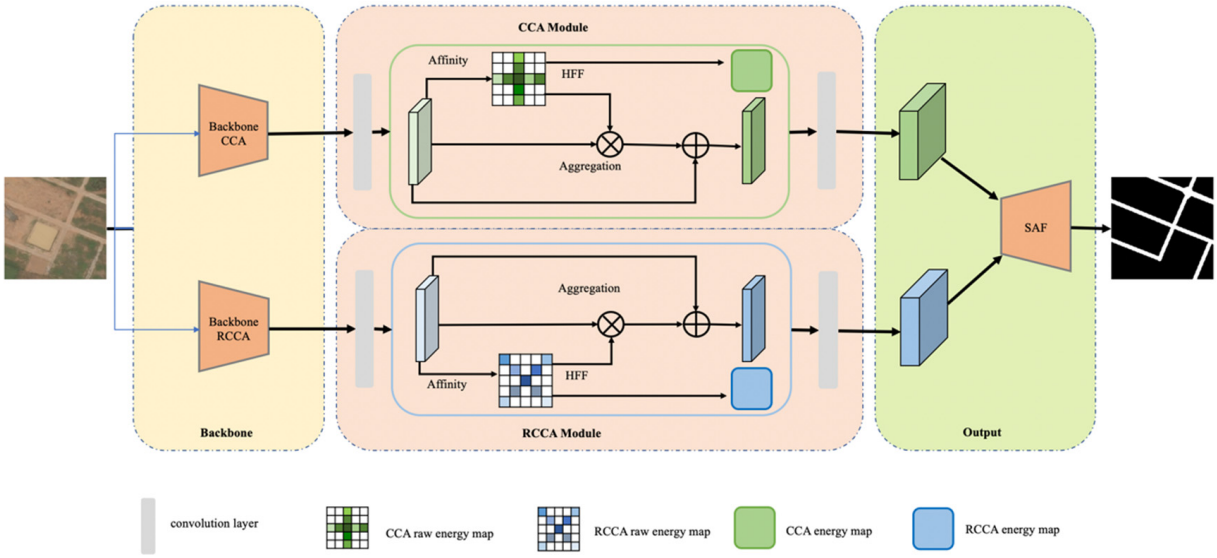

Figure 2. Network Structure.

3.1.1. Backbone

The backbone part consists of two parallel backbones that are used for feature extraction, which are based on the residual network. Each backbone is shown in the Figure 3.

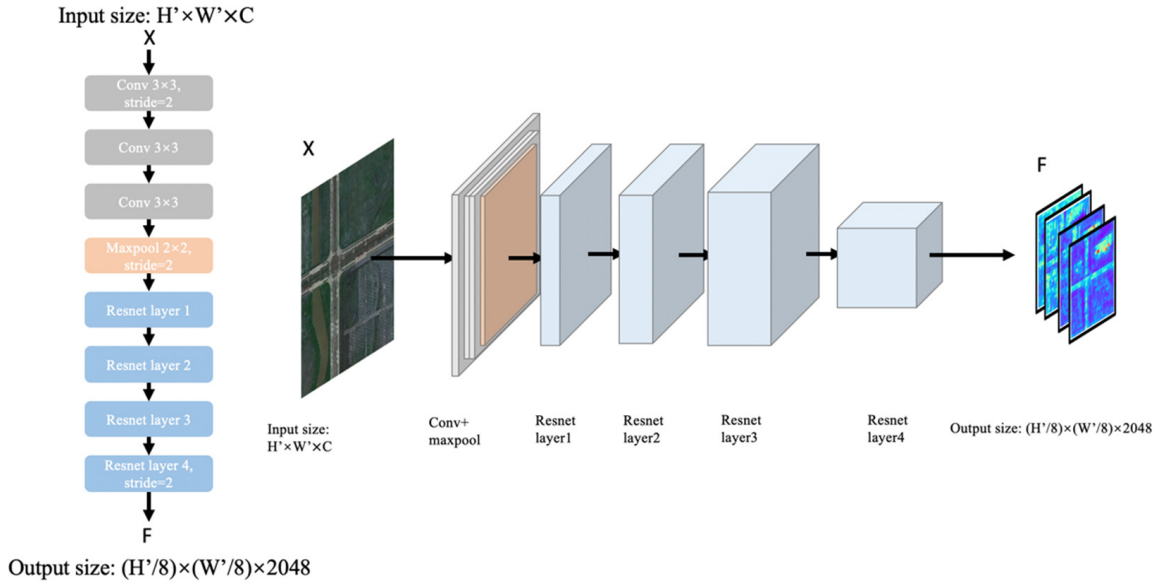

Figure 3. The backbone part.

$X \in \mathbb{R}^{H' \times W' \times C}$ is input at the beginning of the network. The first process is three convolutions and one max-pooling, and then it comes to the feature extraction based on ResNet101. After three downsamplings, the output $F \in \mathbb{R}^{(H'/8) \times (W'/8) \times C}$ is generated.

This process can be expressed by the following formula:

$$F = \begin{cases} f_{CCA}(\ X\ ), & CCA\ Backbone \\ f_{RCCA}(\ X\ ), & RCCA\ Backbone \end{cases} \tag{1}$$

3.1.2. The Implementation of CCA Module

The implementation of the CCA is generally simple compared to that of the RCCA. Therefore, it will be shown firstly in Figure 4.

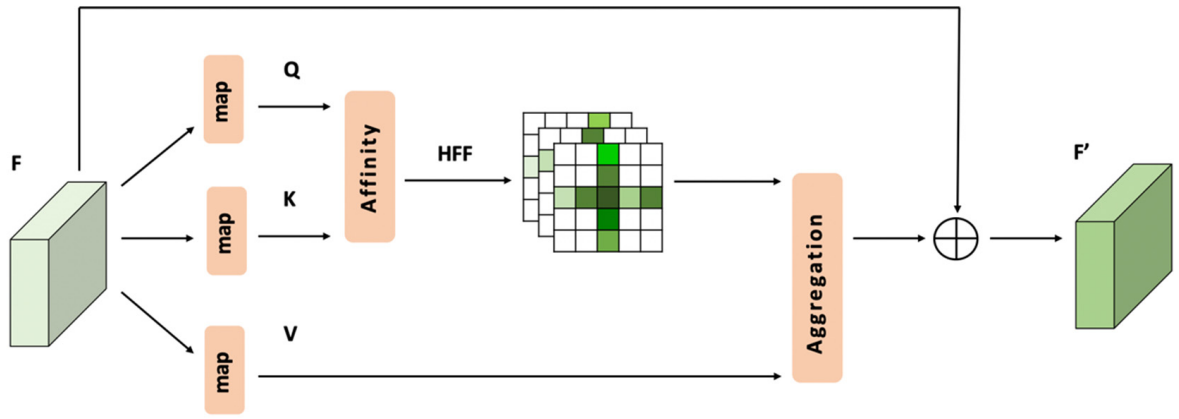

Figure 4. CCA module.

This attention method is from Huang et al. [8], which is used as the CCA module as part of the DCCA module.

First, we represent the dimension of $F \in \mathbb{R}^{(H'/8)\times(W'/8)\times C}$ as $F \in \mathbb{R}^{H\times W\times C}$.

As shown in the Figure 4, the local feature map F, the output of the backbone part, is mapped to $K \in \mathbb{R}^{H\times W\times C}$, $Q \in \mathbb{R}^{H\times W\times C}$ and $V \in \mathbb{R}^{H\times W\times C}$.

$$K, Q, V = F \tag{2}$$

Then, an affinity process is used to process $K \in \mathbb{R}^{H\times W\times C}$ and $Q \in \mathbb{R}^{H\times W\times C}$. For each position (i, j), the centre point channel vector Q_{ij} is multiplied by the channel vectors of crisscross attention area K_{ij}^{Φ}.

$$\begin{gathered} K_{ij}^{\Phi} = K_{ij}\{(x^I, y^I)\ |x^I = i,\ y^I \epsilon\ [1, \cdots, W]\} \cup \\ K_{ij}\{(x^{II},\ y^{II})\ |\ x^{II} \epsilon\ [1, \cdots, H],\ y^{II} = j\} \\ i\ \epsilon\ [1, \cdots, H],\ j\ \epsilon\ [1, \cdots, W\ ] \end{gathered} \tag{3}$$

where $\{(x^I,\ y^I)|\ldots\}$ and $\{(x^{II},\ y^{II})|\ldots\}$ are two point groups that represent the horizontal line and vertical line in the crisscross area. The output of the affinity process is called raw energy map $D \in \mathbb{R}^{(H+W-1)\times H\times W}$.

$$D_{ij} = Q_{ij} \cdot K_{ij}^{\Phi}\ i\epsilon[1, \cdots, H], j\epsilon[1, \cdots, W]$$

For each position (i, j) in the spatial dimension of $D \in \mathbb{R}^{(H+W-1)\times H\times W}$, we use HFF to process the raw energy vector $D_{ij} \in \mathbb{R}^{H+W-1}$. The result is labelled $A_{ij} \in \mathbb{R}^{H+W-1}$, which is the energy vector in position (i, j) of the energy map $A \in \mathbb{R}^{(H+W-1)\times H\times W}$.

$$A_{ij} = HFF(\ D_{ij}\ )$$

For each position (i, j), an aggregation process is used to process the channel vector A_{ij} and the channel vectors of crisscross attention area V_{ij}^{Φ}.

$$\begin{gathered} V_{ij}^{\Phi} = V_{ij}\{(x^I,\, y^I) \mid x^I = i,\, y^I \epsilon\, [1, \cdots, W]\} \cup \\ V_{ij}\{(x^{II},\, y^{II}) \mid x^{II} \epsilon\, [1, \cdots, H],\, y^{II} = j\} \\ i\, \epsilon\, [1, \cdots, H],\, j\, \epsilon\, [1, \cdots, W] \end{gathered} \tag{4}$$

In this way, the residual part between the input $F \in \mathbb{R}^{H \times W \times C}$ and the output $F' \in \mathbb{R}^{H \times W \times C}$ can be obtained as follows:

$$F'_{ij} = \sum V_{ij}^{\Phi} \cdot A_{ij} + F_{ij} \quad i\epsilon[1, \cdots, H], j\epsilon[1, \cdots, W] \tag{5}$$

where F' is the output of this module. In CCA part, the output F' is labelled as F'_{CCA}. The implementation of the RCCA part will be introduced in Section 3.2.

