International Journal of Geo-Information

MDPI

Article

Non-Local Feature Search Network for Building and Road Segmentation of Remote Sensing Image

Cheng Ding [1], Liguo Weng [1,*], Min Xia [1] and Haifeng Lin [2]

[1] Jiangsu Collaborative Innovation Center on Atmospheric Environment and Equipment Technology, Nanjing University of Information Science and Technology, Nanjing 210044, China; dingcheng@nuist.edu.cn (C.D.); xiamin@nuist.edu.cn (M.X.)

[2] College of Information Science and Technology, Nanjing Forestry University, Nanjing 210037, China; haifeng.lin@njfu.edu.cn

* Correspondence: 002311@nuist.edu.cn

Abstract: Building and road extraction from remote sensing images is of great significance to urban planning. At present, most of building and road extraction models adopt deep learning semantic segmentation method. However, the existing semantic segmentation methods did not pay enough attention to the feature information between hidden layers, which led to the neglect of the category of context pixels in pixel classification, resulting in these two problems of large-scale misjudgment of buildings and disconnection of road extraction. In order to solve these problem, this paper proposes a Non-Local Feature Search Network (NFSNet) that can improve the segmentation accuracy of remote sensing images of buildings and roads, and to help achieve accurate urban planning. By strengthening the exploration of hidden layer feature information, it can effectively reduce the large area misclassification of buildings and road disconnection in the process of segmentation. Firstly, a Self-Attention Feature Transfer (SAFT) module is proposed, which searches the importance of hidden layer on channel dimension, it can obtain the correlation between channels. Secondly, the Global Feature Refinement (GFR) module is introduced to integrate the features extracted from the backbone network and SAFT module, it enhances the semantic information of the feature map and obtains more detailed segmentation output. The comparative experiments demonstrate that the proposed method outperforms state-of-the-art methods, and the model complexity is the lowest.

Keywords: semantic segmentation; building and road segmentation; self-attention; deep learning

Citation: Ding, C.; Weng, L.; Xia, M.; Lin, H. Non-Local Feature Search Network for Building and Road Segmentation of Remote Sensing Image. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 245. https://doi.org/10.3390/ijgi10040245

Academic Editors: Stamatis Kalogirou and Wolfgang Kainz

Received: 25 February 2021
Accepted: 3 April 2021
Published: 7 April 2021

1. Introduction

As the material carrier of human survival and development, land resources have the characteristics of fixed location, non-renewable, unbalanced distribution of resources and so on [1]. With the rapid development of population and socio-economic systems, the remaining disposable land resources are decreasing day by day. Therefore, the overall planning and rational planning of land resources has important social value. For urban areas, most of the landforms are composed of buildings and roads, the accurate segmentation of buildings and roads can help realize macro-urban planning. Therefore, the automatic segmentation of buildings and roads in remote sensing images is highly necessary.

In the past decades, many scholars had proposed effective feature engineering remote sensing image segmentation methods. For example, Yuan et al. [2] used the local spectral histograms to calculate the spectral and texture features of the image. Each local spectral histograms linearly combined several representative features, and finally realized the remote sensing image segmentation by weight estimation. Li et al. [3] proposed an improvement on the two key steps of label extraction and pixel labeling in the process of segmentation, which could effectively and efficiently improve the accuracy of high-resolution image edge segmentation. Fan et al. [4] proposed a remote sensing image segmentation method based on prior information. This method used single point iterative weighted fuzzy c-means

clustering algorithm to solve the impact of data distribution and random initialization of clustering center on clustering quality. The above feature engineering segmentation methods could effectively segment remote sensing images. However, they have some problems, such as poor noise resistance, slow segmentation speed and artificial parameter design, and could not competent for the tasks of automatic segmentation of large quantities of data.

In recent years, convolutional neural networks (CNNs) had achieved great success in many fields, such as health care [5,6], marketing [7], power management [8], civil engineering [9], distributed database [10], cyber security [11] and so on. The field of computer vision semantic segmentation is no exception. In the semantic segmentation, CNNs not only has strong noise resistance, but also can realize the automatic segmentation of a large number of data, and has achieved excellent segmentation performance. Full Convolutional Network (FCN) was proposed by Long et al. [12], and it was the first time to use full convolutional neural network to achieve image semantic segmentation, laying a foundation for subsequent segmentation methods. Ronneberger et al. [13] proposed U-shaped structure (U-Net) for semantic segmentation. Based on the FCN framework, U-Net had improved the feature fusion method, and the features of different grades were fused to realize the feature reuse. Fusion of different levels of feature maps enabled the network to contain multi-level semantic information and improve the segmentation accuracy. However, compared with FCN, the calculation amount was increased to a certain extent. Zhao et al. [14] proposed Pyramid Scene Parsing Network (PSPNet) using pyramid structure to aggregate the context information of different regions and can mine the global context information. DeeplabV3+ proposed by Chen et al. [15] used atrous convolution to construct multi-scale pyramid feature map, which enabled subsampling to obtain multi-scale context information and obtain larger receptive field without bringing computational overhead.

Liu et al. [16] proposed a new multi-channel deep convolutional neural network. This network had solved the problem that the spatial and scale features of segmenting objects were lost in some remote sensing images, but it was easy to make mistakes in the case of shadow occlusion. Aiming at the super-high resolution and complex features of remote sensing images, Qi et al. [17] proposed a segmentation model using multi-scale convolution and attention mechanisms. However, the attention mechanism could only capture local receptive field. Therefore, it was necessary to use the self-attention method to obtain important information through its own global receptive field and made effective use of it in remote sensing images. Cao et al. [18] proposed a deep feature fusion method based on self-attention, which performed deep feature fusion for complex objects in remote sensing scene images and emphasized their weight. Sinha et al. [19] used a guided self-attention mechanism to capture the context dependencies of the pixels in the image. Moreover, additional loss was used to emphasize feature correlation between different modules, which guided the attention mechanism to ignore irrelevant information and focused on more discriminant areas of the image. The above self-attention methods [18,19] had achieved initial results in the field of remote sensing images, but there were still more room for exploration, such as the use of self-attention mechanism to achieve hidden layer feature transfer.

In summary, these convolutional neural semantic segmentation networks [12–19] had made significant contributions to the field of semantic segmentation in computer vision. Compared with the feature engineering segmentation method [2–4], it had strong anti-noise performance and could realized end-to-end mass automatic segmentation. FCN [12] and U-Net [13] achieved feature enhancement through feature fusion at different levels and repeated use of feature maps. However, segmentation target lacks scene understanding, so PSPNet [14] built feature pyramid pooling layer, used different size pooling layers to splicing and fusion features, and finally performed feature analysis on the network to obtain scene understanding of segmentation target. In the segmentation process, there were segmentation targets of different scales, Deeplabv3+ [15] used atrous convolution

of different atrous rates to achieve multi-scale fusion. The above convolutional neural network models [12–15] put forward analysis for different problems in the segmentation process, including feature map reused and fusion of different levels of features, feature pyramid pooling layer to realized segmentation target scene understanding, and multi-scale feature fusion of atrous convolution with different atrous rates. However, in the process of feature fusion of these networks, almost all feature maps were directly concatenated and merged in the channel dimension, and the feature information of the hidden layers (the channel dimension of feature map) were not independently developed and utilized. Ignoring the importance level mining of hidden layer features led to the lack of category information of context pixels in pixel classification, resulting in problems such as large area misjudgment of building and road disconnection. In addition, the following semantic segmentation methods [12–15] are high complexity, slow reasoning speed and high cost of model training. To solve these problems, this paper proposes a Non-Local Feature Search Network (NFSNet). The network can improve the segmentation accuracy of buildings and roads from remote sensing images, and help achieve accurate urban planning through high-precision buildings and roads extraction. In general, there are three contributions in our work: (1) the Self-Attention Feature Transfer (SAFT) module is constructed through the self-attention method to effectively explore the feature information of the hidden layer. A feature map containing the category information of each pixel and the category semantic information of the context pixels are obtained. To avoid the problem of large area misjudgement of building and road disconnection. (2) Global Feature Refinement (GFR) module is constructed, and the hidden layer feature information extracted from the SAFT module is effectively integrated with the backbone network. The GFR module guides the backbone network feature map to obtain the feature information in the hidden layer spatial dimension, and enhances the semantic information of the feature map. It helps to restore the feature map with more precise up-sampling, and improves the segmentation accuracy. (3) Experiments are carried out on remote sensing image semantic segmentation dataset and obtain 70.54% mean intersection over union, which outperforms the existing model. In addition, the amount of model parameters and model complexity are the lowest among all comparison models, saving training time and cost.

