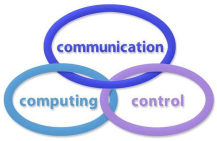
INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
Online ISSN 1841-9844, ISSN-L 1841-9836, Volume: 16, Issue: 4, Month: August, Year: 2021
Article Number: 4245, https://doi.org/10.15837/ijccc.2021.4.4245



CCC Publications

Efficient Building Extraction for High Spatial Resolution Images Based on Dual Attention Network

D. D. Zhao, H. S. Zhao, R. C. Guan, C. Yang

Dandong Zhao and Chen Yang*
College of Earth Sciences,
Jilin University,
Changchun 130061, China
zhaodd19@mails.jlu.edu.cn
*Corresponding author: yangc616@jlu.edu.cn

Haishi Zhao and Renchu Guan
Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education,
College of Computer Science and Technology,
Jilin University,
Changchun 130012, China
zhaohs18@mails.jlu.edu.cn
guanrenchu@jlu.edu.cn

Abstract

Building extraction with high spatial resolution images becomes an important research in the field of computer vision for urban-related applications. Due to the rich detailed information and complex texture features presented in high spatial resolution images, the distribution of buildings is non-proportional and their difference of scales is obvious. General methods often provide confusion results with other ground objects. In this paper, a building extraction framework based on deep residual neural network with a self-attention mechanism is proposed. This mechanism contains two parts: one is the spatial attention module, which is used to aggregate and relate the local and global features at each position (short and long distance context information) of buildings; the other is channel attention module, in which the representation of comprehensive features (includes color, texture, geometric and high-level semantic feature) are improved. The combination of the dual attention modules makes buildings can be extracted from the complex backgrounds. The effectiveness of our method is validated by the experiments counted on a wide range high spatial resolution image, i.e., Jilin-1 Gaofen 02A imagery. Compared with some state-of-the-art segmentation methods, i.e., DeepLab-v3+, PSPNet, and PSANet algorithms, the proposed dual attention network-based method achieved high accuracy and intersection-over-union for extraction performance and show finest recognition integrity of buildings.

Keywords: high spatial resolution images, building extraction, self-attention mechanism, dual attention network, deep learning.

1 Introduction

The goal of automatically extract buildings is efficient recognized building pixels from high spatial resolution images that are widely applicated in the urban planning, terrain analysis and smart cities. The characteristics of high spatial resolution, wide coverage and high timeliness make the high spatial resolution images truly restore the details of ground objects. The high spatial resolution images gradually become the preferred data for building extraction. However, the rich texture, topology and structure in high spatial resolution images increase interference information for buildings in the background, and the diversity of building (e.g. shape, density and distribution) become extremely complex. Therefore, automatic buildings extraction is a challenging task.

Traditional methods focus on extracting artificial and shallow features of buildings [1, 2, 3]by establishing models [4] or using some machine learning classifiers [5, 6] to analyze and extracted building pixels. For example, Huang et al. developed the Morphological Building Index (MBI) based on the inherent spectral information and shape characteristics of buildings [7]. Hu et al. applied morphological reconstruction and decision tree to comprehensively extract the morphological features of buildings and proposed an enhanced morphological construction index (EMBI) building extraction method [8]. Huertas et al. tried a variety of rectangular models of structural shapes to detect the presence of buildings based on projections [9]. Inglada uses a large number of geometric features to automatically identify and extract buildings in high-resolution optical remote sensing images through support vector machine (SVM) [10]. However, for the high spatial resolution images, buildings present large variability in distribution and shapes and tend to be more dispersed and smaller than other individual features such as roads and water bodies. The artificial or fixed morphological features cannot express the higher-level semantic information, affecting the accuracy of building extraction.

With the rapid development of machine learning (ML), deep learning (DL)-based algorithms, especially the convolutional neural networks (CNN) have considerably improved the performance of computer vision and image processing. Many models have been gradually improved from CNN, such as AlexNet [11], VGGNet [12], GoogleNet [13] and ResNet [14]. The CNN-based methods have been continuously presented for building extraction. S. Saito applied a patch CNN method to learn the mapping between the pixel values in the image and the corresponding building labels [15] . Minh et al. introduced a new loss function to train the convolutional neural network and added structure to the output for extracting building block [16]. Lv et al. combined with the idea of GEOBIA and proposed a high-resolution image classification method based on region-based majority voting CNNs [17]. Full convolution neural network (FCN) is a classification network with strong prediction ability. Many improved models based on FCN, with the ability of pixel level prediction, have achieved good results in the task of image segmentation [18, 19, 20, 21, 22, 23, 24, 25]. Therefore, many improved FCN models are used to building extraction [26, 27, 28, 29]. In order to solve the scale changes of buildings in VHR images, R. Davari Maj et al. proposed a new object-based deep OCNN framework [30]. H. Guo et al. performs pixel-level mapping of buildings in different scenarios and proposes a multi-task parallel attentional convolutional network (MTPA-Net) [31]. The above methods demonstrated the advantages of deep convolutiona convolutional neural networks in the field of building extraction. However, there are still two issues should be solved. First, the distribution of buildings is non-proportional. Therefore, the long-distance spatial dependencies of global context information should be combined with local features for considering the spatial characteristic of building. Second, the changes in brightness and scale of buildings in high spatial resolution images result to confused features presented in buildings, and the feature representation must be enhanced.

