Article

A Building Segmentation Network Based on Improved Spatial Pyramid in Remote Sensing Images

Hao Bai ¹, Tingzhu Bai ^{1,*}, Wei Li ² and Xun Liu ²
¹ School of Optics and Photonics, Beijing Institute of Technology, No.5, South Street, Zhongguancun, Haidian District, Beijing 100081, China; baihaohlj@126.com

² Beijing Institute of Space Electricity & Mechanics, Beijing 100081, China; wei_li_cast@163.com (W.L.); liuxun_laby@163.com (X.L.)

* Correspondence: tzhbai@bit.edu.cn

Abstract: Building segmentation is widely used in urban planning, disaster prevention, human flow monitoring and environmental monitoring. However, due to the complex landscapes and high-density settlements, automatically characterizing building in the urban village or cities using remote sensing images is very challenging. Inspired by the recent deep learning methods, this paper proposed a novel end-to-end building segmentation network for segmenting buildings from remote sensing images. The network includes two branches: one branch uses Widely Adaptive Spatial Pyramid (WASP) structure to extract multi-scale features, and the other branch uses a deep residual network combined with a sub-pixel up-sampling structure to enhance the detail of building boundaries. We compared our proposed method with three state-of-the-art networks: DeepLabv3+, ENet, ESPNet. Experiments were performed using the publicly available Inria Aerial Image Labelling dataset (Inria aerial dataset) and the Satellite dataset II (East Asia). The results showed that our method outperformed the other networks in the experiments, with Pixel Accuracy reaching 0.8421 and 0.8738, respectively and with mIoU reaching 0.9034 and 0.8936 respectively. Compared with the basic network, it has increased by about 25% or more. It can not only extract building footprints, but also especially small building objects.

Keywords: CNN; semantic segmentation; super resolution; remote sensing; spatial pyramid; ResNet

Citation: Bai, H.; Bai, T.; Li, W.; Liu, X. A Building Segmentation Network Based on Improved Spatial Pyramid in Remote Sensing Images. *Appl. Sci.* **2021**, *11*, 5069. <https://doi.org/10.3390/app11115069>

Academic Editor: Hyung-Sup Jung

Received: 1 April 2021

Accepted: 26 May 2021

Published: 30 May 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Building segmentation from remote sensing images is the process of segmenting the roof or footprint boundary by grouping pixels which are adjacent with similar feature values (such as brightness, edge, texture, and color). Allowing for fine details in certain areas and very fine details under certain conditions, remote sensing images greatly facilitate the classification and extraction of area-related features, such as roads and buildings. They have also been widely used in many applications, including urban planning [1], urban environmental modeling [2], disaster management [3], land-use change monitoring [4], and digital city evolution [5].

The application of such methods is quite challenging for the following three key reasons: First, the existence of objects are at multiple scales. In a remote sensing image, a pixel may belong to inconspicuous small objects, salient large objects, or stuff. Inconspicuous objects and stuff may be dominated by salient objects and, as such, their information will be somewhat weakened or even disregarded, resulting in an uneven categorical distribution; second, remote sensing images are generally obtained through satellites, airplanes, drones, etc. Therefore, they only provide a top-down view, which does not contain all of the important characteristics of objects. The missing characteristics are usually visible in the ground-based or panoramic view of the object; third, the adjacent buildings belonging to the same category in remote sensing images may have huge differences in appearance

(such as color and shape). In recent years, there have been more and more researches in this field, and some remarkable achievements have been made.

Conventional remote sensing image segmentation methods usually rely on hand-crafted features to extract spatial and texture information of the image, and consider a building as a combination of low-level features, such as spectral information [6], tone [7], texture [8], and geometric shape [9,10]. Many classic powerful feature extractors can also achieve high efficiency, such as support vector machines (SVMs) [11], multinomial logistic regression [12,13], boosted decision trees (DTs) [14], neural networks [15–17], Scale-Invariant Feature Transform (SIFT) [18], Histogram of Oriented Gradient (HOG) [19], and Speeded Up Robust Features (SURF) [20]. These methods do, however, have high requirements for the researchers themselves, and cannot represent the high-level semantic features of the image well.

A deep Convolutional Neural Network (CNN) [21] predicts using only features learned from training data, instead of hand-engineered features. Essential to their success is the built-in invariance of deep CNNs to local image transformations, which allows them to learn increasingly abstract feature representations. The challenge of CNN-based semantic segmentation comes from the inner content, shape, and scale variations of the same objects, as well as the easily confused and fine boundaries among different objects. First, the convolution layer has both shift and spatial-invariant characteristics. While invariance is clearly desirable for high-level vision tasks, it may hamper low-level tasks, such as pose estimation and semantic segmentation, in which precise localization is required, rather than the abstraction of spatial details. Second, CNNs are able to capture informative representations with global receptive fields, by stacking convolutional layers and down-sampling; however, the use of down-sampling to compress data is irreversible, resulting in information loss and further causing translation invariance and smooth results. This also leads to inaccurate object boundary acquisition. Third, the multiple scales of objects are also a challenge for CNNs. The features extracted by CNNs usually have a limited receptive field, where the feature mainly describes the core region and largely ignores the context around the boundary, no matter whether the receptive field is large or small. The extraction of any feature from an object, when conducted at both a certain scale and other scales, will produce dissimilar results.

With the development of deep learning, techniques have been introduced to meet these requirements. To enlarge the receptive field of feature maps, the Pyramid Scene Parsing Network (PSPNet) [22] adopts Spatial Pyramid Pooling, which pools the feature maps into different sizes and concatenates them after up-sampling. DeepLab [23–25] adopts Atrous Spatial Pyramid Pooling (ASPP), which employs dilated/atrous convolution [26], in order to extract at a single scale and to perform accurate and effective classification of regions at any scale. To improve the resolution of the reconstructed image and, thus, object boundaries, Single-Image Super-Resolution (SISR) [27] has been proposed to generate a high-resolution image from low-resolution observations of the same scene. The Super-Resolution Convolutional Neural Network (SRCNN) [28] is the first end-to-end super-resolution algorithm that uses the CNN structure, which has achieved superior performance against traditional methods. Very Deep Convolutional Networks for Super Resolution (VDSR) [29] and Deeply-Recursive Convolutional Network (DRCN) [30] further improve SRCNN by using gradient clipping, skip connection, or recursive-supervision to ease the difficulty of training deep network. Accelerating the Super-Resolution Convolutional Neural Network (FSRCNN) [31] has been proposed to accelerate the training and testing of SRCNN. Super-Resolution Using a Generative Adversarial Network (SRGAN) has been submitted to construct a deeper network with perceptual losses [32] and a generative adversarial network (GAN) [33] for photo-realistic super-resolution. Efficient Sub-Pixel Convolutional Neural Network (ESPCN) [34] uses a shallow network (e.g., FCN) and only upscale the low-resolution image at the very last stage, in order to achieve real-time performance. DeepBack-Projection Networks For Super-Resolution (DBPN) [35] provides an

error feedback mechanism for projection errors at each stage. Recent research has made improvements to the CNN structure, based on the characteristics of remote sensing images.