2. Methodology

In the process of feature fusion, the existing semantic segmentation methods generally used splicing method to fuse the feature map in the channel dimension. The semantic information of the hidden layers (the channel dimension of the feature map) were not been developed separately. Due to the high resolution of remote sensing images and the high complexity of the target, the pixel failed to capture the category of context pixels in the semantic segmentation of remote sensing images, resulting in the misjudgment of large area of building and road disconnection. Secondly, the existing semantic segmentation algorithm models [12–15] had high complexity and high reasoning time cost. In order to solve these two problems, this paper proposes a Non-Local Feature Search Network (NFSNet) for building and road segmentation in remote sensing images. The overall framework of the NFSNet is shown in Figure 1. The NFSNet proposed in this work is an end-to-end training model, and the overall framework is divided into encoding network and decoding network. ResNet [20] is used as the backbone network for feature extraction in the encoding network, the decoding network constructs Self-Attention Feature Transfer (SAFT) module and Global Feature Refinement (GFR) module. The decoding network is the hidden feature search part in Figure 1. The SAFT module explores the feature associations between hidden layers through its self-attention query. The semantic information of the hidden layer is transferred to the original feature map, and a feature map containing the category information of each pixel itself and its context pixels are obtained. So as to improve the problem of large-area misclassification of building and road disconnection in the segmentation process, the GFR module effectively integrates the backbone network feature map and the hidden layer semantic information extracted by SAFT. The GFR

module makes global average pooling of features extracted by SAFT, instructs backbone network feature map to obtain semantic information of hidden layer in spatial dimension, and improves segmentation accuracy. Finally, after the feature fusion of the encoding network and the decoding network, the bilinear interpolation 16 times upsampling is directly used to obtain the segmentation output result.

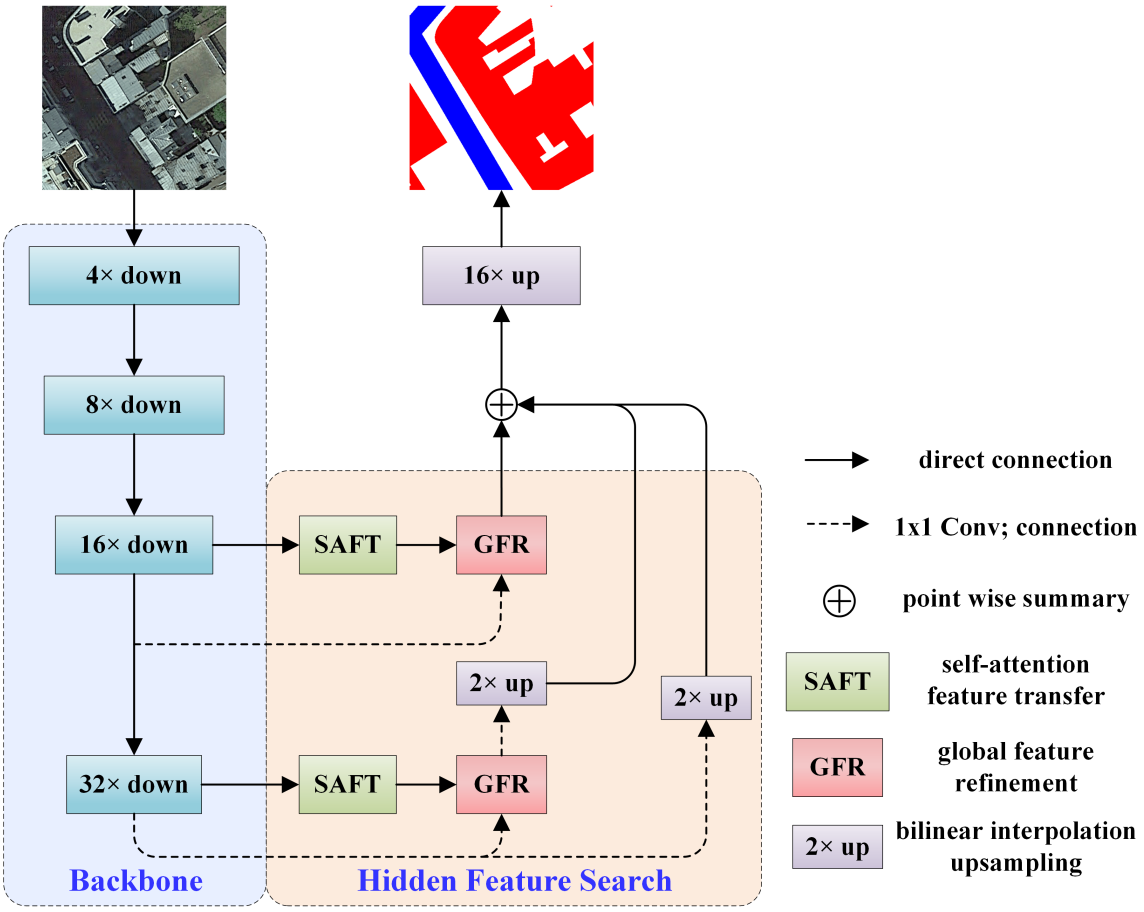


Figure 1. Non-Local Feature Search Network framework.

2.1. Encoding Network

In this paper, CNNs are used as the backbone network to achieve network feature extraction. In recent years, many excellent CNNs have emerged, such as VGG [21], GoogLeNet [22], and ResNet [20]. This work chooses ResNet as the backbone network for feature extraction after weighing the number of network parameters and accuracy. ResNet is the first method to propose the use of skip connections to mitigate model degradation as the network depth increases. ResNet sets different convolution layers for different application scenarios, including 18, 34, 50, 101 and 152 layers respectively. NFSNet proposed in this paper is a lightweight network, so the least number of convolution layers network ResNet-18 is selected as the backbone network. ResNet-18 is sampled layer by layer to obtain the feature map with rich semantic information, the size of the feature map of the last layer is 1/32 of the input image. ResNet-18 is sampled to obtain 1/16 size feature map and 1/32 feature map (hereinafter referred to as CNN), CNN feature maps of different sizes containing rich semantic information are used as the output of the encoding network and provided to the decoding network for semantic information decoding.

2.2. Decoding Network

The decoding network is responsible for decoding the encoded information and restoring the semantic feature information of the feature map. The input of the decoding network

is the feature map of 1/16 and 1/32 sizes of the original image, which is sampled from the backbone network of the encoding network. The decoding network is mainly composed of SAFT module and GFR module. The SAFT module uses the self-attention mechanism to mine the association between hidden layers and transfers the feature information of hidden layers to the original feature map. A feature map containing the category information of each pixel's own category and its context pixels are obtained. The feature map containing the semantic information of hidden layer can alleviate the problems of building misclassification and road disconnection. The GFR module refines the semantic information extracted from SAFT and integrates it with the feature map of the backbone network. GFR module can helps the backbone network feature graph to obtain the semantic information of the hidden layer in the spatial dimension, and improves the segmentation accuracy.