To address the above problem, this paper investigates the self-attention mechanism to optimize the building extraction performance [32]. A build extraction framework based on deep residual neural network with a self-attention mechanism is proposed. This mechanism contains two parts: one is the spatial attention module, which is used to aggregate and relate the local and global features at each position (short and long distance context information) of buildings. The other is channel attention module, in which the representation of comprehensive features (includes color, texture, geometric and high-level semantic feature) are improved. There are two advantages in the proposed framework: (i) It can fully collect the high-level semantic features of buildings in spatial domain and spectrum dimension. (ii) It can not only ensure the division of edges between large buildings and

complex backgrounds, but also ensure the attention and recognition of small buildings to a certain extent by adjusting the characteristic response of channels. Inspired by two modules, the proposed framework is called dual attention-based build extraction method (DANet-based BE). The effectiveness of our method is validated on a wide range high spatial resolution image, i.e., Jilin-1 Gaofen 02A imagery and is compared with some state-of-the-art segmentation methods. Experimental results have demonstrated that the proposed dual attention-based method can makes buildings extracted effectively and completely from the complex backgrounds.

The rest of the paper is organized as follows: Section 2 illustrates the detailed structure of the proposed framework. In the Section 3, the high spatial image resolution data set and preprocessing are described. The experiment results and analysis are arranged to the Section 4. Finally, the paper is summarized in Section 5.

2 Methods

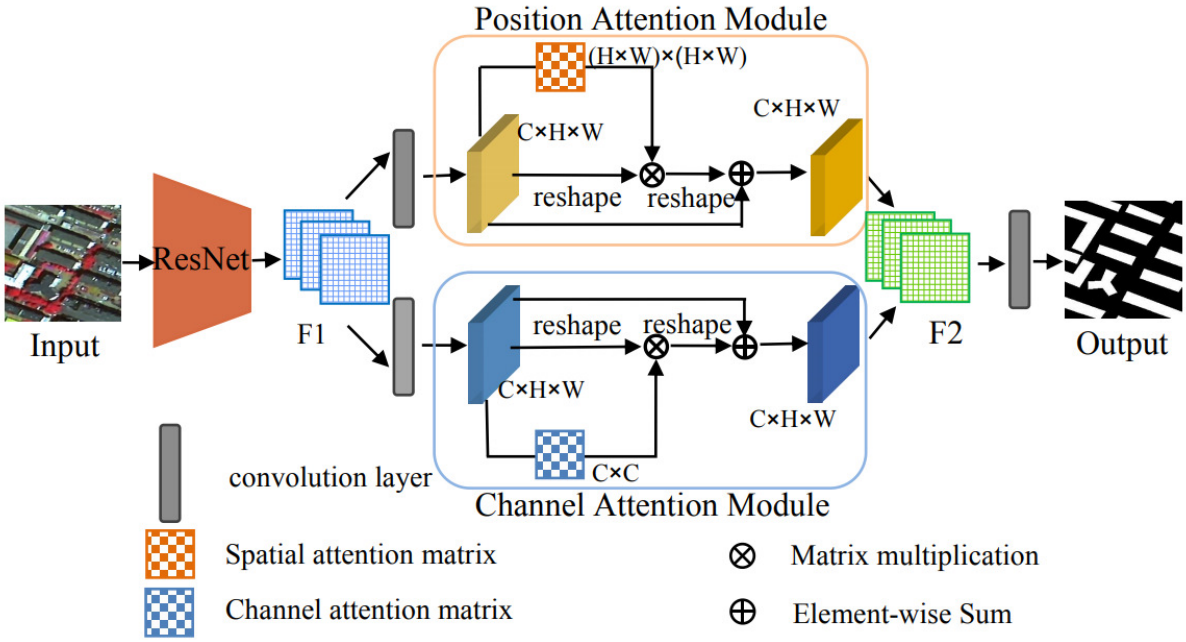


Figure 1: The framework of dual attention-based build extraction

Similar to dual attention-based network, the method for building extraction mainly consists of three parts. First, residual neural network (ResNet) is used as the pre-training backbone network. Second, the generated feature maps by ResNet are sent to two attention modules. For the channel attention module, the feature maps of the convolution layer are input to the channel attention module after dimension reduction. In the channel attention module, the module uses the self-attention mechanism to model the dependencies between different channels and calculate the channel attention matrix. Then, the weighted summation of the mapping between all channels is performed, and each channel graph is updated to generate new features that can reflect the relationship between the channels. Meanwhile, the spatial attention module calculates the weight value according to the feature similarity between pixels, weights and updates the spatial feature of each location, and obtains a new spatial feature that reflects the remote context information. Finally, the new features output by the two attention modules at the pixel level are summarized and merged. Set the convolution layer to convert the results and output, and get the prediction results of the building. The process of building extraction algorithm proposed in this paper is shown in Figure 1.

2.1 Backbone Network

In the ResNet network, the direct correlation channel and the nonlinear identity mapping by shortcut connection learning are increased. The original required learning is transformed for easily-optimization. It improves training speed and reduces the difficulty of model training. Due to the down-sampling operation in the last two blocks of ResNet, the output feature map will be reduced by half. We use two dilated convolutions to replace the down-sampling in these two blocks, thus the feature map of 1 / 8 size of the input image can be obtained and the loss of information in the pre-training process is reduced. Meanwhile, it makes ResNet have stronger pixel-level prediction ability, and no additional parameters need to be added. In this paper, ResNet-50 and ResNet-101 were selected for experiments.

