Improved Fully Convolutional Network for the Detection of Built-up Areas in High Resolution SAR Images

Ding-Li Gao[1,2], Rong Zhang[1,2(✉)], and Di-Xiu Xue[1,2]

[1] Department of Electronic Engineering and Information Science, USTC, Hefei 230027, China
{xy6287,xuedixiu}@mail.ustc.edu.cn, zrong@ustc.edu.cn
[2] Key Laboratory of Electromagnetic Space Information, Chinese Academy of Sciences, Hefei 230027, China

Abstract. High resolution synthetic aperture radar (SAR) images have been widely used in urban mapping and planning, and built-up areas in high resolution SAR images are the key point to the urban planning. Because of the high dynamics and multiplicative noise in high resolution SAR images, it is always difficult to detect built-up areas. To address this matter, we put forward an Improved Fully Convolutional Network (FCN) to detect built-up areas in high resolution SAR images. Our improved FCN model adopt a context network in order to expand the receptive fields of feature maps, and it is because that contextual fields of feature maps which are demonstrated plays a critical role in semantic segmentation performance. Besides, transfer learning is applied to improve the performance of our model because of the limited high resolution SAR images. Experiment results on the TerraSAR-X high resolution images of Beijing areas outperform the traditional methods, Convolutional Neural Networks (CNN) method and original FCN method.

Keywords: High resolutions · SAR images · Built-up areas
Improved Fully Convolution Networks

1 Introduction

High resolution synthetic aperture radar (SAR) is the only imaging system that can generate high resolution imagery anytime, even in inclement weather or darkness. Thus, SAR images are widely used to observe the land, e.g. disaster management, land cover overlapping, etc. With great impact on human's life, especially in urban mapping and planning, the detection of built-up areas is more frequently applied to make rational use of the resource in urban areas. Besides, SAR images processing is commonly recognized as a hard task because of the high dynamics and multiplicative noise.

As for detecting built-up areas in high resolution SAR images, the feature extraction is the key point to get a great detecting performance. There are many methods proposed to extract different features to detect built-up areas in high resolution SAR images, for instance optical texture features [1] and Labeled Co-occurrence Matrix (LCM [2]) improved from Gray Level Co-occurrence Matrix (GLCM [2]). But using these features

Y. Zhao et al. (Eds.): ICIG 2017, Part III, LNCS 10668, pp. 611–620, 2017.
https://doi.org/10.1007/978-3-319-71598-8_54

to detect built-up areas in high resolution SAR images cannot perform well because of the strong speckle noise. In addition, the multiscale CNN method proposed in [13] can extract robust multiscale and hierarchical features directly form images against the strong speckle noise. But the multiscale CNN method [13] cannot form an end-to-end trainable deep network to detect built-up areas in high resolution SAR images, the conversion process of image block labels to pixel labels will limit the detecting performance. Meanwhile, the limited size of image blocks result in the limited size of contextual field of feature maps, adversely affecting the detecting performance.

In this paper, we utilize the FCN model to form an end-to-end trainable network to detect built-up areas in high resolution SAR images, and the FCN model has evolved to be one of the most important architectures in semantic segmentation after it is successfully adopted in [5]. While using the FCN model to detect built-up areas in high resolution SAR images, it is a pixel-to-pixel detecting process and we can directly get the label map of high resolution SAR images so that avoid the conversion process of image block labels to pixel labels, and the limited size of contextual field caused by the image blocks, leading to a better detecting performance.

Generally, it is difficult to train a strong FCN directly, thus, simply adapts the architecture of pre-trained classification network (e.g. VGGNet (VGG-16 [3]) trained on ImageNet [4] etc.) is always a good choice. Although this way can perform well in semantic segmentation, for different resolution data, it has unavoidable architecture defects to limit the performance in semantic segmentation. First and most, the pre-trained CNN is trained with low resolution images (e.g. 224×224 pixels), whereas our input SAR images are in high resolution. The simple adaption techniques adopted in FCN model cannot effectively address the domain gap which lead to a less optimized segmentation performance of FCN [11]. And it is because the feature maps which are used for classification in FCN model have limited contextual fields, causing the inconsistence in predictions for local ambiguous regions.