3.2. Introduction of the RCCA Module

3.2.1. Design and Realization of the RCCA Module

As mentioned in previous chapters, the RCCA module is a complement of the directionality in the attention area. As shown in Figure 5, the RCCA module can let the target pixel receive attention information from oblique directions.

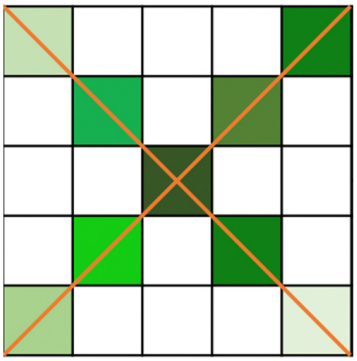

Figure 5. The attention area of RCCA module.

The attention area needs to be extracted. For the pixels at different locations, the size of the attention area is different. The structure of RCCA module is shown in Figure 6.

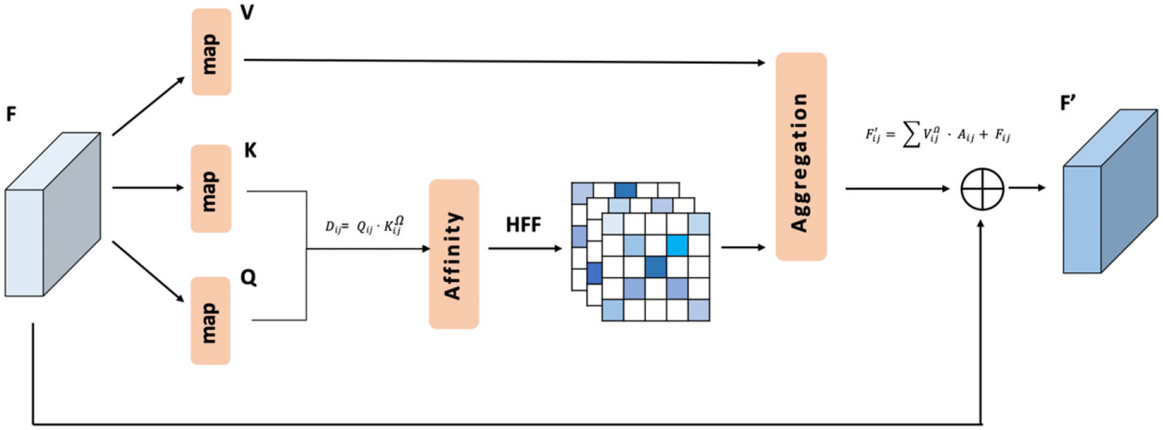

Figure 6. The structure of RCCA module.

First, we represent the dimension of $F \in \mathbb{R}^{(H'/8) \times (W'/8) \times C}$ as $F \in \mathbb{R}^{H \times W \times C}$.

In the RCCA module part, the local feature map $F \in \mathbb{R}^{H \times W \times C}$, the output of the backbone part, is mapped to $K \in \mathbb{R}^{H \times W \times C}$, query $Q \in \mathbb{R}^{H \times W \times C}$ and value $V \in \mathbb{R}^{H \times W \times C}$ (see Equation (2)).

For each position (i, j), in the spatial dimension of Q, we can obtain a vector $Q_{ij} \in \mathbb{R}^{C}$. Then, for this position in feature map K, the channel vectors of the rotated crisscross area can be represented by K_{ij}^{Ω}.

$$\begin{array}{c} K_{ij}^{\Omega} = K_{ij}\{(x^I,\, y^I) \mid x^I = \alpha,\; y^I = i + j - \alpha,\;\; \alpha \,\epsilon\, A(i,j)\} \cup \\ K_{ij}\{(x^{II},\, y^{II}) \mid x^{II} = \beta,\; y^{II} = i - j + \beta,\;\; \beta \,\epsilon\, B(i,j)\} \\ i \,\epsilon\, [1, \cdots, H],\; j \,\epsilon\, [1, \cdots, W] \end{array} \tag{6}$$

$\{(x^I,\, y^I) | \ldots\}$ and $\{(x^{II},\, y^{II}) | \ldots\}$ are two-point groups that represent two oblique lines in the rotated crisscross area. $A(i, j)$ and $B(i, j)$ represent the initial position of the sampling area, which can be obtained as follows:

$$A(i,j) = \begin{cases} \{0, \cdots, i + j\} & if\ i + j \le H \\ \{i + j - H, \cdots,\ W\} & else \end{cases} \tag{7}$$

$$B(i,j) = \begin{cases} \{j - i, \cdots, W\} & if\ i \le j \\ \{0, \cdots,\ j - i + H\} & else \end{cases} \tag{8}$$

For each position (i, j), we can apply an affinity operation to obtain the raw attention map D as follows:

$$D_{ij} = \; Q_{ij} \cdot K_{ij}^{\Omega} \quad i\epsilon[1, \cdots, H], j\epsilon[1, \cdots, W] \tag{9}$$

The $D_{ij} \in D$ represents the correlation between Q_{ij} and $K_{\Omega ij}$.

Then, we use a heterogeneous filter function to process D_{ij} as follows:

$$A_{ij} = HFF(\; D_{ij}\;) \tag{10}$$

where A_{ij} is the feature vector of attention map A at position (i, j).

At each position (i, j) in the spatial dimension of V, we can obtain the channel vectors of the rotated crisscross area, which can be represented by V_{ij}^{Ω}:

$$\begin{array}{c} V_{ij}^{\Omega} = V_{ij}\{(x^I,\, y^I) \mid x^I = \alpha,\; y^I = i + j - \alpha,\;\; \alpha \,\epsilon\, A(i,j)\} \cup \\ V_{ij}\{(x^{II},\, y^{II}) \mid x^{II} = \beta,\; y^{II} = i - j + \beta,\;\; \beta \,\epsilon\, B(i,j)\} \\ i\epsilon[1,\, \cdots, H], j\epsilon[1, \cdots, W] \end{array} \tag{11}$$

For each position (i, j), an aggregation process is used to process the channel vector A_{ij} and the channel vectors of rotated crisscross attention area V_{ij}^{Ω} to obtain the residual part between the input $F \in \mathbb{R}^{H \times W \times C}$ and the output $F' \in \mathbb{R}^{H \times W \times C}$ as follows:

$$F'_{ij} = \sum V_{ij}^{\Omega} \cdot A_{ij} + F_{ij} \quad i\epsilon[1,\, \cdots, H], j\epsilon[1, \cdots, W] \tag{12}$$

F' is the output of this module, which is generated by the combination of F'_{ij}. In RCCA part, the output F' is labelled as F'_{RCCA}.

The pseudocode of the RCCA module is attached in Algorithm 1:

Algorithm 1. The process of RCCA attention.

Input : *feature map F*
Output : *attention feature map F′*
Initialize : $K, Q, V = F$
1: *for i in* $\{1, \cdots, H\}$ *do*
2: *for j in* $\{1, \cdots, W\}$ *do*
/* According to the points groups (x^I, y^I) and (x^{II}, y^{II}), get feature value of the RCCA region positions in K. */
/* $A(i,j)$ and $B(i,j)$ represent the range of index on the horizontal axis of the feature map. */
3: $A(i,j) = \begin{cases} \{0, \cdots, i+j\} & if\ i+j \leq H \\ \{i+j-H, \cdots, W\} & else \end{cases}$
4: $B(i,j) = \begin{cases} \{j-i, \cdots, W\} & if\ i \leq j \\ \{0, \cdots, j-i+H\} & else \end{cases}$
5: *for* α *in* $A(i,j)$ *do*
6: $x^I \leftarrow \alpha$
7: $y^I \leftarrow (i+j) - \alpha$
8: *for* β *in* $B(i,j)$ *do*
9: $x^{II} \leftarrow \beta$
10: $y^{II} \leftarrow (i-j) + \beta$
/* K_{ij}^{Ω} represents the channel vectors sets of the RCCA region in K. */
11: $K_{ij}^{\Omega}\ \epsilon\ K_{ij}(x^I, y^I) \cup K_{ij}(x^{II}, y^{II})$
/* Search Q corresponding to the position in K and then get the energy map E */
12: $E_{ij} = Q_{ij} \cdot K_{\Omega}$
13: **end**
14: **end**
/* Apply activation function HFF to the energy map, HFF is defined in 3.3 */
15: $E \leftarrow HFF(E)$
16: *for i in* $\{1, \cdots, H\}$ *do*
17: *for j in* $\{1, \cdots, W\}$ *do*
/* According to the point set (x^I, y^I) and (x^{II}, y^{II}), get feature value of RCCA region positions in V */
18: $V_{ij}^{\Omega}\ \epsilon\ V_{ij}(x^I, y^I) \cup V_{ij}(x^{II}, y^{II})$
/* Aggregate E and V and F_{ij} represents the residual structure */
19: $F'_{ij} \leftarrow V_{\Omega} \cdot E_{ij} + F_{ij}$
20: **end**
21: **end**

3.2.2. Functional Merits of the RCCA Module

When using the recurrent mechanism to relate position $(x+k, y+k)$ to position (x, y), the process is shown in Figure 7.

