remote sensing

MDPI

Article

DGFNet: Dual Gate Fusion Network for Land Cover Classification in Very High-Resolution Images

Yongjie Guo [1,2,3], Feng Wang [1,2,*], Yuming Xiang [1,2,3] and Hongjian You [1,2,3]

[1] The Key Laboratory of Technology in Geo-Spatial Information Processing and Application System, Chinese Academy of Sciences, Beijing 100190, China; guoyongjie191@mails.ucas.ac.cn (Y.G.); z199208081010@163.com (Y.X.); hjyou@mail.ie.ac.cn (H.Y.)
[2] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China
[3] School of Electronic, Electrical, and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: wangfeng003020@aircas.ac.cn; Tel.: +86-010-58887208

Abstract: Deep convolutional neural networks (DCNNs) have been used to achieve state-of-the-art performance on land cover classification thanks to their outstanding nonlinear feature extraction ability. DCNNs are usually designed as an encoder–decoder architecture for the land cover classification in very high-resolution (VHR) remote sensing images. The encoder captures semantic representation by stacking convolution layers and shrinking image spatial resolution, while the decoder restores the spatial information by an upsampling operation and combines it with different level features through a summation or skip connection. However, there is still a semantic gap between different-level features; a simple summation or skip connection will reduce the performance of land-cover classification. To overcome this problem, we propose a novel end-to-end network named Dual Gate Fusion Network (DGFNet) to restrain the impact of the semantic gap. In detail, the key of DGFNet consists of two main components: Feature Enhancement Module (FEM) and Dual Gate Fusion Module (DGFM). Firstly, the FEM combines local information with global contents and strengthens the feature representation in the encoder. Secondly, the DGFM is proposed to reduce the semantic gap between different level features, effectively fusing low-level spatial information and high-level semantic information in the decoder. Extensive experiments conducted on the LandCover dataset and the ISPRS Potsdam dataset proved the effectiveness of the proposed network. The DGFNet achieves state-of-art performance 88.87% MIoU on the LandCover dataset and 72.25% MIoU on the ISPRS Potsdam dataset.

Keywords: gated convolution; land cover classification; semantic segmentation; remote sensing images

Citation: Guo, Y.; Wang, F.; Xiang, Y.; You, H. DGFNet: Dual Gate Fusion Network for Land Cover Classification in Very High-Resolution Images. *Remote Sens.* **2021**, *13*, 3755. https://doi.org/10.3390/rs13183755

Academic Editor: Markus Immitzer

Received: 17 August 2021
Accepted: 16 September 2021
Published: 19 September 2021

1. Introduction

The rapid development of remote sensing sensors allows diverse access to very high-resolution (VHR) remote sensing images. A pixel-based land cover classification, also known as semantic segmentation, using very high spatial resolution images has significant application value in land resource management [1,2], urban planning [3,4], change detection [5,6], and other fields. Since optical sensors reflect the spectral characteristics of the ground target and show consistent features with the human visual system, optical remote sensing has become the mainstream method of fine land cover mapping. However, the clear and complex spatial structure features exceedingly increase the difficulty of land-cover classification [7]. Typical land-cover classification methods can be roughly separated into three categories: pixel-based classification methods, object-based classification methods, and patch-based classification methods. For the pixel-based method, spectral information provided by the high-resolution images shows prodigious variance for intra-class and the similarity between different classes, leading to lower land-cover mapping accuracies [8]. Furthermore, VHR remote sensing images usually contain a few bands and the pixel-based

classification method only considers the spectral information. It does not take count of the spatial characteristics and the topological relationship of ground objects of original images, making land-cover classification in VHR images more difficult. The object-based method can be divided into two stages: object generation and object determination. Firstly, those methods usually use feature extraction or clustering algorithms, such as simple linear iterative clustering (SLIC) [9], to generate objects. Subsequently, utilizing the spatio-temporal aggregation of multispectral data to determine the attribute of such objects is one of the better choices. The patch-based method is usually proposed in combination with DCNNs, which can capture more robust features. Different from traditional feature-extraction methods, such as SIFT [10], SURF [11], HOG [12], and ORB [13], which are expensive and require a special design, deep convolutional neural networks (DCNNs) can extract features automatically and have more outstanding feature expression abilities. In addition, DCNNs have a stronger non-linear fitting ability, which is better than other classifiers, making land-cover classification more accurate.

DCNN is a well-known model for feature learning, which can automatically learn features of different levels from raw images by stacking convolutional layers and downsampling operators. In 2012, Krizhevsky et al. [14] proposed the AlexNet and won the ILSVRC contest, which plays a significant role in deep learning. Since then, DCNNs have seen an explosive development and have been applied to different tasks, such as object detection [15–17], semantic segmentation [18–20], and image retrieval [21–23], etc. For the semantic segmentation task, Long et al. [18] is a pioneer in building a complete full convolutional network (FCN) to predict pixel-level labels in an end-to-end manner. However, such architecture captures the semantic information by stacking convolution layers through non-linearities and downsampling, reducing the spatial information of original images. Considering this, U-Net [24] adopted the structure of skip connection for feature fusion, which reuses the low-level features to retain the spatial detail to a certain extent. SegNet [25] recorded the corresponding max-pooling index in the process of encoding. In the decoding stage, the recorded pooling index was used to improve the decoding performance. DANet [26] added two types of attention modules to the traditional dilated FCN to simulate semantic interdependence in spatial and channel dimensions separately. PSPNet [27] introduced a pyramid pooling module to aggregate the context information based on different regions to mine the global context information and improve the segmentation effect. HRNet [28] achieves strong semantic information and precise location information through parallel multi-resolution branches and continuous information interaction between different branches.

The method based on DCNNs is introduced into the remote sensing scene naturally. Differing from natural images, the scale of VHR remote sensing images is much larger, as well as the radiometric resolution of such images being much higher, which makes it contain lots of complicated scenes. For example, there are many multi-scale surface objects such as huge buildings and small cars. As for small-scale ground objects, with the decrease in spatial resolution, its structural information may be lost, resulting in poor segmentation effects. According to the characteristics of remote sensing images, some researchers have established the network based on multi-scale feature fusion. Nogueira et al. [29] used extended convolution [30–32] to enhance the context information of feature aggregation. Li et al. [33] designed an additional branch that uses the boundary information of original images as input to improve image segmentation. Marmanis et al. [34] combined multi-scale features of different layers and used auxiliary data digital surface model (DSM) data to improve land-cover classification accuracy. In [35], Wang et al. proposed a gated convolutional neural network named GSN using the entropy of low-level features as a gate to refine the high-level features. The core idea of the above research is to fuse different level features directly. It is worth noting that low-level features in the shallow layers of DCNN can provide more detailed structural information, and high-level features in the deeper layers of DCNN contain more discriminative semantic information. Regardless of differences in the semantic information, direct fusion will inevitably embed background

noise of low-level features and thus affect the robustness of features, which may lead to the loss of detailed spatial information. We consider the difference between different level features as a semantic gap and propose an end-to-end network named the Dual Gate Fusion neural Network (DGFNet). In detail, DGFNet consists of two main components: Feature Enhancement Module (FEM) and Dual Gate Fusion Module (DGFM). The FEM combines local information with global contents and strengthens the feature representation in the encoder. Secondly, the DGFM is proposed to reduce the semantic gap between different level features, effectively fusing low-level spatial information and high-level semantic information in the decoder. In general, the main contributions of this paper can be summarized as follows:

1. We propose a simple but efficient encoder–decoder segmentation network, which effectively captures the global content and fuses different multi-level features, improving the performance of land-cover classification in VHR images.
2. We propose a novel feature enhancement module (FEM). It combines local information and global context information, enhancing the representation of different layer features.
3. A dual-gate fusion module (DGFM) with the gate mechanism is proposed, which promotes the fusion of low-level spatial features and high-level semantic features effectively.
4. Exhaustive experiments are conducted to prove the effectiveness of the proposed network. We also achieve the state-of-art performance of 88.87% MIoU on the LandCover dataset and 72.25% MIoU on the ISPRS Potsdam dataset.

The remainder of this paper is organized as follows: related works are presented in Section 2. In Section 3, we introduce the proposed DGFNet in details. Section 4 presents the experimental details and experimental results to validate our approach, followed by the conclusions in Section 5.