2.2.1. Self-Attention Feature Transfer Module

The prototype of self-attention mechanism was proposed by Vaswani [23], which usually used for information extraction in the encoding and decoding process of natural language processing. When a text message is entered, the relationship between each character in the text and its context is extracted to obtain the importance degree of each character in the text [24]. Inspired by this idea, the self-attention mechanism is embedded into the hidden layers of convolutional neural network. The association between each hidden layer and its context hidden layers are obtained through self-attention, so as to realize the transfer of hidden layer feature information to the original feature map. When the feature maps containing the semantic information of the hidden layers are obtained, the category of the current pixel and its context pixels can be captured during pixel classification, which can effectively reduce pixel misclassification and avoid large area building misjudgment and road disconnection.

The self-attention feature transfer module proposed in this paper is shown in Figure 2. Firstly, the query matrix, key value matrix and numerical value matrix are obtained by three 1×1 convolutions and mapping functions of φ, ζ, η; secondly, after multiplying the query matrix and the key value matrix, softmax is calculated in the first channel dimension; finally, depth separable convolution is used to enhance features. The input of this module is a feature map of 1/32 (or 1/16) size from the original image after the backbone network down-sampled. The dimension of feature map X (CNN in Figure 2) is $C' \times H \times W$. Due to the number of channels $C' = 512$ (or $C' = 256$) too large, the calculation amount in the parameter transfer process is relatively large. In order to reduce the computational burden, 1×1 convolution is used to reduce the dimensionality of features, get the feature map with $C = C'/2$ channels. The three branches go through 1×1 convolution, and the batch normalization (BN) [25] and ReLU activation [26] layers get $\hat{X}_q, \hat{X}_k, \hat{X}_v$ with dimension $C \times H \times W$ respectively. The calculation process is shown in Equation (1):

$$\hat{X} = \sigma(\beta(Conv_{1\times1}(X))), \tag{1}$$

where $Conv_{1\times1}$ is 1×1 convolution, β is BN, σ is ReLU activation function.

Next, we need to calculate the attention information between channels, mining the semantic information between channels, so as to capture the category information of each pixel and its context pixels. Three mapping functions φ, ζ, η are used to map $\hat{X}_q, \hat{X}_k, \hat{X}_v \in \mathbb{R}^{C\times H\times W}$to the query matrix $\hat{X}_q$, key matrix $\hat{X}_k$ and value matrix $\hat{X}_v$ of the channel respectively. The purpose of feature mapping is to facilitate matrix multiplication. Matrix multiplication can transfer the extracted feature information of hidden layer to the original feature map [27].

Through flattening function F_s, mapping function φ flattens the last two dimensions of feature map $\hat{X}_q$ into $X_q \in \mathbb{R}^{C\times(HW)}$. The calculation process is shown in Equation (2).

$$X_q = F_s(\hat{X}_q), \tag{2}$$

where ζ is similar to φ. Firstly, the last two dimensions of the feature map $\hat{X}_k$ are flattened into $X_k{}' \in \mathbb{R}^{C\times(HW)}$ by using the flattening function F_s. Then transpose $X_k{}'$ using the function T_s to get $X_k \in \mathbb{R}^{(HW)\times C}$. The transpose operation is to match the dimensions when multiplying $\hat{X}_q$ and $\hat{X}_k$ matrices. See Equation (3) for the calculation process.

$$X_k = T_s(F_s(\hat{X}_k)), \tag{3}$$

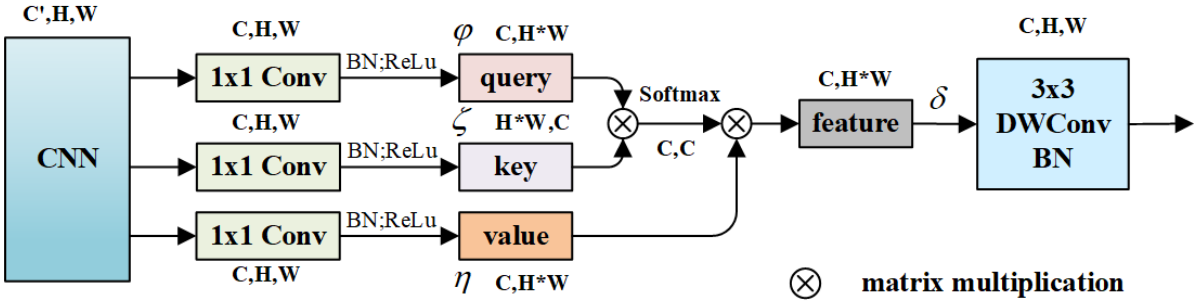


Figure 2. Structure diagram of self-attention feature transfer module. The input convolutional neural network (CNN) of the module corresponds to 16× down and 32× down in Figure 1, φ, ζ, η are mapping functions respectively, DWConv represents the depth separable convolution, BN represents the Batch Normalization, C is the number of channels, H is the height of the feature map, and W is the width of the feature map.

The value matrix of the channel X_v is obtained by mapping function η in the same way as the channel query matrix X_q, and Equation (4) is obtained by referring to Equation (2).

$$X_v = F_s(\hat{X}_v), \tag{4}$$

The query matrix X_q, key value matrix X_k and value matrix X_v are obtained. Query matrix is used to query the feature information between channels by the key matrix. The key matrix is multiplied by the query matrix, which can get the feature matrix of dimension $C \times C$. Softmax is performed on the first dimension of the obtained feature matrix, and normalized scores are generated for each channel to obtain the feature matrix $\bar{X}$. The calculation process is shown in Equation (5):

$$\bar{X} = \Omega(X_k \times X_q), \tag{5}$$

where $\times$ is matrix multiplication, Ω is calculated softmax in the first dimension.

The importance of each channel of the eigenmatrix $\bar{X}$ is distinguished. Multiply the value matrix X_v with the containing the degree of channel importance matrix $\bar{X}$, the eigenmatrix $\tilde{X}' \in \mathbb{R}^{C\times(HW)}$ can be obtained. The mapping function δ decomposes the second dimension of the feature matrix $\tilde{X}'$ into two dimensions through the flattening function $F_s{}'$, a two-dimensional matrix $\tilde{X}' \in \mathbb{R}^{C\times(HW)}$ maps to a three-dimensional matrix $\tilde{X} \in \mathbb{R}^{C\times H\times W}$. The calculation process is shown in Equation (6):

$$\begin{aligned} \tilde{X}' &= (\bar{X} \times X_v), \\ \tilde{X} &= F_s{}'(\tilde{X}'). \end{aligned} \tag{6}$$

where $\times$ is matrix multiplication, $F_s{}'$ is flattening function.

The attention information between each channel is extracted in $\tilde{X}$, which can capture the category of its context pixels and search for the characteristics of the hidden layer. The hidden layer feature information is transferred to the original feature map, and the feature map containing the category information of each pixel and its context pixels are obtained. Thus, the problems of large area misclassification of building and road disconnection in the segmentation process can be improved.

Finally, the feature map $\tilde{X}$ obtained by the feature search of the hidden layer is feature-enhanced to extract the effective information of the feature map. Considering the

computational efficiency of the model, deep separable convolution is used for feature enhancement, and feature enhancement can be achieved without introducing more calculation parameters. Sets the groups of depth separable convolution to the number of channels [28]. After the depth separable convolution, the connection is Batch Normalization. The forward propagation is shown in Equation (7):

$$X_{out} = \beta(DWConv_{3\times3}(\tilde{X})). \tag{7}$$

where $DWConv_{3\times3}$ is the depth separable convolution of convolution kernel 3×3, β is Batch Normalization, $X_{out} \in \mathbb{R}^{C\times H\times W}$ is output.