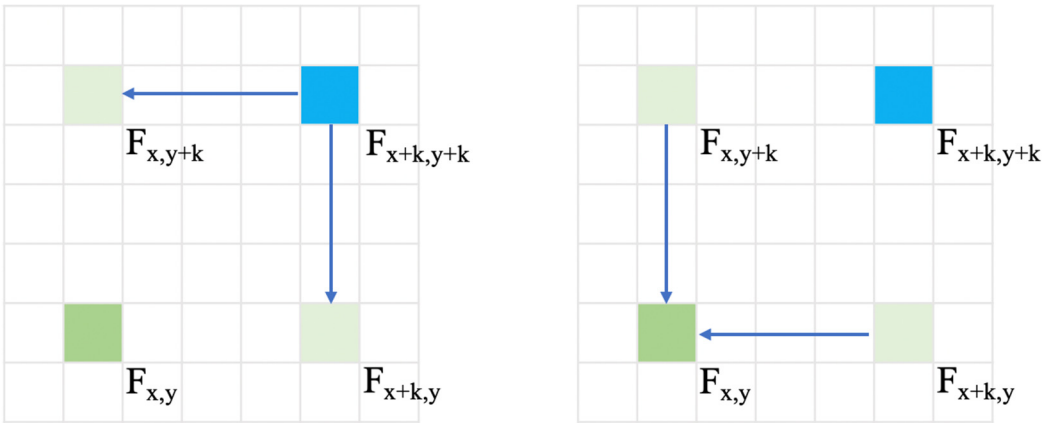

Figure 7. The recurrent mechanism.

In the first attention process, the value change of each position in the recurrent mechanism can be shown:

$$\begin{aligned} F_{x+k,y}(CCA) = \; & Q_{x+k,y} \cdot K_{x+k,y}(x+k,y+k) \cdot F_{x+k,y}(x+k,y+k) + Q_{x+k,y} \\ & \cdot \Big(K_{x+k,y}\{(x^I,\, y^I) \mid x^I = x+k,\; y^I \epsilon\, [1,\cdots,W],\; y^I \neq y+k\} \\ & \cup K_{x+k,y}\{(x^{II},\, y^{II}) \mid x^{II} \epsilon\, [1,\cdots,H],\; y^{II} = y\}\Big) \\ & \cdot \Big(F_{x+k,y}\{(x^I,\, y^I) \mid x^I = x+k,\; y^I \epsilon\, [1,\cdots,W],\; y^I \neq y+k\} \\ & \cup F_{x+k,y}\{(x^{II},\, y^{II}) \mid x^{II} \epsilon\, [1,\cdots,H],\; y^{II} = y\}\Big) \\ = \; & Q_{x+k,y} \cdot K_{x+k,y}(x+k,y+k) \cdot V_{x+k,y}(x+k,y+k) + \delta_1 \end{aligned} \tag{13}$$

$$\begin{aligned} F_{x,y+k}(CCA) = \; & Q_{x,y+k} \cdot K_{x,y+k}(x+k,y+k) \cdot F_{x,y+k}(x+k,y+k) + Q_{x,y+k} \\ & \cdot \Big(K_{x,y+k}\{(x^I,\, y^I) \mid x^I = x,\; y^I \epsilon\, [1,\cdots,W]\} \\ & \cup K_{x,y+k}\{(x^{II},\, y^{II}) \mid x^{II} \epsilon\, [1,\cdots,H],\; y^{II} = y+k,\; x^{II} \neq x+k\}\Big) \\ & \cdot \Big(F_{x,y+k}\{(x^I,\, y^I) \mid x^I = x,\; y^I \epsilon\, [1,\cdots,W]\} \\ & \cup F_{x,y+k}\{(x^{II},\, y^{II}) \mid x^{II} \epsilon\, [1,\cdots,H],\; y^{II} = y+k,\; x^{II} \neq x+k\}\Big) \\ = \; & Q_{x,y+k} \cdot K_{x,y+k}(x+k,y+k) \cdot V_{x,y+k}(x+k,y+k) + \delta_2 \end{aligned} \tag{14}$$

$Q_{x+k,y} \cdot K_{x+k,y}(x+k,y+k)$ is the similarity parameter of positions $(x+k,y+k)$ and $(x+k,y)$ and $Q_{x,y+k} \cdot K_{x,y+k}(x+k,y+k)$ is the similarity parameter of positions $(x+k,y+k)$ and $(x,y+k)$. δ_1 and δ_2 are the redundancy information which is caused by feature vectors of other positions.

In the second attention process, the value change of each position in the recurrent mechanism can be shown:

$$\begin{aligned} F_{x,y}(CCA) = \; & Q_{x,y} \cdot K_{x,y}(x+k,y) \cdot F_{x,y}(x+k,y) + Q_{x,y} \cdot K_{x,y}(x,y+k) \cdot F_{x,y}(x,y+k) + \\ & Q_{x,y} \cdot (\, K_{x,y}\{(x^I,\, y^I) \mid x^I = x,\; y^I \epsilon\, [1,\cdots,W],\; y^I \neq y+k\} \cup \\ & K_{x,y}\{(x^{II},\, y^{II}) \mid x^{II} \epsilon\, [1,\cdots,H],\; y^{II} = y+k,\; x^{II} \neq x+k\}\,) \cdot \\ & (F_{x,y}\{(x^I,\, y^I) \mid x^I = x,\; y^I \epsilon\, [1,\cdots,W],\; y^I \neq y+k\} \cup \\ & F_{x,y}\{(x^{II},\, y^{II}) \mid x^{II} \epsilon\, [1,\cdots,H],\; y^{II} = y+k,\; x^{II} \neq x+k\} \\ = \; & Q_{x,y} \cdot K_{x,y}(x+k,y) \cdot F_{x,y}(x+k,y) + Q_{x,y} \cdot K_{x,y}(x,y+k) \cdot F_{x,y}(x,y+k) + \; \delta_3 \end{aligned} \tag{15}$$

$Q_{x,y} \cdot K_{x,y}(x+k,y)$ is the similarity parameter of positions $(x+k,y)$ and (x,y) and $Q_{x,y} \cdot K_{x,y}(x,y+k)$ is the similarity parameter of positions $(x,y+k)$ and (x,y). δ_3 is the redundancy information in this step.

Therefore, when we generate two processes together, we can obtain $f_{x,y}(CCA)$ as follows:

$$\begin{aligned} & F_{x,y}(CCA) = Q_{x,y} \cdot K_{x,y}(x+k,y) \times \\ & (Q_{x+k,y} \cdot K_{x+k,y}(x+k,y+k) \times F_{x+k,y+k}(CCA) + \delta_1) + \\ & Q_{x,y} \cdot K_{x,y}(x,y+k) \times \\ & (Q_{x,y+k} \cdot K_{x,y+k}(x+k,y+k) \times F_{x+k,y+k}(CCA) + \delta_2) + \delta_3 \end{aligned} \tag{16}$$

After summing all the redundancy information to $\sum \delta$, we can obtain the following:

$$\begin{aligned} F_{x,y}(CCA) = \; & Q_{x,y} \cdot K_{x,y}(x+k,y) \times Q_{x+k,y} \cdot K_{x+k,y}(x+k,y+k) F_{x+k,y+k}(CCA) + \\ & Q_{x,y} \cdot K_{x,y}(x,y+k) Q_{x,y+k} \cdot K_{x,y+k}(x+k,y+k) F_{x+k,y+k}(CCA) + \sum \delta \end{aligned} \tag{17}$$

When using the RCCA module, the relation between $(x+k,y+k)$ and (x,y) is shown in Figure 8:

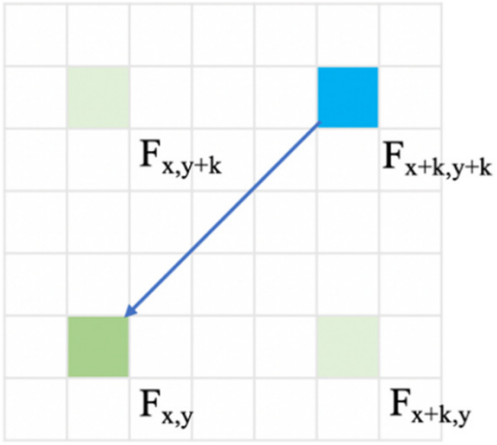

Figure 8. The relation information transference in RCCA.

The value change can be represented as follows:

$$\begin{aligned} F_{x,y}(RCCA) = \quad & Q_{x,y} \cdot K_{x,y}(x+k,y+k) \cdot F_{x,y}(x+k,y+k) + \\ & Q_{x,y} \cdot (\, K_{x,y}\{\,(x^I,\, y^I) \mid x^I = \alpha,\; y^I = x + y - \alpha,\;\; \alpha \,\epsilon\, A(x,y)\,\} \cup \\ & K_{x,y}\{\,(x^{II},\, y^{II}) \mid x^{II} = \beta,\; y^{II} = x - y + \beta,\;\; \beta \,\epsilon\, B(x,y), \beta \neq x + k\,\}\,) \cdot \\ & (F_{x,y}\{\,(x^I,\, y^I) \mid x^I = \alpha,\; y^I = x + y - \alpha,\;\; \alpha \,\epsilon\, A(x,y)\,\} \cup \\ & F_{x,y}\{\,(x^{II},\, y^{II}) \mid x^{II} = \beta,\; y^{II} = x - y + \beta,\;\; \beta \,\epsilon\, B(x,y), \beta \neq x + k\,\}\,) \\ = \quad & Q_{x,y} \cdot K_{x,y}(x+k,y+k) \cdot F_{x+k,y+k}(RCCA) + \delta_4 \end{aligned} \tag{18}$$

$Q_{x,y} \cdot K_{x,y}(x+k,y+k)$ is the similarity parameter of positions $(x+k,y+k)$ and (x,y). δ_4 is the information redundancy caused by feature vectors of other positions.

When we compare the similarity parameters of $F_{x,y}(CCA)$ and $F_{x,y}(RCCA)$, the ratio r can be represented as follows:

$$r = \frac{Q_{x,y} \cdot K_{x,y}(x+k,y) \times Q_{x+k,y} \cdot K_{x+k,y}(x+k,y+k) + Q_{x,y} \cdot K_{x,y}(x,y+k) Q_{x,y+k} \cdot K_{x,y+k}(x+k,y+k)}{Q_{x,y} \cdot K_{x,y}(x+k,y+k)} \tag{19}$$

When positions $(x+k,y+k)$ and (x,y) are homogeneous and $(x+k,y)$ and $(x,y+k)$ are heterogeneous compared to (x,y), the similarity parameters of $(x+k,y+k)$ and (x,y) will be extremely low in the recurrent mechanism. When using the RCCA module under the same conditions, which can let these two positions be related directly, the similarity parameter can be high, corresponding to the real situation. The ratio r can sometimes be very low because the recurrent mechanism will cause similarity distortion.