2. Related Work

2.1. DCNNs in Land-Cover Classification

Land-cover classification (also known as semantic segmentation) in VHR remote sensing images is difficult due to the large scale of original images in such a pixel-level task, which results in significant variation for intra-class and the similarity between different classes (e.g., trees and low vegetation). Since Long et al. [18] first built a full convolutional network (FCN) to achieve state-of-art performance in the pixel-level task, there has been a considerable number of works focussing on land-cover classification. For example, Mou and Zhu [36] proposed RiFCN, which recursively embedded different scale features into the learning framework to achieve accurate boundary inference and land cover classification. To increase the representation capacity of the framework such as FCN for land-cover classification in high-resolution remote sensing images, they also designed a relation module [37] to describe the relationships between observations in convolved images and augmented feature representations. Liu et al. [38] improved the classification result by integrating a spatial and channel relation-enhanced block with neural networks, which increases the variety of receptive field sizes. Chen et al. [39] fused different layer features of DCNN to restore the spatial resolution and improve the performance of land-cover classification. Sun and Wang [40] used an additional digital surface model (DSM) to restore the black spatial areas, such as shadows. They integrated the spectral information of color images with the geometry information of DSM, which improve the accuracy of land-cover classification. Diakogiannis et al. [41] proposed a novel DCNN architecture named ResUNet-a, which uses a UNet encoder/decoder as their backbone, in combination with residual connections, atrous convolutions, pyramid scene parsing pooling, and multi-tasking inference. ResUNet-a inferred a series of tasks sequentially, including the boundary of the objects, the distance transform of the segmentation mask, the segmentation mask, and a colored reconstruction of the input. Each of the tasks were conditioned on the inference of the previous ones, thus establishing a conditioned relationship between the various missions, which improve the final performance of land-cover classification.

The above work fully demonstrates the powerful feature extraction ability of DCNNs in land cover classification. With the increase in spatial resolution, VHR remote sensing images can capture more diverse scenes, which provide rich geometric information and feature information, increasing the difficulty of land-cover classification and reducing the classification accuracy of VHR images to certain extents. Due to the superior feature ability of DCNNs, the model based on DCNNs can be applied to different complex scenes, including VHR images. At present, the method based on DCNNs for land-cover classification in VHR images has become one of the mainstream methods.

2.2. Gate Mechanism in Neural Networks

Long short-term memory (LSTM) [42] is a famous framework for processing sequence data. LSTM can selectively transmit the previous information to the current state by the gate mechanism so LSTM can deal with the long-distance dependence problem effectively. Recently, Dauphin et al. [43], as well as Gehring et al. [44] extend the definition of gate mechanisms in conjunction with convolutional networks. They regard a convolutional layer without a non-linear function, followed by a unit with a sigmoid function as a "gate" unit. Li and Kameoka [45] used a gated CNN architecture instead of LSTM, which was introduced to model word sequences for language modeling and was shown to outperform LSTM language models trained in a similar setting. Subsequently, the gate mechanism is applied to various computer vision tasks. Yang et al. [46] combined the gate mechanism with hybrid connectivity for image classification to retain the capability of feature re-exploitation to some extent, which improves the accuracy of classification. Yu et al. [47] used gated convolution instead of partial convolution to obtain a better restoration effect for image inpainting, as well as Chang et al. [48], who proposed 3D gated convolutions to tackle the uncertainty of free-form masks for video inpainting. Rayatdoost et al. [49] utilized the gate mechanism to fuse the features among different models for emotion recognition from facial behaviors. Cao et al. [50] built a classification network with a linear skip gated connection which can benefit information propagation for action recognition. For aspect–category sentiment analysis (ACSA) and aspect–term sentiment analysis (ATSA) tasks, Xue and Li [51] proposed an efficient convolutional neural network with gating mechanisms to achieve state-of-art performance in related fields.

The mentioned research above verifies that the gate mechanism can promote the fusion and transmission of feature information. For pixel-level land-cover classification tasks, DCNNs usually adopt encoder–decoder architecture. The encoder captures semantic representation by stacking convolution layers and shrinking image spatial resolution. Furthermore, the decoder restores the spatial information by an upsampling operation and combines it with different level features through summation or skip connection. Due to the semantic gap between different level features, simple addition and skip-connection operations can not fully fuse the feature maps of different levels. Considering the character of the gate mechanism, we use the gate mechanism instead of the summation and skip-connection operation to integrate different level features. In short, embedding the gate mechanism in neural networks is a simple and effective method for feature learning and fusion.

3. Methods

With the increase in remote sensing image resolution, the distinct and complex spatial structure characteristics of remote sensing images become more visible. DCNNs capture the semantic representations with global contents by stacking convolution layers through non-linearities and downsampling. That operation reduces the spatial information of original images, which may affect the land-cover classification accuracy vastly. At present, the existing methods cannot recover the lost spatial information well. Therefore, the architecture based on DCNNs still needs to be improved in the decoder through an effective fusion approach.

In this paper, we will introduce the overall framework of DGFNet as shown in Figure 1, firstly. Then, two main modules in the DGFNet, including FEM and DGFM, are described in detail after the overall framework.

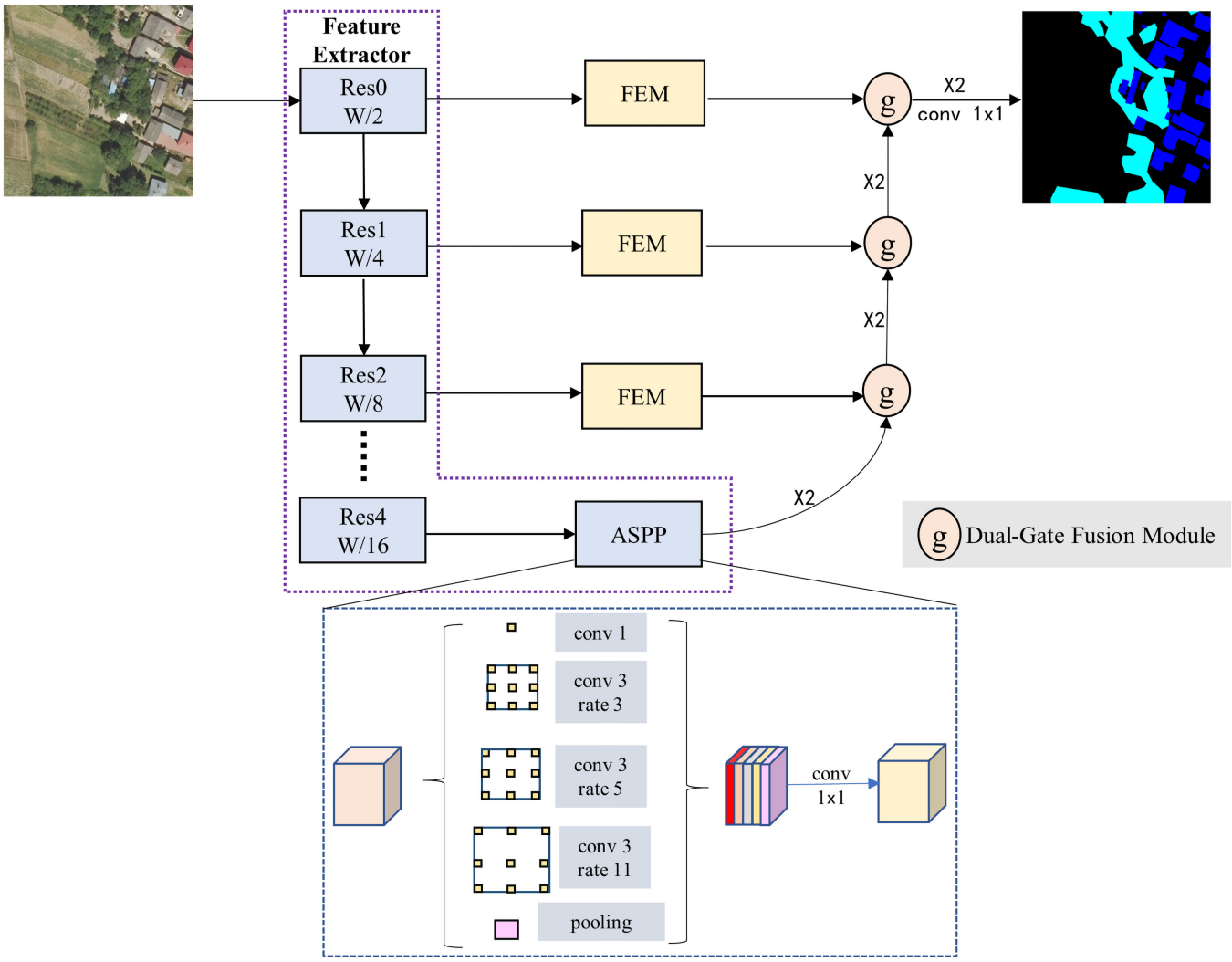


Figure 1. Overall framework about the proposed DGFNet, which consists of three parts: one-stream feature extractor, feature enhancement module (FEM) and dual gate fusion module (DGFM).

3.1. Overall Framwork

Our overall segmentation model is shown in Figure 1. We adapt the encoder–decoder architecture proposed in FCN [18] as the semantic segmentation framework. The encoder is composed of a one-stream feature extractor and feature enhancement module. In the encoder stage, we use ResNet-50 [52] combined with an atrous spatial pyramid pooling (ASPP) module proposed in [30], which consists of different rate-dilated convolution, as the feature extractor to obtain different-level feature maps. Following the feature extractor, we design the FEM to combine the local information with global contents, making the extracted feature more robust. In the decoder stage, we propose a dual gate fusion module (DGFM) to fuse the low-level (shallow-layers) features and high-level (deeper-layers) features, making the information fusion more sufficient, which is beneficial to the recovery of spatial details in the decoding phase. The details about the FEM and DGFM are described in the subsequent subsection.