2.2 Dual Attention Module

2.2.1 Spatial attention module

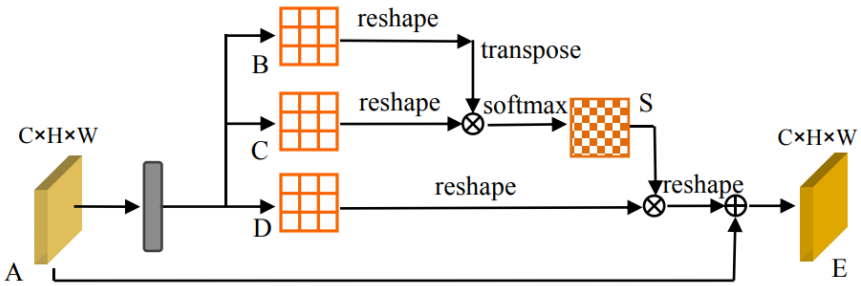


Figure 2: Spatial Attention Module

The spatial attention module adaptively aggregates the similar features in the global spatial range and uniformly encodes them into local features, which increases the representation range of local features and improves the ability of the model to distinguish features. In this paper, the module learns the detailed features (including the shape, proportion and distribution of buildings) between the spatial positions of all buildings in the image, which improves the semantic consistency and compactness within the buildings in the local features, and the resulting spatial feature map has stronger feature expression. The local features generated by the module have good classification effect on fuzzy pixels and can reduce the error classification of edge pixels.

As shown in Figure 2, the local features $A \in R^{C\times H\times W}$ sent to the spatial attention module generate three new feature maps B, C, D $(B, C, D, \in R^{C\times H\times W})$ through the convolution layer, then reshape B, C, D as the number of pixels of the feature map $R^{C\times N}, N = H \times W,$. The matrix $R^{N\times N}$ obtained by multiplying the transpose of C and B, performs softmax processing on each point (i, j) in row i and column j of the matrix to obtain a spatial attention map $S \in R^{N\times N}$, s_{ij}represents the effect of the $i-th$ pixel on the $j-th$ pixel. The larger the value is, the closer the two pixels are.

$$S \in R^{N\times N}; \quad s_{ji} = \frac{\exp(B_i \cdot C_j)}{\sum_{i=1}^{N} \exp(B_i \cdot C_j)} \tag{1}$$

Then, the transpose of the feature map D and S are multiplied and reshaped to $R^{C\times H\times W}$. Finally, it is multiplied by the scale parameter α (the initial value is 0, and gradually learn to assign more weights in training learning), and perform element-wise summation with A to obtain the final output as the feature matrix $E \in R^{C\times H\times W}$. The above operation is as follows:

$$E \in R^{C\times H\times W}; \quad E_j = \alpha \sum_{i=1}^{N} (s_{ji} D_i) + A_j \tag{2}$$

The resulting feature at each location is the weighted sum of the features E at all locations and the original features. Therefore, the spatial attention module can selectively aggregate the context information and enhance the semantic consistency within the class.

2.2.2 Channel attention module

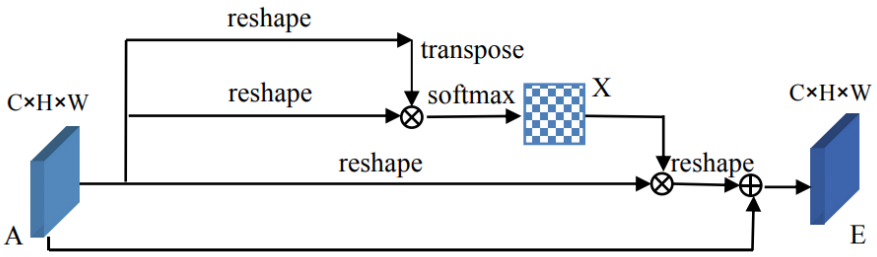


Figure 3: Channel Attention Module

For high resolution remote sensing images, each channel map with advanced features is a response to a specific class of objects on the image. There is a part of information fusion between the channel graphs. Fully exploiting the interdependence among them can generate feature graphs that highlight the interdependence and improve the representation ability of specific semantics. Channel attention module builds the interdependence between channels in an explicit way. Buildings and some non-buildings have similar material and reflection conditions, resulting in little difference in channel characteristics of pixels. The channel attention module can analyze and integrate the characteristic relationship between different channels on the basis of learning the single channel characteristics of the image. Therefore, the different types of pixels with similar channel characteristics have good classification ability, which can distinguish the pixels of buildings and other objects more clearly.

As shown in Figure 3, the original feature $A \in R^{C\times H\times W}$ is reshaped into $R^{C\times N}$, matrix A_i and its transpose matrix A_j to perform matrix multiplication, and then through the softmax layer, the channel attention matrix $X \in R^{C\times C}$ is obtained, x_{ji} represents the influence of the $i-th$ channel on the $j-th$ channel.

$$X \in R^{C\times C}; \quad x_{ji} = \frac{\exp(A_i \cdot A_j)}{\sum_{i=1}^{C} \exp(A_i \cdot A_j)} \tag{3}$$

Then, X and the transpose matrix of A is matrix multiplied, and the resulting matrix is reshaped into $R^{C\times H\times W}$. Finally, the matrix is multiplied by the parameters *beta*, and then the element by element sum operation is performed with the A matrix, and the final characteristic matrix $E \in R^{C\times H\times W}$ of each channel is finally output. The calculation process is as follows:

$$E \in R^{C\times H\times W}; \quad E = \beta \sum_{i=1}^{C} (x_{ji} A_i) \tag{4}$$

The learnable parameter β is initialized to 0. The above two formulas show that the final feature of a single channel is the weighted sum of the features of all channels and the original channel. Therefore, the channel attention module models the dependency between channels, which helps to improve the long-term semantic dependency between feature maps.