Recent studies have also applied CNN networks for remote sensing images, such as FCN, Deeplabv3+, ENet [36], and ESPNet [37], which have been used to achieve excellent results in building detection. Bittner et al. used FCN to extract buildings [38]. Xu et al. used the improved UNet [39] network with the guided filter to extract buildings [40]. Wu [41] used SegNet for Unmanned Aerial Vehicle-Based Images in Riverbank Monitoring. Zhang [42] proposed a multi-constraint fully convolutional network (MC-FCN) model to perform end-to-end building segmentation. Sanjeevan [43] improved the FCN using an exponential linear unit (ELU), in order to reduce the noise (i.e., falsely classified buildings) and sharpen the boundaries of the buildings. SPP-PCANet [44] has been designed to combine PCANet and spatial pyramid pooling, in order to reduce the amount of false positives and to improve the detection rate. Wang [45] applied the ASPP and superpixel-based DenseCRF to perform dense semantic labeling on remote sensing images. Ma [46] used an MR-SSD model to detect different objects in large-scale SAR images. WSF-Net [47] can achieve comparable results to fully supervised methods, when using only image-level annotations with weakly supervised binary segmentation. SPMF-Net [48] takes the image-level labels as supervision information in a classification network that combines superpixel pooling and multi-scale feature fusion structures for building segmentation. Wang [49] combined semi-supervised training, Average Update of Pseudo-label (AUP) with pseudo-labels, and strong labels to improve the segmentation performance. Gergelova [50] proposes a processing procedure in a geographic information system (GIS) environment which advocates the identification of roof surfaces based on the LiDAR point cloud. Lyu [51] proposes the statistical region merging (SRM) and a shape context similarity model for Light Detection and Ranging (LiDAR) data.

Inspired by the methods above, we propose a novel building segmentation network, which combines semantic segmentation and super-resolution for accurate building segmentation in remote sensing images. Our main contributions can be summarized as follows:

1. Detail Enhancement (DE) module based on SISR is proposed to enhance the boundaries of buildings. Deep Residual Block is designed to improve the learning ability of the network. This design is to reduce the local information loss caused by the use of convolutional layers and down-sampling in a single Deep CNN which will make negatively impact on the details of the image around the small object.
2. Widely Adapted Spatial Pyramid (WASP) is designed to enlarge the receptive fields for extracting more context information. This design is to multiple scale feature maps with pooling layer and dilated convolution, and integrates feature of different scales step-by-step, which can incorporate neighbor scales of context features more precisely and introducing better feature learning and fitting capabilities. We further apply the improved ResNet50 and Global Feature Guidance to improve the segmentation accuracy.

The paper is organized as follows: Section 2 presents our building segmentation network. Section 3 presents the experimental results on two public datasets along with the discussion. Our conclusions are presented in Section 4.

2. Proposed Methods

In this paper, the Semantic Segmentation module and DE modules are designed. The Semantic Segmentation module is used to achieve the coarsely segmentation results building objects in the remote sensing images. Among them, the DE module is independent of the Semantic Segmentation module, using an independent feature extraction module (Deep Residual Block) and a feature transfer method (Feature Fusion Connection) to decrease the information loss caused by the pooling layer and down-sampling in the segmentation network. As shown via the overall architecture in Figure 1, we apply the improved ResNet50 (LResNet50) to extract dense features. We used pooling operation to obtain spatial distribution information in the image and use this as a guide for intact

bodies and accurate boundaries, by embedding the results of the Detail Enhancement and Semantic Segmentation modules.

Figure 1. Overview of the building segmentation network. WASP module and Detail Enhancement module are our most important contributions in this paper.

2.1. Detail Enhancement (DE) Module

A super-resolution reconstruction algorithm can reconstruct a corresponding high-resolution image from a low-resolution image. In a neural network, the convolutional layer can directly learn low-frequency information through a non-linear filter; the low-frequency information is well-preserved, such that the low-resolution image input to the network and the high-resolution image output are similar, considering the outlines of ground objects. However, the loss of high-frequency information will still cause the reconstructed image to lack details. Therefore, existing super-resolution networks generally use a shallow network to ensure the quality of the reconstructed image and the shallow network means limited learning ability. As remote sensing images contain a large amount of information, the network can be fully utilized only when the depth of the network is guaranteed. Meanwhile, in the Deep Learning community, many theoretical studies [52–54] have shown that deep convolutional neural networks (number of layers > 5) can recursively identify patches of the images at lower layers, resulting in exponential increases in many linear regions of the input space. The deep layers typically extract more complex features from the input image than the shallow layers (i.e., layers < 5).

Based on the above analysis, we designed a Detail Enhancement (DE) module to realize the detail enhancement of the buildings in the remote sensing image, based on ESPCN. The module, as shown in Figure 2, contains a Deep Residual Block and Feature Fusion Connections. ESPCN uses a pixel shuffle layer, which uses the sub-pixel up-sampling method. By using sub-pixel up-sampling differs from the deconvolution layer and interpolation-based up-sampling, the interpolation function is implicitly included in the previous convolutional layer which can be learned automatically. The image size is transformed only in the last layer, the previous convolution operation is performed on low-resolution images, the efficiency of sub-pixel up-sampling will be higher than the bicubic up-sampling and deconvolution. In this module, we retain the original sub-pixel upsampling layer. For the feature extraction part, we designed a feature enhancement module, in order to increase the depth of the network, and we used a feature transfer module to ensure that the network retains more high-frequency information.

In Deep Residual Block, we apply the Residual Block proposed by EDSR [55]. Compared with the original ResNet [56], this Residual Block proposed by EDSR does not apply the pooling layer and the batch normalization (BN) [57] layer; correspondingly, the structure is simpler and less calculation is required, which makes it suitable for processing pictures with rich information. As shown in Figure 2a, every Group Block contains three Residual Blocks, in Figure 2b the Deep Residual Block (DRblock) contains two Group Blocks. Since

the image super resolution is an image-to-image translation task where the input image is highly correlated with the target image, the residuals between them is only we need to learn the feature which we named global residual feature. In this case, it avoids learning a complicated transformation from a complete image to another, instead only requires learning a residual map to restore the missing high-frequency details. The local residual learning is similar to the residual learning in ResNet and used to alleviate the degradation problem caused by increasing network depths, reducing training difficulty and improving the learning ability.

Figure 2. Deep Residual Block and Feature Fusion Connection. (a) The Group Block. (b) The Deep Residual Block. In the Group Block, the black lines represents Group Feature Connection. In the Deep Residual Block, the blue line and thick black line represent Deep Residual Feature Connection.

Between Deep Residual Blocks and Group Blocks, we applied skip connections for feature transfer, using the Feature Fusion Connection inspired by ResNet. The Feature Fusion Connections are both implemented by shortcut connections and element-wise addition. The difference is that the former directly connects the input and output images as black line can transport the global residual feature to the pixel shuffle layer, the latter one usually adds multiple shortcuts between layers with different depths inside the network which are represented as black line and transfer the feature extracted from every block step by step to the sub-pixel up-sampling layer. In this case, the Feature Fusion Connection part of the module is intended to achieve mitigation of the dense network, which is used to address the gradient vanishing problem.