3.2. Feature Enhancement Module

The semantic segmentation problem could be divided into the pixel-wise classification as well as location tasks [53], where the classification task requires global contents by stacking small-sized convolution kernels, while the location task needs large-sized convolution kernels. The requirement of the convolution kernel size is contradictory. DCNNs stack the small kernel size convolution through downsampling to obtain the global context, which reduces the spatial resolution of original images. Inspired by the SE module proposed in [54], we design the FEM to combine the local information with the global context, instead of using large-sized convolution kernels. Differing from the SE module, we define the context aggregation for the input feature maps, making the representation of global context more plain. As shown in Figure 2, the FEM consists of two substructures: channel regulation structure (Figure 2a) and context aggregation structure (Figure 2b). We model the channel structure as a simple residual layout, whose main purpose is to regulate the channel of different level feature maps from the feature extractor. More specifically, we define the regulated feature map as V, calculated as:

$$V = WU + \mathcal{R}(WU) \tag{1}$$

where W is the parameters of a linear transformation, U is a coarse feature map, and $\mathcal{R}(\cdot)$ is the residual branch. We observed that Equation (1) is similar to the formula of image sharpening so that the channel regulation structure can help enhance the spatial information to a certain extent. Following the channel regulation structure, we designed the context aggregation structure to combine global contents with the local information. Let $X \in R^{C\times H\times W}$ be the feature map of one input instance, where C is the channel of feature maps, $x_i \in R^N$ is the i th channel of the feature map X and flat into a vector, $N = H \times W$ is the number of positions in the feature map, $y \in R^{C\times 1\times 1}$ and $Z \in R^{C\times H\times W}$ denote intermediate variable, and the output of the context aggregation structure can be formulated as:

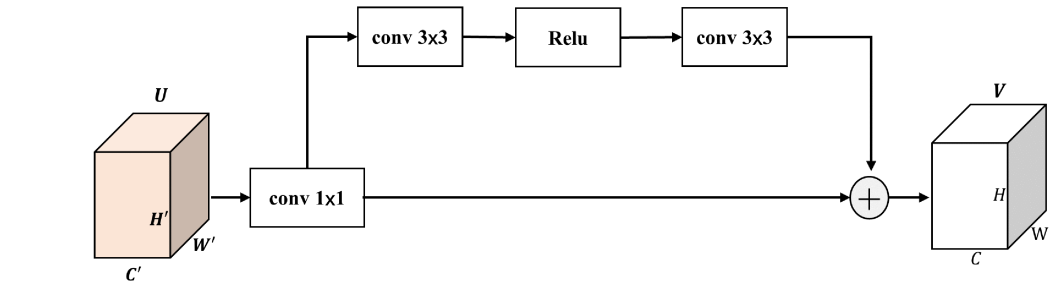
$$y_i = \mathcal{G}\left(\sum_{j=1}^{N} \frac{e^{W_k x_j}}{\sum_{m=1}^{N} e^{W_k x_m}} x_j\right) \tag{2}$$

$$z_i = x_i \oplus \mathcal{H}(W_l y)_i \tag{3}$$

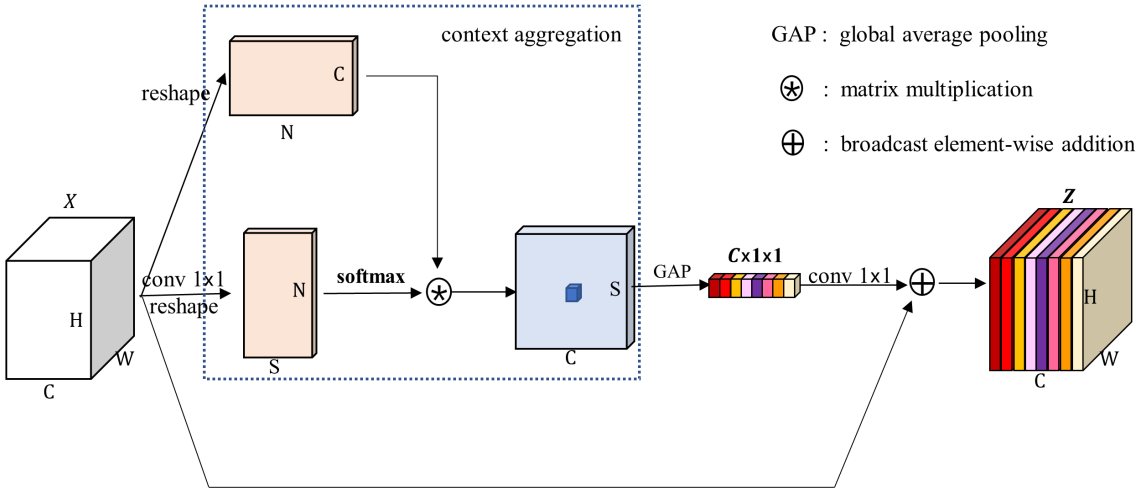
where $\mathcal{G}(\cdot)$ is the global average pooling (GAP) operation, can be defined as:

$$\mathcal{G}(x_i) = \frac{1}{N}\sum_{j=1}^{N} x_i(j) \tag{4}$$

$\oplus$ is broadcast element-wise product operator, $\mathcal{H}(\cdot)$ is ReLU function and a batch-normal operator, and W_k and W_l are the parameters of linear transformation. Based on Equation (2), the formula $\frac{e^{W_k x_j}}{\sum_{m=1}^{N} e^{W_k x_m}}$ indicates the weight of each position in the feature map, so we can obtain the context information through $\sum_{j=1}^{N} \frac{e^{W_k x_j}}{\sum_{m=1}^{N} e^{W_k x_m}} x_j$. Finally, the final output z_i combines the global context y_i with the local information x_i to strengthen the representation of different level features. Compared to using large-sized convolution kernels to capture the global contents, the context aggregation structure needs fewer parameters and computing resources. When the input features $X \in \mathcal{R}^{C\times H\times W}$, the number of parameters is $C^2 \times H \times W$ through using large size kernels to cover the full feature maps and $C \times (S+2)$ $(S << C)$ for making use of the context aggregation structure.



(**a**) channel regulation structure



(**b**) context aggregation structure

Figure 2. The Feature Enhancement Module (FEM) consists of two substructures: (**a**) channel regulation structure and (**b**) context aggregation structure.

3.3. Dual Gate Fusion Module

As is well known, high-level (deeper layers) features contain more discriminative information. Utilizing the rich discriminative information of high-level features can identify the category of objects (including the background) more accurately. Low-level (shallow layers) features contain more spatial information which can help to restore spatial details better. Most DCNNs directly fuse those different-level features using simple element-wise addition, e.g., FCN [18], or skip connection, e.g., U-Net [24]. As there exists a semantic gap between shallow layer features and deeper layer features, direct fusion will inevitably embed the background noise of deeper layer features. Considering that information is transferred from high-level features to low-level features in the decoding stage, we designed a dual-gate fusion module to combine the high-level semantic information with low-level spatial structural information, which makes the information fusion more effective. As shown in Figure 3, X_l, X_h represent the low-level features, and for high-level features, the "position gate" p_t can be defined as:

$$p_t = \sigma(W_p X_l + b_p) \tag{5}$$

where W_p, b_p are the weights and bias of a linear transformation, and σ is the sigmoid function which ensures the "position gate" p_t between 0 and 1. The gate1 "position gate" p_t indicates how important each spatial position is at the high-level features through the

low-level features. To make full use of the discriminant information of high-level features, we define gate2 "filter gate" f_t as:

$$f_t = \sigma\Big(W_f[p_t \otimes X_h + X_l] + b_f\Big) \tag{6}$$

The same as W_p and b_p, W_f and b_f are the weights and bias of a linear transformation, σ is the sigmoid function, and $\otimes$ is broadcast element-wise multiplication operator. The "filter gate" f_t decides how to refine the low-level features and suppresses the background noise of low-level features to a certain extent through high-level features. Finally, the output feature X_o can be expressed as:

$$X_o = p_t \otimes X_h + (1 - p_t) \otimes tanh(X_l \otimes f_t) \tag{7}$$

Based on Equation (7), we use "position gate" p_t to refine high-level features X_h and use "filter gate" f_t to update the low-level feature X_l. For doing so, we make the different scales feature mutual constraint, which promotes the integration between low-level spatial features and high-level semantic features.

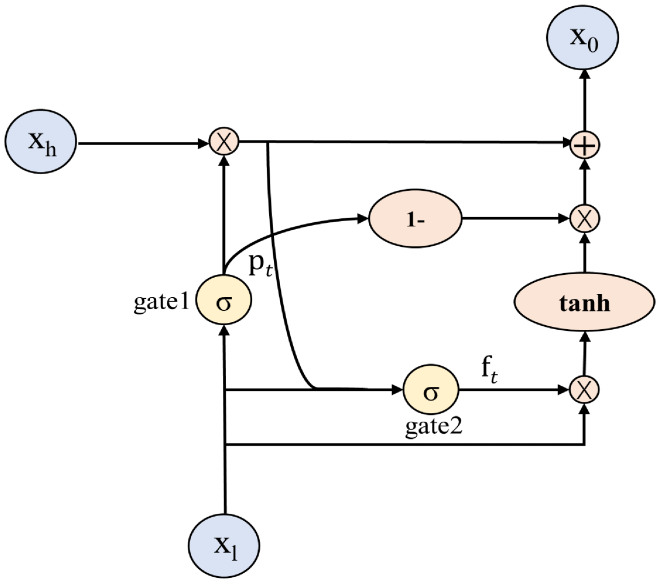


Figure 3. Dual Gate Fusion Module.