2.3 Module Embedding

In order to aggregate building features from two dimensions and prepare for pixel prediction, we embed the spatial attention module and channel attention module into the existing FCN pipeline. First, we set up the convolutional layer to convert and output the new features generated by the two attention modules. Then, feature fusion is performed on the two types of new features after conversion in a way of element-by-element accumulation. Finally, the final prediction map is generated through the convolutional layer. The embedding process effectively enhances the feature representation under simple operations and make the network achieve a better prediction effect for building extraction.

3 Experimental preparation

3.1 Dataset

To evaluate the proposed method, the Jilin-1 Gaofen 02A imagery was used for building extraction validation. Jilin-1 Gaofen 02A imagery is a push-broom image with a panchromatic resolution better than 0.75m, a multi-spectral resolution better than 3m, and a width greater than 40km. There are four spectral bands, i.e., blue, green, red and near-infrared spectrum channels. The orbit height of the Jilin-1 Gaofen 02A imagery is 535km, the sub-satellite point is 21.5km, and the uncontrolled positioning accuracy is 20m. The Jilin-1 Gaofen 02A imagery has high spatial resolution, and the

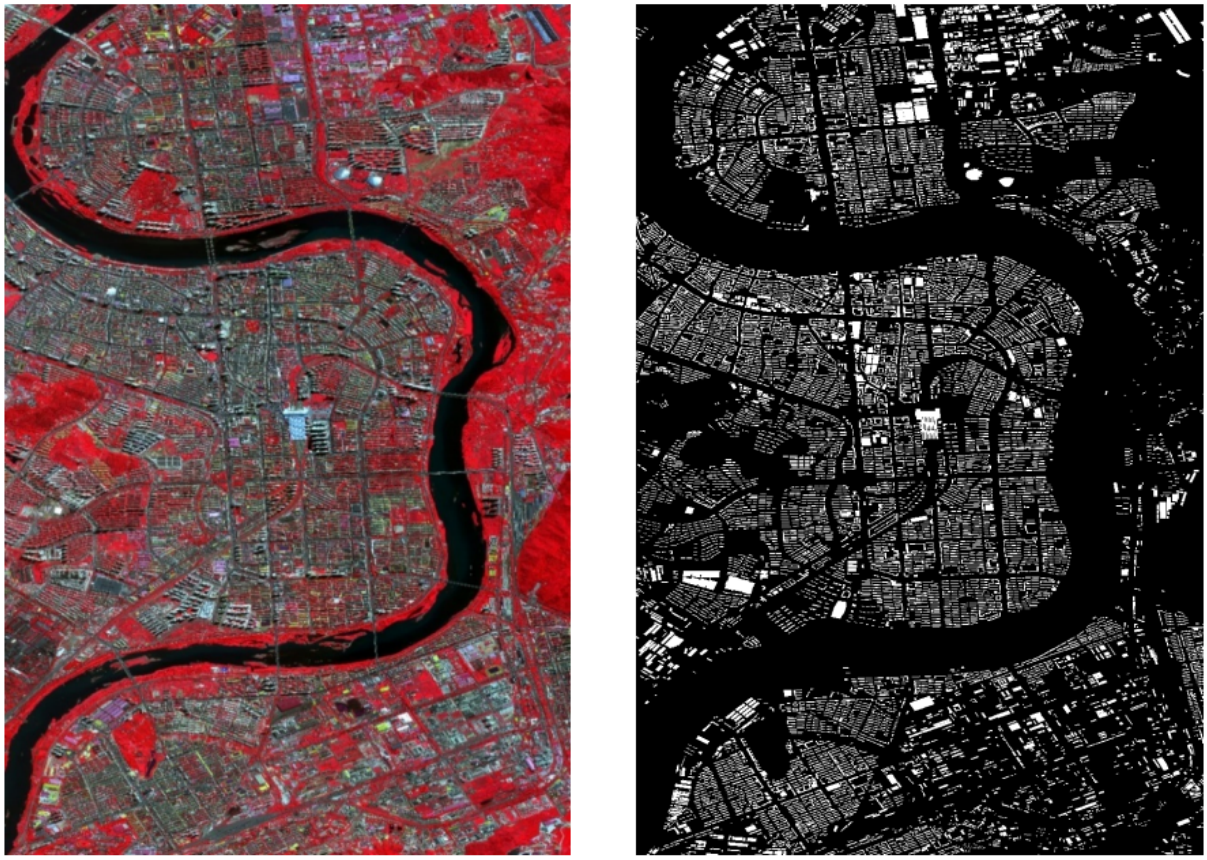


Figure 4: Channel Attention Module

edges of buildings are relatively clear. The characteristic of the considered dataset is that the image contains both the city center with dense buildings and the suburbs with few buildings. There are many green bushes between large and small buildings. Buildings have rich changes in density, height, top surface area and shape, degree of inclination, light and shadow. The roads and rivers show confused features with at the border of buildings. Therefore, this data set can be used to verify the ability of different algorithms for building extraction.

3.2 Image preprocessing

The extraction of buildings is manually drawn by visual interpretation through comprehensive direct interpretation, comparative analysis and logical reasoning. In the rendering process, due to the influence of satellite sampling angle and illumination angle, there is partial overlap in the roof of adjacent buildings with large height difference in the image, and the edge of the building under the shadow is too dark to be clear. After comprehensive comparison of real ground objects and image enhancement, the building boundary information in complex situations is drawn. Auxiliary references include Google Earth and Baidu Maps. In this paper, three spectral channels, i.e., the near infrared, red and green are selected for false color synthesis of the image. The false color composition of the image and the corresponding ground truth map are shown in Figure 4. The image covers about 80% of the central area of Jilin City, Jilin Province, China.