2.2. Widely Adapted Spatial Pyramid (WASP) Module

In a deep neural network, the receptive field [58] size can roughly indicate how much we use contextual information. Due to the convolutional nature of a CNN, local convolutional features usually have a limited receptive field. Even with a large receptive field, the feature mainly describes the core region and ignores the contextual information around boundary. Due to the arbitrary sizes of objects in the image, purely using single-scale convolution for image semantic segmentation will lead to the incorrect classification of pixels between categories. Some small-size objects, such as streetlight, are hard to distinguish but may be of great importance. In addition, large objects may exceed the receptive field of the CNN, thus causing discontinuous prediction. These issues demand processing at multiple scales. As strided convolutions are used in convolution layers and spatial pooling is used in pooling layers, the resolution of the feature maps is reduced. The use of a pooling layer also leads to reduced localization accuracy in the labelled images. As a result, many recent works have attempted to deal with the problems mentioned above, in one way or the other.

The DeepLab series utilizes Atrous spatial pyramid pooling (ASPP), with image-level features responsible for encoding the global context. Differing from the conventional approaches, which tackle objects at multiple scales, it applies rescaled versions of an input image, then aggregates the feature maps. The ASPP applies dilated/atrous convolution to integrate the visual context at multiple scales. With the dilated convolution, the size of the feature map captured will avoid reducing. The structure can also probe an incoming feature map using multiple-sampling rate filters with effective fields-of-view encoding the

context at multiple scales. Aggressively increasing the dilation rate may, however, cause an inherent problem called “grid” [59]. PSPNet abstracts different sub-regions by adopting varying-size pooling kernels, and uses global pooling to extract global context information, combining it with the spatial pyramid. The structure combines global pooling with the spatial pyramid [60–63], fuses features under four different pyramid scales, and up-samples the low-dimensional feature maps to obtain the same size features as the original feature map by bilinear interpolation. The use of pooling layers reduces the resolution of the feature map and abandons local information, which further makes it difficult to distinguish objects, especially small-sized objects.

Current semantic segmentation architectures perform feature learning by extracting multi-scale features and merging them, but lack selective extraction of features at different scales. ENet base SegNet uses two blocks with maxpooling to heavily reduce the input size, and use only a small set of feature maps. The idea behind it, is that visual information is highly spatially redundant, and thus can be compressed into a more efficient representation. Additionally, our intuition is that the initial network layers should not directly contribute to classification. Instead, they should rather act as good feature extractors and only preprocess the input for later portions of the network. ESPNet uses point-wise convolutions to be helped in reducing the computation, while the spatial pyramid of dilated convolutions re-samples the feature maps to learn the representations from large effective receptive field. EncNet [64] employs the dilation strategy and proposes a Context Encoding Module (CEM) incorporating the Semantic Encoding Loss (SE-loss), in order to capture contextual information and to selectively highlight class-dependent feature maps. The proposed SE-Loss, unlike per-pixel loss, is capable of taking the global context into consideration, resulting in context-aware training. The Selective Kernel Network (SENet) [65] proposes a dynamic selection mechanism to adaptively adjust the receptive field size of each neuron, based on multiple scales of the input information, and multiple branches with different kernel sizes are fused using the softmax attention function, guided by the information in these branches. Although these branches can extract receptive fields with different scales in feature maps, they lack global context prior attention to select the features in a channel-wise manner. The Pyramid Attention Network (PAN) [66] exploits the impact of global contextual information to combine an attention mechanism and spatial pyramid to extract precise dense features for pixel labeling, instead of complicated dilated convolutions and artificially designed decoder networks.

With Pyramid Attention Network (PAN) [66], we consider how to exploit high-level feature maps to guide low-level features in recovering pixel localization. Global context features and sub-region context are helpful, in this regard, to distinguish among various categories. This global prior is designed to remove the fixed-size constraint of CNNs for image classification.

Considering the above observations, the Widely Adapted Spatial Pyramid (WASP) is proposed, as shown in Figure 3. In this module, we concatenate dilated convolution and pooling layers in the Multi-Scale Feature Extraction block. Each branch first uses a global pooling operation to capture global contextual information at the image level, where the contextual information can use spatial statistical data to interpret the entire image scene. Then we use dilated convolution to expand the receptive field of the feature map without reducing the resolution of the feature map, where each pixel belongs to the feature map from the different image levels and corresponds different sub-regions in the original image. When the level size of the pyramid is N , we reduce the size of the context representation to $1/N$ of the original size. Specifically, the design in the pyramid is $1/8$, $1/4$, $1/2$ and 1 of the four branches. Compared with a single branch, this setting can more effectively restore the spatial information damaged by the pooling operation [67], so that objects in the final prediction have a sharper boundary. We use a spatial pyramid structure similar to DeepLabv3, where the dilated convolutions we selected were 6, 12, 18, and 1. This design is based on the formula for the receptive field shown as Equation (1). This structure can avoid the scale of the receptive field obtained by one branch being equal or an integral

multiple of that of another branch, under the premise of obtaining a receptive field as large as possible.

$$r_n = r_{n-1} + (k_n - 1) \prod_{i=1}^{n-1} S_i \quad (1)$$

where n represents the number of layers, r_n represents the receptive field size of the n th layer, k_n represents the convolution kernel size of the n th layer, d represents the dilated convolution size of the n th layer, and S_n represents the convolution stride size of the n th layer.

Figure 3. Widely Adapted Spatial Pyramid (WASP). (a) The Multi-Scale Feature Extraction. (b) The Progressive Upsample structure. “d” represents the dilated rate of convolution.

The calculation formula of the dilated convolutional kernel is:

$$k_d = k_n + (k_n - 1)(d - 1) \quad (2)$$

where k_d represents the equivalent convolution kernel for dilated convolution of the n th layer. In our network design, each branch of the pyramid has a pooling layer and a dilated convolution layer. Then the receptive field of each branch can be calculated as:

$$r_n = r_{n-1} + (k_p - 1) \prod_{i=1}^{n-1} S_i + (k_d - 1) \prod_{i=1}^{n-1} S_{i+1} \quad (3)$$

where k_p represents the convolution kernel of pooling layer. In this paper, we set the stride and convolution size in the pooling layer to be equal to the pooling rate. Assuming that the pooling rate = p , the formula can be expressed as:

$$r_n = r_{n-1} + (p - 1) \prod_{i=1}^{n-1} S_i + (k_d - 1) \prod_{i=1}^{n-1} S_{i+1} \quad (4)$$

It can also be expressed as:

$$r_n = r_{n-1} + [(k_n - 1)(d - 1) + k_n]p \prod_{i=1}^{n-1} S_i - \prod_{i=1}^{n-1} S_i \quad (5)$$

Then, we upsample the feature maps to obtain features with the same size as the original feature map. Benefiting from the Progressive Upsample structure, the feature maps at multi-scales are fused and learned through the convolution layers in the progressive process. Compared with the structures that directly concatenate the feature maps after the parallel branches, as in DeepLab or PSPNet, it can lead to better feature learning and fitting capabilities, as well as relatively dense feature maps. The Progressive Upsample structure

integrates features at different scales step-by-step, and can incorporate neighboring scales of context features more precisely. At the output end of the final multi-scale branch, we also adopt a progressive convolution structure design. After the feature processing of each parallel branch is completed, we use bilinear interpolation to upsample the feature map of the current branch, and feature maps of the same size are added in a step-by-step manner. Finally, a feature map of the same size as the original input can be obtained. In this way, the feature maps at multiple scales are fused and learned through the convolution operator in the progressive process. Compared with structures that only use a single-layer convolution to process the feature maps after parallel branches are merged, it can introduce more convolution operations, better feature learning and fitting capabilities, and relatively dense feature maps.