2.2.2. Global Feature Refinement Module

After the SAT module explores the hidden layer feature information, this work builds the GFR module to fuse the hidden layer feature information with the backbone network feature map. The GFR module can guide the backbone network feature graph to obtain the rich semantic information of the hidden layer. The feature map with rich semantic information can help to restore the details better in the process of upsampling. The GFR module proposed in this work is shown in Figure 3. The GFR module integrates the backbone network feature map and the hidden layer feature map extracted by the SAFT module. Similar to the idea of SENet [29], the hidden layer feature map extracted by the SAFT module is globally averaged pooling to obtain the feature information of the hidden layer in the spatial dimension. The corresponding multiplication with the backbone network feature map can guide the backbone network feature map to obtain the semantic information of the hidden layer in the spatial dimension [30–32]. Finally, the backbone network feature map and the feature map extracted by the SAFT module are merged to improve the segmentation accuracy.

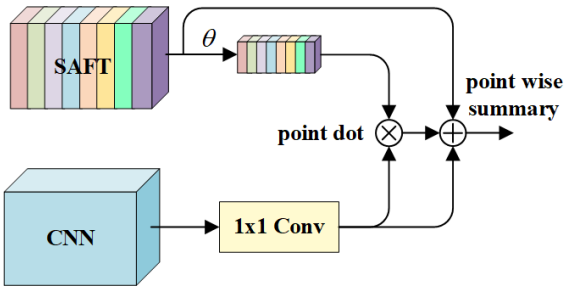


Figure 3. Structure diagram of global feature refinement module, θ is global average pooling.

GFR can fuse feature maps of different scales. As shown in Figure 1, GFR is used to fuse the backbone network feature map of 1/32 (or 1/16) size of the original image and SAFT feature map. Feature maps of different scales provide semantic information of different receptive fields. The number of output channels of the SAFT module is reduced to 1/2 of its input channels, and the input of the SAFT module is the backbone network feature map. Therefore, before the GFR module integrates SAFT module feature map and backbone network feature map, the channel number of backbone network feature map and SAFT module feature map should be standardized to the same level [33]. This work reduce the dimensionality of the channel of the backbone network feature map X (CNN in Figure 3) to 1/2 of the original backbone network feature map by 1×1 convolution, which matches the channel dimension of the SAFT module feature map. The output of the SAFT module X_{out} is globally averaged pooling by θ, and the feature map of the original dimension $C \times H \times W$ is mapped to $C \times 1 \times 1$, which can obtain the SAFT modules spatial dimension information. The reduced dimensionality of the backbone network feature map is multiplied by the spatial information of the SAFT module feature map in the channel dimension, and the backbone network feature map is guided to obtain spatial semantic information in the channel dimension [34]. Finally, the backbone network feature map containing the spatial

semantic information of the hidden layer, the original backbone network feature map and the SAFT module feature map are combined and fused. In this way, not only the original backbone network feature information and the hidden layer feature information extracted from the original SAFT module are retained, but also the backbone network feature map containing the spatial semantic information of the hidden layer is added. Through the GFR module, different types of feature images can be fused [35], which can help to further improve the segmentation accuracy. The calculation and derivation process of GFR is shown in Equation (8):

$$X_{GFR} = \theta(X_{out}) \cdot Conv_{1\times1}(X) + X_{out} + Conv_{1\times1}(X). \quad (8)$$

where θ is global average pooling on a channel dimension, $Conv_{1\times1}$ is 1×1 convolution, $\cdot$ is corresponding multiplication, + is corresponding addition, X_{out} is the output of the SAFT module, X_{GFR} is the output of GFR module.

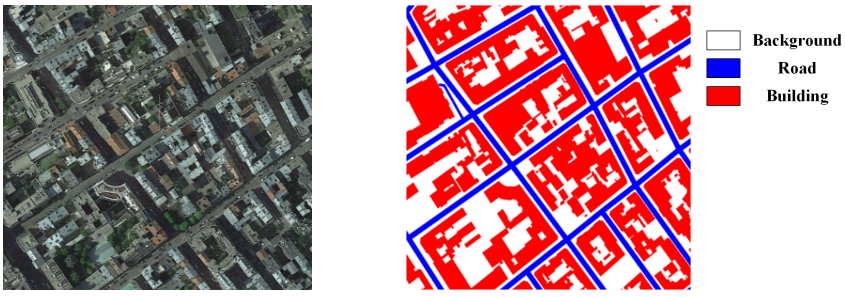
3. Experiments and Results

In order to verify the effectiveness of the proposed NFSNet, experiments were carried out on the open dataset Aerial Image Segmentation Dataset (AISD) [36] and ISPRS 2D Semantic Labeling Contest (ISPRS) [37]. The quantitative analysis indicators of the experiment adopted the overall accuracy rate (OA), recall rate (Recall), F1-Score and mean intersection over union (MIoU). The model proposed in this paper was compared with the current excellent semantic segmentation models FCN-8S [12], U-Net [13], DeeplabV3+ [15] and PSPNet [14]. The experimental results showed that the NFSNet proposed in this paper exceeded the comparison model in multiple evaluation indicators, which proved the effectiveness of the model proposed in this paper.

3.1. Datasets

3.1.1. AISD Dataset

The original images of AISD dataset were collected from OpenStreetMap online remote sensing image data, and the semantic segmentation dataset of high resolution remote sensing images were constructed by manual annotation. AISD included image data from six regions: Berlin, Chicago, Paris, Potsdam, and Zurich. In this paper, the Potsdam regional data were selected for the experiment, and the data set was named Potsdam-A. The Potsdam-A dataset contained a total of 24 original images and labels of 3000×3000 average size. A schematic diagram of training data is shown in Figure 4. Figure 4a is the original image and Figure 4b is the label. Potsdam-A consisted of three categories: building, road and background, corresponding to red, blue and white in Figure 4b.



(**a**) (**b**)

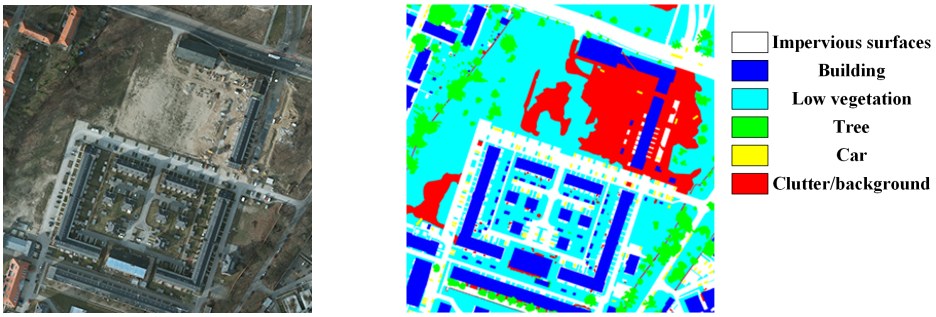
Figure 4. Potsdam-A data presentation; (**a**) original image; (**b**) label image.

Since the original picture size of Potsdam-A was too large for model training, we cropped the large size picture of 3000×3000 into the small size picture of 512×512, and finally obtained 1728 pictures of 512×512 size. When the amount of data was small, the learning feature ability of the model was weak and the generalization effect was poor.

In order to let the model have reliable learning capability, data enhancement was essential. We performed random horizontal flips, vertical flips, and 90-degree rotations on the original data set to expand to 4307 pictures. Finally, the data set was divided into 4000 training sets and 307 test sets.

3.1.2. ISPRS Dataset

ISPRS 2D Semantic Labeling Contest dataset is a high-resolution aerial image dataset with complete Semantic Labeling published by the International Society for Photogrammetry and Remote Sensing (ISPRS). The ISPRS dataset contained semantic segmentation images of the Potsdam region in the AISD dataset, so the Potsdam region in the ISPRS dataset was selected to verify the generalization performance of the model, and this dataset was named Potsdam-B. Potsdam-B contained a total of 38 finely labeled remote sensing images, there were five types of foreground: impervious surfaces, building, low vegetation, tree and car. The data display is shown in Figure 5, Figure 5a is the original image and Figure 5b is the label. In Figure 5b, a total of six categories are shown, including five foreground categories and one background category.