For training and prediction, the high spatial resolution image is cut into two sub-regions by 7: 3. The 70% image is randomly cut into 1415 non-overlapping 256×256 pixels pictures as training dataset. The cropping principle is that the number of building pixels in a single 256×256 pixels image accounts for 60% and more than 60% of it. Another 30% of the image is cut into 585 256×256 pixels pictures as test set. Some edge images without buildings are discarded.

4 Experiments and discussion

4.1 Experimental conditions

The experiment is based on the Ubuntu 16.04.5 LTS system, using a single GPU acceleration, the server processor is 12 Intel (R) core (TM) i7-5930K CPU @ 3.50GHz, the graphics card is NVIDIA TITAN X, the running environment version is Python 3.8.5, torch 1.6, CUDA 9.2. For the high spatial resolution data set used in this paper, the basic learning rate is set to 0.01, and the momentum and weight attenuation coefficients are 0.9 and 0.0001, respectively.

4.2 Evaluate indicators

This paper evaluates the building extraction results of high spatial resolution images from both qualitative and quantitative aspects. For the qualitative evaluation, the extraction results in test set is compared with the ground truth by visual inspection of the integrity and the continuity for the building. For the quantitative evaluation, the pixel accuracy (PA), overall accuracy (ACC) and mean intersection over union (MIOU) were used to compare and evaluate the building extraction results. The specific calculation is as follows:

$$PA = \frac{\sum_i n_{ii}}{\sum_i t_{ii}} \tag{5}$$

$$ACC = \frac{\sum_i^n n_{ii}}{N} \tag{6}$$

$$MIOU = \frac{1}{n_c} \sum_i \frac{n_{ii}}{t_i + \sum_i n_{ji} - n_{ii}} \tag{7}$$

where represents the total number of pixels whose category i is predicted to be category j, $t_i = \sum_j n_{ij}$, n_c represents the number of building categories in the data set, and N represents the total number of pixels.

4.3 Results and analysis

4.3.1 Parameter influence analysis

In order to analyze the influence of different parameters on building extraction results for the consider high spatial resolution data set, different convolution levels of the DANet-based BE, i.e., DANet-50 and DANet-101 are compared. In the experiment, 1415 remote sensing images were trained,

Table 1: Time and Evaluation Parameters of DANet-50 and DANet-101

Model	Backbone	PA	ACC	MIOU	Training time (min)	Test time (s/image)
DANet-50	ResNet-50	0.9431	0.9515	0.9074	244	14
DANet-101	ResNet-101	0.9408	0.9488	0.8594	288	22

and 525 images were tested. The number of iterations is 40000. The training time and test time are list in Table 1.

A qualitative assessment of the various methods on the Jilin-1 Gaofen 02A dataset can be seen in Figure 5. As can be seen from Figure 5 (c1) and (d1), the DANet-50 completely segments the boundary of the "I"-shaped building, while the DANet-101 miss the middle area of the building. For the image (d2), the DANet-101 recognizes parts of the river pixels as buildings. In the dense buildings in original image (a3), the gaps, i.e., adjacent or separated relationship between the buildings can be reflected more accurate using the DANet-50. For complex buildings in original image (a4), the DANet-50 is also better than the DANet-101 at discriminating non-buildings. As can be seen from table 1, the PA, ACC and MIOU of the DANet-50 are slightly higher than the DANet-101.

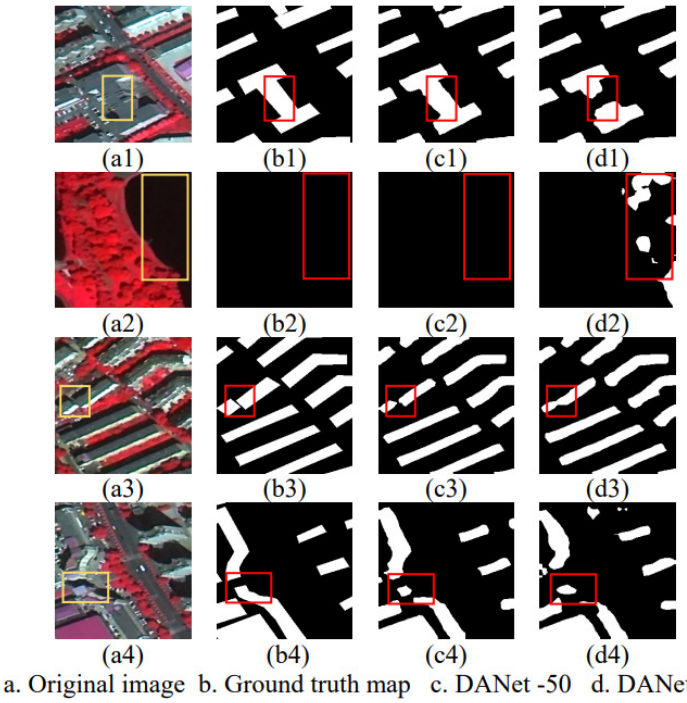


Figure 5: Segmentation Results of DANet-50 and DANet-101 on Jilin-1 Building Data Set

From the above qualitative and quantitative evaluation, the DANet-50 with a low number of convolutional layers shows more accurate results than the DANet-101 which have a high number of layers. It may be due to the fact that the increased depth in DANet-101 leads to an increase in the number of parameters and features that causes overfitting to the model. Therefore, the DANet-50 network has a better performance. In this paper, we use the DANet-50 for buildings extraction in the following experiment.

