IRSNET: An Inception-Resnet Feature Reconstruction Model for Building Segmentation

Kepeng Xu[1], Li Nie[1], Zhiqiang Zhang[1], Wenxin Yu[1](✉), Yunye Zhang[1], Wei Chen[1], Siyuan Li[1], Xinxin Zhou[4,5], Shangwei Deng[1], Pengfei Yu[1], Yibo Fan[2], Hui Zhang[1], and Valentin Bouillon[3]

[1] Southwest University of Science and Technology, Mianyang, China
star_yuwenxin27@163.com
[2] State Key Laboratory of ASIC and System, Fudan University, Shanghai, China
[3] École internationale des sciences du traitement de l'information, Cergy, France
[4] College of Geographical Science, Nanjing Normal University, Nanjing, China
[5] Key Laboratory of Virtual Geographic Environment, Ministry of Education, Nanjing Normal University, Nanjing, China

Abstract. Effective analysis of remote sensing images remains a challenging topic in deep learning research field. This paper proposes a semantic segmentation approach based on the inception architecture. Specifically, the combination of residual network and Inception improves the feature extraction capability of encoder network. In addition, this paper discusses the problem of unbalanced information allocation caused by concatenated operation in the up-sampling process of network models such as Unet, and the RCSE module is proposed to complete feature reconstruction to solve this problem. The approach ensures accurate assignment of semantic labels to the buildings in the aerial images. Experiments based on the dataset proposed by CrowdAi certify the effectiveness of our approach with a 3.7% IoU improvement compared to Unet.

Keywords: Aerial image · Building extraction · Convolution network · Semantic segmentation

1 Introduction

Remote sensing technology has developed rapidly in recent decades, which enabled us to obtain a very large amount of information on the surface of the Earth. Remote sensing image is one of the most important information obtained by remote sensing equipment. At present, the task of labeling building objects in remote sensing images in the industry is still at a very low level of technology, relying on a large amount of manpower for manual labeling. So building a soft system that generates semantic annotations for maps is very important.

T. Gedeon et al. (Eds.): ICONIP 2019, CCIS 1143, pp. 41–48, 2019.
https://doi.org/10.1007/978-3-030-36802-9_5

The goal of semantic segmentation is to assign a corresponding semantic annotation to each pixel in the image. In order to achieve this goal, most of the current network architecture is mainly based on two points to improve: 1. Improve the global information extraction capability of the model, that is, to establish a model that can obtain the global view and detect the building objects. 2. Improve the model's ability to save pixel-level location information.

Remote sensing image building extraction tasks are more complicated than semantic segmentation tasks in conventional scenarios, which can be concluded in the following five points: 1. The shape of a building is varied and very complicated. 2. The styles of buildings in different areas vary greatly. 3. The chromaticity of the acquired images is caused by different atmospheric lighting conditions. 4. The size of the building caused by the difference in sensor height is too large.

This paper proposes an end-to-end semantic segmentation model for building footprint segmentation. The encoder side of the model is based on the inception architecture [9], using the residual connection, as well as the atrous convolution is applied to the encoder. Among them, the receptive field of the convolution kernel is expanded in the last part of the encoder, while in the decoder part of the model, the feature layer is restored to the original image size by stepping up-sampling. Furthermore, in the recovery process, the feature layer in the model encoder is concatenated to the corresponding shape feature layer in the decoder using the shortcut connection [4], which contributes the decoder part of the proposed model to effectively utilize the high-level information and low-level information of the aerial image. Unlike previously, this paper considers a problem that connecting directly to the high and low level of information through the shortcut, though, to some extent, makes the decoder obtain more accurate pixel texture information, the feature layer obtained by the decoder is directly concatenated from the high and low levels and does not undergone a corrective process. And such feature information will interferes with the decoder of the entire semantic recovery to some extent. Considering what has been mentioned above, this paper adds the operation of channel weighting after the up-sampling. Moreover, to concatenate the feature of the layer after each channel gives different weights, RCSE module is proposed in this paper, which can help decoder to complete information reconstruction of the feature layer after upsampling and concatenating. And through such module, the proposed model can reduce the dependence on the information of the redundant feature layer, and finally increase the information recovery ability of decoder, so as to improve the accuracy of the segmentation in the model.

The main contributions of this paper include the following:

- A semantic segmentation model based on inception-resnet for building footprint segmentation is proposed.
- The application of shortcut in the model based on encoder-decoder and its shortcomings are discussed. RCSE module is proposed to reduce the feature redundancy and interference caused by shortcut.
- The proposed model can effectively complete the aerial image building footprint segmentation and achieve 91.2% Iou in the validation set.

2 Related Work

The purpose of image semantic segmentation is to assign pixel-level semantic annotations to images, which is an important part of computer vision. The full convolutional neural network applies convolutional neural networks to semantic segmentation tasks for the first time, and is very effective.