4. Experiment

4.1. Dataset Description

LandCover Dataset: the dataset proposed in [55] contains 41 tiles of RGB images covering the whole of Poland, which can be grouped into four common land-cover categories: building, woodland, water, and background. In detail, the dataset contains 33 tiles with resolution 25 cm (9000 × 9500 pixels) and 8 tiles with resolution 50 cm (ca. 4200 × 4700 pixels), which gives 176.76 km^2 and 39.51 km^2, respectively, and 216.27 km^2 overall. In [55], those images are partitioned into non-overlapping patches by a grid with a size of 512 × 512, so that we can obtain 10,674 patches. Among all these patches, 7470 patches are available for training, 1602 patches for validation, and 1602 patches for testing. Examples of the LandCover dataset are shown in Figure 4.

ISPRS Potsdam Dataset: the ISPRS Potsdam dataset [56] consists of 38 tile aerial images (6000 × 6000 pixels). Each image has a corresponding DSM with the same spatial resolution of 5 cm. The dataset contains the six most common land-cover categories, namely impervious surfaces (e.g., roads), buildings, low vegetation, tree, car, and clutter/background. Among them, 24 patches are available for training, and the remaining 14 for testing. Those images are partitioned into non-overlapping patches by a grid with a size

of 400 × 400, so we can obtain 5400 patches for training and 3150 for testing. In addition, the image in the ISPRS Potsdam dataset has different channel compositions, including IRRG, RGB, and RGBIR. In this paper, we only use the RGB channel for training and testing. Examples of the ISPRS Potsdam dataset are shown in Figure 5.

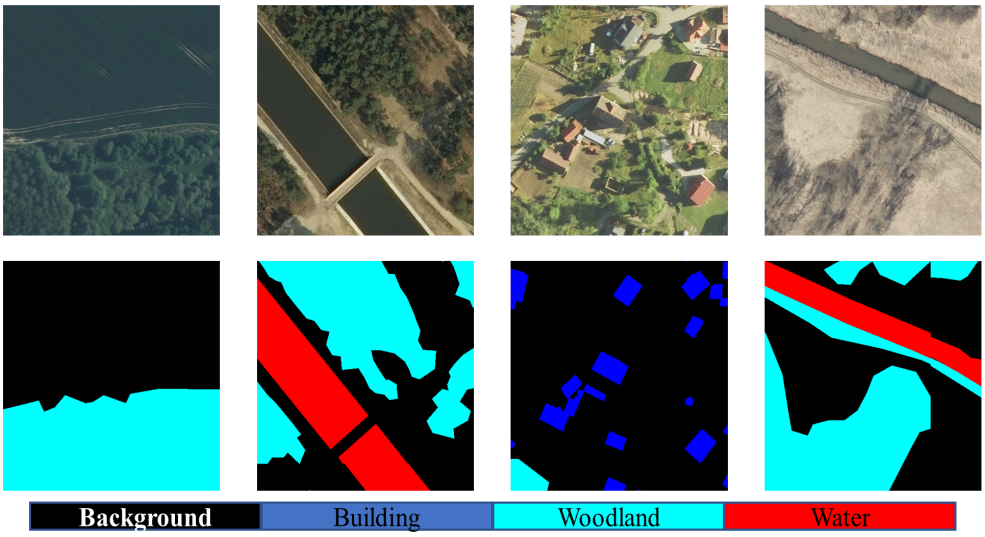


Figure 4. The images and corresponding reference mask. Buildings are blue, woodlands are cyan, water is red, and background is black.

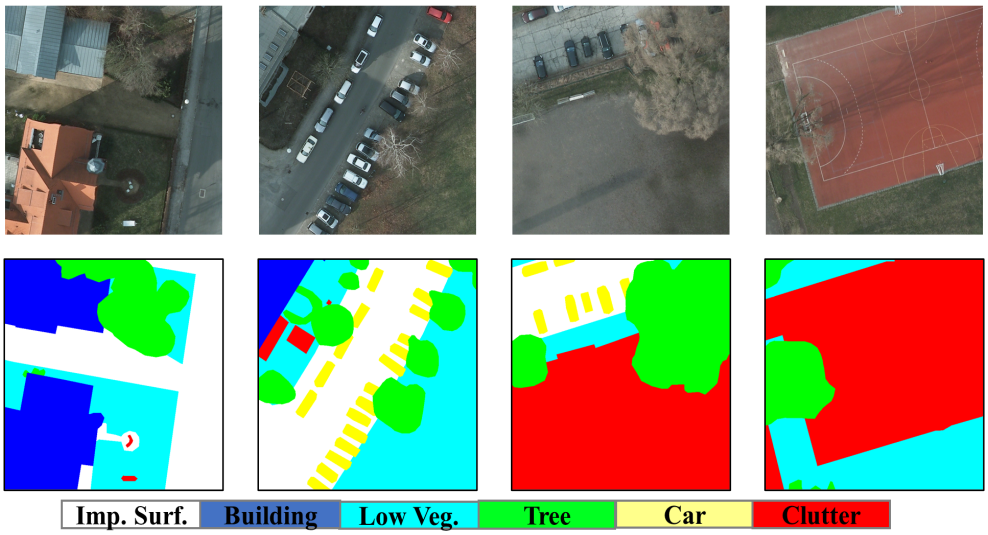


Figure 5. The images and corresponding reference mask. Impervious surfaces are white, buildings are blue, low vegetation is cyan, trees are green, cars are yellow, and background is red.

4.2. Implementation Settings

4.2.1. Parameter Setting

We employ pre-trained ResNet-50 as our backbone network of the DGFNet, implemented in PyTorch. We use a standard stochastic gradient descent (SGD) optimizer with 0.9 momentum and weight decay 0.001. Data augmentation with random-Gaussian blur and random-flipping operations are adopted on each iteration in the training phase. Our learning rate is scheduled by poly, starting with 7×10^{-3} for the LandCover dataset and 5×10^{-3} for the ISPRS Potsdam dataset. All the comparative experiments are trained with a batch size of 20 for the LandCover dataset and a batch size of 16 for the ISPRS Potsdam

dataset. We retrain all models by using one NVIDIA A100 GPU and 500 epochs for both datasets. For the loss function, we only use cross-entropy loss, which can be defined as:

$$CE = -\sum_{c=1}^{K} t_c log(p_c) \tag{8}$$

where t_c is a one-hot vector, and p_c indicates the probability that the prediction sample belongs to class c.

4.2.2. Evaluation Metrics

To assess the quantitative performance, four mainstream metrics for semantic segmentation are used, including pixel accuracy (PA) and mean pixel accuracy (MPA), mean intersection over union (MIoU), and frequency weighted intersection over union (FWIoU). Suppose there are K different land type classes. Let m_{ij} be the number of pixels belonging to class i predicted to belong to class j, $m_i = \sum_{j=1}^{K} m_{ij}$ is the total number of pixels belonging to class i, and $n_i = \sum_{j=1}^{K} m_{ji}$ is the total number of pixels predicted to class j. Those metrics can be defined as:

$$PA = \frac{\sum_{i=1}^{K} m_{ii}}{\sum_{i=1}^{K} m_i} \tag{9}$$

$$MPA = \frac{1}{K} \sum_{i=1}^{K} \frac{m_{ii}}{m_i} \tag{10}$$

$$MIoU = \frac{1}{K} \sum_{i=1}^{k} \frac{m_{ii}}{m_i + n_i - m_{ii}} \tag{11}$$

$$FWIoU = \frac{1}{\sum_{i=1}^{K} m_i} \left(\sum_{i=1}^{K} \frac{m_{ii}}{m_i + n_i - m_{ii}} m_i \right) \tag{12}$$

where the MIoU, FWIoU, MPA, and PA can describe the global land-cover classification performance. For example, the PA with 0.1% improvement indicates that millions of samples identified correctly for a pixel-level task. In addition, the MIoU, FWIoU, and MPA can avoid land-cover classification bias because of the imbalance of different classes on the LandCover and ISPRS Potsdam datasets. In addition to the mainstream metrics, we use the classical metrics, consisting of precision (P), recall (R), and $F1$, as the auxiliary metrics to evaluate classification results.The P, R, and $F1$ can be defined as:

$$P = \frac{1}{K} \sum_{i=1}^{K} \frac{TP_i}{TP_i + FP_i} \tag{13}$$

$$R = \frac{1}{K} \sum_{i=1}^{K} \frac{TP_i}{TP_i + FN_i} \tag{14}$$

$$F1\text{-}score = 2 * \frac{P * R}{P + R} \tag{15}$$

where TP is the true positive, FP is the false positive, TN is the true negative, FN is the false negative, and the index i indicates that the sample belongs to class i.