(**a**) (**b**)

Figure 5. Potsdam-B data presentation; (**a**) original image; (**b**) label image.

The average size of the pictures in the dataset Potsdam-B was 6000 × 6000, and the same cropping strategy of the Potsdam-A dataset was adopted to obtain 5184 pictures of 512 × 512 size. Finally, the data set was divided into 4684 training sets and 500 test sets.

3.2. Implementation Details

This work used overall accuracy rate (OA), recall rate (Recall), F1-Score and intersection over union (IoU) as the evaluation indicators of the model to verify the learning effect of the model, the calculation process is shown in Equations (9)–(13). OA is the proportion of predicted correct pixels in all pixels. Recall refers to the proportion of pixels in the actual positive sample predicted to be positive sample to pixels in the original positive sample. F1-score is the harmonic mean of recall and precision. Among them, precision is the proportion of the pixels predicted as positive samples to the pixels predicted as positive samples. IoU is the proportion of pixels that are predicted to be positive samples to all pixels. MIoU is the cumulative average of IoU of all categories.

$$OA = \frac{TP + TN}{TP + FP + FN + TN}, \tag{9}$$

$$Recall = \frac{TP}{TP + FN}, \tag{10}$$

$$Precision = \frac{TP}{TP + FP}, \tag{11}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}, \tag{12}$$

$$IoU = \frac{TP}{TP + FP + FN}. \tag{13}$$

The model in this work was a supervised learning method. At the end of the model, a loss function needed to be set to evaluate the gap between the predicted value and the true value. Cross entropy was mainly used to measure the difference between two probability distributions in information theory, and was often used as a loss function in deep learning. In this paper, the cross-entropy loss function (CE_{loss}) was used to measure the difference between the predicted value and the true value, and the difference value was used to guide the model to conduct back propagation and learn the optimal parameters. The derivation process of CE_{loss} is shown in Equation (14):

$$CE_{loss}(p, q) = -\frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{n} p(x_{ij}) log(q(x_{ij})). \tag{14}$$

where m is the number of samples, n represents the number of categories, $p(x_{ij})$ is variable (If the category j and the sample i are the same, it is 1, otherwise it is 0), $q(x_{ij})$is the probability sample i is predicted to be class j.

The network training parameters were as follows: using a single GTX1080TI graphics card for inference calculation on the Ubuntu16.04 platform. The model was built using the deep learning framework Pytorch, the model converged with 300 epochs, the initial learning rate was 0.001, and every 10 epochs was multiplied by the attenuation coefficient 0.85. Using adam as the optimizer to optimize the model, we set the weight_decay of the adam optimizer to 0.0001, and the other parameters as default values.

3.3. Analysis of Implementation Results

3.3.1. Comparison of Model Test Evaluation Index and Visualization Effect

(1) Main experimental dataset Potsdam-A experimental results

In order to verify the effectiveness of our proposed model, this work conducted comprehensive experiments on the Potsdam-A dataset, and various indicators on the test set exceeded the existing model. The specific quantitative experimental results are shown in Table 1, and the visual comparison effect is shown in Figure 6. The comparison models were U-Net, FCN-8S, DeeplabV3+ and PSPNet, the backbone networks for which were as consistent as possible with the original paper; the backbone networks of FCN-8S, DeeplabV3+ and PSPNet were VGG16, ResNet-50 and ResNet-50 respectively. In order to verify the effectiveness of the proposed GFR module, ablation experiments were carried out. The network without GFR module was tested, and the network with sat module was named NFSNet-1.

Table 1. Experimental results of Potsdam-A test set. The highest values for the different metrics are highlighted in bold, ↑ means the higher the better.

Methods	Backbone	Recall (%)↑	F1 (%)↑	OA (%)↑	MIoU (%)↑
U-Net	-	83.29	81.78	77.35	63.61
FCN-8S	VGG16	85.35	84.01	79.90	66.99
DeeplabV3+	ResNet-50	85.70	85.06	81.12	68.61
PSPNet	ResNet-50	85.48	85.70	81.41	69.11
NFSNet-1	ResNet-18	86.64	86.00	82.25	70.17
NFSNet	ResNet-18	**86.96**	**86.31**	**82.43**	**70.54**

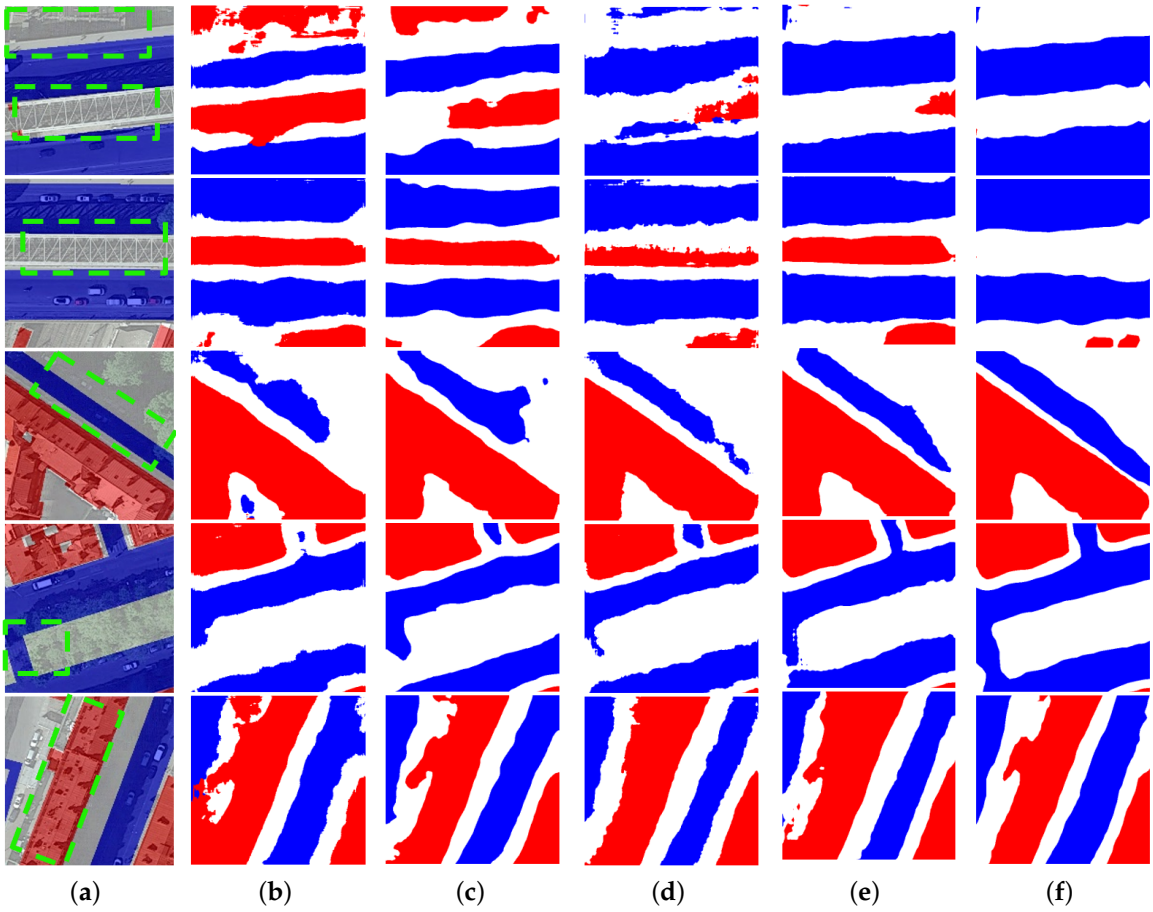


Figure 6. Visual effect comparison of Potsdam-A test set; (**a**) the superposition of the original image and the label; (**b**) U-Net; (**c**) FCN-8S; (**d**) DeeplabV3+; (**e**) PSPNet; (**f**) NFSNet.