In the past few decades, remote sensing devices have been widely used, and a large amount of remote sensing image data has been collected. The aerial image building footprint extraction has been extensively studied. In the past traditional computer vision technology, researchers extracted features by manually designing feature extractors such as texture and color features, and then through the machine learning classifiers (such as AdaBoost, support vector machine, random forest, etc.) to complete the rough Scene classification, which requires a lot of subsequent processing to optimize the results. With the development of remote sensing equipment production technology, researchers have been able to obtain a large number of high-resolution remote sensing images, however, the traditional methods can't effectively process high-precision images and can't achieve pixel-level segmentation accuracy, but only complete scene-level classification. Therefore, in recent years, scholars have focused on the research of high-resolution image building footprint extraction tasks with convolutional neural networks.

Recently, convolutional neural networks have been rapidly developed in the extraction of building footprints. Zhang et al. proposed a method based on CNN classification, which calculates and classifies scores for each sliding window by using sliding windows of different scales, and then optimizes by using non-maximum suppression result.

Contrast to them, this paper proposes an end-to-end encoder-decoder model. In the encoder, the inception architecture is effectively combined with resnet, and the dilated convolution method is used to enhance the receptive field of the convolution kernel to enhance the coding capability of the model encoder. In the decoder of the model, the squeeze excitation module is combined with the shortcut connection, so that the decoder of the model can assign weights to the feature layer after the shortcut, which enables the model to allocate limited computing resources to the feature layer with the largest amount of information.

3 Methodology

3.1 IRSNET

This section will introduce the model in detail. Figure 1 is the overall architecture of the proposed model, which is divided into two parts: the encoder (Fig. 1 left) decoder (Fig. 1 Right side). The encoder side effectively extracts and abstracts the image pixel information by combining the Inception architecture with the residual connection. In the decoder part, the model gradually restores the feature map to the size of the input picture by upsampling. In order to ensure that the decoder can have pixel-level segmentation accuracy in the case of extracting global information, the proposed method connects the different sizes of the encoder layer directly to the decoder through the shortcut in the upsampling stage.

After each up-sampling process of the model decoder, the feature layer is further modified by the proposed RCSE module. Thus, the interference of the feature layer with less information can be reduced and the feature recovery accuracy of the decoder can be enhanced, which improves the expressive power of the model.

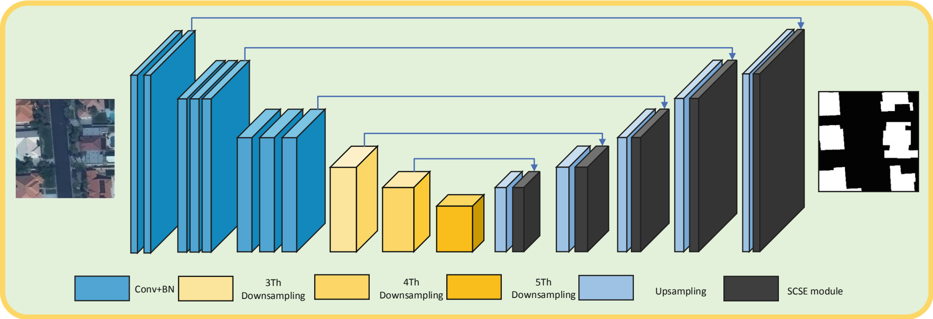

Fig. 1. Model architecture diagram

3.2 Encoder

At the encoder end of the model, for the input of the model, the feature is extracted using three convolution processes in the first two downsampling stages of the encoder. In the first convolution process, we take step size of 2 and complete the down-sampling. To reduce the size of the feature map, the third convolution process uses a hole convolution kernel of rate = 2 for convolution to expand the receptive field of the convolution kernel. During the last three sampling processes of the model encoder, this paper adds the Inception-Resnet structure to the model to improve the feature aggregation ability of the encoder. The Inception-Resnet structure of the three stages is shown in the Fig. 2.

3.3 Decoder

In the decoder part of the model, the feature layer is restored step by step to the original image size by upsampling. In order to enable the decoder to simultaneously acquire the image high-level semantic information and the image pixel position information, the provided method connect directly the corresponding size feature layer in the encoder to the decoder during the upsampling process. And the shape of the feature layer is also directly connected to the decoder. The research process reveals that simply connecting the feature layer and then convoluating allow the model to save both high-level information and low-level pixel information, however, decoder can not effectively distinguish the importance of different feature layers. Therefore, this paper introduces the Squeeze-and-Excitation mechanism into the feature layer concatenation, with different

