



Deep Learning for Building Extraction from High-Resolution Remote Sensing Images

Abderrahim Norelyaqine¹(✉) and Abderrahim Saadane²

¹ Department of Mineral Engineering, Mohammedia School of Engineers, Rabat, Morocco

² Department of Geology, Faculty of Sciences of Rabat, Rabat, Morocco

Abstract. The extraction of buildings from satellite images of very high spatial resolution is an important issue for many applications mainly related to the implementation of public policies, urban development, land use planning, and the updating of geographic databases, and during the last two decades, it has been the subject of much research. Several existing classical techniques have been proposed in remote sensing images, but they have several limitations that prevent segmenting buildings with high accuracy. We propose, in this paper, a U-net architecture with a ResNet50 encoder for the extraction of buildings from Massachusetts building datasets, in order to automate the production chain of urban three-dimensional models. The results obtained in this study show a very promising segmentation with huge accuracy, it outperforms many presented models with 82.63% of intersection over union (IoU).

Keywords: Deep learning · Building extraction · Remote sensing · Very high resolution

1 Introduction

Feature extraction from remote sensing images is a crucial and important topic in the field of remote sensing. How to accurately identify and draw features from remotely sensed images is extremely important for change detection, disaster assessment, land cover detection and other areas. However, due to the complex and diverse features contained in real scenes, the extraction of specific features from remotely sensed images is sensitive to interference from background factors.

In remote sensing images of the urban area, the building area is more than 80%, so the extraction of buildings plays an irreplaceable role in various human activities. In modern society, buildings are also important identification objects in maps and geographic information systems. With the construction of geographic information systems, the technology of automatic building extraction appears and continues to develop. An urban area information system can play an essential role in many areas such as urban development, urban change monitoring, urban planning, population estimation, and topographic map production. The renewal of human observation methods and the rapid growth of urban construction put forward higher requirements for automatic building extraction technology. On the one hand, today's human construction activities are rapidly

changing the information about the city. On the other hand, humans have more abundant means to observe the Earth, and the demand for digital maps is also very high.

With the construction of geographic information systems, the technology of automatic building extraction appears and continues to develop. An urban area information system can play an essential role in many areas such as urban development, urban change monitoring, urban planning, population estimation, and topographic map production. The renewal of human observation methods and the rapid growth of urban construction put forward higher requirements for automatic building extraction technology. On the one hand, today's human construction activities are rapidly changing the information about the city. On the other hand, humans have more abundant means to observe the Earth, and the demand for digital maps is also very high.

The application of high-resolution satellites, aerial drones, radar, and other equipment has led to a massive increase in available data. In this case, how to use massive amounts of data to quickly and thoroughly extract urban information to update. Geographic information systems have become a focal point of research in this area.

Traditionally, the main work of extracting buildings from aeronautical, aerospace, and remote sensing images have focused on empirically designing an appropriate feature to express building and build a set of features to better understand and automatically detect buildings, with indicators that are commonly used such as pixel [1], spectrum [2], the texture [3], and shadow [4]. However, these indicators will change considerably with the season, light, atmospheric conditions, quality of sensors, style of building and environment. Therefore, this feature method can often only process specific data, but cannot be genuinely automated. And using traditional methods to analyze and extract building features from remote sensing imagery has some complexity and does not effectively extract spatial structure.

Artificial intelligence and specifically modern deep learning computer vision technology has made significant progress over the past ten years. It is used in several applications such as video analysis, image classification, image processing for robots and autonomous vehicles, and object recognition. Several computer vision applications need intelligent segmentation to understand features images content and facilitate extraction of each part.

Segmentation of geospatial objects plays an important role in the object extraction task. It can provide location and semantic information for objects of interest. It is part of a special semantic segmentation process. The objective is to separate the pixels in the image into two subsets of background regions and foreground objects. At the same time, it must additionally attribute a unified semantic label to each pixel of the foreground object region.