We add the multi-scale feature map which is the coarsely obtained from the Semantic Segmentation module and the feature map obtained by the Detail Enhancement module, for which average pooling is used for the image. The global context information of the entire image is obtained by pooling operation. Assuming that the input image is $h * w * c$, a tensor of size $1 * 1 * C$ is obtained after pooling. It is more native to the convolution structure, by enforcing correspondences between feature maps and categories, such that it can generate one feature map for each corresponding category of the classification task and sums the spatial information. Therefore, it is more robust to spatial translations of the input. We apply the global context information to directly perform pixel-by-pixel weighted selection of the input feature map. Here we call it Global Feature Guidance (abbreviated as GFG) to facilitate subsequent experiments.

2.3. Other Settings in Our Method

In our method, ResNet50 was chosen to extract the feature map. The original ResNet50 uses the Rectified Linear Unit (ReLU) as the activation function. ReLU (which can be calculated as in Equation (2)) retains the biological inspiration of the step function (i.e., the neuron is activated only when the input exceeds the threshold). When the input is positive, the derivative is non-zero, allowing gradient-based learning (although, at $x = 0$, the derivative is undefined). Using this function can make the calculation process very fast, as neither the function nor its derivative contains complex mathematical operations. When the input is less than zero or the gradient is zero, its weight cannot be updated and, therefore, the learning speed of ReLU may become very slow, or it may even make the neuron directly invalid. The Leaky ReLU function (LReLU) (which can be calculated as in Equation (3)) is a variant of the classic ReLU activation function. The output of this function still has a small slope when the input is negative. When the derivative is non-zero, it can reduce the appearance of silent neurons, allowing for gradient-based learning (although it will be slow), thus solving the problem of neurons not learning after the ReLU function enters a negative interval. Compared with ReLU, LReLU has a larger activation range. Based on the above, we replaced all ReLU functions in ResNet50 with LReLU (Leaky Rectified Linear Unit) as the basic module. In this paper, the ResNet50 with LReLU, instead of ReLU, is termed “LResNet50”.

$$f(y_i) = \begin{cases} y_i & \text{if, } y_i > 0 \\ 0 & \text{if, } y_i \leq 0 \end{cases} \quad (6)$$

$$f(y_i) = \begin{cases} y_i & \text{if, } y_i > 0 \\ 0.01 & \text{if, } y_i \leq 0 \end{cases} \quad (7)$$

3. Results and Discussion

3.1. Datasets Description and Training Details

To verify the effectiveness of our proposed method, we conducted various experiments using two publicly available datasets: the Inria Aerial Image Labeling dataset (Inria aerial dataset) [68] and the Satellite dataset II (East Asia) [69]. Considering building shapes are usually regular and mostly rectangular, we use thousands of images in 2 databases as the training set to ensure the generalization performance of our method.

The Inria Aerial Image Labeling Dataset, released by Maggiori et al., covers an area of 810 square kilometers, with a total of 360 images, each with 5000×5000 pixels. The dataset includes 10 densely populated cities and remote villages (Austin, Bellingham, Bloomington, Chicago, Innsbruck, Kitsap, San Francisco, Western and Eastern Tyrol and Vienna). The spatial resolution of the images reaches 0.3 m and the shape and structure of the buildings are clearly visible. The images cover dissimilar urban settlements, ranging from densely populated areas (e.g., San Francisco's financial district) to alpine towns (e.g., Lienz in Austrian Tyrol). Instead of splitting adjacent portions of the same images into the training and test subsets, different cities are included in each of the subsets. The datasets are constructed by combining public domain imagery and public domain official building footprints. The Satellite dataset II (East Asia) consists 17,388 images which are 512×512 , of six neighboring satellite images covering 860 km² on East Asia with 0.45 m ground resolution. In this experiment, we use 3135 images in the training set and 903 images in the test set. Unlike the Satellite dataset II, the testing set in the Inria aerial dataset does not have the labels that be used for segmentation comparative experiment, so we chose four group images and their labels from training set as the testing set of Inria aerial dataset in our experiments.

Due to GPU memory limitations, we resize the images from the datasets. We cropped the image and its corresponding label to 512×512 size. In order to improve the image utilization of the datasets, the images and corresponding labels in the original dataset were randomly cropped. In order to facilitate the observation of the experimental results, the test images were cropped to a size of 512×512 pixels with an overlap of 10 pixels between the images. These operations can ensure that the network has enough data and prevent the model from overfitting and improve its robustness.

3.2. Evaluation Metrics

To verify the effectiveness of the Detail Enhancement module, we used the Peak Signal-to-Noise Ratio (PSNR) as an evaluation metric. In the experiments, the original images from the dataset were used as I_{HR} images. The corresponding I_{LR} were obtained by downscaling the images with the factor U through bicubic interpolation. A low spatial resolution image I_{LR} with W columns, H rows, and C bands can be expressed by a tensor of size $W \times H \times C$. Its corresponding I_{HR} had a size of $UW \times UH \times UC$. We compared our module with some super-resolution state-of-the-art methods. MSE represents the mean square error between the reconstructed image F and the original high-resolution image I . Assume the size of the reconstructed image and original high-resolution image is $m \times n$, MSE can be calculated as:

$$MSE = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [I(i, j) - F(i, j)]^2 \quad (8)$$

Let k represent the number of bits per pixel. The PSNR can be calculated as:

$$PSNR = 10 * \log_{10} \left(\frac{(2^k - 1)^2}{MSE} \right) \quad (9)$$

To verify the effectiveness of the Semantic Segmentation module, we used the Pixel Accuracy (PA) and Mean Intersection over Union (mIoU) as evaluation metrics. PA is used to indicate the proportion of correctly classified pixels to the total pixels. Assuming that

there are a total of $k + 1$ types of targets in the image, p_{ij} represents the number of pixels that belong to class i but are predicted to be class j , while p_{ii} represents the number of pixels that are actually classified correctly.

$$PA = \frac{\sum_{i=0}^k p_{ii}}{\sum_{i=0}^k \sum_{j=0}^k p_{ij}} \quad (10)$$

The mIoU is used to measure the accuracy of an object detector on the datasets. The mIoU can therefore be calculated as:

$$mIoU = \frac{1}{k+1} \sum_{i=0}^k \frac{p_{ii}}{\sum_{j=0}^k p_{ij} + \sum_{j=0}^k p_{ji} - p_{ii}} \quad (11)$$

3.3. Comparison Experiments

In the experiments, we decomposed and combined the proposed method, in order to verify the effectiveness of each module. In Section 3.3.1, ESPCN is used as the basic network. In Section 3.3.2, ResNet50 with ASPP (similar to DeepLabV3) is used as the basic network. The feature map extracted by ResNet50 was fed into the WASP. In Section 3.3.3, we integrated the two modules together, with some other optimization operations, to obtain our proposed method (LResNet50 + DE + WASP + GFC). The results are summarized in the associated Tables and Figures.