4.3.2 Efficiency Comparison

In this experiment, the building extraction performances of the proposed DANet-50 are compared with those of other state-of-the-art methods, i.e., DeepLab-v3+, PSPNet and PSANet. During the experiment, each model uses the same training data set and test data set, the learning rate is set to

0.01, and ResNet-50 is selected as the backbone network. Figure 6 shows the visual performances of different models on Jilin-1 building data set. In the visualization results, the white area represents the building, the black represents the non-building area and the yellow and red rectangles circle the details that need attention.

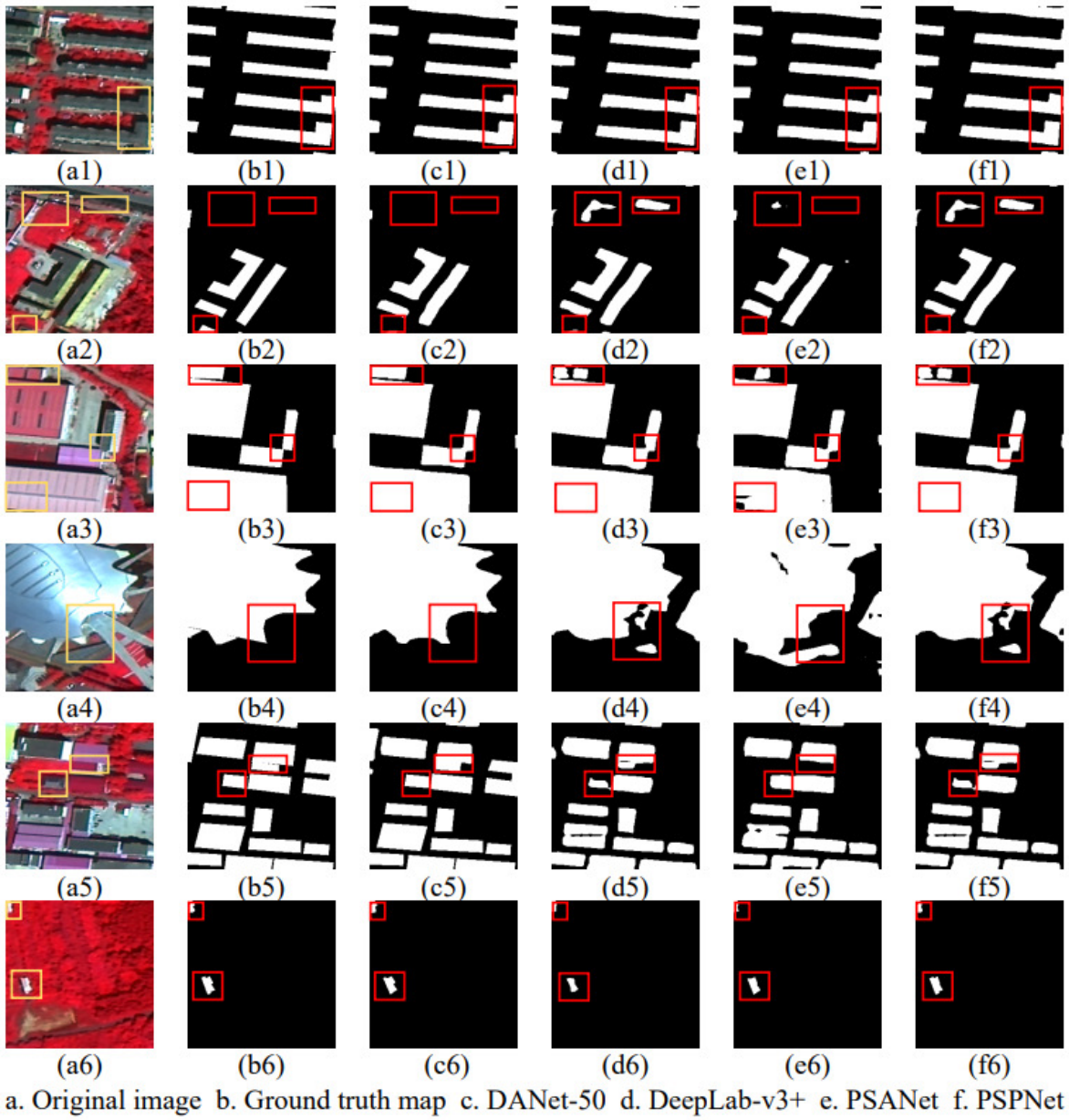


Figure 6: Visual Performances of Different Models on Jilin-1 Building Data Set

From Figure 6, one can observe that the building boundary and integrity segmented by proposed DANet-50 approach is more consistent with the ground truth map and achieved the best building extraction performances compared with other methods. However, for the other three compared methods, i.e., the DeepLab-v3+, PSPNet and PSANet, the extracted building pixels relatively dispersed and confused with non-building pixels. The original image (a1) presents the buildings with regular distribution, the proposed DANet-50 and the three considered methods performed well. There is objects similar to asphalt pavement in the original image (a2), the DANet-50 has the best recognition results, the DeepLab-v3+ and the PSPNet misclassified some of the non-buildings parts, and the PSANet cannot extracts small buildings at the edge of the image. In the original image (a3), the shape and color of buildings are quite different. As can be seen from Figure 6 (d3) (e3) and (f3), the roofs of buildings are "broken" and "missing". However, as shown in figure (c3), the buildings result performed by the proposed DANet-50 is excellent in integrity and boundary of buildings. Image