For this problem, we put forward an Improved Fully Convolutional Network to improve the detecting performance, we introduce a context network to replace the fully connected (fc) layers in the original FCN model so that expanding the receptive fields of corresponding feature maps, reducing such a domain gap so that get a better performance in semantic segmentation. The context network is stacked with several convolutional blocks (conv blocks), and we assemble Convolution, Batch Normalization, Rectified Linear Units (ReLU) as a conv block. Experimental results show that our improved FCN model performances better than traditional methods, the multiscale CNN method [13] and the original FCN method.

2 Fully Convolutional Network

Deep learning model can learn different hierarchies of features, among features, high-level features is building upon low-level features. Take the classic deep learning model CNN for demonstration, the CNN model can learn different hierarchies of semantic information in images, when trained with regularization, the CNN model can perform well in visual object recognition and image classification. Based on CNN

model, Long et al. [5] extend the classic deep classification architecture (CNN) by replacing the fully connected layers with fractional convolutional layers to learn per-pixel labels end-to-end. Besides, Long et al. [5] describe the fractional convolutional layer as useful learning filters to upsample so that mapping or connecting the coarse outputs to the dense pixel space.

Comparing with the CNN model to get semantic information in images, FCN model can form an end-to-end trainable network to learn per-pixel labels rather than per-block labels. Thus, the prediction process of per-pixel labels avoid the conversion of image block labels to image pixel labels so that improve the performance of semantic performance. Besides, there is not size limitation in input images when using FCN model while the input images should have the same size if using CNN model.

In this paper, we employ the FCN model to detect built-up areas in high resolution SAR images. Empirically, it is hard to train a strong FCN directly. The usual method is to adapt the architecture of pre-trained classification network trained on ImageNet [4]. And for us, we choose the VGG-16 [3] model as our basic model, then we use our dataset of high resolution SAR images to conduct fine-tune process.

Figure 1 illustrates the architecture of FCN model (the black vertical line denotes the convolution layer). Each convolutional layer is followed by ReLU activation function. Such a high-dimensional model is prone to overfit on the relatively small high resolution SAR images dataset. Under consideration, we take great care to mitigate overfitting through dropout and regularization during training. We plug in a batch normalization layer after every three convolutional layers from the fifth convolutional layer in the FCN model. Besides improving the training speed of the experiment, it can also mitigate overfitting which works similiarly as a regularization way. In addition, we add a dropout layer after the last convolutional layer in the FCN model which can also have an effect on mitigating overfitting.

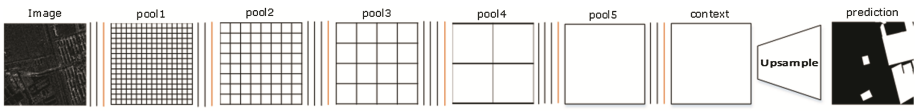

Fig. 1. Architecture of the FCN model

3 Improved Fully Convolutional Network

The pre-trained CNN in FCN model is optimized for the classification of low-resolution images, leading to a less competitive performance when used for classification or semantic segmentation in high resolution SAR images. It is because that the simple adaptation techniques adopted in FCN model cannot effectively address the domain gap, and the feature maps which used for classification in FCN model have limited contextual fields. Take the well-known FCN model (VGG-16 [3]) as instance, the feature map used for prediction (fc7, c.f. Table 1) don't have the matching contextual view with high resolution SAR images, the size of corresponding contextual view is only 404×404 pixels, and it is not large enough to make sure that nearly all elements in feature maps

are accessible to full-image context of high resolution. Meanwhile, the role of contextual filed plays in the final semantic segmentation performance is considered pretty important on the basis of previous literatures [6, 7]. In order to close such a domain gap, we adopt a context network in our FCN model to improve the contextual fields as for the feature maps which used for classification part right after the pre-trained CNN model.