As can be seen from Table 1, the NFSNet network proposed in this paper, recall, F1, OA and MIoU obtained 86.96%, 86.31%, 82.43% and 70.54% respectively. The network proposed in this work strengthened the importance search between hidden layer channels, effectively integrated the hidden layer feature information with the backbone network feature map, reducing the large-area misjudgement of building and road disconnection in remote sensing images. All four indicators exceeded comparison networks [12–15]. The U-Net network with the lowest indicators, OA and MIoU achieved 77.35% and 63.61% respectively. FCN-8S, which used VGG16 as the backbone network, had slightly improved indicators, with 79.90% OA and 66.99% MIoU. DeeplabV3+, which used dilated convolution to obtain a larger receptive field, had a certain improvement in segmentation accuracy compared with FCN-8S, with OA and MIoU of 81.12% and 68.61%, respectively. Compared with Deeplabv3+, PSPNet used deep convolutional network to extract high-level feature information and

feature pyramid module for multi-scale fusion, was 0.29 higher than that of OA and 0.5 higher than that of MIoU. The above advanced semantic segmentation network achieved satisfactory segmentation accuracy. However, in the process of feature fusion of these networks, almost all feature maps were directly concatenated and merged in the channel dimension, and the feature information of the hidden layers (the channel dimension of feature map) was not independently developed and utilized, leading to the neglect of the category of context pixels in pixel classification, resulting in problems such as large area misjudgment of building and road disconnection. Compared with the PSPNet with the highest index in the comparison network, the NFSNet proposed in this paper outperformed the PSPNet in both OA and MIoU by making full use of the characteristics of hidden layer, 1.02 higher on OA and 1.43 higher on MIOU. NFSNet-1 without GFR module achieved the highest accuracy except NFSNet, with OA reaching 82.25% and MIOU reaching 70.17%, the effectiveness of the proposed module was verified.

The IOU results of different model categories on the Potsdam-A test set are shown in Table 2. The IOU indexes of building, road and background categories proposed by NFSNet on the test set were 59.50%, 71.19% and 80.91% respectively, exceeding the four existing excellent comparison [12–15]. NFSNet was compared with other models, the category background IoU was 2.02 higher than the highest DeeplabV3+, the category road IoU was 1.74 higher than the highest PSPNet, the category building IoU was 0.73 higher than the highest PSPNet. It can be seen from the experimental results that the NFSNet proposed in this work effectively improved the segmentation accuracy of roads and buildings. The improvement of segmentation accuracy could effectively identify buildings and roads in remote sensing images, which is of great significance to realize accurate urban planning.

Table 2. Potsdam-A test set category IoU results. The highest values for the different metrics are highlighted in bold, ↑ means the higher the better.

Methods	Backbone	Background (%)↑	Road (%)↑	Building (%)↑	MIoU (%)↑
U-Net	-	51.76	61.95	77.13	63.61
FCN-8S	VGG16	55.59	66.02	79.36	66.99
DeeplabV3+	ResNet-50	57.48	68.84	79.49	68.61
PSPNet	ResNet-50	56.99	69.45	80.18	69.11
NFSNet-1	ResNet-18	59.37	70.85	80.29	70.17
NFSNet	ResNet-18	**59.50**	**71.19**	**80.91**	**70.54**

In order to facilitate intuitive comparison of model prediction results, this work visualized the prediction results of different models and obtained Figure 6. Figure 6 shows a total of five prediction maps, and each row in Figure 6 represents a comparison map of one image. Figure 6 is divided into six columns, column (a) is the superposition of the original image and the label, while column (b)–(f) respectively correspond to the visualization diagram of the predicted results of U-Net, FCN-8S, DeeplabV3+, PSPNet and NFSNet. The green boxes in column (a) are the prominent effect area of NFSNet. From the first row in Figure 6, it can be seen that the NFSNet proposed in this paper had the best performance in segmentation noise control. The segmentation result realized the accurate extraction of roads and greatly reduced the misclassification of buildings. This achievement was attributed to the NFSNet proposed in this work, which made up for the existing network to ignore the use of hidden layer feature information, and fully excavated the hidden layer feature information. The feature map included the category of its context pixels during classification, helping to achieve accurate classification. From the second row of Figure 6, it can be seen that there were large areas of background misclassified as buildings from column (b) to column (e). The f-column network this work proposed could overcome this difficulty and accurately classify the background by exploring the semantic features of the hidden layer. The third row and the fourth row in Figure 6 reflect the situation of extracting disconnections from the comparison network roads.

Columns (b) to (f) show the effect of reducing road disconnection in order. The NFSNet proposed by us could basically extract the outline of the road, which was the result of fusing the hidden layer feature information extracted by the SAFT module through the GFR module and the backbone network feature map. The fused feature map contained not only the rich location information of the backbone network but also the spatial dimension information of the hidden layer feature map, which effectively solved the problem of road disconnection. The last line in Figure 6 shows the problem of unclear buildings outlines. The f-column network this paper proposed fully explored the hidden layer feature information, provided rich semantic information feature maps, and achieved effective extraction of buildings outline.

(2) Generalization experimental dataset Potsdam-B experimental results

Since it was difficult for a single datum set to reflect the generalization performance of the model, this work used the Potsdam-B data set to test the generalization performance of the model. The results of the foreground experiment on the Potsdam-B test set are shown in Table 3. It can be seen from Table 3 that the recall, F1, OA and MIoU of NFSNet reached 89.12%, 87.41%, 87.52% and 78.09% respectively. All indicators achieved the highest value, which could prove the effectiveness and good generalization performance of the model proposed in this paper.

This work quantified each category on the Potsdam-B test set. Through experiments, the NFSNet proposed in this paper could achieve good segmentation effect in different categories. Among them, the IoU index of impervious surfaces (Imp_sur), building, low vegetation (Low_veg), tree and car were the highest values in the comparison model, which could prove that the NFSNet proposed by us had good generalization ability. The IoU results for each category on the Potsdam-B test set are shown in Table 4.

Table 3. Experimental results of Potsdam-B test set. The highest values for the different metrics are highlighted in bold, ↑ means the higher the better.

Methods	Backbone	Recall (%)↑	F1 (%)↑	OA (%)↑	MIoU (%)↑
FCN-8S	VGG16	86.06	86.96	85.58	75.69
U-Net	-	87.69	86.70	86.11	76.80
DeeplabV3+	ResNet-50	88.26	86.81	86.18	76.99
PSPNet	ResNet-50	87.53	86.85	86.73	77.09
NFSNet	ResNet-18	**89.12**	**87.41**	**87.52**	**78.09**

Table 4. Potsdam-B test set category IoU results. The highest values for the different metrics are highlighted in bold, ↑ means the higher the better.

Methods	Backbone	Imp_sur (%)↑	Building (%)↑	Low_veg (%)↑	Tree (%)↑	Car (%)↑	MIoU (%)↑
FCN-8S	VGG16	77.98	88.09	66.65	69.52	76.19	75.69
U-Net	-	79.61	88.82	69.83	69.51	76.23	76.80
DeeplabV3+	ResNet-50	80.15	89.80	70.39	68.69	75.92	76.99
PSPNet	ResNet-50	80.02	90.93	71.10	69.06	74.35	77.09
NFSNet	ResNet-18	**81.14**	**91.11**	**71.68**	**70.15**	**76.35**	**78.09**

In order to visually compare the segmentation effect of the model, this paper shows three renderings in Figure 7. Through comparison, it can be found that, because of the deep mining of hidden semantic information in the network proposed by us, the classification feature map contained the category of its context pixels, which greatly reduced the situation of large-area misclassification and continuous category disconnection. The second and third lines of Figure 7 well illustrate the advantages of our proposed model.