Table 2: Evaluation Accuracy and Execution Time by the Proposed DANet-50 and the DeepLab-v3+, PSPNet and PSANet

Model	**PA**	**ACC**	**MIOU**	**Training time(min)**	**Test time(s/image)**
PSPNet	0.9132	0.9216	0.8547	225	14
PSANet	0.9224	0.9305	0.8701	252	15
DeepLab-v3+	0.9243	0.9329	0.8742	264	15
DANet-50	0.9431	0.9515	0.9074	244	14

(a4) shows buildings with irregular shapes, from (c4) the extraction results of proposed DANet-50 is most close to the ground truth map. The other three considered methods were prone to incorrect exactions [see (d4) (e4) and (f4)]. For the original image (a5), the DANet-50 can better distinguish individual building in a dense region, but the other three considered methods have different degrees of omissions. In the original image (a6), all four methods can find small buildings hidden by plants, but the DANet-50 can extract more accurate building boundary.

Table 2 lists the evaluation accuracy and execution time of different methods. For the building extraction accuracy, the proposed DANet-50 achieves the highest PA, ACC and MIOU compared with the DeepLab-v3+, PSPNet and PSANet. Among the compared methods, the accuracy of PSANet and DeepLab-v3+ is close, and the PSPNet accuracy is relatively low. It is worth noting that the ACC and the MIOU obtained by the DANet-50 reached more than 95This is expected due to the spatial attention module and channel attention module applied in the DANet-50 which have strong adaptability in the scale context feature and each channel feature and can learn the details of the building in different scales, shapes and distribution. The comprehensive function of the two modules summarizes the high-level features of buildings and enhances the extraction effect of the model on buildings. Compared with the DeepLab-v3+, PSPNet and PSANet network models, the extraction results are more accurate. Moreover, the DANet-50 can ensure the integrity of complex building extraction and the consideration of global building information. It shows that the proposed DANet-50 has high adaptability and generalization ability for building extraction. For the execution time, under the same training samples, the training time of the proposed DANet-50 is slight longer than that of PSPNet network, and faster than that of PSANet network and DeepLab-v3+ network. This can be attributed to the fact that the method has a strong ability to integrate features and predict, the initial loss value is relatively small, and the frequency of modifying the weight value is low.

5 Conclusion

To efficiently extract building on high spatial resolution images, this paper presents a self-attention mechanism, i.e., DANet for building extraction. In the proposed building extraction method, the spatial attention module and the channel attention module are used to learn the dependency relationship between architectural spatial features and channels at different scales, respectively. Meanwhile, the global feature fusion and the correlation between semantic features are enhanced that lead to the spatial characteristic of building and the complex spectral feature representation were captured. A high spatial resolution imagery, Jilin-1 Gaofen 02A dataset, which contains complex building information was used to verify the effectiveness of the DANet. Experimental results have demonstrated that the DANet can maintain the integrity of the extracted building and distinguish the boundary of confused building-like objects easily in different scales and shapes, and dense urban areas or sparse suburban areas. Compared with other state-of-the-art methods, i.e., DeepLab-v3+, PSPNet and PSANet, the proposed DANet can aggregate and relate the local and global features at each position from the spatial attention module and represent comprehensive features with channel attention module. The

building extraction of visual performances and evaluation accuracy indicators were improved from complex background information of high spatial resolution images. In the future, the integrity of building extraction cloud be further improved by combination with the superpixel technique.

Funding

This work was supported in part by the National Natural Science Foundation of China under Grant 61572228, the Science-Technology Development Plan Project of Jilin Province of China under Grant 20190303006SF and 20190302107GX, the Industrial Innovation Special Funds Project of Jilin Province under Grant 2019C053-5 and 2019C053-7, and the Open Funds Project of Key Laboratory of Lunar and Deep Space Exploration LDSE201906.

Author contributions

The authors contributed equally to this work.

Conflict of interest

The authors declare no conflict of interest.

References

[1] Zhao, L.; Zhou, X.; Kuang, G. (2013). Building detection from urban SAR image using building characteristics and contextual information, *Journal on Advances in Signal Processing*, 56(1), 1-16, 2013.

[2] Aytekin, O.; Ulusoy, I.; Erener, A.; Duzgun, H. (2009). Automatic and unsupervised building extraction in complex urban environments from multi spectral satellite imagery, In *International Conference on Recent Advances in Space Technologies*, IEEE, 287-291, 2009.

[3] Chen, D.; Shang, S.; Wu, C. (2014). Shadow-based building detection and segmentation in high-resolution remote sensing image, *Journal of Multimedia*, IEEE, 287-291, 2009.

[4] Mohammad, A.; Clive, F. (2014). Automatic segmentation of raw lidar data for extraction of building roofs, *Remote Sensing*, 6(5), 3716-3751, 2014.

[5] Ok, A. O.; Senaras, C.; Yuksel, B. (2013). Automated detection of arbitrarily shaped buildings in complex environments from monocular VHR optical satellite imagery, *IEEE Transactions on Geoscience and Remote Sensing*, 51(3), 1701-1717, 2013.

[6] Meng, Y.; Peng, S. (2009). Object-oriented building extraction from high-resolution imagery based on fuzzy SVM, *International Conference on Information Engineering and Computer Science*, IEEE, 1-6, 2009.

[7] Huang, X.; Zhang, L.; Zhu, T. (2013). Building change detection from multitemporal high-resolution remotely sensed images based on a morphological building index, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(1), 105-115, 2013.