4.3. Model Analysis

4.3.1. Influence of Different Modules on Classification

To verify the effectiveness of the proposed FEM and DGFM, we conducted a series of ablation experiments on the LandCover dataset. In the ablation experiment, we used 1×1 convolution combined with SE [54] instead of FEM and a simple addition operation instead of DGFM as our baseline. As the numerical result shown in Table 1, when we used the FEM instead of 1×1 convolution and SE alone, the MIoU and MPA increase by 1.97

and 1.48 percentage points, respectively. It shows that FEM can combine local information and global features effectively to some extent. When we utilized DGFM instead of the addition operation, the MIoU increases by 2.55 percentage points, and the other metrics also have different degrees of improvement, which means the effectiveness of the proposed DGFM. It indicates that the DGFM is helpful to fuse the low-level spatial feature and high-level semantic feature to some extend. Finally, we integrate the FEM and DGFM into the baseline, the MIoU has an obvious increase of 3.15 percentage points. More specifically, from the per-class IoU results shown in Table 2, the IoU of category "buildings" has a 8.54 percentage points improvement. Compared to other categories, the size of category "buildings" is small in the VHR remote sensing images, which means that our DGFNet can get global contents to correctly detect small objects.

Table 1. Numerical comparisons with ablation experiments on the LandCover testing set. Baseline represents that DGFNet uses conv 1 × 1 combined with SE instead of FEM and uses simple addition operator instead of DGFM.

Method	FEM	DGFM	MIoU (%)	FWIoU (%)	MPA (%)	PA (%)
Baseline	×	×	85.72	91.89	91.41	95.75
	✓	×	87.69	92.54	92.89	96.10
	×	✓	88.27	92.89	92.63	96.30
	✓	✓	**88.87**	**94.02**	**93.28**	**96.41**

Table 2. Per-class IoU results with ablation experiments on LandCover testing set. Baseline represents that DGFNet uses conv 1 × 1 combined with SE instead of FEM and uses simple addition operator instead of DGFM.

Method	FEM	DGFM	Build. (%)	Wood. (%)	Water (%)	Back. (%)	MIoU (%)
Baseline	×	×	67.18	90.33	92.19	93.17	85.72
	✓	×	71.69	90.96	93.22	93.71	87.69
	×	✓	73.94	91.35	93.82	94.00	88.27
	✓	✓	**75.72**	**91.56**	**94.03**	**94.16**	**88.87**

4.3.2. Influence of Different Training Size on Classification

In addition to the ablation experiment about the proposed feature enhancement module (FEM) and dual gate fusion module (DGFM), we also conduct a series of experiments using different image sizes as the input to train DGFNet and test on the LandCover test set. As shown in Table 3, we crop the image size into 3 × 288 × 288, 3 × 320 × 320, 3 × 352 × 352, and 3 × 384 × 384 to train DGFNet, respectively. It is observed that we achieve the best results when we crop the training image size into 3 × 384 × 384 to train the DGFNet. The classification results improve with the increase in training image size to a certain extent. Owing to the feature enhancement module, the DGFNet can capture more global contexts with the increase in training image size to improve the performance of land-cover classification.

Table 3. Classification results of the Landcover test set. The bold numbers represent the best score.

Method	Size	MIoU (%)	FWIoU (%)	MPA (%)	PA (%)
DGFNet	3 × 288 × 288	87.89	92.86	92.51	96.27
	3 × 320 × 320	88.55	93.00	93.04	96.36
	3 × 352 × 352	88.40	92.99	93.16	96.35
	3 × 384 × 384	**88.87**	**94.02**	**93.28**	**96.41**

4.4. Comparisons with Other Networks

4.4.1. Comparison of LandCover Dataset

To show the effectiveness of DGFNet, the proposed method is compared with the state-of-the-art methods, as listed in Table 4. The neural network DANet [26], PSPNet [27], FCN [18], and deeplabv3+ [32] are using pre-trained ResNet-50 as their backbone, DenseASPP [57] takes DenseNet [58] as its and all of them are implemented with the PyTorch framework. In addition, our method was compared with other published research on the same dataset recently, such as DFFAN [59] and MFANet [60]. As the results of Table 4 show, our DGFNet outperforms other methods in terms of the MIoU, FWIoU, MPA, and PA. Specifically, comparisons with DANet, DGFNet has an increase of 13.09, 5.41, 10.18, and 2.54 percentage points in MIoU, FWIoU, MPA, and PA, respectively. Comparing to MFANet, our model also has an increase of 2.42, 4.13, 1.19, and 0.91 percentage points in those four metrics, separately. This result validates the effectiveness of the FEM and DGFM in our network.

Table 4. Numerical comparisons with state-of-the-art methods on LandCover testing set. (Bold numbers represent the best score for the testing set).

Method	MIoU (%)	FWIoU (%)	MPA (%)	PA (%)
DAnet [26]	75.78	88.61	83.10	93.87
PSPNet [27]	80.69	90.54	86.79	95.00
FCN-8s [18]	83.64	91.35	89.29	95.46
HRNet [28]	84.08	91.20	89.43	95.38
Deeplabv3+ [32]	84.99	91.89	90.66	95.75
DenseASPP [57]	85.02	91.56	91.10	95.56
U-Net [24]	85.65	91.70	90.83	95.66
SegNet [25]	85.69	92.02	90.96	95.82
DFFAN [59] (report)	84.81	89.21	90.64	-
MFANet [60] (report)	86.45	89.89	92.09	95.50
DGFNet (ours)	**88.87**	**94.02**	**93.28**	**96.41**

In Table 5, per-class IoU is computed to estimate the performance of recognizing distinct objects. The result indicates that our network has a better ability to distinguish objects with a small scale, such as buildings, because the feature enhancement module can combine the local information with global contents, strengthening the feature representation. From another perspective, our network can better distinguish the complex background information owing to the existence of DGFM to some extent. The DGFM makes different levels feature mutual constraint and promotes the integration between low-level spatial information and high-level semantic information. In Table 6, we also get the best performance about metrics P, R, and F1-score. Figures 6 and 7 show several example effects of remote sensing images within different scenarios.As the more complex scene in Figure 6, our prediction result (Figure 6g) is closest to the real land cover classification result (Figure 6b). In more detail, as displayed in the first row and second row of Figure 6, other methods mistakenly classify woodland as background. However, our proposed model can distinguish each category correctly, which is largely due to the designed feature enhancement module. The feature enhancement module (FEM) integrates the local information and global information to enhance the representation ability of features so it can recognize different objects very well. In addition, our model can restore the spatial resolution more accurately, e.g., the boundary of water in the third row. Different from Figure 6, the scene in Figure 7 is simple, but the spatial information is more clear, which is convenient for us to observe the restoration of spatial structure of original images, such as the boundary of woodland. Compared with other methods, our model can better recover the spatial edge details of the object. For example, the comparison models, such as deeplabv3+ and SegNet, cannot recover the edge information of woodland as shown in the

first row of Figure 7 and the large area of woodland is misclassified as background in the third row, which makes the boundary discontinuous. In contrast to the other method, our DGFNet shows the most complete and accurate land-cover mapping results. This result indicates that our network has an advantage in distinguishing complex scenes and has a better recovery of spatial information. That is due to the dual gate fusion module reducing the semantic gap between different levels, making the low-level spatial features and high-level semantic features more effectively fused. Compared with other networks, we can get more precise land-cover classification results, which verify the above conclusion again.

Table 5. Per-class IoU results with state-of-the-art methods on LandCover testing set.

Method	Build. (%)	Wood. (%)	Water (%)	Back. (%)	mIoU (%)
DAnet [26]	37.74	87.12	87.90	90.37	75.78
PSPnet [27]	51.11	89.10	90.54	90.03	80.69
FCN-8s [18]	60.36	89.80	91.67	92.73	83.64
HRNet [28]	62.85	89.61	91.27	92.58	84.08
Deeplabv3+ [32]	64.12	90.39	92.30	93.16	84.99
DenseASPP [57]	64.90	90.02	92.32	92.82	85.02
U-Net [24]	67.34	90.04	92.20	93.02	85.65
SegNet [25]	66.43	90.50	92.57	93.26	85.69
DFFAN [59] (report)	-	-	-	-	84.81
MFANet [60] (report)	75.09	87.78	91.66	91.25	86.45
DGFNet (ours)	75.72	91.56	94.03	94.16	88.87

Table 6. Metrics P, R, F1-score with state-of-the-art methods on LandCover testing set.