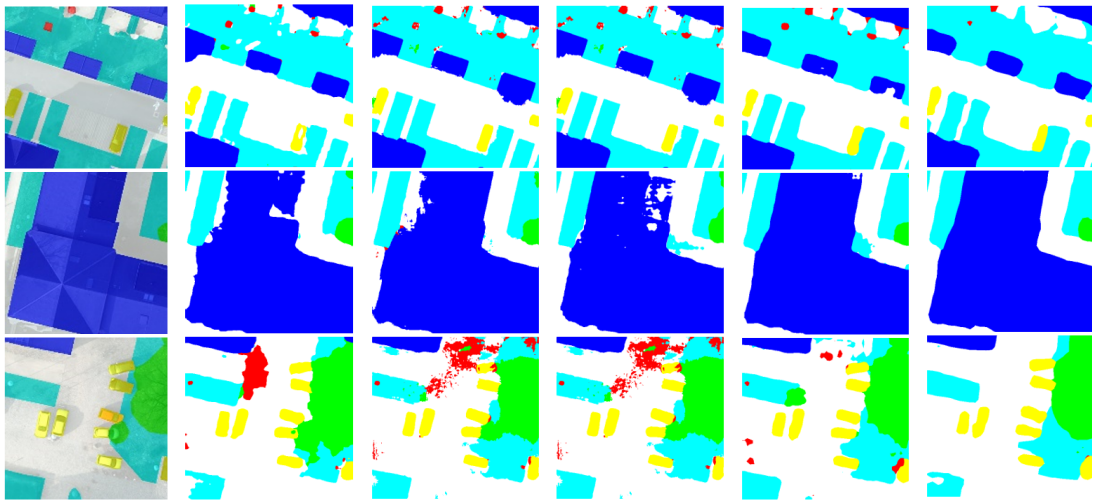


Figure 7. Visual effect comparison of Potsdam-B test set; (**a**) the superposition of the original image and the label; (**b**) FCN-8S; (**c**) U-Net; (**d**) DeeplabV3+; (**e**) PSPNet; (**f**) NFSNet.

3.3.2. Model Parameters and Complexity Experiments

The NFSNet this paper proposed not only had a high level of segmentation accuracy, but also had good advantages in model parameters, model complexity and inference speed. The number of parameters, model complexity and inference speed of different networks are shown in Table 5. Generally, floating point operations (FLOPs, GFLOPs is equal to 10^9 FLOPs) were used to measure the complexity of the model, and frames per second (FPS) was used to measure the reasoning speed. The inference speed test equipment was a single GTX1080TI, the input is a three-channel size picture, a total of three categories. When NFSNet used ResNet-18 as the backbone network, it had the least amount of parameters and GFLOPs, and the model inference speed was the fastest. The model parameter quantity was 11.91 M, which was only 24% of PSPNet. The model complexity was 9.82 GFLOPs, which was only 0.05% of U-Net. The inference speed was 116.26 FPS, which was 17.43 times that of PSPNet.

Table 5. Comparison of Parameters, floating point operations (FLOPs) and frames per second (FPS) of different networks. The highest values for the different metrics are highlighted in bold, ↑ means the higher the better, ↓ means the lower the better.

Methods	Backbone	Parameters↓	GFLOPs↓	FPS↑
U-Net	-	19.52M	184.01	11.03
FCN-8S	VGG16	15.12M	80.70	26.13
DeeplabV3+	ResNet-50	40.35	69.22	8.88
PSPNet	ResNet-50	48.94	177.46	9.54
NFSNet	ResNet-18	**11.91M**	**9.82**	**116.26**

In order to see the comparison of model segmentation accuracy (MIoU) and inference speed more intuitively (FPS), this paper provide a visual comparison chart of different models on the Potsdam-A data set, as shown in Figure 8. The abscissa of Figure 8 is the model name, and the ordinate is the segmentation accuracy MIoU and FPS. It can be seen intuitively from Figure 8 that NFSNet ranks first with the highest accuracy and fastest inference speed.

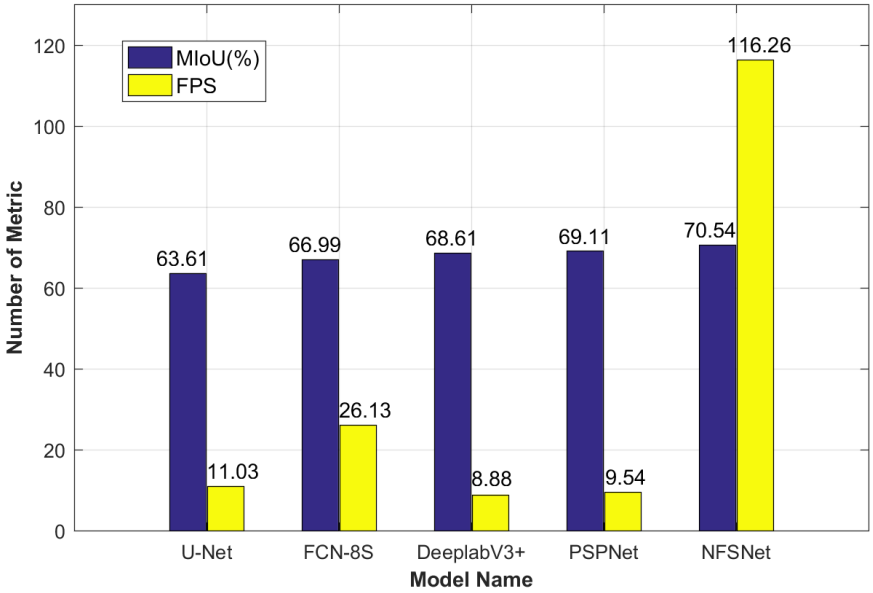


Figure 8. Visualized comparison of model segmentation accuracy and inference speed.

3.3.3. Backbone Network Quantification Experiment

Since the backbone network of the comparison model used ResNet-50, in order to reflect the fairness of the experiment, the backbone network was replaced with ResNet-50 with a deeper ResNet layer for comparison experiments, and the network was named NFSNet*. The quantitative comparison results of the backbone network are shown in Table 6. Although NFSNet* OA and MIoU were both 0.24 higher than NFSNet, the model parameters of 35.08M and model complexity of 26.25 GFLOPs were about three times that of NFSNet. Moreover, NFSNet had 116.26 FPS in inference speed, which was 80 FPS faster than NFSNet*.

This proves the advantage of using ResNet-18 as the backbone network. Without losing too much accuracy, our proposed NFSNet saved a lot of training costs with lower model complexity and parameter amount, and had a good speed performance in predictive inference.

Table 6. Quantitative comparison results of backbone network based on Potsdam-A dataset. The highest values for the different metrics are highlighted in bold, ↑ means the higher the better, ↓ means the lower the better.

Methods	Backbone	OA (%)↑	MIoU (%)↑	Parameters↓	GFLOPs↓	FPS↑
NFSNet	ResNet-18	82.43	70.54	**11.91M**	**9.82**	**116.26**
NFSNet*	ResNet-50	**82.67**	**70.78**	35.08M	26.25	36.00

4. Conclusions

In this paper, NFSNet is proposed for building and road segmentation of high resolution remote sensing images. Compared with existing semantic segmentation networks, NFSNet has the following advantages: (1) SAFT module is constructed to enhance the importance search between hidden layer channels and obtain the correlation between channels. The semantic information of the hidden layer is transferred to the original feature map, which contains the category semantic information of each pixel and its context pixels. Thus, the problems of large area misclassification of building and road disconnection in the segmentation process can be improved. (2) Using the GFR module, the hidden layer feature information extracted from the SAFT module is effectively fused with the backbone network feature map. In this way, the backbone network can obtain the feature information of the hidden layer in the spatial dimension, enhance the up-sampling feature information

and improve the segmentation accuracy. (3) The model has the lowest complexity but achieves the highest precision index.