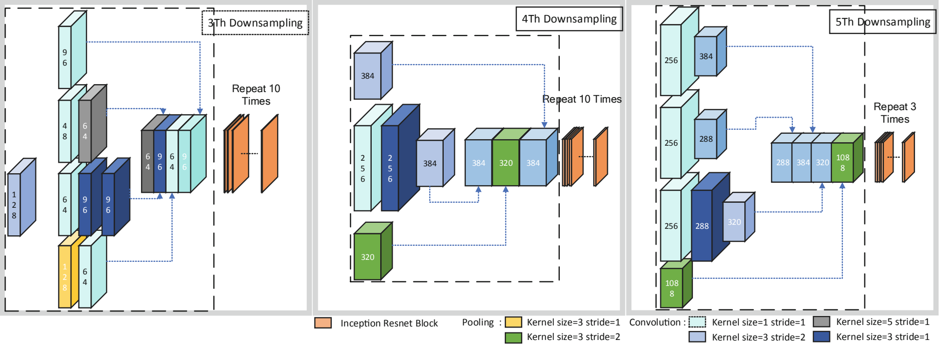

Fig. 2. These are the last three down-sampling model structure diagrams, and the black boxes of each diagram represent the Inception-Resnet structures of each process. These corresponding modules are superimposed several times after each down-sampling, with The Times of (10, 10, 3) respectively. Since in this article, in view of the remote sensing image segmentation task, the image size is small, the down-sampling size is small after the last stage. Although these feature layers extracte the high-level semantic information, lots of convolution padding operations also increase a lot of noise. The last Inception-Resnet has limited effect on improving the performance of the model, which leads that the last model of encoder only had three stages superposition Inception-Resnet structure.

weights assigned to different channels of the concatenated feature layer, which enhances the expressive ability of the decoder and thereby effectively improves the performance of the model.

3.4 RCSE: Refactoring ShortCut Squeeze Excitation Block

Unet segmentation model based on encoder and decoder, in the process of upsampling by Shortcut, can achieve high and low level of information fusion, and the decoder in the process of completing image information recovery can not only complete advanced level of the image object level detection, but be able to complete the accurate using low-level pixel characteristic information to the pixel level distribution of the label.

However, the decoder has the ability of capturing both high and low levels of image feature information at the same time, such simple directly concatenating of feature layers at different levels restricts the expretion ability of decoder. Therefore, we consider a problem that how much higher or lower levels of information contribute to overall model performance improvement after the operation of concatenation. Upon this thinking, we design the RCSE module and add it to the model upsampling process. In the concrete implementation of RCSE each channel of the shortcut feature layer was assigned different weights by SE block [3]. Making the decoder in the process of upsampling different features can extract more directional layer information, improve the efficiency of the decoder. The specific structure of RCSE is shown in Fig. 3.

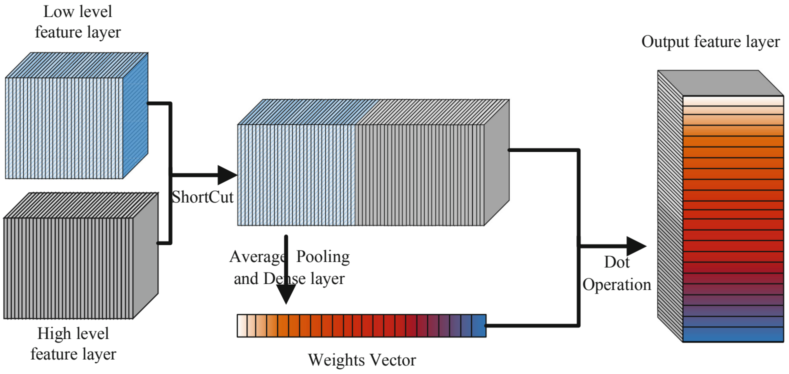

Fig. 3. RCSE architecture. As shown in the diagram, considering the different functions between high and low level feature layers, this paper designs the feature layer after the completion of feature scaling of the RCSE module after concatenation. The proposed module improved model performance with small additional parameters.

4 Experiments

The experiment is based on the Map Challenge data mining competition proposed by Crowdai. The dataset consists of a training set (280,000 300 * 300 aerial images and corresponding manual annotations) and a validation set (60,000 images and annotations). The data set has many advantages, and the semantic labeling is clear. This paper evaluates the experiment based on the dataset and compares it with SegNet, pspnet [11] and Unet, and proves the effectiveness of the proposed method.

4.1 Metrics

Our experiment takes Acc, Iou and F1 Score for evaluation metric.