[8] Hu, R.; Huang, X.; Huang, Y. (2014). An enhanced morphological building index for building extraction from high-resolution images, *Acta Geodaetica et Cartographica Sinica*, 3(5), 514-520, 2014.

[9] Huertas, A.; Nevatia, R. (1988). Detecting buildings in aerial images, *Computer Vision Graphics and Image Processing*, 41(2), 131-152, 1988.

[10] Inglada, J. (2007). Automatic recognition of man-made objects in high resolution optical remote sensing images by SVM classification of geometric image features *ISPRS Journal of Photogrammetry and Remote Sensing*, 62(3), 236-248, 2007.

[11] Krizhevsky, A.; Sutskever, I.; Hinton, G.E. (2017). ImageNet Classification with Deep Convolutional Neural Networks, *Communications of the ACM*, 1097–1105, 2017.

[12] Simonyan, K.; Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition, *Computer Science*, 2014.

[13] Szegedy, C.; Wei, L.; Jia, Y.; Sermanet, P.; Rabinovich, A. (2014). Going deeper with convolutions, *IEEE Computer Society*,7, 1-9, 2014.

[14] He, K.; Zhang, X.; Ren, S. (2016). Deep Residual Learning for Image Recognition, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770-778, 2016.

[15] Saito, S.; Yamashita, T.; Aoki, Y. (2016). Multiple object extraction from aerial imagery with convolutional neural networks, *Journal of Imaging Science and Technology*, 60(1), 104021-104029, 2016.

[16] Mnih, V. (2013). Machine Learning for Aerial Image Labeling, *(Doctoral dissertation, University of Toronto (Canada).)*, 2013.

[17] Lv, X.; Ming, D.; Lu, T.; Zhou, K.; Wang, M.; Bao, H. (2018). A new method for region-based majority voting CNNs for very high resolution image classification, *Remote Sensing*, 10(12), 2072-4292, 2018.

[18] Chen, L. C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A. L. (2018). DeepLab: semantic image segmentation with deep convolutional nets, Atrous convolution, and fully connected CRFs, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4), 834-848, 2018.

[19] Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. (2017). Pyramid Scene Parsing Network. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 17, 6230-6239, 2017.

[20] Zhao, H.; Zhang, Y.; Liu, S., Shi, J.; Loy, C. C.; Lin, D. et al. (2018). Psanet: Point-wise spatial attention network for scene parsing, *European Conference on Computer Vision*, 11213, 270-286, 2018.

[21] Chen, L. C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation, *Lecture Notes in Computer Science*, 11211, 833-851, 2018.

[22] Ronneberger, O.; Fischer, P.; Brox, T. (2015). U-net: convolutional networks for biomedical image segmentation, *Lecture Notes in Computer Science*, 9351, 234-241, 2015.

[23] Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; Sang, N. (2018). Learning a Discriminative Feature Network for Semantic Segmentation, *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018

[24] Badrinarayanan, V.; Kendall, A.; Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation, *IEEE transactions on pattern analysis and machine intelligence*, 39, 2481–2495, 2017.

[25] Cheng, B.; Chen, L.C.; Wei, Y.; Zhu, Y.; Huang, Z.; Xiong, J.; Huang, T.S.; Hwu, W.M.; Shi, H. (2019). Spgnet: Semantic prediction guidance for scene parsing, *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 5218–5228, 2019.

[26] Y. Tan.; S. Xiong.; Y. Li. (2018). Automatic extraction of built-up areas from panchromatic and multispectral remote sensing images using double stream deep convolutional neural networks, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(11), 3988–4004, 2018.

[27] Sun, W.; Wang, R. (2018). Fully Convolutional Networks for Semantic Segmentation of Very High Resolution Remotely Sensed Images Combined With DSM, *IEEE Geoscience and Remote Sensing Letters*,15(3), 474-478, 2018.

[28] Zhang, R.; Li, G.; Li, M. et al. (2018). Fusion of images and point clouds for the semantic segmentation of large-scale 3D scenes based on deep learning, *ISPRS Journal of Photogrammetry and Remote Sensing*, 143, 85-96, 2018.

[29] Kampffmeyer, M.; Salberg, A. B.; Jenssen, R. (2016). Semantic Segmentation of Small Objects and Modeling of Uncertainty in Urban Remote Sensing Images Using Deep Convolutional Neural Networks. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 1, 680-688, 2016.

[30] R. Davari Majd,; M. Momeni.; P. Moallem. (2019). Transferable object-based framework based on deep convolutional neural networks for building extraction, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(8), 2627–2635, 2019.

[31] Guo, H.; Shi, Q.; Du, B.; Zhang, L.; Ding, H. (2020). Scene-driven multitask parallel attention network for building extraction in high-resolution remote sensing images, *IEEE Transactions on Geoscience and Remote Sensing*, 1-20, 2020.

[32] Fu, J.; Liu, J.; Tian, H; Li, Y.;Bao, Y.; Fang, Z. et al. (2020). Dual Attention Network for Scene Segmentation, *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3141-3149, 2020.

This journal is a member of, and subscribes to the principles of,
the Committee on Publication Ethics (COPE).
https://publicationethics.org/members/international-journal-computers-communications-and-control

Cite this paper as:
Zhao D.D.; Zhao H.S.; Guan R.C.; Yang C. (2021). Efficient Building Extraction for High Spatial Resolution Images Based on Dual Attention Network, *International Journal of Computers Communications & Control*, 16(4), 4245, 2021.
https://doi.org/10.15837/ijccc.2021.4.4245