Segmentation of geospatial objects in HSR images is more difficult compared to natural scenes, for three main reasons:

- In HSR remote sensing images, objects always have large-scale changes, which causes problems at multiple scales, making it difficult to locate and identify objects [5].
- The foreground ratio is much smaller than the natural image, causing the problem of foreground to background imbalance.
- The background of HSR remote sensing image is more complicated, and due to the large intra-class difference, it is easy to cause serious false alarms [6].

The target object segmentation task for natural images is directly considered as a semantic segmentation task in computer vision, and due to the multiscale problem its performances are mainly limited. Consequently, the latest current general semantic segmentation methods focus on multi-scale modeling and scale-sensitive [7].

Today's image segmentation technology uses computer vision deep learning models to analyze each pixel of the image that represents the real object. Deep learning can learn patterns from specific inputs to predict object classes, which was unimaginable. The task of extracting buildings from remote sensing images can be studied as a subset of the semantic segmentation task of images, i.e. developing a segmentation algorithm to separate buildings of the background.

In recent years, deep learning (DL) technology has developed rapidly. Semantic image segmentation based on deep learning methods is also evolving every day, and more and more researchers are trying to solve feature extraction problems through deep learning. Many studies have proven that the use of deep learning methods can significantly improve the accuracy of soil feature extraction. [8] proposed a deep neural network (DNN)-based neural dynamic tracking framework for road network extraction. Study results show that this method is more efficient and accurate than traditional methods.

Since early 2015, deep learning convolution neural networks (CNN) show excellent potential for image classification and target detection, it has been gradually introduced in the fields of remote sensing [9]. The success of CNN is that it can automatically learn a multilayer feature expression that maps the original input to a continuous vector (a regression problem) or a binary label (a classification problem). This self-learning feature-capability gradually outperforms and replaces the traditional manual feature method; in particular, it provides a more automated and robust solution for automatic building extraction on which this paper focuses. A series of general CNN architectures have been gradually developed on this basis, such as VGGNet [10], ResNet [11], etc. Among them, ImageNet is composed of 10 million natural images covering 1,000 categories, which indirectly supports the explosion of deep learning methods.

However, building extraction is a semantic segmentation task at the pixel level. Using CNN will result in a large increase in memory overhead, low computational efficiency, and limited perceptual areas. According to this, Fully Convolutional Network (FCN) [12] removed the fully connected layer in the traditional CNN and deconvoluted the feature map at the end to generate a segmentation result consistent with the resolution of the input image; it has been used in the field of remote sensing target extraction [13]. FCN is specially developed and widely used for semantic segmentation: all pixels in the image are given category labels, including SegNet [14], U-Net [15] and many other variants. Among them, U-net is based on FCN, adopts symmetric structure design, and achieves high extraction accuracy in medical image segmentation. Many neural network architectures have adopted a jump connection structure similar to U-Net and achieved good classification results in practice [16]. Abderrahim et al. [17] mounted the use of the U-net model for road segmentation of Massachusetts roads dataset.

This paper uses the U-net deep learning method to perform automatic intelligent building detection in Massachusetts aerial images.