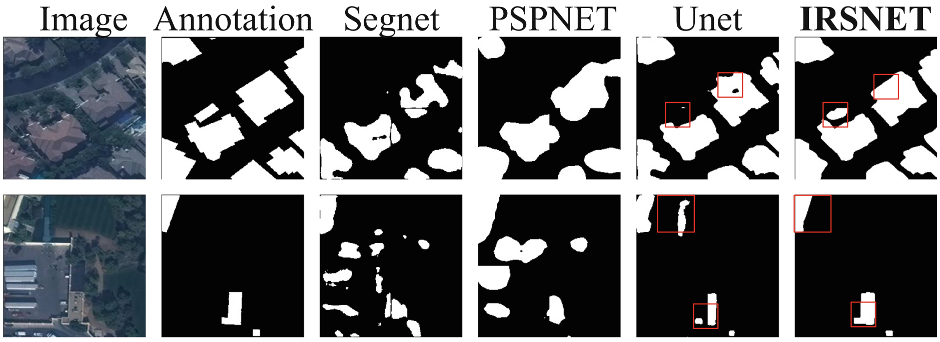

Fig. 4. Experimental result diagram. As shown in the figure, the rightmost side is the output part of the model proposed in this paper. The proposed model improves the extraction accuracy of small buildings and the segmentation effect of special edge structures such as right-angle and straight-line structures (as shown in the red box). (Color figure online)

4.2 Results

The results of the proposed model on the validation set are as follows Fig. 4 and Table 1. From the table, we can see that the performance of our model based on Inception-Resnet is higher to some extent compared with other models, and the RCSE module proposed significantly improves the performance of the model.

Table 1. Comparison table of experimental results

Model	Acc	Iou	F1score
IRSNET	**96.8**	**91.2**	**92.7**
IRSNET (Without RCSE)	**95.7**	**90.2**	**91.7**
Unet (With RCSE)	96	89.0	90.2
Unet	95.1	87.5	89.7
Segnet	87.0	69.13	69.24
pspnet	86.3	68.4	68.9
FCN	83.8	65.8	67.1

4.3 Analysis

The proposed model adds Inception-Resnet to the encoder side of the model, which improves the feature extraction capabilities of the model. In addition, the proposed RCSE block can reconstruct the feature layer formed by short cut after the up-sampling process. This operation enables the model to modify the large difference of eigenvalues between the two ends of short cut, so as to enhance the expressive ability of the model.

Because remote sensing images have the characteristics of great scale changes, PSPNET is chosen as the contrast benchmark to compare the contrast model in the conventional scene. The results show that the model proposed in this paper is much better than the conventional computer vision model.

Compared to Unet, which performs well in the two-category segmentation task, the Inception-resnet based architecture replaces the original simple convolution pooled downsampling process in the encoder. In the decoder, the extrusion excitation module is added after each shortening process, which improves the feature aggregation capability of the model. These allow the model to selectively combine high-level information, rather than simply concatenating high-level and low-level information like traditional shortcut modules, which are the reasions why the performance of the model has been improved.

5 Conclusion

This paper proposes a semantic segmentation model based on Inception-Resnet for aerial image building segmentation in order to solve the problem of unbalance between two ends after the shortcut process. This paper also proposes the RCSE module which can re-correct the feature layer, improving the expressive ability of the model decoder. Furthermore, experiments show that, compared with the previous methods, the proposed method significantly improves the quality of building segmentation in remote sensing images.

Acknowledgement. This research was supported by [2018GZ0517] [2019YFS0146] [2019YFS0155] which supported by Sichuan Provincial Science and Technology Department, [2018KF003] Supported by State Key Laboratory of ASIC & System, Science and Technology Planning Project of Guangdong Province [2017B010110007]. Project S201910619048 supported by Sichuan's Training Program of Innovation and Entrepreneurship for Undergraduate.

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**, 2481–2495 (2017)
2. Chollet, F.: Xception: Deep Learning with Depthwise Separable Convolutions. arXiv e-prints, October 2016
3. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. CoRR abs/1709.01507 (2017). http://arxiv.org/abs/1709.01507
4. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
5. Huang, Z., Cheng, G., Wang, H., Li, H., Shi, L., Pan, C.: Building extraction from multi-source remote sensing images via deep deconvolution neural networks. In: 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 1835–1838, July 2016. https://doi.org/10.1109/IGARSS.2016.7729471
6. Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U.: Classification With an Edge: Improving Semantic Image Segmentation with Boundary Detection. arXiv e-prints, December 2016
7. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
8. Shelhamer, E., Long, J., Darrell, T.: Fully convolutional networks for semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(4), 640–651 (2017). https://doi.org/10.1109/TPAMI.2016.2572683
9. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. CoRR abs/1512.00567 (2015). http://arxiv.org/abs/1512.00567
10. Yang, H., Wu, P., Yao, X., Wu, Y., Wang, B., Xu, Y.: Building extraction in very high resolution imagery by dense-attention networks. Remote Sensing **10**(11) (2018). https://doi.org/10.3390/rs10111768. http://www.mdpi.com/2072-4292/10/11/1768
11. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)