Table 1. Receptive fields of feature maps in FCN model (VGG-16 [3])

Layer	Conv1_1	Conv1_2	Pool1	Conv2_1	Conv2_2	Pool2
Receptive field(px)	3 × 3	5 × 5	6 × 6	10 × 10	14 × 14	16 × 16
Conv3_1	Conv3_2	Conv3_3	Pool3	Conv4_1	Conv4_2	Conv4_3
24 × 24	32 × 32	40 × 40	44 × 44	60 × 60	76 × 76	92 × 92
Pool4	Conv5_1	Conv5_2	Conv5_3	Pool5	Fc6	Fc7
100 × 100	132 × 132	164 × 164	196 × 196	212 × 212	404 × 404	404 × 404

3.1 Context Network

According to the principles mentioned above, we attach a context network on top of the pre-trained CNN in order to improve the final performance of detecting built-up areas in high resolution SAR images. And the main effect which context network brought is to enlarge the receptive fields of the feature maps matching with the high resolution SAR images. Besides, our improved FCN model discards the last two fc layers (fc6, fc7) which are specific to image classification [8]. And replace it with our context networks. Through this way, we can efficiently reduce the size of our improved FCN model and gain a larger contextual field for high resolution SAR images. As for the context network, the components consist of the following three structures:

1. Convolution: Conv can efficiently aggregate neighborhood context. From a mathematical point of view $y(i) = \sum_{j=i+m}^{j=i-m} k(i)x(i)$ where k means the Conv kernel with size $2m + 1$, x and y are input and output signals. And it illustrates that signal y can be interpreted as a contextualized x. Meanwhile, Conv has been broadly and successfully devised and used in many literatures [6, 9, 15] to perform context aggregation.
2. Batch Normalization: Batch Normalization [10] is important to accelerate training process and improve the semantic segmentation performance. In the context network, we plug in the batch normalization layer right after the convolution layer.
3. ReLU: ReLU is the activation function we plug in the ReLU layer right after the batch normalization layer.

We assemble convolution, batch normalization, ReLU mentioned above as a conv block, and we constitute our context network by stacking multiple conv blocks.

3.2 Receptive Field

We know that the receptive field of convolution layers is related to the size of convolutional kernel and the stride size of former layers. When calculating the receptive filed of each layers, first, we can get the stride size of each layer, it can be calculated as (1),

$$\text{stride(t)} = \text{stride(1)} \times \text{stride(2)} \times \ldots \times \text{stride(t} - 1) \tag{1}$$

stride(t) means the stride size of each layer l_t.

The way to calculate the size of receptive field is from the deep layer to shallow layer. And it can be calculated as (2)

$$V'_{receptivefield} = \left(\left(V_{receptivefield} - 1\right) \times stride\right) + Size_{conv} \tag{2}$$

In (2), $V'_{receptivefield}$ means the receptive field size of the layer l_t which we want, and $V_{receptivefield}$ means the receptive field size of the layer l_t mapped to the previous layer l_{t-1}, *stride*. means the stride size of each layer, $Size_{conv}$ means the size of convolutional kernel.

From the calculating process in (2), we can get the size of receptive field which we want layer by layer.

According to the calculating process above, we can conveniently enlarge the size of receptive field when stacking multiple conv blocks to form context network with proper number of conv blocks used so that improve the detecting performance efficiently. And the size of receptive field we want will be the key rule to design the proper context network (determine the number of conv blocks used to be stacked with.)

3.3 Architecture of the Improved FCN Model

Figure 2 illustrates the overall framework of the improved FCN model employed in this paper. Comparing with the original FCN model, the improved FCN model replaces the fc6, fc7 layer in original FCN model as the context network which is stacked with several conv blocks. As for the high resolution SAR images, the designed context network in the improved FCN model can enlarge the receptive filed to a proper size corresponding to the input images.