3.3.1. Comparison Experiments for DE Module

In this subsection, the super-resolution results under different factors are discussed. The performance of the proposed DE module was evaluated over the two selected datasets. For each dataset, a 512×512 sub-region was randomly selected from the image, in order to train the network, and another 512×512 sub-region, divided equally from the image, was used to validate the performance of our proposed method. In order to simulate a low resolution image, we first downsampled by the factor. The corresponding image size was changed to the original $1/N$, and used as the input to the super-resolution module. For all the compared models in this experiment, we used the PSNR as the metric.

Comparison experiment of the proposed DE module with different settings: the factors we set were 2, 3, and 4. The results are summarized in Tables 1 and 2. It was observed that: (1) Our proposed DE module obviously outperformed all of the other methods over these two datasets, showing the highest PSNR; (2) the PSNR value improved when we used the Deep Residual Block (DRblock); (3) when we further used the Feature Fusion Connection (FFC), the PSNR further improved. Based on the use of these two blocks, the module can extract features more accurately, and can fuse global and local residual features efficiently.

Table 1. Comparison experiment of the proposed DE module, when different settings were adopted, using the Inria aerial dataset.

U	ESPCN	DRblock	DE(DRblock + FFC)
4	26.0611 dB	26.2111 dB	26.8953 dB
3	27.9854 dB	28.8974 dB	29.6001 dB
2	31.4781 dB	32.3776 dB	33.3720 dB

Table 2. Comparison experiment of the proposed DE module, when different settings were adopted, using the Satellite dataset II (East Asia).

U	ESPCN	DRblock	DE(DRblock + FFC)
4	26.3522 dB	27.0531 dB	27.5876 dB
3	27.6836 dB	28.8912 dB	29.7317 dB
2	31.9551 dB	32.7683 dB	33.9959 dB

Comparison Experiment of the proposed DE module with different methods: in this experiment, we compared our DE module with four other state-of-the-art methods (ESPCN, EDSR, SRCNN, and DBPN). Among these methods, SRCNN and ESPCN have the simplest structure, consisting of only a few convolutional layers in the feature extraction part, where a sub-pixel upsampling layer is used in the up-sampling part of ESPCN. EDSR uses the Residual Block module, but uses an interpolation method in the up-sampling part. The structure of DBPN is the most complicated. The DE module we designed uses the same sub-pixel upsampling layer as ESPCN, and uses the Residual Block in EDSR as the feature extraction module. The factors we set were 2, 3, and 4. The results are summarized in Tables 3 and 4. It was observed that: (1) The PSNR of DE was very close to the value of DBPN and much higher than those of EDSR and SRCNN on the two datasets; and (2) when the upsampling factor increased, the superiority of our proposed DE became less obvious. This was due to the problem of SR being more difficult with a higher upsampling factor.

Table 3. Comparison of different methods using the Inria aerial dataset. Among them, ESPCN and our DE module were mentioned above; their PSNR values are also listed here.

U	ESPCN	SRCNN	EDSR	DBPN	DE
4	26.0611 dB	27.0830 dB	23.3839 dB	27.2730 dB	27.3529 dB
3	27.9854 dB	28.7625 dB	29.2732 dB	29.3326 dB	29.6001 dB
2	31.4781 dB	32.9577 dB	33.2932 dB	33.3457 dB	33.3720 dB

Table 4. Comparison of different methods using the Satellite dataset II (East Asia). Among them, ESPCN and our DE module were mentioned above; their PSNR values are also listed here.

U	ESPCN	SRCNN	EDSR	DBPN	DE
4	26.3522 dB	27.5891 dB	23.7169 dB	27.3679 dB	27.5876 dB
3	27.6836 dB	29.0024 dB	29.6821 dB	29.7826 dB	29.7317 dB
2	31.9551 dB	32.8070 dB	33.3641 dB	33.9941 dB	33.9959 dB

The visual results of our proposed DE module were compared with those of five other state-of-the-art methods (bicubic, ESPCN, EDSR, SRCNN, and DBPN) on the two datasets, as shown in Figures 4 and 5. In order to facilitate comparison, we used the bicubic method (which is commonly used for up-sampling) to up-sample the low-resolution images obtained through downsampling to the size of the original image, in order to compare the visual results. It can be observed that the details in the image reconstructed using the DE module were richer. In Figure 4, compared with the SRCNN and ESPCN algorithms, objects with different sizes in the image reconstructed by DE module can be easily distinguished from the surrounding environment. Some small-sized objects (such as cars in shadows) can also be distinguished from their surroundings. However, due to a lack of details, the distinction between the outline of the car in the images reconstructed by SRCNN and ESPCN and the surrounding scenery was not high. The reconstruction effects of DBPN and our algorithm were similar. In the red boxes of Figure 5, the ridges in the fields can be clearly seen in the image reconstructed by our DE module. The DBPN algorithm performed slightly better, but the fields in the red box area of the reconstructed image of the SRCNN and ESPCN algorithms are mixed and difficult to distinguish. This demonstrates that the proposed module can effectively enrich the details of the reconstructed image; especially the boundaries between some insignificant adjacent objects, which can also be clearly distinguished.

Figure 4. Comparison of reconstruction results using the Inria aerial dataset: (a) bicubic; (b) ESPCN; (c) SRCNN; (d) EDSR; (e) DBPN; and (f) DE.

Figure 5. Comparison of reconstruction results using the Satellite dataset II(East Asia): (a) bicubic; (b) ESPCN; (c) SRCNN; (d) EDSR; (e) DBPN; and (f) DE.

3.3.2. Comparison Experiments for Semantic Segmentation Module

In this experiment, we applied ResNet50 with ASPP (which is a similar structure to the ASPP in DeepLab V3) as the basic structure. In detail, we conducted experiments

with many settings, including max pooling or average pooling, WASP, and Global Feature Guidance. The accuracy results of the proposed structure are summarized in Tables 5 and 6. The segmented results of different settings are shown in Figures 7–10.

Table 5. Comparison experiments of the Semantic Segmentation module with different settings, using the Inria aerial dataset.

Methods	mIoU	PA
ResNet50 + ASPP	0.6412	0.6826
ResNet50 + WASP(MAX)	0.7209	0.7201
ResNet50 + WASP(AVE)	0.7527	0.7433
LResNet50 + WASP(AVE)	0.8218	0.7713
LResNet50 + WASP(AVE) + GFG	0.8367	0.8124

Table 6. Comparison experiments of the Semantic Segmentation module with different settings, using the Satellite dataset II (East Asia).