DCCA module consists of CCA module and RCCA module. It can also distribute the weight of each attention module according to the statistics of the energy map in the SAF module. The shape of the attention area allows each pixel to receive contextual information from eight directions, as shown in Figure 9.

Considering the local information extracted by 3×3 convolution, the DCCA module can let each pixel perceive the local information and eight-direction nonlocal information, as shown in Figure 10.

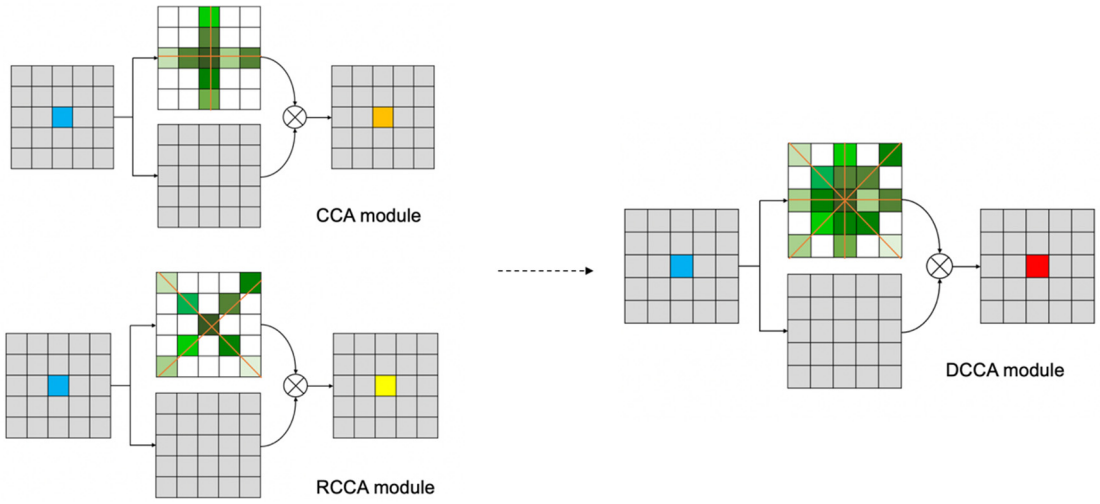

Figure 9. The relation information transference in RCCA.

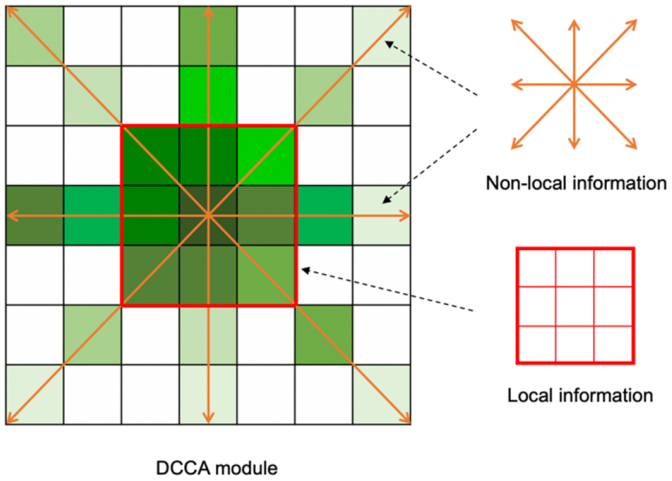

Figure 10. The local and non-local information in DCCA module.

3.2.3. Computational Advantages of the RCCA Module

From the aspect of the theory of the algorithm, we will analyze the computational cost based on the attention information received by each pixel. In CCNet, each pixel receives the information from the pixels in the same column and row. The total computational cost can be described as follows:

$$\begin{aligned} O(CCA) &= \iint_D^a f(i,j)didj \\ D &= \{(i,\ j) \mid i\ \epsilon\ [1,\cdots,H],\ j\ \epsilon\ [1,\cdots,W]\} \\ f(i,j) &= H+W-1 \end{aligned} \tag{20}$$

So, we can conclude that:

$$O(CCA) = H \times W \times (H+W-1) \tag{21}$$

The computational cost of RCCA module can be calculated as follows:

$$\begin{aligned} O(RCCA) &= \iint_D^a f(i,j)didj \\ D &= \{(i,\ j) \mid i\ \epsilon\ [1,\cdots,H],\ j\ \epsilon\ [1,\cdots,W]\} \\ A(i,j) &= \begin{cases} \{0,\cdots,i+j\} & if\ i+j \le H \\ \{i+j-H,\cdots,\ W\} & else \end{cases} \\ B(i,j) &= \begin{cases} \{j-i,\cdots,W\} & if\ i \le j \\ \{0,\cdots,\ j-i+H\} & else \end{cases} \\ f(i,j) &= card(A \cup B) \end{aligned} \tag{22}$$

The function card( ) represents the number of elements in the set. So, we can conclude that:

$$\begin{aligned} O(RCCA) &= H \times W \times (H+W-1) \times \tfrac{2}{3} = \tfrac{2}{3}\, O(CCA) \\ O(DCCA) &= O(CCA) + O(RCCA) = \tfrac{5}{3} \times H \times W \times (H+W-1) \end{aligned} \tag{23}$$

That is to say, theoretically, the computational cost of the RCCA module is two-thirds of that of a single CCA. The computational cost of the DCCA module is five-thirds of that of a single CCA. The computational cost of Non-local Network [2] can be described as:

$$O(Non\ local) = H \times W \times (H \times W) \gg O(DCCA) \tag{24}$$

Through comparison, it can be concluded that the computational cost of DCCA module is much smaller than that of Non-local Network.

3.3. Heterogeneous Filter Function and Output Part

3.3.1. Heterogeneous Filter Function

In energy processing, there is an unevenly distributed sequence of energy values. These energy groups come from the values taken in different attention areas. In the DCCA module, we need to process these energy values to assign weights. The value of energy represents the similarity between point pairs. In such a process, we expect that this filter function can benefit our classification problem. That is, the filter function needs to remove the influence of the heterogeneous part, which can let each part receive more relation information from the homogeneous part.

We design the Heterogeneous Filter Function (HFF) to help us remove the influence of a part of the energy values that are relatively low in one energy group. The implementation details are shown in Figure 11:

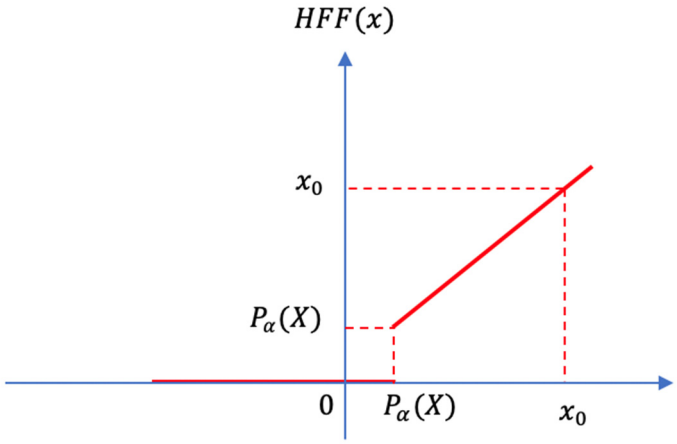

Figure 11. HFF.

The mathematical formula of HFF can be expressed as follows:

$$HFF(x) = \begin{cases} x & if\ x \geq P_\alpha(X) \\ 0 & else \end{cases} \tag{25}$$

We introduce a position parameter α to determine how large the lost part is. We arrange the value of energy from small to large in a sequence. The parameter α is a percentage. The position function $P_\alpha(X)$ refers to the value at the α position in sorting from small to large, as shown in Figure 12.

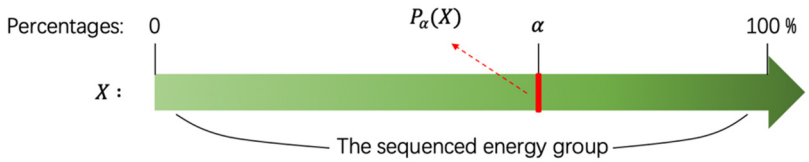

Figure 12. The position function $P_\alpha(X)$.

The energy value, which is smaller than $P_\alpha(X)$, will be fixed to 0. Consequently, the lost part of the energy will have no influence in the attention process. The existence of the position function $P_\alpha(X)$ ensures that the DCCA model still has the ability to identify homogeneous and heterogeneous regions for the change of energy distribution. In the experiment, it can be found that the most proper position parameter α is often related to the road pixel percentages of the images in the training dataset, as shown in Figure 13.

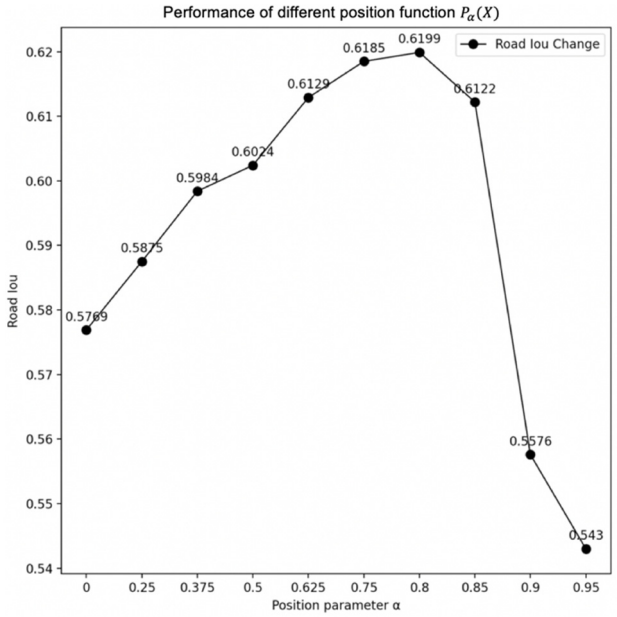

Figure 13. The performance of different position function $P_\alpha(X)$.