Method	P (%)	R (%)	F1-Score (%)
DAnet [26]	85.22	83.10	84.15
PSPnet [27]	89.77	86.62	88.17
FCN-8s [18]	91.74	89.29	90.50
HRNet [28]	92.42	89.43	90.90
Deeplabv3+ [32]	92.16	90.66	91.40
DenseASPP [57]	91.79	91.10	91.45
U-Net [24]	93.04	90.83	91.92
SegNet [25]	92.86	90.97	91.90
DGFNet (ours)	**94.60**	**93.28**	**93.93**

4.4.2. ISPRS Potsdam Dataset

To further illustrate the effectiveness of the proposed DGFNet, we also conducted comparative experiments on the ISPRS Potsdam dataset. Compared with the LandCover dataset, the scene in the ISPRS Potsdam dataset is more complex, which includes more small targets, such as vehicles. As the experimental results showing in Table 7, our network achieves the best results in terms of MIoU, FWIoU, MPA, and PA. Comparing to DANet, our model has an increase of 12.04, 7.73, 10.08, and 5.16 percentage points in MIoU, FWIoU, MPA, and PA, respectively. Compared with other state-of-art methods, the MIoU and MPA using DGFNet increases at least 1.24 and 1.79 percentage points, and other metrics have different degrees of improvement. For per-class IoU results, as shown in Figure 8, DGFNet has a better capability to detect small objects, such as cars. The success of detecting small targets is owed to the FEM, and can combine the local information with global content to strengthen the representation of different level features. In Table 8, we also get the best performance in terms of R and F1-score.

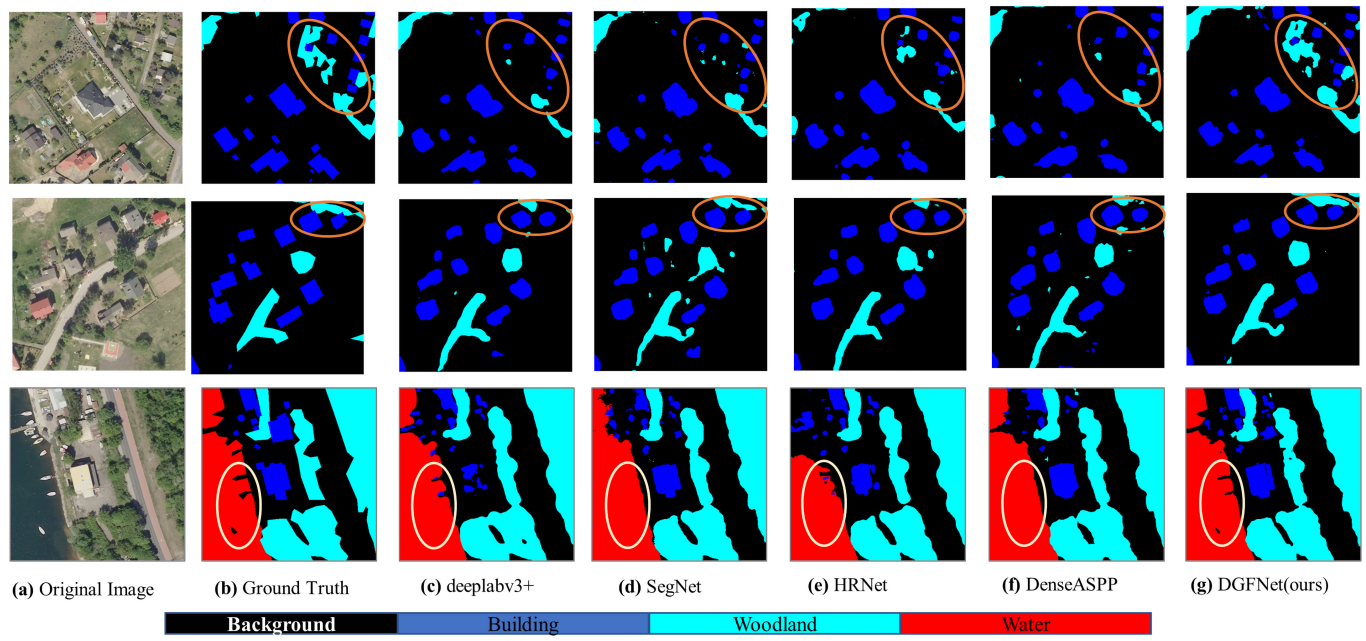


Figure 6. Qualitative comparisons with complex scenes between our method and several comparison methods on LandCover dataset. (**a**) Original image, (**b**) Ground Truth, (**c**) prediction map of deeplabv3+, (**d**) prediction map of SegeNet, (**e**) prediction map of HRNet, (**f**) prediction map of DenseASPP, (**g**) prediction map of DGFNet (ours).

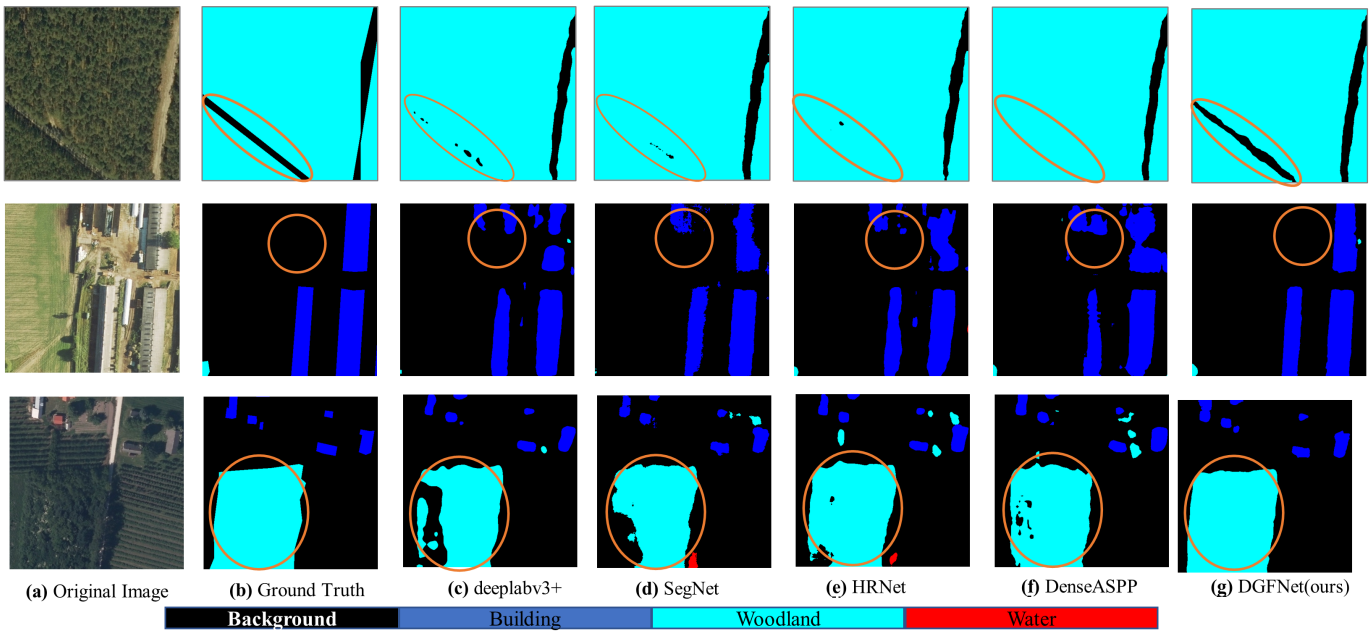


Figure 7. Qualitative comparisons with simple scenes between our method and several comparison methods on LandCover dataset. (**a**) Original image, (**b**) Ground Truth, (**c**) prediction map of deeplabv3+, (**d**) prediction map of SegeNet, (**e**) prediction map of HRNet, (**f**) prediction map of DenseASPP, (**g**) prediction map of DGFNet (ours).

Table 7. Numerical comparisons with state-of-the-art methods on ISPRS Potsdam testing set. (Bold numbers represent the best score for the testing set).

Method	MIoU (%)	FWIoU (%)	MPA (%)	PA (%)
DAnet [26]	60.21	70.23	71.31	82.25
PSPnet [27]	67.73	74.71	77.77	85.31
FCN-8s [18]	68.30	74.97	78.31	85.46
HRNet [28]	67.52	73.82	77.48	84.71
Deeplabv3+ [32]	70.34	76.24	79.84	86.28
SegNet [25]	70.36	76.17	79.55	86.23
DenseASPP [57]	71.01	77.33	79.58	87.05
DGFNet (ours)	**72.25**	**77.96**	**81.37**	**87.41**

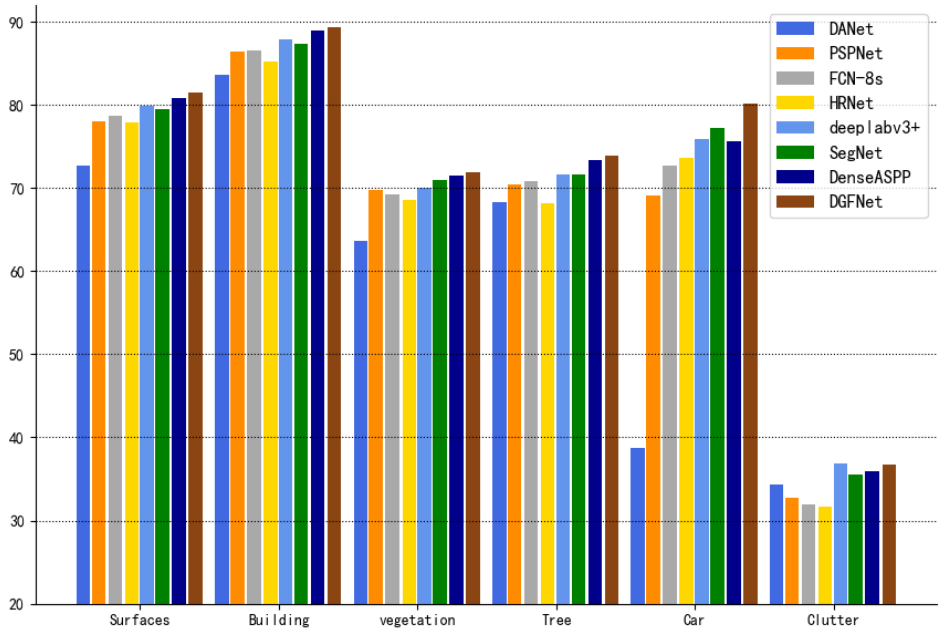


Figure 8. Per-class IoU results with related mthods on ISPRS Potsdam testing set.