2 Related Work

Automation of map production based on aerial image analysis was first studied in [18] when they used neural networks to facilitate a manual tracing process. The authors [18, 19] improved it by using another neural network to classify texture patterns. Both networks are fed by features extracted from satellite images, as CNN's are not yet used. Ahmadi et al. [20] propose an active contour model to extract buildings. They test their method on a single image. And to capture the desired features of the image, the model involves dynamic surfaces that move in an image domain. Vakalopoulou et al. [21] designed models trying to extract the outlines of buildings from satellite images. In the first case, they used a CNN to classify the patches as containing buildings or not; then, they performed patch segmentation with an SVM. They extract the contours after segmentation and evaluate their model on Quickbird and Worldview 2 images for Attica, Greece. They do not specify what a true positive is, i.e. how to consider that an extracted contour corresponds to the manually extracted contour (ground truth). This lack of clarification prevents their model from being compared to others. Their model is evaluated on three images, containing a total of only 1438 buildings. Wu Guangming et al. [22] proposed an improved U-net with dual constraints, which optimized the parameter update process that improves the accuracy of building extraction. In [23] authors used an innovative non-local residual U-shaped method (ENRU-Net) of image processing that uses an encoder-decoder structure. It presents a very remarkable improvement when compared with other semantic segmentation methods in the extraction of buildings on the Massachusetts data with an overall accuracy of (94.12%). Cai et al. [24] have proposed a fully convolutional MHA-Net network which consists of multipath HDC, the encoding network and DUC operation. To detect and extract building using high-resolution Aerial images. In [25] authors introduced a lightweight and efficient memory model, RCA-Net, that can accurately capture inter-channel connections using pre-trained layers of ResNet-50 and the ECA module, for extracting building Footprints. They showed satisfactory results on high resolution aerial images (Massachusetts, Inria).

3 Methodology

3.1 Data Preprocessing

Since the original remote sensing image is too large with a resolution of 1500×1500 pixels, the direct input of the original image into the neural network requires a large amount of memory and video memory, so it must be preprocessed as an input layer is 256×256 pixels, to simplify the training process.

3.2 U-net Architecture Overview

In this paper, the U-net architecture [15] was used. In the field of semantic segmentation, this architecture has proven to be a state-of-the-art solution. The input belongs to the encoder/decoder architecture, where the encoder is responsible of extracting features from the image, while the decoder is in charge of reconstructing the segmentation map from the features obtained from the original image. The decoder and encoder can be composed of any number of blocks. A decoder block composed of two convolutional layers and a transposed convolution layer, while an encoder block composed of two convolutional layers and a max pool layer. Number of filters can vary in the convolution layers, but it is usually divided by two. Also, the corresponding blocks of encoders and decoders are connected by skip links, which transmit some information provided by the encoder to the decoder in the image reconstruction phase.

Convolution Layer

Convolution layer allows to drag a matrix over an image, and for each pixel it uses the sum of the multiplication of that pixel by the value of the matrix. This technique allows us to find parts of the image that could be interesting to us. and each convolution kernel has particular parameters that are learned by the algorithm from the training data. In particular, through the gradient backpropagation technique, which allows the adjustment of the parameters according to the gradient value of the loss function (Fig. 1).

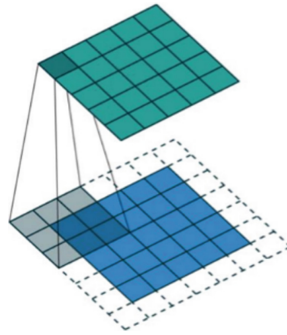


Fig. 1. Convolution with a kernel of size 3×3 and a step of 2

Max Pooling

Similar to the convolution layer, the sampling layer is responsible for reducing the spatial size of the feature maps, but it retains the most important information. There are different types of sampling including Max Pooling, Average Pooling, etc. Sampling consists in applying a kernel of size $n \times n$ on the activation map by dragging it with a previously defined step (the step is usually equal to the size of the kernel n to avoid the overlapping effect). Max Pooling returns the maximum value of the part of the image covered by the kernel, he allows us to remove the noise (Fig. 2).