In Figure 13 we choose the road intersection over union (IOU) as a statistical index to show the performance of each different α. When the position parameter α is approximately 0.8, corresponding to the average percentage of nonroad parts in images, the performance will be the best. Therefore, α needs to be matched with the percentage of non-road parts in the dataset.

For such a type of data processing problem, the ReLU function is often used in traditional methods, which is shown in Figure 14.

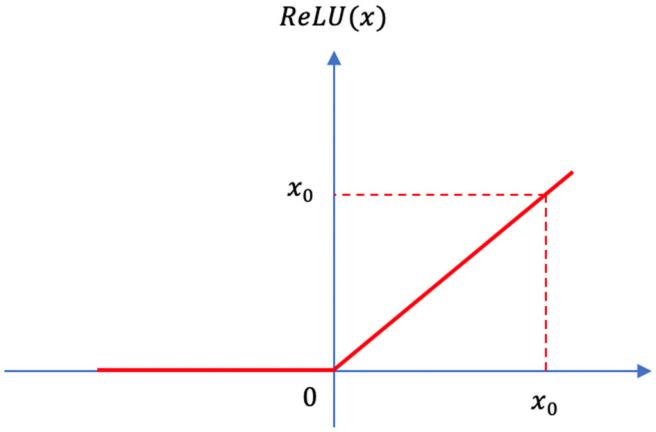

Figure 14. ReLU function.

The function can be expressed in the following form:

$$ReLU(x) = \begin{cases} x & if\ x \geq 0 \\ 0 & else \end{cases} \tag{26}$$

In the traditional ReLU function, there is an absolute threshold for numerical filtering. In most cases, this threshold is 0, similar to the example in Figure 14. However, in the road extraction problem, the threshold for judging the energy cannot be a fixed number. In different samples, the value of energy will vary greatly; therefore, we need to perform the filter process according to the energy value ranking in the sampling group, which is the reason why the ReLU function is not suitable for use in the DCCA module.

3.3.2. Output Part: SAF Module

In the output part, the SAF module can recognize the road shape through the mean values of energy in two different attention intervals. The mean value of the energy reflects the homogeneity and heterogeneity in the sampling area, which will be illustrated in the experimental section.

SAF module consists of these steps:

For each position u, we take the mean value of the energy vector $A_u^{CCA} \in \mathbb{R}^{H+W-1}$ and $A_u^{RCCA} \in \mathbb{R}^{L}$ (which has been introduced in Section 3.2) of the crisscross and rotated crisscross attention modules: E_u^{CCA} and E_u^{RCCA}.

The softmax function is used to process the two mean values and obtain the weights of the two attention modules, ω_{CCA} and ω_{RCCA}. This can be shown as follows:

$$(\omega_{CCA}, \omega_{RCCA}) = SoftMax(E_u^{CCA}, E_u^{RCCA}) \tag{27}$$

The weights of the two attention modules represent the prediction of road shape from the SAF module, which is the key part of the self-adaptive mechanism. If one attention module obtains a higher weight in one point, the road shape around this point is more likely to be the shape of this attention module. Under this condition, in the SAF module, a higher weight can let this point receive more relation information from the road area.

Multiply the weights by the outputs of the two attention modules to obtain the vector of fusion result in position u.

$$F_u^{output\ \prime} = \omega_{CCA} \times F_u^{CCA\ \prime} + \omega_{RCCA} \times F_u^{RCCA\ \prime} \tag{28}$$

$F_u^{output\ \prime}$ is combined in the spatial dimension to obtain F'_{output}, the fusion result. The function g represents the upsampling process. The structure of the output part is shown in the Figure 15.

$$F'_{output} = g(F'_{output}) \tag{29}$$

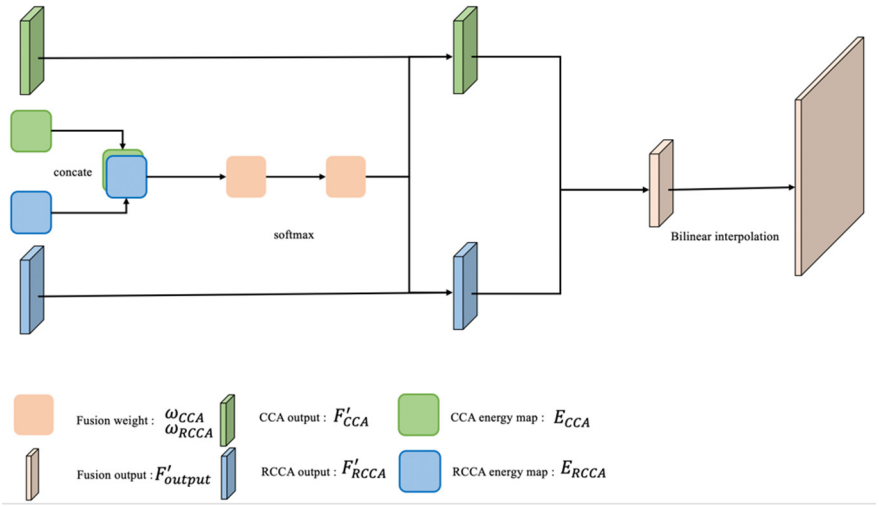

Figure 15. The structure of the output part.

4. Experiment

4.1. Dataset and Implement Details

The dataset of this experiment is based on the DeepGlobe Road Extraction Challenge [43]. The size of the road image has been changed from 1024 pixels to 256 pixels by downsampling. To emphasize the function of our special attention mechanism, we select those images that have distinctive direction features. The road directions on these images are evenly distributed in space. Then, we create two subdatasets, which are both parts of the entire dataset. The first subdataset consists of all the images that mainly include horizontal and vertical roads, whose angles between their direction and the horizontal or vertical direction are less than 22.5°.

This subdataset is called the crisscross dataset (CC dataset). In Figure 16, an example image is shown, which contains two horizontal roads.

Figure 16. The image in CC dataset.

Another subdataset is the residual part in the total dataset, which contains images with slope roads inside, called the rotated crisscross dataset (RCC dataset). In Figure 17, an example image is shown, and it contains two slope roads.

Figure 17. The image in RCC dataset.

We implement different attention methods and compare the results to analyse the advantages of our special attention method to a great extent. The difference between the CCA module and RCCA module is the attention area. For each pixel in one feature map, we will generate two energy maps for two different attention areas. In these energy maps, we will analyse the energy distribution to determine whether the energy can reflect the homogeneity and heterogeneity of pixels. Then, we use the SAF module to distribute the weight of the CCA module and RCCA module according to the energy map. In this way, we can obtain the result of the DCCA module and compare it with the results of the CCA module and RCCA module. In detail, we will also select some typical examples to illustrate the merits of our DCCA module.

4.2. Results of Different Attention Methods

In this part, we select the IOU as the statistical index. In this two-classification problem, we set the classification labels as road and background.

The road IOU and mean IOU are indices that we choose to compare the results of different attention modules. Mean IOU is the mean value of the Road IOU and the Background IOU.

When applying the CCA module to the complete dataset and two subdatasets, the CCA module shows the highest result when being used in the crisscross dataset from the perspectives of both mean IOU and road IOU in Table 1. The road IOU and mean IOU of the complete dataset are at the mean level of the results of the CC dataset and RCC dataset. From these two characteristics, we can conclude that the CCA module has a better processing effect on crisscross-shaped roads. It seems powerless for the road in the rotated crisscross shape. The reason for its relatively poor performance on the RCC dataset is that the CCA module cannot effectively extract the geometric information of rotated crisscross-shaped roads.

Table 1. Applying CCA module to different datasets.

Indicators	Complete Dataset	CC Dataset	RCC Dataset
Road IOU	0.4948	0.5442	0.4498
Mean IOU	0.7030	0.7327	0.6759

Table 2 shows that the road IOU and mean IOU of the CC dataset are still the highest. Although these 2 indicators of the RCC dataset are still not as good as those of the CC dataset, there is still a rise compared to the result of the CCA module on the RCC dataset. This result shows that the RCCA module can extract the geometric information of the rotated crisscross shape better than the CCA module.

Table 2. Applying RCCA module to different datasets.

Indicators	Complete Dataset	CC Dataset	RCC Dataset
Road IOU	0.5067	0.5242	0.4909
Mean IOU	0.7117	0.7229	0.7015

When applying the DCCA module to three different datasets in Table 3, the road IOU and mean IOU are the best. The performance of the CCA module, RCCA module and DCCA module on the complete dataset can show the properties of each attention method. The results of three different attention modules on the complete dataset can reveal the high performance of the DCCA module.

Table 3. Applying DCCA module to different datasets.

Indicators	Complete Dataset	CC Dataset	RCC Dataset
Road IOU	0.5414	0.5722	0.5140
Mean IOU	0.7256	0.7454	0.7077

In Table 4, the performance of the DCCA module on the complete dataset is significantly ahead of that of the CCA module and RCCA module. Focusing on the road IOU, the advantages of the DCCA module are shown in Figure 18 as follows:

Table 4. The results of the different attention module on complete dataset.

Indicators	CCA	RCCA	DCCA
Road IOU	0.4948	0.5067	0.5414
Mean IOU	0.7030	0.7117	0.7256

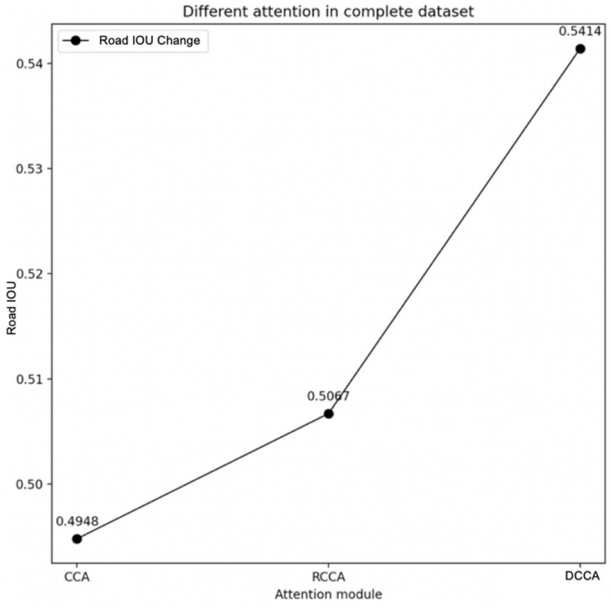

Figure 18. The result of different attention on complete dataset.