Table 8. Metrics P, R, F1-score with state-of-the-art methods on the Potsdam testing set.

Method	P (%)	R (%)	F1-Socre (%)
DAnet [26]	77.10	71.31	74.09
PSPnet [27]	82.20	77.76	79.93
FCN-8s [18]	82.10	78.31	80.16
HRNet [28]	82.17	77.48	79.75
Deeplabv3+ [32]	83.77	79.84	81.76
SegNet [25]	83.97	79.55	81.71
DenseASPP [57]	**85.58**	79.58	82.48
DGFNet (ours)	85.12	**81.37**	**83.20**

Figure 9 shows the visualization results of three different scenarios on the test set. In the first line, the scene is about objects with occlusion, such as some low vegetation being disturbed by shadows. SegNet cannot distinguish the low vegetation in the shadow area, while deeplabv3+, HRNet, and DenseASPP can only discriminate a part of low vegetation. DGFNet can completely recognize the low vegetation in the whole shadow area. The second scenario is focused on small objects. As shown in the second line of

Figure 9, other models miss detecting the small car, while DGFNet has a better capability to detect the category "car". The success of detecting small objects represents that the FEM in DGFNet can effectively combine local information with global content to enhance the ability to identify small targets. The last scenario is the case of different objects with the same spectrum. As shown in the third line, the building targets and roads are similar in outward appearance, which means a slight bias between the different classes. In this case, other models have worse performance, resulting in lots of misjudgment. However, our DGFNet can distinguish the different categories correctly and restore the spatial details completely. For example, our model can recover the boundary of building better as shown in the third row of Figure 9. It indicates that the DGFM can suppress the semantic gap from different scales and fuse the low-level features and the high-level features effectively.

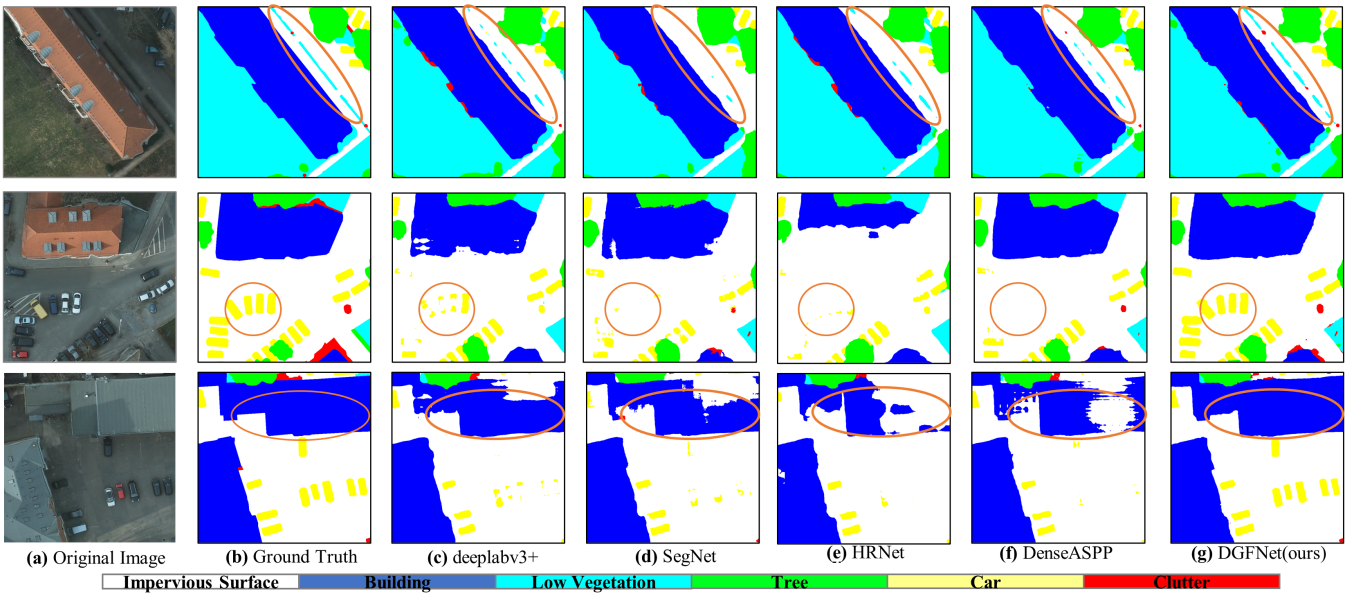


Figure 9. Qualitative comparisons between our method and several comparison methods on LandCover dataset. (**a**) Original image, (**b**) Ground Truth, (**c**) prediction map of deeplabv3+, (**d**) prediction map of SegeNet, (**e**) prediction map of HRNet, (**f**) prediction map of DenseASPP, (**g**) prediction map of DGFNet (ours).

4.4.3. Model Size and Efficiency Analysis

To analyze the size and efficiency of the proposed model, we calculate the number of trainable parameters and the average inference time of a single image based on the land-cover dataset. The size of all network input images is $3 \times 320 \times 320$. As shown in Table 9, the parameter of our network is equivalent to DANet, but the MIoU is increased by 13.09 percentage points. At the same time, compared with other models, our network also achieved the best performance in terms of MIoU. However, the mean inference time of a single image is higher than the other model and only lower than the DenseASPP model. That is because we adopt the DGFM in the decoding stage, and the DGFM promotes the fusion of low-level spatial features and high-level semantic features to improve the accuracy of land-cover classification. On the other hand, it increases the inference time due to the existence of multiple branches of DGFM. On the premise of ensuring the accuracy of land cover classification, our future work will focus on designing a lightweight network model to improve computational efficiency.

Table 9. Model size and efficiency analysis on LandCover testing set, including parameters, inference time, and MIoU.

Method	Backbone	Parameters (M)	Time (ms)	MIoU (%)
DAnet [26]	ResNet 50	45.24	10.46	75.78
PSPnet [27]	ResNet 50	50.85	12.58	80.69
FCN-8s [18]	ResNet 50	22.81	11.83	83.64
HRNet [28]	-	1.46	12.68	84.08
Deeplabv3+ [32]	ResNet 50	38.48	9.83	85.02
U-Net [24]	-	32.93	6.05	85.69
SegNet [25]	ResNet 50	28.08	7.35	84.99
DenseASPP [57]	DenseNet 201	20.47	22.91	85.65
DGFNet (ours)	ResNet 50	44.53	15.94	**88.87**

5. Conclusions

In this paper, a simple but efficient segmentation network named DGFNet is proposed for land-cover classification in VHR remote sensing images. The proposed DGFNet contains two novel modules: FEM and DGFM. The FEM can combine local information with global contents, strengthening the representation ability of feature maps in the encoder. The DGFM with the gate mechanism makes the different level features restrain each other, which promotes the fusion of multi-scale features and the restoration of spatial structure information. With those well-motivated modules, the DGFNet can capture more robust features by combining the local information and integrating multi-scale features, improving the performance of land-cover classification in VHR images. Exhaustive experiments prove the effectiveness of the proposed DGFNet. We also achieve the state-of-art performance of 88.87% MIoU on the LandCover dataset and 72.25% MIoU on the ISPRS Potsdam dataset.

Author Contributions: Conceptualization, Y.G.; methodology, Y.G.; validation, Y.X. and F.W.; formal analysis, Y.G.; investigation, Y.G.; resources, F.W. and Y.X.; writing—original draft preparation, Y.G.; writing—review and editing, Y.X., F.W. and H.Y; visualization, Y.G.; supervision, F.W. and H.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China under Grant 61901439, and the Key Research Program of Frontier Sciences, Chinese Academy of Science, under Grant ZDBS-LY-JSC036.

Acknowledgments: The authors would like to thank the anonymous reviewers for their constructive comments and suggestions.