Methods	mIoU	PA
ResNet50 + ASPP	0.6981	0.7408
ResNet50 + WASP(MAX)	0.7322	0.7609
ResNet50 + WASP(AVE)	0.7608	0.8130
LResNet50 + WASP(AVE)	0.8015	0.8234
LResNet50 + WASP(AVE) + GFG	0.8427	0.8562

In Tables 5 and 6, “AVE” represents average pooling. “LResNet50” represents ResNet50 with LReLU instead of ReLU. “MAX” and “AVE” represent max pooling and average pooling, respectively. “GFG” represents Global Feature Guidance. It can be observed that: (1) the higher amount of contextual information extracted by WASP was very effective in improving the segmentation accuracy. WASP could also effectively solve the problem of large differences in object scales. The proposed structure could combine the features of every parallel branch and improve the segmentation effect; (2) average pooling and LReLU improved the segmentation effect, but the effect was limited. In Table 5, they improved the performance by almost 5% on the Inria aerial dataset and almost improved the mIoU by 7% on the Satellite dataset II (East Asia); (3) the use of GFG did not obviously increase the mIoU (less than 1%), but increased the PA (by more than 4% on the Inria aerial dataset and by almost 7% on the Satellite dataset). This indicates that the use of global features can preserve the spatial information of the original image, in order to improve the segmentation effect, such that our method could better distinguish categories and reduce false positives.

The visual results of our proposed Semantic Segmentation module are shown, using the two datasets, in Figures 6 and 7. The roofs of buildings in remote sensing images (which are from a top view) may be very close to the square, road, etc. Therefore, some adjoining building objects may have been mixed together and their boundaries could not be distinguished. In Figures 6c and 7c, which show the segmented results with the basic structure, small building objects had a relatively good segmentation effect, but the shapes of the segmented buildings were quite different from the shapes in the ground-truth or the original image. Some adjoining buildings still had mixed boundaries. In Figures 6d and 7d, the number of building objects was the same as the number in the original image and label. The shapes were also very close, with clear boundaries between adjacent building targets. In Figures 7e and 8e, the building object boundary in the segmentation result was clear and the segmented object shape was close to the target shape in the original image. In Figures 6f and 7f, the original image and most of the building targets in the ground-truth were segmented. The shapes of the building objects were very close to those in the ground-truth, and their outlines were very clear; however, the boundaries and shapes of some adjoining buildings in the image were still different from the buildings in the original image.

Different from Figures 6 and 7, which use the Inria aerial dataset, the buildings and backgrounds in Figures 8 and 9 are very similar in color (green); some adjacent buildings

were very small and difficult to distinguish. While ensuring the accurate segmentation of the building shape in the ground-truth, the proposed Semantic Segmentation module also segmented small-size buildings that were not marked in the original image. This demonstrates that our proposed method can be especially useful for small building object when the difference between the sizes of objects is too large.

Figure 6. Comparison of reconstruction results using Inria aerial dataset. (a) the input image; (b) ground truth; (c) ResNet50 + ASPP; (d) ResNet50 + WASP(AVE); (e) LResNet50 + WASP(AVE); (f) Our method(LResNet50 + WASP(AVE) + GFG).

Figure 7. Comparison of reconstruction results using Inria aerial dataset. (a) the input image; (b) ground truth; (c) ResNet50 + ASPP; (d) ResNet50 + WASP(AVE); (e) LResNet50 + WASP(AVE); (f) Our method(LResNet50 + WASP(AVE) + GFG).

Figure 8. Comparison of reconstruction results using Satellite dataset II (East Asia). (a) The input image; (b) ground truth; (c) ResNet50 + ASPP; (d) ResNet50 + WASP(AVE); (e) LResNet50 + WASP(AVE); (f) Our method (LResNet50 + WASP(AVE) + GFG).

Figure 9. Comparison of reconstruction results using Satellite dataset II (East Asia). (a) The input image; (b) ground truth; (c) ResNet50 + ASPP; (d) ResNet50 + WASP(AVE); (e) LResNet50 + WASP(AVE); (f) Our method (LResNet50 + WASP(AVE) + GFG).

3.3.3. Comparison Experiments for Different Semantic Methods

In this experiment, ResNet50 with ASPP was set as the basic method. The quantitative results of this experiment, using the two datasets, are listed in Tables 7 and 8. Compared with other state-of-the-art semantic methods, our method had a better segmentation effect. Compared with the basic method, our method improved the performance, in terms of the mIoU, from 0.6412 to 0.9034 (i.e., constituting an improvement of mIoU by almost 28.2%) and the PA was improved from 0.6826 to 0.8421 (i.e., improved by 24.5%) when using the Inria aerial dataset. When using the Satellite dataset II (East Asia), our method improved the performance, in terms of the mIoU, from 0.6981 to 0.8936 and the PA from 0.7408 to 0.8738, thus improving the mIoU by almost 28% and the PA by 17%. The reason for this is that the connection method of dilated convolution and pooling in the parallel branch of the proposed pyramid structure could extract richer contextual information. At the same time, due to the use of the super-resolution algorithm in the DE module, the low- and high-frequency information in the network could be better utilized. Therefore, it was more suitable for remote sensing images.

Table 7. Comparison quantitative results of different methods using Inria aerial dataset.

Methods	mIoU	PA
ResNet50 + ASPP	0.6412	0.6826
DeepLabV3+	0.7967	0.8342
ENet	0.7127	0.8121
ESPNet	0.7467	0.8042
Proposed Method(LResNet50 + WASP(AVE) + DE + GFG)	0.9034	0.8421

Table 8. Comparison of quantitative results using Satellite dataset II (East Asia).

Methods	mIoU	PA
ResNet50 + ASPP	0.6981	0.7408
DeepLabV3+	0.7416	0.8042
ENet	0.7758	0.8221
ESPNet	0.8167	0.8442
Proposed Method(LResNet50 + WASP(AVE) + DE + GFG)	0.8936	0.8738

From the visualization results, we determined that our method had an ideal segmentation effect, especially for small building objects, as can be seen in Figures 10 and 11. The methods used in the experiment used pooling operations, dilated convolutions, and pyramid structures, in order to obtain contextual information. Therefore, almost all the ground objects and boundaries were correctly identified when using our method. Similar to the results in Figures 10 and 11, our method also outperformed the other methods on the Satellite dataset II (East Asia). It can be seen that our method could segment accurate contours for complex-shaped building objects. The building objects in Figures 12 and 13, especially the adjacent buildings, were accurately segmented, regardless of whether they were large- or small-sized buildings. In particular, the shape and location of small-sized buildings were consistent with the ground-truth. These findings show that our method can effectively improve the performance of remote sensing image semantic segmentation.

Figure 10. Comparison results results of semantic segmentation methods. (a) Input Image (b) Ground Truth; (c) ResNet50 + ASPP; (d) DeepLabV3+; (e) ENet; (f) ESPNet; (g) Our method (LResNet50 + WASP(AVE) + GFG).

Figure 11. Comparison results results of semantic segmentation methods. (a) Input image; (b) ground truth; (c) ResNet50 + ASPP; (d) DeepLabV3+; (e) ENet; (f) ESPNet; (g) Our method (LResNet50 + WASP(AVE) + GFG).

Figure 12. Comparison results of semantic segmentation methods using Satellite dataset II (East Asia). (a) Input image; (b) ground truth; (c) ResNet50+ASPP; (d) DeepLabV3+; (e) ENet; (f) ESPNet; (g) Our method(LResNet50+WASP(AVE)+GFG).