In Figure 18, from the perspective of the road IOU indicator, the result of the DCCA module generally leads the CCA and RCCA by 4 percent.

In order to illustrate the performance of the DCCA module, we also incorporate a 30° module into the DCCA module for experimentation. At the same time, we also tested the performance of Non-local Network [2], CBAM [9] and PSPNet [17] mentioned above. The final results are shown in the Table 5.

Table 5. The comparison results of other models.

Models	Road IOU	Mean IOU
DCCA	0.5414	0.7256
DCCA (with 30°)	0.5275	0.7282
Non-local	0.5411	0.7372
CBAM	0.5325	0.7318
PSPNet	0.5357	0.7322

We conclude that adding more directional modules and letting each pixel perceive the information of all pixels (Non-local) will improve the Mean IOU of the road and the background. But the Road IOU will not necessarily improve.

From the above comparison, we can also conclude that the DCCA module is in the leading position on the Road IOU. Only the Non-local method is similar to the result of the DCCA module. The computational cost we got in Section 3.2.3 shows that the computational cost of Non-local is much greater than that of the DCCA module. The performance of DCCA module is significantly better than CBAM and PSPNet. This reflects the advantages of the DCCA module in road extraction problems.

4.3. Explanation of the Result

The raw energy, D_{ij} $i\epsilon[1, \cdots, H]$, $j\epsilon[1, \cdots, W]$, is calculated by the inner product of the 2-pixel channel vector, which reflects the similarity of the centre pixel and the pixels in the sampling area. In the road extraction, since the similarity between the road points is higher than the similarity between the road and other features, the energy calculated between the points on the road will be very high in this way.

Based on this principle, the calculation of energy helps us distinguish roads from nonroads. In the DCCA module, for one point, two attention areas will be considered, the crisscross area and rotated crisscross area, which regard this point as the centre point. The energy inside these two areas is calculated, and then the weight distribution is performed according to the calculated energy in these two areas.

The position of energy calculation processing is in front of the attention process; therefore, whether the energy can correctly reflect the similarity between the two points becomes the key point. The following will take several typical samples to analyse the numerical characteristics of energy in the similarity between road points.

4.3.1. Example with Only Crisscross Shaped Roads

In Figure 19, there is an image in the dataset and its corresponding ground truth. The selected red points on the image are used to analyse the energy. To generate the energy map, we calculate the inner product of this pixel and each pixel in the image. There are two sampling areas in the energy map: crisscross area and rotated crisscross area. The crisscross area is represented by the blue lines, while the rotated crisscross area is represented by the green lines.

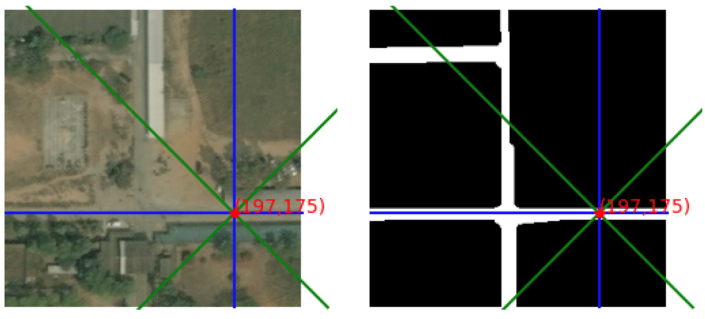

Figure 19. The image and its corresponding ground truth.

(1) The energy heat map:

The energy heat map for these two areas is shown as follows:

In Figure 20, the heat map of the energy distribution of two different sampling areas is shown.

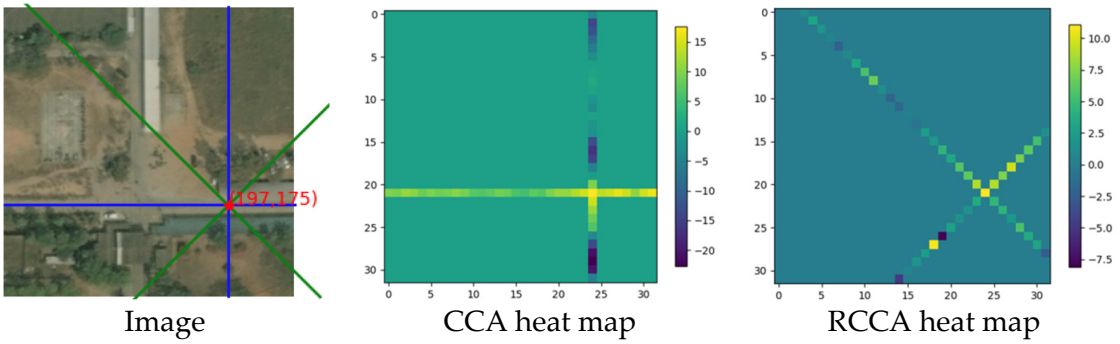

Figure 20. The heat map of energy distribution.

The fixed point is in the road; therefore, most of the points that are homogeneous with the fixed point are distributed on the road. From the energy map, we can see clearly that when the sampling area comes to the road area, the energy value becomes higher than that of the non-road area. This rule can be seen from the brightness of the pixels in the image.

(2) The energy line chart:

The energy value of each sampled pixel can be displayed more clearly through the curve. In Figure 21, the abscissa represents the pixel points in two-point groups of the CCA module: $\{(x^I, y^I)|\ldots\}$ and $\{(x^{II}, y^{II})|\ldots\}$. For the red line, the ordinate is used to distinguish roads from non-roads, which is explained in the legend. The ordinate is the value of energy for the blue line. The abscissa represents the pixel label, and each pixel in the sampling interval has a label number. It can be seen from these two figures that when energy is used to express the similarity of two homogeneous points, the value will become larger compared to the case of heterogeneous points. The distribution of the entire energy value is highly similar to the distribution of the ground truth.

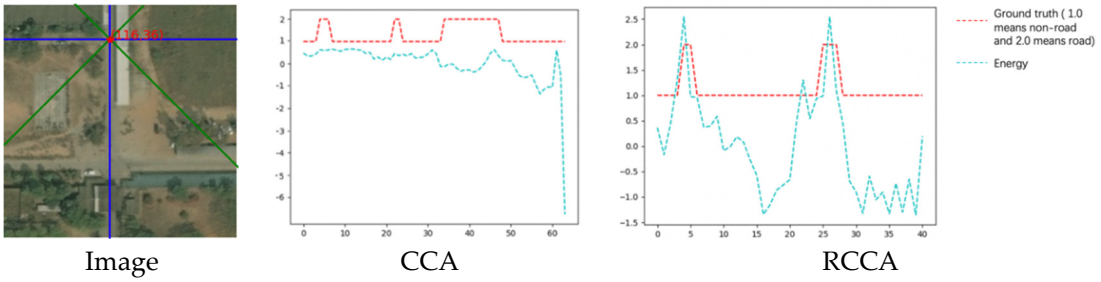

Figure 21. The energy line chart of two sampling area.

(3) The energy table:

In the two sampling areas, the mean value of energy for road points and non-road points can be displayed in Table 6:

Table 6. The energy comparison of CCA and RCCA sampling.

Classification	CCA	RCCA
Road	10.21	7.96
Non-road	7.95	−0.14

The fixed point is on the road. The energy between this point and other road points is higher than the energy between this point and other non-road points. It can be concluded that the energy map is very distinguishable for the homogeneity and heterogeneity of points.

Such an energy map can help the attention process play an important role in determining to what extent the point pair needs to be focused.

(4) The output collection:

The outputs of this image are shown below:

Figure 22 shows that the result of the CCA module is better than the result of the RCCA module. The CCA module can help to focus on more areas of roads because roads are crisscross-shaped. Due to the distinguishability of the energy map, the CCA module can help each pixel pay more attention to the road area. Because of the SAF module inside the DCCA module, the result of the DCCA module remains the advantage of the CCA module.

Figure 22. Image, ground truth and the output compilation.

4.3.2. Example with Only Rotated Crisscross Shaped Roads

The next example is for the road in a rotated crisscross shape:

Figure 23 shows an image with a rotated crisscross-shaped road and its corresponding ground truth. The selected red points on the image are used to analyse the energy.

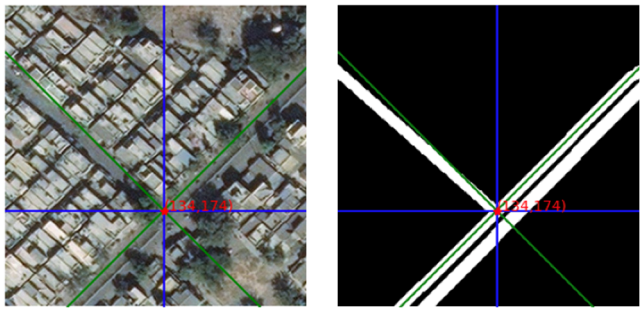

Figure 23. The image and its ground truth.

(1) The energy line chart:

By the same way to calculate the energy, the energy value of each sampled pixel can be displayed clearly through the curve:

In Figure 24, the abscissa also represents the pixel points in two-point groups of the RCCA module: $\{(x^I, y^I)|\ldots\}$ and $\{(x^{II}, y^{II})|\ldots\}$. In this rotated crisscross-shaped road example, the rotated crisscross-shaped sampling area contains more road pixels. We can obtain the same conclusion as the last example that the distribution of the entire energy value is highly similar to the distribution of the ground truth.