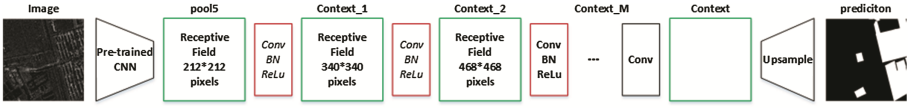

Fig. 2. Architecture of the context network, which is composed with a stack of M conv blocks. After each conv block (5 × 5 conv kernel in above example), the receptive fields of feature maps are expanded. Note that the spatial dimensionality of these feature maps keeps unchanged [11].

4 Transfer Learning

In this paper, we also explore the benefits of transfer learning in training our FCN model and the improved FCN model with limited high resolution SAR images.

In transfer learning, the method adapted to initialize our FCN model and the improved FCN model is sharing the learned weights from the source model. Specifically, the weights selected from convolution and upsampling layers in the source model will be copied or transferred to the corresponding layers in our own model, and the remaining layers of our model are then randomly initialized and trained using the high resolution SAR images dataset via supervised fine-tuning. For instance, we initialize our FCN model and the improved FCN model with the learned weights of selected convolution layers from the VGG-16 [3] model.

When training a large target model on the small target dataset, transfer learning can significantly decrease the probability of the occurrence of severe overfitting. Specifically, the benefits that transfer learning offered are listed as follows [14]:

1. Domain adaption –It will be allowed that the target task is different but related to the source task in transfer learning.
2. The difference between the distribution of source and target datasets are allowed.
3. It will be achievable to gain better convergence and accuracy performance when trained with limited training data.

In addition, when using transfer learning to do the semantic segmentation task, we can do some changes in FCN model for different resolution dataset in order to get a proper size of receptive field. Because it is known that the size of contextual field of feature maps plays a great role in the performance of semantic segmentation.

5 Experiments and Results

In our experiment, we use the high resolution TerraSAR-X SAR images collected on November 25, 2011 of Beijing areas to verify our method. The SAR images have range resolution of 2.3 m and azimuth resolution of 3.3 m. Building types in images include Dot villa district, residential quarter buildings, squatter settlement, etc. And 7900 samples were extracted in eastern areas of Beijing according to the high resolution optical maps to train the FCN model and the improved FCN model.

As for the context network in the improved FCN model, notice that the off-centering pixels require significantly larger receptive fields than centering pixels, in order to get the same actual contextual views, the context network is expected to be designed to expand the receptive field twice larger than the input image [11]. And the hidden dimension of Conv blocks is fixed to 512. In the training process, we initialized the new parameters engaged from the context work by Gaussian distribution with variance 0.01. And for our FCN model and the improved FCN model. The deconvolution operation [5] is used to perform upsampling operation. Both two models mentioned above are trained with SGD, and the value of momentum is set to 0.9. At the beginning of the training process, the learning rate is initialized to be 0.001, the value will decreased by 10 times

after 10 epochs (total epochs is 25). Besides, the FCN model and the improved FCN model are trained with the provided segmentation masks, the size of images for the phase of train and validation is 500 × 500 pixels. And the test set is a SAR image of (2500 × 4000) pixels from the northern areas of Beijing, the test image is separate from the train and validation dataset. As for the batch normalization layer, the statistics (mean and variance) is updated after the network is converged.

In our experiment, we use detection rate, false alarm rate and accuracy of classification to evaluate our result [12]. GLCM is the detecting result when using GLCM texture features consulting [1, 13], and LCM is the result of [2, 13], Multiscale CNN is the result of [13]. FCN, Improved FCN represent the result using our FCN model and improved FCN model discussed in this paper. From the comparing of results mentioned in Table 2, we can find that our improved model performs better to detect built-up areas in high resolution SAR images.