Figure 13. Comparison results of semantic segmentation methods using Satellite dataset II (East Asia). (a) Input image; (b) ground truth; (c) ResNet50 + ASPP; (d) DeepLabV3+; (e) ENet; (f) ESPNet; (g) Our method (LResNet50 + WASP(AVE) + GFG).

4. Conclusions

In this paper, we proposed a notable building segmentation network for remote sensing images combining the characteristics of image semantic segmentation and super-resolution methods. Through extensive experiments on two remote sensing image datasets, our method was shown to significantly improve the segmentation performance, compared to the state of the arts, it can achieve the accurate segmentation of building objects at multiple scales, especially considering small objects. The method we designed can also achieve very effective segmentation in complex scenes facing complex-shaped buildings.

At present, remote sensing technology is developing rapidly, especially for autonomous driving and artificial intelligence applications. The recognition of multi-scale targets in remote sensing images has been the focus of a large body of research. These applications also have requirements, in terms of calculation speed and the computational complexity of the associated algorithms, which will be the focus of our future work.

Author Contributions: conceptualization, H.B.; methodology, H.B.; software, H.B.; validation, H.B.; formal analysis, H.B.; investigation, H.B.; resources, H.B.; data curation, H.B.; writing—original draft preparation, H.B.; writing—review and editing, T.B.; visualization, H.B.; supervision, H.B.; project administration, X.L., W.L.; funding acquisition, X.L., W.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We thank the providers of Inria aerial dataset and the Satellite dataset II (East Asia). We also thank Li Hanchao for his guidance during the writing of the paper.

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

References

1. Ahmadi, S.; Zoj, M.J.V.; Ebadi, H.; Abrishami, H.; Mohammadzadeh, A. Automatic urban building boundary extraction from high resolution aerial images using an innovative model of active contours. *Int. J. Appl.* **2010**, *12*, 150–157. [\[CrossRef\]](#)
2. Liu, C.; Huang, X.; Zhu, Z.; Chen, H.; Tang, X.; Gong, J. Automatic extraction of built-up area from ZY3 multi-view satellite imagery: Analysis of 45 global cities. *Remote Sens. Environ.* **2019**, *226*, 51–73. [\[CrossRef\]](#)
3. Awrangjeb, M. Using Point Cloud Data to Identify, Trace, and Regularize the Outlines of Buildings. *Int. J. Remote Sens.* **2016**, *37*, 551–579. [\[CrossRef\]](#)
4. Li, J.; Huang, X.; Gong, J. Deep neural network for remote sensing image interpretation: Status and perspectives. *Natl. Sci. Rev.* **2019**, *6*, 1082–1086. [\[CrossRef\]](#)
5. Huang, X.; Cao, Y.; Li, J. An automatic change detection method for monitoring newly constructed building areas using time-series multi-view high-resolution optical satellite images. *Remote Sens. Environ.* **2020**, *244*, 111802. [\[CrossRef\]](#)
6. Peng, J.; Zhang, D.; Liu, Y. An improved snake model for building detection from urban aerial images. *Pattern Recognit. Lett.* **2005**, *26*, 587–595. [\[CrossRef\]](#)
7. Müller, S.; Zaum, D. Robust Building Detection in Aerial Images. In Proceedings of the International Archives of Photogrammetry and Remote Sensing, Vienna, Austria, 29–30 August 2005; pp. 143–148.
8. Liu, Z.; Cui, S.; Yan, Q. Building extraction from high resolution satellite imagery based on multi-scale image segmentation and model matching. In Proceedings of the International Workshop on Earth Observation and Remote Sensing Applications, Beijing, China, 30 June–2 July 2008.
9. Shackelford, A.K.; Davis, C.H.; Wang, X. Automated 2-D Building Footprint Extraction from High-Resolution Satellite Multi-spectral Imagery. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, Honolulu, HI, USA, 20–24 September 2004; pp. 1996–1999.
10. Zhang, Q.; Huang, X.; Zhang, G. Urban Area Extraction by Regional and Line Segment Feature Fusion and Urban Morphology Analysis. *Remote Sens.* **2017**, *9*, 663. [\[CrossRef\]](#)
11. Melgani, F.; Bruzzone, L. Classification of hyperspectral remote sensing images with support vector machines. *IEEE Trans. Geosci. Remote Sens.* **2004**, *42*, 1778–1790. [\[CrossRef\]](#)
12. Li, J.; Bioucas-Dias, J.M.; Plaza, A. Semisupervised hyperspectral image segmentation using multinomial logistic regression with active learning. *IEEE Trans. Geosci. Remote Sens.* **2010**, *48*, 4085–4098. [\[CrossRef\]](#)

13. Li, J.; Bioucas-Dias, J.M.; Plaza, A. Spectral-spatial hyperspectral image segmentation using subspace multinomial logistic regression and Markov random fields. *IEEE Trans. Geosci. Remote Sens.* **2011**, *50*, 809–823. [\[CrossRef\]](#)
14. Pal, M.; Mather, P.M. An assessment of the effectiveness of decision tree methods for land cover classification. *Remote Sens. Environ.* **2003**, *86*, 554–565. [\[CrossRef\]](#)
15. Atkinson, P.M.; Tatnall, A.R. Introduction neural networks in remote sensing. *Int. J. Remote Sens.* **1997**, *18*, 699–709. [\[CrossRef\]](#)
16. Foody, G.; Arora, M. An evaluation of some factors affecting the accuracy of classification by an artificial neural network. *Int. J. Remote Sens.* **1997**, *18*, 799–810. [\[CrossRef\]](#)
17. Zhong, Y.; Zhang, L. An adaptive artificial immune network for supervised classification of multi-/hyperspectral remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2011**, *50*, 894–909. [\[CrossRef\]](#)
18. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 90–110. [\[CrossRef\]](#)
19. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the Computer IEEE Computer Society Conference on Vision and Pattern Recognition, San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893.
20. Herbert, B.; Andreas, E.; Tinne, T.; Luc, V.G. Speeded-up robust features (SURF). *Comput. Vis. Image Underst.* **2008**, *110*, 346–359.
21. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436. [\[CrossRef\]](#)
22. 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 (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
23. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv* **2015**, arXiv:1412.7062.
24. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv* **2016**, arXiv:1606.00915.
25. Chen, L.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**, arXiv:1706.05587.
26. Yu, F.; Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. *arXiv* **2016**, arXiv:1511.07122v3.
27. Yang, W.; Zhang, X.; Tian, Y.; Wang, W.; Xue, J.-H. Deep Learning for Single Image Super-Resolution: A Brief Review. *arXiv* **2018**, arXiv:1808.03344v1.
28. Dong, C.; Loy, C.C.; He, K.; Tang, X. Learning a deep convolutional network for image super-resolution. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2014; pp. 184–199.
29. Kim, J.; Lee, J.K.; Lee, K.M. Accurate image super-resolution using very deep convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1646–1654.
30. Kim, J.; Lee, J.K.; Lee, K.M. Deeply-recursive convolutional network for image super-resolution. In Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence, Las Vegas, NV, USA, 27–30 June 2016; pp. 1637–1645.
31. Dong, C.; Loy, C.C.; He, K.; Tang, X. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 295–307. [\[CrossRef\]](#)
32. Johnson, J.; Alahi, A.; Li, F.-F. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016.
33. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Bengio, Y. Generative adversarial nets. *arXiv* **2014**, arXiv:1406.2661.
34. Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A.P.; Bishop, R.; Wang, Z. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1874–1883.
35. Haris, M.; Shakhnarovich, G.; Ukita, N. Toyota Technological Institute at Chicago, United States. DeepBack-Projection Networks For Super-Resolution. Toyota Technological Institute, Japan. *arXiv* **2018**, arXiv:1803.02735v17.
36. Paszke, A.; Chaurasia, A.; Kim, S.; Culurciello, E. ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation. *arXiv* **2016**, arXiv:1606.02147v1.
37. Huang, T.; Liu, Z.; Chen, X.; Bai, X. EPNet: Enhancing Point Features with Image Semantics for 3D Object Detection. *arXiv* **2020**, arXiv:2007.08856.
38. Bittner, K.; Cui, S.; Reinartz, P. Building Extraction from Remote Sensing Data Using Fully Convolutional Networks. *ISPRS Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2017**, *XLII-1/W1*, 481–486. [\[CrossRef\]](#)
39. 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, 5–9 October 2015; pp. 234–241.
40. Xu, Y.; Wu, L.; Xie, Z.; Chen, Z. Building Extraction in Very High Resolution Remote Sensing Imagery Using Deep Learning and Guided Filters. *Remote Sens.* **2018**, *10*, 144. [\[CrossRef\]](#)
41. Wu, G.; Guo, X.S.Z.; Chen, Q.; Yuan, W.; Shi, X.; Xu, Y.; Shibasaki, R. Automatic Building Segmentation of Aerial Imagery Using Multi-Constraint Fully Convolutional Networks. *Remote Sens.* **2018**, *10*, 407. [\[CrossRef\]](#)
42. Zhang, P.; Ke, Y.; Zhang, Z.; Wang, M.; Li, P.; Zhang, S. Urban Land Use and Land Cover Classification Using Novel Deep Learning Models Based on High Spatial. *Sensors* **2018**, *18*, 3717. [\[CrossRef\]](#)
43. Shrestha, S.; Vanneschi, L. Improved Fully Convolutional Network with Conditional Random Fields for Building Extraction. *Remote Sens.* **2018**, *10*, 1135. [\[CrossRef\]](#)