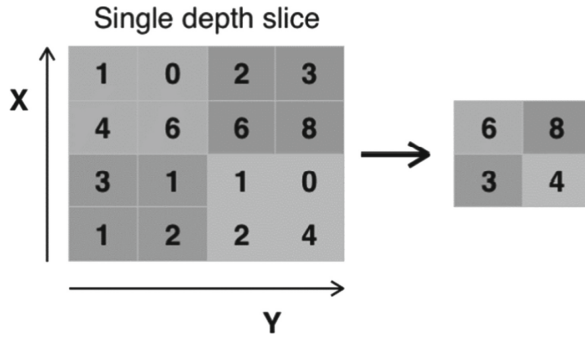


Fig. 2. Pooling operation with size 2 * 2 and a step of 2

The most common form is a pooling layer with 2×2 size kernels applied with a step of 2 thus reducing the input dimension by 2, eliminating 75% of the activations and leaving the depth dimension unchanged. In Fig. 5, the number of activations is reduced from 16 to 4 by applying pooling.

3.3 ResNet

The residual neural network, better known by the name ResNet [6], appears as a result of the problem of backpropagation of the gradient and the increase of the learning error. Indeed, when the neural network is too deep, the gradient reduces to zero and becomes very low to update the weights of the different layers, hence the difficulty of learning.

With ResNet as encoder, the optimization of deep networks is ensured by using residual connections. This allows gradients to pass through two convolution layers, but also to pass directly through a jump to the next layers by connecting the input of the nth layer to the output of the (n + a) th layer (Fig. 3).

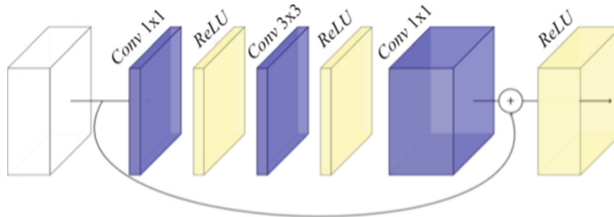


Fig. 3. Residual convolutional block

This residual block can be modeled by the following equation:

$$x_1 = H_1(x_1 - 1) + x_1 - 1$$

With x_1 the output of the residual block and H_1 a compound function which represents all the layers present in the residual block.

3.4 U-net with Resnet Encoder

Since the encoder is part of an architecture that is structurally identical to the architectures used to classify images, it can be replaced by a classification model by removing the last layers used to make the output to the classification model. This results in a model that generates a feature vector, which is connected to the decoder part. The corresponding blocks of this model are connected to the corresponding blocks in the decoder to obtain the U-net architecture. The interest of using other models as encoders in U-net is to use a model that has already been searched on a larger dataset (like imagenet), to get better feature vectors at the output of the encoder. In this case, the model weights can be used without modification, or these weights can be used as a starting point for new training. In this project, we used a model of different classification architectures called Resnet50 (Fig. 4).