Table 2. Performance comparisons of different methods (corresponding to different models)

Method	Detection rate	False alarm rate	Accuracy of classification
GLCM	84.38%	15.82%	88.78%
LCM	89.39%	23.40%	86.16%
Multiscale CNN	92.14%	10.71%	92.86%
FCN	92.20%	9.16%	92.48%
Improved FCN	**92.53%**	**8.63%**	**93.05%**

In addition, we design several different context networks with different architectures. And all the context networks designed have the same size of receptive field. Specifically, all the improved FCN model with different context networks have a significantly large receptive field than the origin FCN model, and the size of receptive fields make sure that nearly all elements in feature maps are accessible to full-images context.

Table 3. Performance comparisons of improved FCN model with different setup of context networks. The format k × k (m) denotes that the context network is stacked with m consecutive k × k convolution blocks

Context network	Receptive Field (px)	Detection rate	False alarm rate	Accuracy of classification
3 × 3(12)	980 × 980	92.45%	8.72%	92.95%
5 × 5(6)	980 × 980	**92.53%**	**8.63%**	**93.05%**
7 × 7(4)	980 × 980	92.38%	8.59%	92.84%
FCN	404 × 404	92.20%	9.16%	92.48%

The performance in Table 3 clearly shows the context network can improve the performance to detect built-up areas in high resolution SAR images. From Table 3, the format k × k (m) denotes that the context network is stacked with m consecutive k × k convolution blocks. All context networks have the same size receptive field which is carefully designed to make sure that all elements in feature maps are

accessible to full –image context. And their detecting performances are very similar, almost the same which illustrates the size of receptive field rather than the architecture of context network plays a critical role in the final performance when detecting built-up areas in high resolution SAR images.

In addition, the new parameters play a great role in filling the domain gap during the fine-tuning of improved FCN. Comparing the magnitude of parameters of improved FCN and the original FCN model, the magnitude of parameters of improved FCN is larger than the original FCN model. It illustrates that the detecting performance improvement of built-up areas in high resolution SAR images is not simply originate from engaging more parameters. Besides, we conduct a controlled experiment to validate that the larger receptive field rather than the larger magnitude of parameters to improve the detecting performance built-up areas in high resolution SAR images.

In order to make the comparsion of context networks with different magnitude of parameters, we design two different context networks, the only differences between them is the hidden dimension. The hidden dimension of context network (*) is increased from 512 to 1024. Thus, the context network with mark (*) involves larger magnitude of parameters. And the result in Table 4 illustrates that the larger magnitude of parameters don't significantly improve the performance when comparing with the original context network (hidden dimension is 512). Therefore, it can be inferred that the magnitude of parameters doesn't play a great role in improving the detecting performance of built-up areas in high resolution SAR images.

Table 4. Performance comparisons of context networks with different magnitudes of parameters.

Context network	Detection rate	False alarm rate	Accuracy of classification
5*5(6)	92.53%	8.63%	93.05%
5*5(6)(*)	92.46%	8.57%	93.13%

Figure 3 shows our detection result on the high resolution SAR images, Fig. 3(a) shows us the origin high resolution SAR image, it illustrates that the size and the type of buildings in this area are complex. Both large span of built-up areas and scattered building appeared in this areas. Figures 3(b) shows the label of built-up areas which is manually labeled according to optical maps, and the regions in red are the built-up areas. Figure 3(c) shows the detecting result with FCN model, the regions in blue are the built-up areas. Figure 3(d) shows the detecting result of built-up areas with the improved FCN model. According to the detecting result (c), (d), we can find that the improved FCN model performs better than the FCN model.