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

Abbreviations

The following abbreviations are used in this manuscript:

VHR	Very High-Resolution
DCNN	Deep Convolutional Neural Networks
DGFNet	Dual Gate Fusion Network
FEM	Feature11Enhancement Module
DGFM	Dual Gate Fusion Module
DSM	Digital Surface Mode
MIoU	Mean Intersection over Union
FWIoU	Frequency Weighted Intersection over Union
MPA	Mean Pixel Accuracy
PA	Pixel Accuracy

References

1. López, J.A.; Verdiguier, E.I.; Chova, L.G.; Marí, J.M.; Barreiro, J.R.; Valls, G.C.; Maravilla, J.C. Land cover classification of VHR airborne images for citrus grove identification. *ISPRS J. Photogram. Remote Sens.* **2011**, *66*, 115–123. [CrossRef]
2. Hegazy, I.R.; Kaloop, M.R. Monitoring urban growth and land use change detection with GIS and remote sensing techniques in Daqahlia governorate Egypt. *Int. J. Sustain. Built Environ.* **2015**, *4*, 117–124. [CrossRef]
3. Zhang, X.; Du, S.; Wang, Q. Hierarchical semantic cognition for urban functional zones with VHR satellite images and POI data. *ISPRS J. Photogram. Remote Sens.* **2017**, *132*, 170–184. [CrossRef]
4. Stefanov, W.L.; Ramsey, M.S.; Christensen, P.R. Monitoring urban land cover change: An expert system approach to land cover classification of semiarid to arid urban centers. *Remote Sens. Environ.* **2001**, *77*, 173–185. [CrossRef]
5. Bayarsaikhan, U.; Boldgiv, B.; Kim, K.R.; Park, K.A.; Lee, D. Change detection and classification of land cover at Hustai National Park in Mongolia. *Int. J. Appl. Earth Obs. Geoinf.* **2009**, *11*, 273–280. [CrossRef]
6. Hussain, M.; Chen, D.; Cheng, A.; Wei, H.; Stanley, D. Change detection from remotely sensed images: From pixel-based to object-based approaches. *ISPRS J. Photogram. Remote Sens.* **2013**, *80*, 91–106. [CrossRef]
7. Tuia, D.; Ratle, F.; Pacifici, F.; Kanevski, M.F.; Emery, W.J. Active learning methods for remote sensing image classification. *IEEE Trans. Geosci. Remote Sens.* **2009**, *47*, 2218–2232. [CrossRef]
8. Zhao, W.; Du, S. Spectral–spatial feature extraction for hyperspectral image classification: A dimension reduction and deep learning approach. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 4544–4554. [CrossRef]
9. Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; Süsstrunk, S. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 2274–2282. [CrossRef] [PubMed]
10. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
11. Bay, H.; Tuytelaars, T.; Van Gool, L. Surf: Speeded up robust features. In Proceedings of the European Conference on Computer Vision, Graz, Austria, 7–13 May 2006; pp. 404–417.
12. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893.
13. Rublee, E.; Rabaud, V.; Konolige, K.; Bradski, G. ORB: An efficient alternative to SIFT or SURF. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2564–2571.
14. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
15. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
16. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 91–99. [CrossRef]
17. Cheng, G.; Han, J. A survey on object detection in optical remote sensing images. *ISPRS J. Photogram. Remote Sens.* **2016**, *117*, 11–28. [CrossRef]
18. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
19. Hong, S.; Oh, J.; Lee, H.; Han, B. Learning transferrable knowledge for semantic segmentation with deep convolutional neural network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3204–3212.
20. Kampffmeyer, M.; Salberg, A.B.; Jenssen, R. Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1–9.
21. Babenko, A.; Lempitsky, V. Aggregating local deep features for image retrieval. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1269–1277.
22. Gordo, A.; Almazán, J.; Revaud, J.; Larlus, D. Deep image retrieval: Learning global representations for image search. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 241–257.
23. Zhou, W.; Newsam, S.; Li, C.; Shao, Z. PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 197–209. [CrossRef]
24. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 4–8 October 2015; pp. 234–241.
25. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef] [PubMed]
26. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 3146–3154.
27. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.

28. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 5693–5703.
29. Nogueira, K.; Dalla Mura, M.; Chanussot, J.; Schwartz, W.R.; dos Santos, J.A. Dynamic multicontext segmentation of remote sensing images based on convolutional networks. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 7503–7520. [CrossRef]
30. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv* **2015**, arXiv:1511.07122.
31. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
32. 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 15th European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 801–818.
33. Li, A.; Jiao, L.; Zhu, H.; Li, L.; Liu, F. Multitask Semantic Boundary Awareness Network for Remote Sensing Image Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2021**. [CrossRef]
34. Marmanis, D.; Schindler, K.; Wegner, J.D.; Galliani, S.; Datcu, M.; Stilla, U. Classification with an edge: Improving semantic image segmentation with boundary detection. *ISPRS J. Photogram. Remote Sens.* **2018**, *135*, 158–172. [CrossRef]
35. Wang, H.; Wang, Y.; Zhang, Q.; Xiang, S.; Pan, C. Gated convolutional neural network for semantic segmentation in high-resolution images. *Remote Sens.* **2017**, *9*, 446. [CrossRef]
36. Mou, L.; Zhu, X.X. RiFCN: Recurrent network in fully convolutional network for semantic segmentation of high resolution remote sensing images. *arXiv* **2018**, arXiv:1805.02091.
37. Mou, L.; Hua, Y.; Zhu, X.X. Relation matters: Relational context-aware fully convolutional network for semantic segmentation of high-resolution aerial images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 7557–7569. [CrossRef]
38. Liu, C.; Zeng, D.; Wu, H.; Wang, Y.; Jia, S.; Xin, L. Urban land cover classification of high-resolution aerial imagery using a relation-enhanced multiscale convolutional network. *Remote Sens.* **2020**, *12*, 311. [CrossRef]
39. Chen, X.; Wang, G.; Guo, H.; Zhang, C.; Wang, H.; Zhang, L. Mfa-net: Motion feature augmented network for dynamic hand gesture recognition from skeletal data. *Sensors* **2019**, *19*, 239. [CrossRef]
40. Sun, W.; Wang, R. Fully convolutional networks for semantic segmentation of very high resolution remotely sensed images combined with DSM. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 474–478. [CrossRef]
41. Diakogiannis, F.I.; Waldner, F.; Caccetta, P.; Wu, C. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS J. Photogramm. Remote Sens.* **2020**, *162*, 94–114. [CrossRef]
42. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
43. Dauphin, Y.N.; Fan, A.; Auli, M.; Grangier, D. Language modeling with gated convolutional networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 933–941.
44. Gehring, J.; Auli, M.; Grangier, D.; Yarats, D.; Dauphin, Y.N. Convolutional sequence to sequence learning. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 1243–1252.
45. Li, L.; Kameoka, H. Deep clustering with gated convolutional networks. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 16–20.
46. Yang, C.; An, Z.; Zhu, H.; Hu, X.; Zhang, K.; Xu, K.; Li, C.; Xu, Y. Gated convolutional networks with hybrid connectivity for image classification. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12581–12588.
47. Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; Huang, T.S. Free-form image inpainting with gated convolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 4471–4480.
48. Chang, Y.L.; Liu, Z.Y.; Lee, K.Y.; Hsu, W. Free-form video inpainting with 3d gated convolution and temporal patchgan. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 9066–9075.
49. Rayatdoost, S.; Rudrauf, D.; Soleymani, M. Multimodal gated information fusion for emotion recognition from EEG signals and facial behaviors. In Proceedings of the 2020 International Conference on Multimodal Interaction, Utrecht, The Netherlands, 11–15 October 2020; pp. 655–659.
50. Cao, C.; Lan, C.; Zhang, Y.; Zeng, W.; Lu, H.; Zhang, Y. Skeleton-based action recognition with gated convolutional neural networks. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *29*, 3247–3257. [CrossRef]
51. Xue, W.; Li, T. Aspect based sentiment analysis with gated convolutional networks. *arXiv* **2018**, arXiv:1805.07043.
52. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
53. Chao, P.; Zhang, X.; Gang, Y.; Luo, G.; Jian, S. Large Kernel Matters—Improve Semantic Segmentation by Global Convolutional Network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4353–4361.
54. 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.
55. Boguszewski, A.; Batorski, D.; Ziemba-Jankowska, N.; Zambrzycka, A.; Dziedzic, T. LandCover. ai: Dataset for Automatic Mapping of Buildings, Woodlands and Water from Aerial Imagery. *arXiv* **2020**, arXiv:2005.02264.

56. Rottensteiner, F.; Sohn, G.; Jung, J.; Gerke, M.; Baillard, C.; Benitez, S.; Breitkopf, U. The ISPRS benchmark on urban object classification and 3D building reconstruction. In Proceedings of the ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Melbourne, Australia, 25 August–1 September 2012; pp. 293–298.
57. Yang, M.; Yu, K.; Zhang, C.; Li, Z.; Yang, K. Denseaspp for semantic segmentation in street scenes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3684–3692.
58. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
59. Huang, J.; Weng, L.; Chen, B.; Xia, M. DFFAN: Dual Function Feature Aggregation Network for Semantic Segmentation of Land Cover. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 125. [CrossRef]
60. 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]