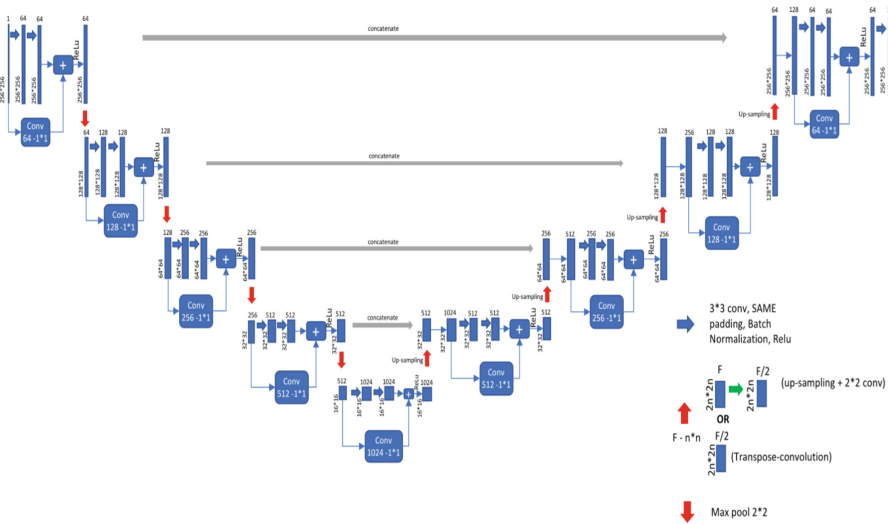


Fig. 4. U-net architecture with Resnet encoder

4 Experiments and Analysis

4.1 Dataset Description

In this paper, we have used aerial imagery dataset of Massachusetts buildings [26] especially the City of Boston. The collection is available online by the University of Toronto, based on OpenStreetMap data. Each image has a pixel size of 1500×1500 . The data covers rural and cities areas, covering an area of over 2600 square kilometres; the dataset was divided into a training set (1108 images), a test set (49 images) and a validation set (14 images). As shown in Fig. 5, each image contains a corresponding binary map on which background are marked in black and the buildings in white.

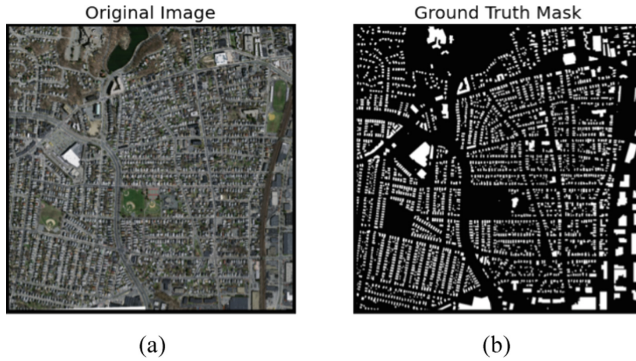


Fig. 5. (a) Original RGB image, (b) Ground Truth Mask of Massachusetts buildings dataset

4.2 Data Augmentation

To train our models, we need huge amounts of data. Indeed, the quantity and especially the quality of the dataset will have a major role in the elaboration of a good model. It is logical to have data that are comparable between them, and that have the same format, the same size and length, etc. And it is from these constraints that the problems begin. Having specific data according to our problem with the above mentioned points can often be impossible. This is where data augmentation will be very useful.

The principle of data augmentation is based on the principle of artificially increasing our data, which is a solution to the problem of limited data by applying transformations. We will be able to increase the diversity and therefore the learning field of our model, which will be able to better adapt to predict new data.

There are a series of simple and effective methods for data enhancement among which we have chosen Horizontal Flip, Vertical Flip and Random Rotate as shown in Fig. 6.

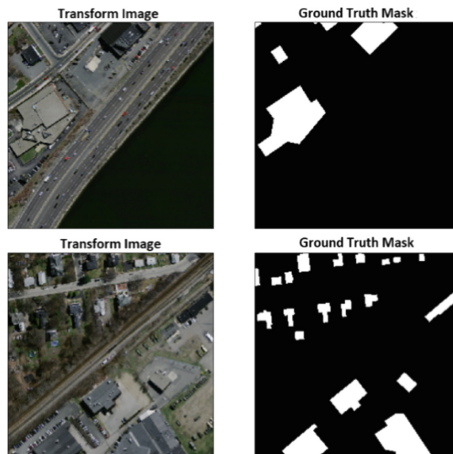


Fig. 6. Example of data augmentation

4.3 Evaluation Metrics

To evolve our model we used the IoU score - or Intersection over Union - is a way to measure the quality of object detection by comparing, in a training dataset, the known position of the detected object in the image with the prediction made by the model. The IoU is the ratio between the area of the union of the sets and the area of the intersection of the considered bounding boxes (Fig. 7).


$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$


Fig. 7. Computing intersection over union

From this measurement, an acceptable tolerance threshold will be defined to determine if the prediction is correct or not. It is therefore the IoU that will determine if a prediction is:

- True positive: The object is correctly detected.
- False positive: The object is detected when it is not present in the ground truth.
- False negative: the model does not detect an object when one is present in the ground truth (Fig. 8).