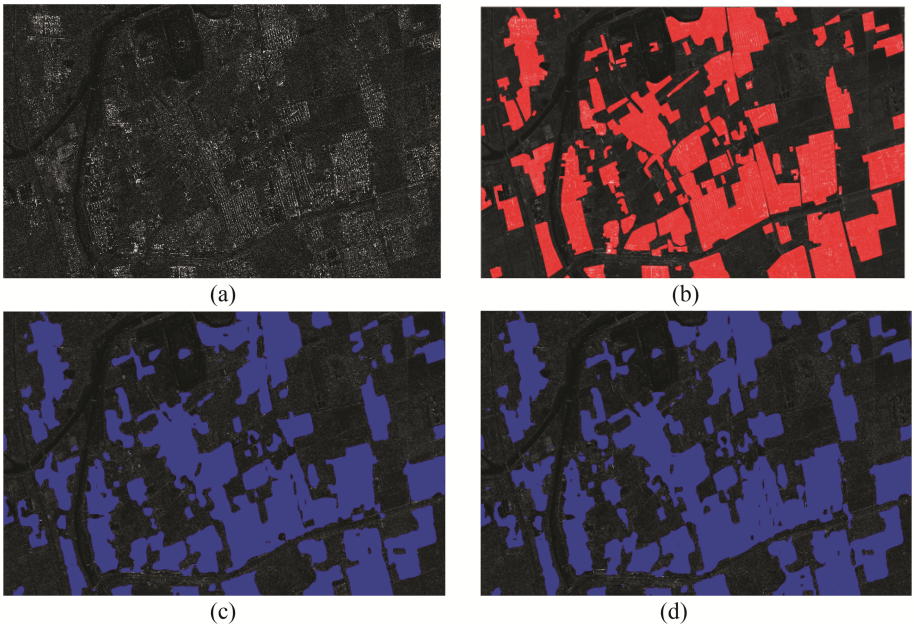

Fig. 3. Experimental results (a) SAR Image of northern areas of Beijing. (b) Manually Labeled Image. (c) Detection Result of FCN model. (d) Detection Result of improved FCN model.

6 Conclusion

In this paper, we proposed an improved FCN model to solve the problem of built-up areas detection in high resolution SAR images. By the approach of introducing a context network to expand the receptive fields of feature maps in the FCN model, we improve the performance of detecting built-up areas in high resolution SAR images. Experimental results on TerraSAR-X SAR images get a detection rate of 92.53%, false alarm rate of 8.63%, accuracy of classification of 93.05%. It indicates that the improved FCN model is effective to detect built-up areas in high resolution SAR images.

Acknowledgement. This work was supported in part by the National Nature Science Foundation of China (No. 61331020).

References

1. Yang, W., et al.: Supervised land-cover classification of TerraSAR-X imagery over urban areas using extremely randomized clustering forests. In: 2009 Joint Urban Remote Sensing Event (2009). IEEE
2. Li, N., et al.: Labeled co-occurrence matrix for the detection of built-up areas in high-resolution SAR images. In: SPIE Remote Sensing. International Society for Optics and Photonics (2013)

3. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
4. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
5. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015). 1, 2, 3, 4, 5, 6, 7, 8
6. 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 preprint arXiv:1606.00915 (2016). 2, 3, 7, 8
7. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated con-volutions. arXiv preprint arXiv:1511.07122 (2015). 2, 3, 7, 8
8. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: Advances in Neural Information Processing Systems, pp. 3320–3328 (2014). 1
9. Lin, G., Shen, C., van dan Hengel, A., Reid, I.: Efficient piecewise training of deep structured models for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016) (2016). 2, 3, 7, 8
10. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: ICML (2015). 5
11. Shuai, B., Liu, T., Wang, G.: Improving fully convolution network for semantic segmentation. arXiv preprint arXiv:1611.08986 (2016)
12. Shufelt, J.A.: Performance evaluation and analysis of monocular building extraction from aerial imagery. IEEE Trans. Pattern Anal. Mach. Intell. **21**(4), 311–326 (1999)
13. Li, J., Zhang, R., Li, Y.: Multiscale convolutional neural network for the detection of built-up areas in high-resolution SAR images. In: 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS). IEEE (2016)
14. Tran, P.V.: A fully convolutional neural network for cardiac segmentation in short-axis MRI. arXiv preprint arXiv:1604.00494 (2016)
15. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015)