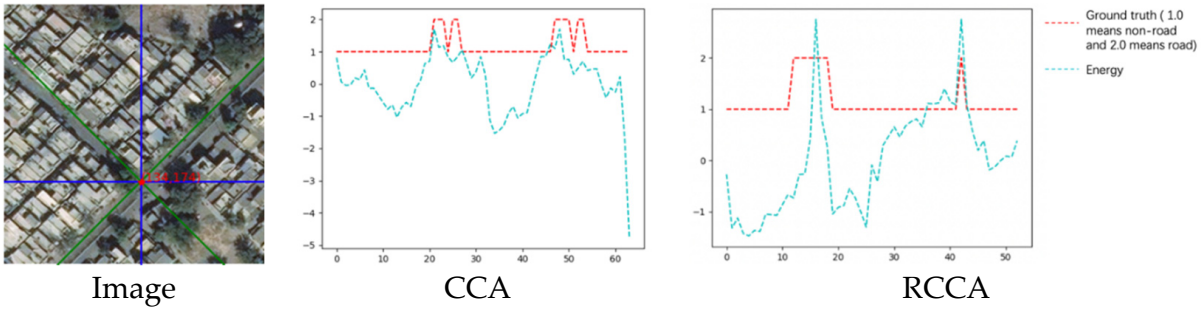

Figure 24. The energy line chart of two sampling area.

(2) The energy table:

In the two sampling areas, the mean value of energy for road points and non-road points can be displayed in Table 7:

Table 7. The energy comparison of CCA and RCCA sampling.

Classification	CCA	RCCA
Road	6.97	5.75
Non-road	1.12	3.43

The distinguishability of the energy map is still clear in this rotated crisscross-shaped road sample.

(3) The output collection:

The output of this sample by the CCA module and RCCA module is shown in Figure 25:

Figure 25. The image, ground truth and the output compilation.

Figure 25 shows that the result of the RCCA module is better than that of the CCA module. The RCCA module truly predicts more road parts, which means that in this case, the RCCA module can focus on more road areas. In this rotated crisscross-shaped road image, the DCCA module can distribute a higher weight to the RCCA module according to the energy map, which can let each road pixel receive more relation information from the road part. In this way, the geometric information can be considered largely in the attention process, which is the reason why the result of the DCCA module can retain the advantages of the RCCA module.

4.3.3. Example with Both Crisscross Shaped Roads and Rotated Crisscross Shaped Roads

We put an image containing both a rotated crisscross-shaped road and a crisscross-shaped road as an example.

Figure 26 shows that the CCA module has a relatively poor performance when dealing with sloping roads and rotated-crisscross shaped roads. Meanwhile, the RCCA module will miss some road parts when the roads are crisscrossed. The output of the DCCA module

complements the advantages and disadvantages of the CCA module and RCCA module. Through the SAF module inside the DCCA module, a larger weight can be assigned to the RCCA module on inclined roads, while a larger weight can be assigned to the CCA module on horizontal and vertical roads to obtain the best results in the end, which is the reason why the DCCA module can adaptively focus on roads of different shapes, and the result is the best.

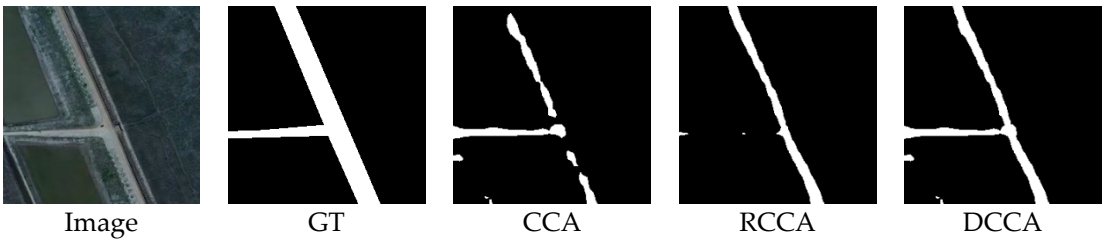

Figure 26. The image, ground truth and the output compilation.

In Figure 27, the yellow part is the common prediction of the RCCA and CCA, and the red part is predicted by the RCCA module but not by the CCA module, while the blue module is the opposite. From the comparison of these two modules, we can see clearly that the CCA module plays an important role in the extraction of crisscross-shaped roads, while the RCCA module also has a better result in the extraction of rotated crisscross-shaped roads.

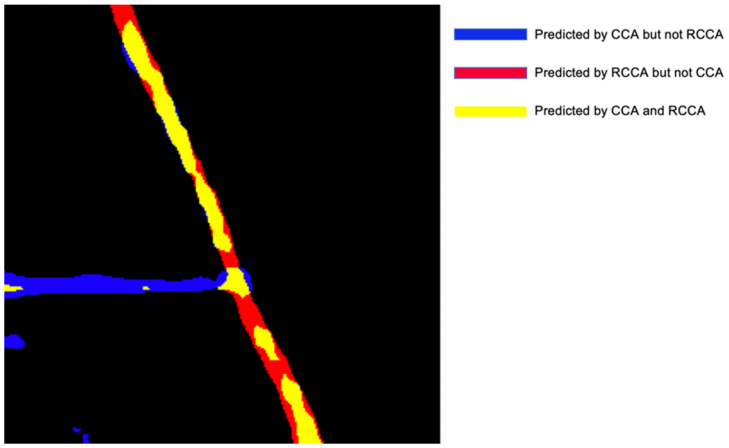

Figure 27. The comparison between CCA module and RCCA module.

This result shows that the CCA module and RCCA module have their own directional advantages. According to this, the DCCA module can let each module maximize its advantages to obtain better results.

4.4. Some Typical Examples

In Figure 28, the first column are the input images and the figures in the second column are the ground truths. The figures in the third column are the results of the CCA module, while the figures in the fourth column are the results of the RCCA module. The figures in the fifth column are the results of the DCCA module.

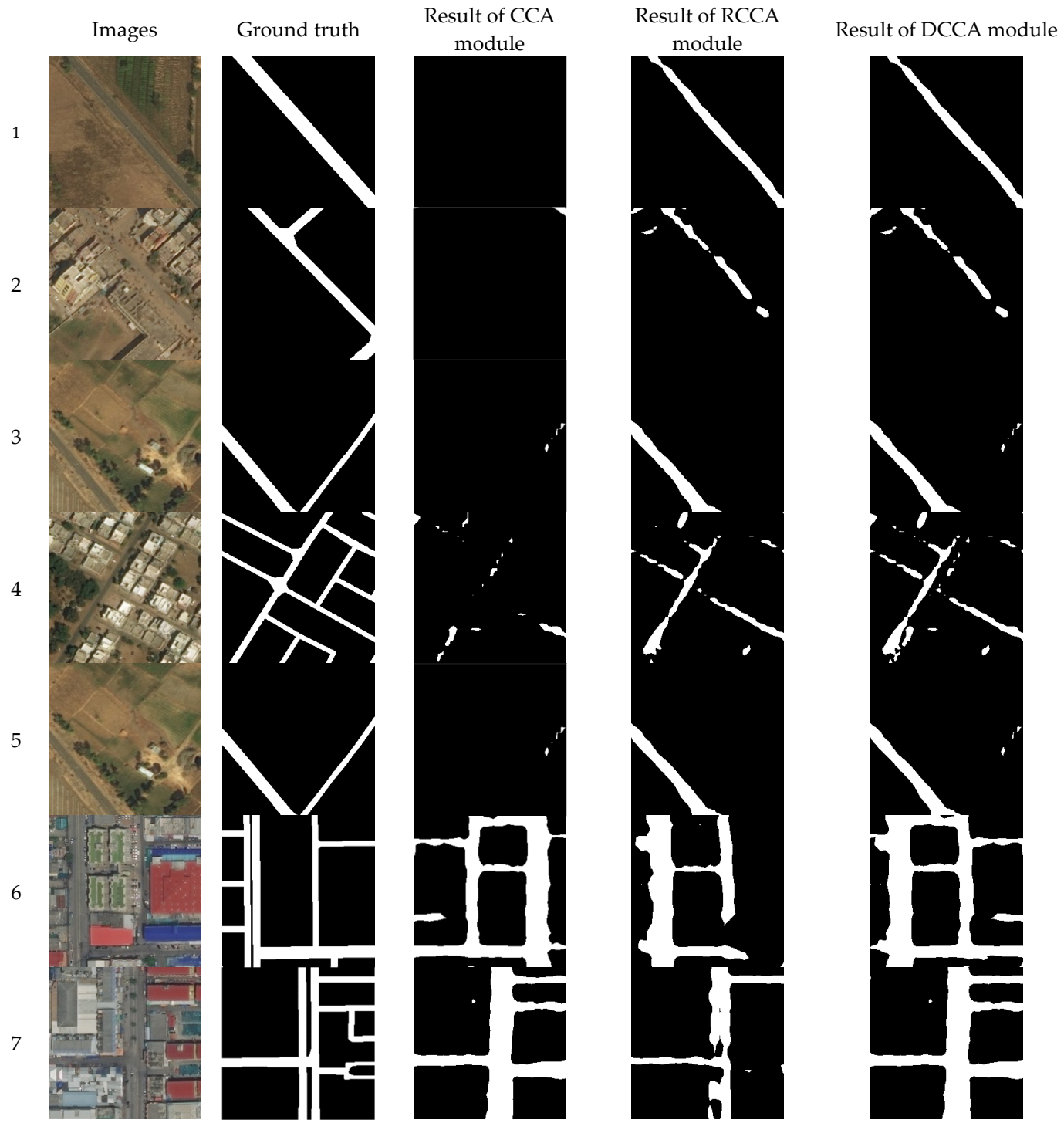

Figure 28. *Cont*.