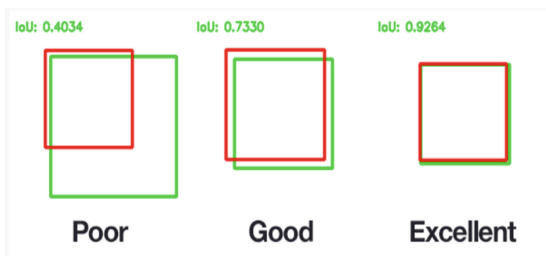


Fig. 8. Eexample of computing Intersection over Unions

4.4 Comparison and Analysis

Figure 9 represents a representative graph when training at 60 epochs the U-net model on the Massachusetts remote sensing images, the IoU score on the training set is about

82.63%, obtaining good results. And it represents the loss function curve, and the model trains at about 60 epochs and tends to be stable and close to 0.09 (Fig. 10).

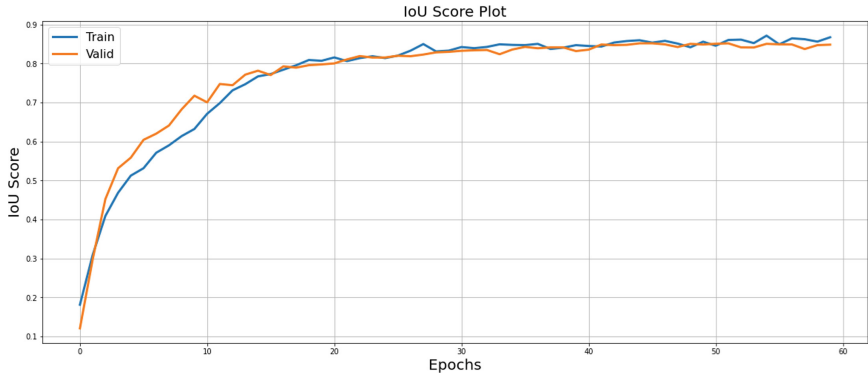


Fig. 9. IoU score plot

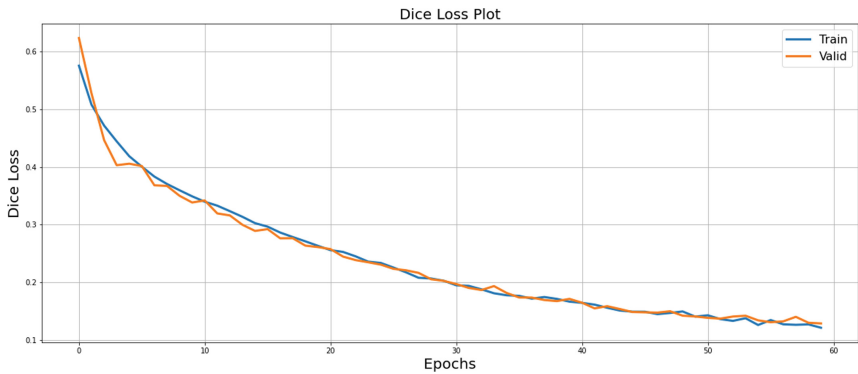


Fig. 10. Dice-coefficient loss function plot

Table 1. Comparison of our results (%) with four other methods in testing data.

Method	IoU
SegNet [14]	67.98
U-Net [15]	70.99
ENRU-Net [23]	73.02
MHA-Net [24]	74.46
Ours	82.63

Almost all the most advanced building extraction methods are based on FCN structures. Table 1 shows that our model proposed in this paper on the testing data for the

Massachusetts dataset obtained its best results, better than the other basic models Segnet, U-net and 11% better than the last results obtained in 2021 [24]. 11% can be considered a significant improvement.

At the same time, these results also reflect that compared to traditional methods (which are often difficult to exceed 50% accuracy), deep learning-based methods have pushed building extraction to a new level of automation. And as shown in Fig. 11 our model fully reflects the IoU measure obtained in the test, it can better distinguish buildings and backgrounds with high accuracy.