However, there are still some defects in the segmentation of building and road: (1) There is room for improvement in the accuracy of edge segmentation of building and road. (2) When there is a lot of noise in the remote sensing image, the segmentation accuracy will decrease. We will continue to optimize NFSNet to improve the edge segmentation accuracy of building and road, and overcome the reduction of segmentation accuracy caused by large amounts of noise in remote sensing images. (3) The module structure proposed in this paper can be easily transplanted to other models, and we will experiment on more benchmark networks to expand richer application scenarios.

Author Contributions: Conceptualization, Cheng Ding, Min Xia and Liguo Weng; methodology, Cheng Ding and Min Xia; software, Cheng Ding; validation, Cheng Ding, Min Xia and Haifeng Lin; formal analysis, Cheng Ding, Min Xia and Liguo Weng; investigation, Cheng Ding and Min Xia; resources, Min Xia and Liguo Weng; data curation, Liguo Weng; writing—original draft preparation, Cheng Ding; writing—review and editing, Liguo Weng and Haifeng Lin; visualization, Cheng Ding; supervision, Min Xia; project administration, Liguo Weng and Min Xia; funding acquisition, Min Xia. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of PR China of grant number 42075130.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data and the code of this study are available from the corresponding author upon request.

Acknowledgments: The authors would like to thank the Assistant Editor of this article and anonymous reviewers for their valuable suggestions and comments.

Conflicts of Interest: No potential conflict of interest was reported by the author.

References

1. Pham, H.M.; Yamaguchi, Y.; Bui, T.Q. A case study on the relation between city planning and urban growth using remote sensing and spatial metrics. *Landsc. Urban Plan.* **2011**, *100*, 223–230. [CrossRef]
2. Yuan, J.; Wang, D.; Li, R. Remote sensing image segmentation by combining spectral and texture features. *IEEE Trans. Geosci. Remote Sens.* **2013**, *52*, 16–24. [CrossRef]
3. Li, D.; Zhang, G.; Wu, Z.; Yi, L. An edge embedded marker-based watershed algorithm for high spatial resolution remote sensing image segmentation. *IEEE Trans. Image Process.* **2010**, *19*, 2781–2787.
4. Fan, J.; Han, M.; Wang, J. Single point iterative weighted fuzzy C-means clustering algorithm for remote sensing image segmentation. *Pattern Recognit.* **2009**, *42*, 2527–2540. [CrossRef]
5. Sarki, R.; Ahmed, K.; Zhang, Y. Early detection of diabetic eye disease through deep learning using fundus images. *EAI Endorsed Trans. Pervasive Health Technol.* **2020**, *6*, e1. [CrossRef]
6. Sharma, M.; Kaur, P. A comprehensive analysis of nature-inspired meta-heuristic techniques for feature selection problem. *Arch. Comput. Methods Eng.* **2020**, 1–25. [CrossRef]
7. Sarkar, M.; De Bruyn, A. LSTM response models for direct marketing analytics: Replacing feature engineering with deep learning. *J. Interact. Mark.* **2021**, *53*, 80–95. [CrossRef]
8. Elbes, M.; Alrawashdeh, T.; Almaita, E.; AlZu'bi, S.; Jararweh, Y. A platform for power management based on indoor localization in smart buildings using long short-term neural networks. *Trans. Emerg. Telecommun. Technol.* **2020**, e3867. [CrossRef]
9. Ni, F.; Zhang, J.; Noori, M.N. Deep learning for data anomaly detection and data compression of a long-span suspension bridge. *Comput.-Aided Civ. Infrastruct. Eng.* **2020**, *35*, 685–700. [CrossRef]
10. Sharma, M.; Singh, G.; Singh, R. Design and analysis of stochastic DSS query optimizers in a distributed database system. *Egypt. Inform. J.* **2016**, *17*, 161–173. [CrossRef]
11. Ferrag, M.A.; Maglaras, L.; Moschoyiannis, S.; Janicke, H. Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study. *J. Inf. Secur. Appl.* **2020**, *50*, 102419. [CrossRef]
12. 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.
13. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.

14. 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.
15. 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.
16. Liu, W.; Zhang, Y.; Fan, H.; Zou, Y.; Cui, Z. A New Multi-Channel Deep Convolutional Neural Network for Semantic Segmentation of Remote Sensing Image. *IEEE Access* **2020**, *8*, 131814–131825. [CrossRef]
17. Qi, X.; Li, K.; Liu, P.; Zhou, X.; Sun, M. Deep attention and multi-scale networks for accurate remote sensing image segmentation. *IEEE Access* **2020**, *8*, 146627–146639. [CrossRef]
18. Cao, R.; Fang, L.; Lu, T.; He, N. Self-attention-based deep feature fusion for remote sensing scene classification. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 43–47. [CrossRef]
19. Sinha, A.; Dolz, J. Multi-scale self-guided attention for medical image segmentation. *IEEE J. Biomed. Health Inform.* **2020**. [CrossRef]
20. 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, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
21. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
22. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
23. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 5998–6008.
24. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
25. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv* **2015**, arXiv:1502.03167.
26. 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.
27. Xia, M.; Zhang, X.; Weng, L.; Xu, Y. Multi-stage feature constraints learning for age estimation. *IEEE Trans. Inf. Forensics Secur.* **2020**, *15*, 2417–2428. [CrossRef]
28. Xia, M.; Tian, N.; Zhang, Y.; Xu, Y.; Zhang, X. Dilated multi-scale cascade forest for satellite image classification. *Int. J. Remote Sens.* **2020**, *41*, 7779–7800. [CrossRef]
29. 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.
30. Xia, M.; Cui, Y.; Zhang, Y.; Xu, Y.; Liu, J.; Xu, Y. DAU-Net: A novel water areas segmentation structure for remote sensing image. *Int. J. Remote Sens.* **2021**, *42*, 2594–2621. [CrossRef]
31. Li, H.; Xiong, P.; An, J.; Wang, L. Pyramid attention network for semantic segmentation. *arXiv* **2018**, arXiv:1805.10180.
32. Xia, M.; Wang, T.; Zhang, Y.; Liu, J.; Xu, Y. Cloud/shadow segmentation based on global attention feature fusion residual network for remote sensing imagery. *Int. J. Remote Sens.* **2021**, *42*, 2022–2045. [CrossRef]
33. Qian, J.; Xia, M.; Zhang, Y.; Liu, J.; Xu, Y. TCDNet: Trilateral Change Detection Network for Google Earth Image. *Remote Sens.* **2020**, *12*, 2669. [CrossRef]
34. Xia, M.; Wang, K.; Song, W.; Chen, C.; Li, Y. Non-intrusive load disaggregation based on composite deep long short-term memory network. *Expert Syst. Appl.* **2020**, *160*, 113669. [CrossRef]
35. Chen, B.; Xia, M.; Huang, J. MFANet: A Multi-Level Feature Aggregation Network for Semantic Segmentation of Land Cover. *Remote Sens.* **2021**, *13*, 731. [CrossRef]
36. Kaiser, P.; Wegner, J.D.; Lucchi, A.; Jaggi, M.; Hofmann, T.; Schindler, K. Learning aerial image segmentation from online maps. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 6054–6068. [CrossRef]
37. Rottensteiner, F.; Sohn, G.; Gerke, M.; Wegner, J.D. *ISPRS Semantic Labeling Contest*; ISPRS: Leopoldshöhe, Germany, 2014.