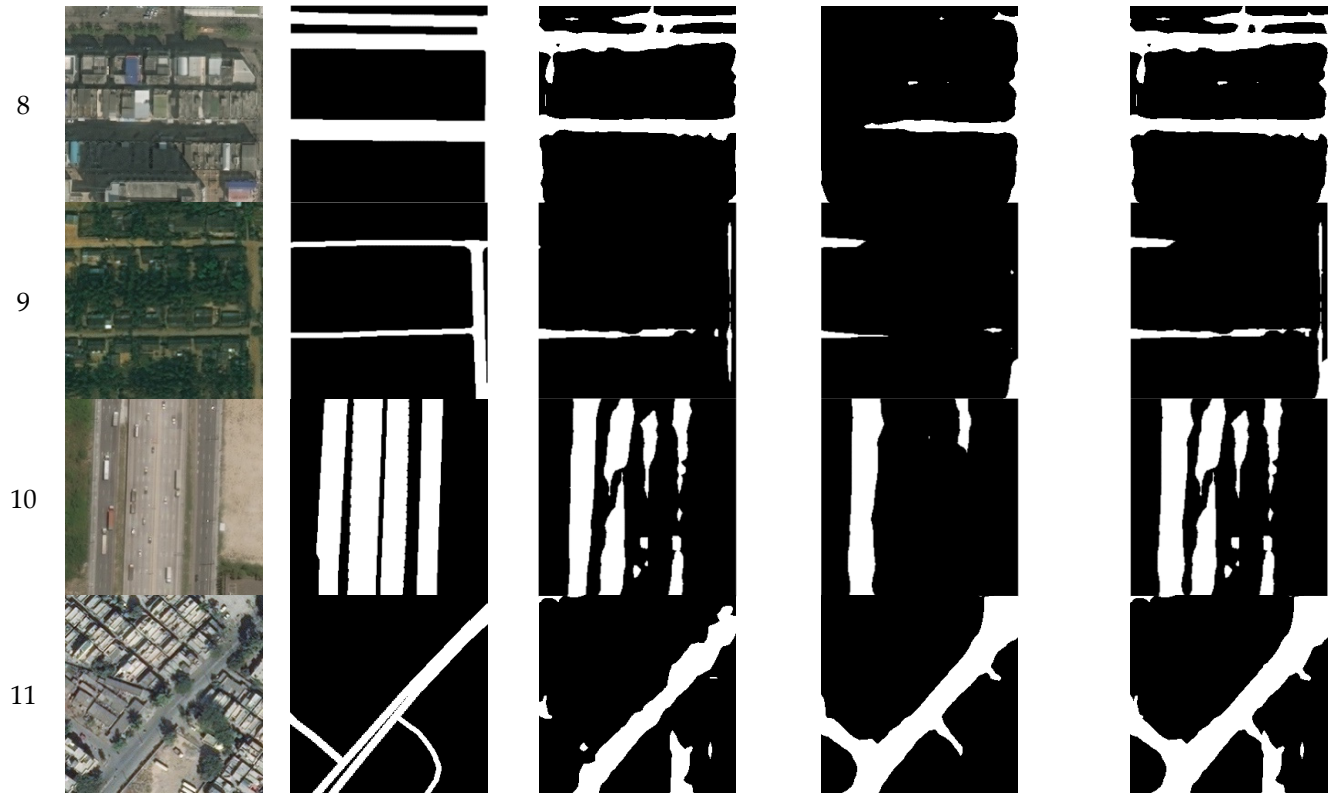

Figure 28. Typical examples.

In examples 1 to 5, the CCA module cannot extract sloping roads in many cases. However, in these cases, roads can be extracted by the RCCA module. In examples 6 to 10, the output of the RCCA module misses parts of the roads, while the CCA module has a better performance on these crisscross-shaped roads. In these 10 examples, the DCCA module can allow each module to play its own advantage to obtain a better result. This is the reason why the DCCA module can always have the best result. When there are more roads in crisscross shape, the DCCA module will distribute more weights to the CCA module. In contrast, the same situation will come to the RCCA module. Typically, Example 11 contains both crisscross-shaped roads and rotated crisscross-shaped roads. The CCA and RCCA modules make up for each other, and then the result of the DCCA module is the best.

5. Conclusions

This paper proposes a special attention mechanism for road extraction, the DCCA module. This module is designed from two attention modules based on different directions, the CCA module and the RCCA module. It also contains the SAF module, a module that can assign weights to the CCA module and RCCA module based on pixel similarity. Since the recurrent dilemma caused by the recurrent mechanism is avoided, the DCCA module can effectively reduce the similarity distortion in the relation between each point pair in the attention process.

In the experiment, the indicators of the DCCA module are 4% higher than those of the CCA and RCCA modules. By analysing specific samples, the DCCA module combines the merits of the CCA module and RCCA module and has better results for road extraction with a richer directionality. In general, the DCCA module has a breakthrough in road extraction by exploiting geometric road directionality information.

In the future, we still have relevant research that needs further exploration. The directionality of the DCCA module is richer than that of the CCA module, but it is still confined to eight directions. Roads in actual situations are often in more than eight

directions. There are some circular roads whose direction changes continuously. How to deal with the directional geometric information of these roads is a question worth considering in the future.

Author Contributions: C.C. contributed toward creating the original ideas of the paper. W.C. conceived and designed the experiments. C.C., H.Z., X.H. prepared the original data, performed the experiments and analysed the experimental data with the help of W.C., C.C. wrote and edited the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Key R & D Program of China (Grant No. 2018YFC0810600, 2018YFC0810605).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

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

References

1. Geng, K.; Sun, X.; Yan, Z.; Diao, W.; Gao, X. Topological Space Knowledge Distillation for Compact Road Extraction in Optical Remote Sensing Images. *Remote Sens.* **2020**, *12*, 3175. [CrossRef]
2. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.
3. He, H.; Wang, S.; Yang, D.; Wang, S.; Liu, X. An road extraction method for remote sensing image based on Encoder-Decoder network. *Acta Geod. Cartogr. Sin.* **2019**, *48*, 330.
4. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
5. Luo, W.; Li, Y.; Urtasun, R.; Zemel, R. Understanding the effective receptive field in deep convolutional neural networks. *arXiv* **2016**, arXiv:1701.04128.
6. Liu, W.; Rabinovich, A.; Berg, A.C. Parsenet: Looking wider to see better. *arXiv* **2015**, arXiv:1506.04579.
7. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv* **2015**, arXiv:1511.07122.
8. Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; Liu, W. Ccnet: Criss-cross attention for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 603–612.
9. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European conference on computer vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
10. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
11. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
12. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Springer: Cham, Switzerland, 2015.
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. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef]
15. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
16. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
17. 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, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
18. Zhao, H.; Zhang, Y.; Liu, S.; Shi, J.; Loy, C.C.; Lin, D.; Jia, J. Psanet: Point-wise spatial attention network for scene parsing. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 267–283.
19. Mnih, V.; Heess, N.; Graves, A. Recurrent models of visual attention. *arXiv* **2014**, arXiv:1406.6247.
20. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
21. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762.
22. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.

23. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154.
24. Ulku, I.; Akagunduz, E. A Survey on Deep Learning-based Architectures for Semantic Segmentation on 2D images. *arXiv* **2019**, arXiv:1912.10230.
25. Cordonnier, J.B.; Loukas, A.; Jaggi, M. On the Relationship between Self-Attention and Convolutional Layers. *arXiv* **2019**, arXiv:1911.03584.
26. Huang, X.; Zhang, L. Road centreline extraction from high-resolution imagery based on multiscale structural features and support vector machines. *Int. J. Remote Sens.* **2009**, *30*, 1977–1987. [CrossRef]
27. 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.
28. Unsalan, C.; Sirmacek, B. Road network detection using probabilistic and graph theoretical methods. *IEEE Trans. Geosci. Remote. Sens.* **2012**, *50*, 4441–4453. [CrossRef]
29. Cheng, G.; Wang, Y.; Gong, Y.; Zhu, F.; Pan, C. Urban road extraction via graph cuts based probability propagation. In Proceedings of the 2014 IEEE International Conference on Image Processing (ICIP), Paris, France, 27–30 October 2014; pp. 5072–5076.
30. Saito, S.; Yamashita, T.; Aoki, Y. Multiple object extraction from aerial imagery with convolutional neural networks. *Electron. Imaging* **2016**, *2016*, 1–9. [CrossRef]
31. Alshehhi, R.; Marpu, P.R. Hierarchical graph-based segmentation for extracting road networks from high-resolution satellite images. *ISPRS J. Photogramm. Remote Sens.* **2017**, *126*, 245–260. [CrossRef]
32. Liu, B.; Wu, H.; Wang, Y.; Liu, W. Main road extraction from zy-3 grayscale imagery based on directional mathematical morphology and vgi prior knowledge in urban areas. *PLoS ONE* **2015**, *10*, e0138071. [CrossRef]
33. Sujatha, C.; Selvathi, D. Connected component-based technique for automatic extraction of road centerline in high resolution satellite images. *EURASIP J. Image Video Process.* **2015**, *2015*, 1–16. [CrossRef]
34. 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]
35. Cheng, G.; Zhu, F.; Xiang, S.; Pan, C. Road centerline extraction via semisupervised segmentation and multidirection nonmaximum suppression. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 545–549. [CrossRef]
36. Qiaoping, Z.; Couloigner, I. Automatic road change detection and GIS updating from high spatial remotely-sensed imagery. *Geo-Spat. Inf. Sci.* **2004**, *7*, 89–95. [CrossRef]
37. Song, M.; Civco, D. Road extraction using SVM and image segmentation. *Photogramm. Eng. Remote Sens.* **2004**, *70*, 1365–1371. [CrossRef]
38. Das, S.; Mirnalinee, T.T.; Varghese, K. Use of salient features for the design of a multistage framework to extract roads from high-resolution multispectral satellite images. *IEEE Trans. Geosci. Remote Sens.* **2011**, *49*, 3906–3931. [CrossRef]
39. Alvarez, J.M.; Gevers, T.; LeCun, Y.; Lopez, A.M. Road scene segmentation from a single image. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 376–389.
40. Zhong, Z.; Li, J.; Cui, W.; Jiang, H. Fully convolutional networks for building and road extraction: Preliminary results. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; pp. 1591–1594.
41. Zhang, Z.; Liu, Q.; Wang, Y. Road extraction by deep residual u-net. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 749–753. [CrossRef]
42. Peng, B.; Li, Y.; Fan, K.; Yuan, L.; Tong, L.; He, L. New Network Based on D-Linknet and Densenet for High Resolution Satellite Imagery Road Extraction. In Proceedings of the IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 3939–3942.
43. 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 the2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018; pp. 172–181. [CrossRef]