We can also notice that in some places the buildings are not well segmented. Namely, in the dataset, there are images in which the buildings are not marked in output as. All this affects the fact that the model can not converge in the learning process to a given objective, which translates into such results in the validation and learning set. It can also

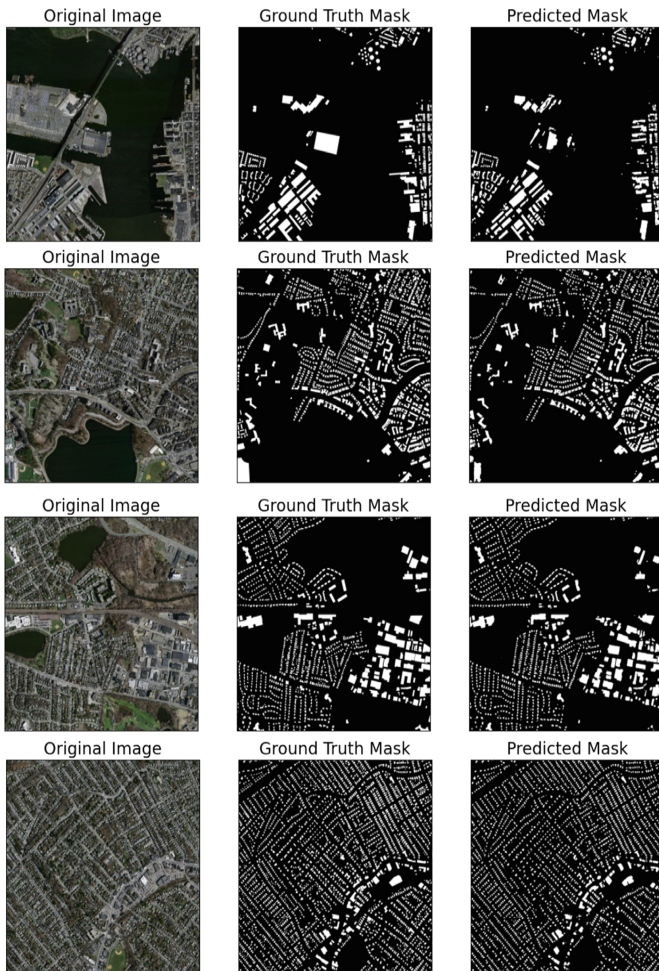


Fig. 11. Some samples of predictions in massachusetts building dataset.

be noticed that the resolution so high for the original images is not sufficient to make some buildings clearly visible.

5 Conclusion

U-net with a ResNet50 encoder is applied to building extraction from remote sensing images to accurately extract buildings in the target area. Moreover, when buildings are extracted in urban areas in a complex environment, environmental information and building information are easily confused, making the extraction poor. In response to the above problems, an improved U-net neural network is proposed, improving the detailed information in the network transmission process and improving the model's ability to obtain details in the 2600 km² coverage area. After training and testing on Massachusetts remote sensing image datasets and comparing with other methods, the results show that the average values of IoU coefficients of the U-net model proposed in this paper will reach 82.60, respectively, which is better than the other models compared.

The solar shading and the difference in the characteristics of the building itself will impact the integrity of the building extraction. In future work, it will be necessary to study the shadow and more and more building characteristics in the image to improve the building extraction effect.