44. Wang, N.; Li, B.; Xu, Q.; Wang, Y. Automatic Ship Detection in Optical Remote Sensing Images Based on Anomaly Detection and SPP-PCANet. *Remote Sens.* **2019**, *11*, 47. [CrossRef]
45. Wang, Y.; Liang, B.; Ding, M.; Li, J. Dense Semantic Labeling with Atrous Spatial Pyramid Pooling and Decoder for High-Resolution Remote Sensing Imagery. *Remote Sens.* **2019**, *11*, 20. [CrossRef]
46. Ma, M.; Chen, J.; Liu, W.; Yang, W. Ship Classification and Detection Based on CNN Using GF-3 SAR Images. *Remote Sens.* **2018**, *10*, 2043. [CrossRef]
47. Fu, K.; Lu, W.; Diao, W.; Yan, M.; Sun, H.; Zhang, Y.; Sun, X. WSF-NET: Weakly Supervised Feature-Fusion Network for Binary Segmentation in Remote Sensing Image. *Remote Sens.* **2018**, *10*, 1970. [CrossRef]
48. Chen, J.; He, F.; Zhang, Y.; Sun, G.; Deng, M. SPMF-Net: Weakly Supervised Building Segmentation by Combining Superpixel Pooling and Multi-Scale Feature Fusion. *Remote Sens.* **2020**, *12*, 1049. [CrossRef]
49. Wang, J.; HQ Ding, C.; Chen, S.; He, C.; Luo, B. Semi-Supervised Remote Sensing Image Semantic Segmentation via Consistency Regularization and Average Update of Pseudo-Label. *Remote Sens.* **2020**, *12*, 3603. [CrossRef]
50. Gergelova, M.B.; Labant, S. Identification of Roof Surfaces from LiDAR Cloud Points by GIS Tools: A Case Study of Lučenec, Slovakia. *Sustainability* **2020**, *12*, 6847. [CrossRef]
51. Lyu, X.; Hao, M.; Shi, W. Building Change Detection Using a Shape Context Similarity Model for LiDAR Data. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 678. [CrossRef]
52. Montufar, G.F.; Pascanu, R.; Cho, K.; Bengio, Y. On the number of linear regions of deep neural networks. *arXiv* **2014**, arXiv:1402.1869.
53. Xu, Z.; Wei, X.; Luo, X.; Liu, Y.; Mei, L.; Hu, C.; Chen, L. Knowle: A semantic link network based system for organizing large scale online news events. *Future Gener. Comput. Syst.* **2015**, *4344*, 40–50. [CrossRef]
54. Song, W.; Liu, P.; Wang, L.; Lv, K. Intelligent processing of remote sensing big data: Status and challenges. *J. Eng. Stud.* **2014**, *6*, 6.
55. Lim, B.; Son, S.; Kim, H.; Nah, S.; Lee, K.M. Enhanced deep residual networks for single image super-resolution. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; Volume 1, p. 3.
56. 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.
57. Sergey, I.; Christian, S. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine learning, Lille, France, 6–11 July 2015.
58. Zhang, H.; Dana, K.; Shi, J.; Zhang, Z.; Wang, X.; Tyagi, A.; Agrawal, A. Context encoding for semantic segmentation. *arXiv* **2018**, arXiv:1803.08904.
59. Wang, P.; Chen, P.; Yuan, Y.; Liu, D.; Huang, Z.; Hou, X.; Cottrell, G. Understanding Convolution for Semantic Segmentation. *arXiv* **2018**, arXiv: 1702.08502v3.
60. Dai, S.; Han, M.; Xu, W.; Wu, Y.; Gong, Y.; Katsaggelos, A.K. Soft-cuts: A soft edge smoothness prior for color image super-resolution. *IEEE Trans. Image Process.* **2009**, *18*, 969–981.
61. Sun, J.; Xu, Z.; Shum, H.-Y. Image super-resolution using gradient profile prior. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
62. Yan, Q.; Xu, Y.; Yang, X.; Nguyen, T.Q. Single image superresolution based on gradient profile sharpness. *IEEE Trans. Image Process.* **2015**, *24*, 3187–3202.
63. Marquina, A.; Osher, S.J. Image super-resolution by tv- regularization and bregman iteration. *J. Sci. Comput.* **2008**, *37*, 367–382.
64. Zhang, Y.; Tian, Y.; Kong, Y.; Zhong, B.; Fu, Y. Residual dense network for image super-resolution. *arXiv* **2018**, arXiv:1802.08797.
65. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. *arXiv* **2018**, arXiv:1709.01507.
66. Li, H. Pyramid Attention Network for Semantic Segmentation. *arXiv* **2018**, arXiv:1805.10180.
67. 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]
68. Inria Aerial Image Labeling Dataset. Available online: <https://project.inria.fr/aerialimagelabeling/> (accessed on 11 August 2018).
69. Satellite Dataset II (East Asia). Available online: <https://study.rsgis.whu.edu.cn/pages> (accessed on 10 September 2019).