References

1. Sirmacek, B., Unsalan, C.: Building detection from aerial images using invariant color features and shadow information. In: 2008 23rd International Symposium on Computer and Information Sciences, pp. 1–5. IEEE (2008)
2. Zhong, S.-H., Huang, J.-J., Xie, W.-X.: A new method of building detection from a single aerial photograph. In: 2008 9th International Conference on Signal Processing, pp. 1219–1222. IEEE (2008)
3. Zhang, Y.: Optimisation of building detection in satellite images by combining multispectral classification and texture filtering. *ISPRS J. Photogr. Remote Sens.* **54**(1), 50–60 (1999)
4. Chen, D., Shang, S., Wu, C.: Shadow-based building detection and segmentation in high-resolution remote sensing image. *J. Multim.* **9**(1), 181–188 (2014)
5. Deng, Z., Sun, H., Zhou, S., Zhao, J., Lei, L., Zou, H.: Multi-scale object detection in remote sensing imagery with convolutional neural networks. *ISPRS J. Photogr. Remote Sens.* **145**, 3–22 (2018)
6. Deng, Z., Sun, H., Zhou, S., Zhao, J.: Learning deep ship detector in SAR images from scratch. *IEEE Trans. Geosci. Remote Sens.* **57**(6), 4021–4039 (2019)
7. Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 801–818 (2018)
8. Wang, J., Song, J., Chen, M., Yang, Z.: Road network extraction: a neural-dynamic framework based on deep learning and a finite state machine. *Int. J. Remote Sens.* **36**(12), 3144–3169 (2015)
9. Guo, J., Pan, Z., Lei, B., Ding, C.: Automatic color correction for multisource remote sensing images with Wasserstein CNN. *Remote Sens.* **9**(5), 483 (2017)
10. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014)

11. 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, pp. 770–778 (2016)
12. 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)
13. Qingsong, S., Chao, Z., Yu, C., Xingli, W., Xiaojun, Y.: Road segmentation using full convolutional neural networks with conditional random fields. *J. Tsinghua Univ.* **58**(8), 725–731 (2018)
14. Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(12), 2481–2495 (2017)
15. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W., Frangi, A. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
16. Rakhlin, A., Davydov, A., Nikolenko, S.: Land cover classification from satellite imagery with U-net and lovász-softmax loss. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 262–266 (2018)
17. Abderrahim, N.Y.Q., Abderrahim, S., Rida, A.: Road segmentation using u-net architecture. In: 2020 IEEE International conference of Moroccan Geomatics (Morgeo), pp. 1–4. IEEE (2020)
18. Hunt, B.R., Ryan, T.W., Sementilli, P.J., DeKruger, D.: Interactive tools for assisting the extraction of cartographic features. In: Image Understanding and the Man-Machine Interface III, pp. 208–218. International Society for Optics and Photonics (1991)
19. DeKruger, D., Hunt, B.R.: Image processing and neural networks for recognition of cartographic area features. *Pattern Recogn.* **27**(4), 461–483 (1994)
20. Ahmadi, S., Zojj, M.V., Ebadi, H., Moghaddam, H.A., Mohammadzadeh, A.: automatic urban building boundary extraction from high resolution aerial images using an innovative model of active contours. *Int. J. Appl. Earth Observ. Geoinf.* **12**(3), 150–157 (2010)
21. Vakalopoulou, M., Karantzalos, K., Komodakis, N., Paragios, N.: Building detection in very high resolution multispectral data with deep learning features. In: 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 1873–1876. IEEE (2015)
22. Guangming, W., Qi, C., Shibasaki, R., Zhiling, G., Xiaowei, S., Yongwei, X.: High precision building detection from aerial imagery using a U-Net like convolutional architecture. *Acta Geodaetica Cartogr. Sinica* **47**(6), 864 (2018)
23. Wang, S., Hou, X., Zhao, X.: Automatic building extraction from high-resolution aerial imagery via fully convolutional encoder-decoder network with non-local block. *IEEE Access* **8**, 7313–7322 (2020)
24. Cai, J., Chen, Y.: MHA-net: multipath hybrid attention network for building footprint extraction from high-resolution remote sensing imagery. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* (2021)
25. Das, P., Chand, S.: Extracting building footprints from high-resolution aerial imagery using refined cross AttentionNet. *IETE Tech. Rev.* 1–12 (2021)
26. Mnih, V.: Machine learning for aerial image labeling. University of Toronto (Canada) (2013)