A SHORT-CUT CONNECTIONS-BASED NEURAL NETWORK FOR BUILDING EXTRACTION FROM HIGH RESOLUTION ORTHOIMAGERY

Zhimeng He[1]*, Hongjie He[2], Jonathan Li[2], Michael A. Chapman[3], Haiyong Ding[1]

[1] Nanjing University of Information Science and Technology, Nanjing, 210044, China
[2] Department of Geography and Environmental Management, University of Waterloo, Waterloo, ON N2L 3G1, Canada
[3] Department of Civil Engineering, Geomatics Engineering, Ryerson University, Toronto, NO M5B 2K3, Canada

Commission I, WG I/2

KEY WORDS: Deep Learning; Building Extraction; Dilated Convolution; Short-cut Connections; High Resolution Orthoimagery

ABSTRACT:

Extracting building footprints utilizing deep learning-based (DL-based) methods for high-resolution remote sensing images is one of the current research interest areas. However, the extraction results suffer from blurred edges, rounded corners and detail loss in general. Hence, this article presents a detail-oriented deep learning network named eU-Net (enhanced U-Net). The method adopted in this study, imagery send into the pre-module, which consists of the Canny edge detector, Principal Component Analysis (PCA) and the inter-band ratio operations, before feeding them into the network. Then, process skips connections used in the network to reduce the loss of details during edge and corner detection. The encoding and decoding modules, in this network, are redesigned to expand the perceptual field with shortcut connections and stacked layers. Finally, a Dropout module is added in the bottom layer of the network to avoid the over-fitting problem. The experimental results indicate that the methods used in this study outperform other commonly used and state-of-the-art methods of FCN-8s, U-net, DeepLabv3 and Fast SCNN.

1. INTRODUCTION

Building maps are an important task for urban planning, disaster monitoring, traffic management and scientific planning of the ecological environment (Wei and Liu, 2020). While traditional in-situ mapping and surveying can generate high accuracy maps, they tend to be inefficient. With the development of sensor technologies, remote sensing-based object extraction provides an efficient way to map buildings. Considering all remote sensing-based methods for building footprint extraction, deep learning-based methods (Xin et al., 2012) show a higher performance compared with visual interpretation and traditional machine learning based methods. For instance, He et al. (2020) utilized Mask R-CNN with the attention mechanism for building footprint extraction. The extraction results revealed that deep learning-based methods were better than traditional machine learning methods, such as Support Vector Machines (SVM) (Cortes and Vapnik, 1995). Therefore, the focus of research for extracting building footprints turned to deep learning-based methods.

With the revival of deep learning techniques, various commonly used networks have been widely used in remote sensing applications and building footprint extraction. In 2015， Ronneberger. O (2015) proposed U-Net, which used an Encoder-Decoder structure. This structure was effective and many researchers chose the structure as a module combined with their existing works. Wang et al. (2020) proposed a network named Efficient Non-local Residual U-shape Network (ENRU-Net), which improved non-local block with U-net structure to capture contextual information. This network also achieved better result compared to Fully Convolution Networks (FCN)-8s (Wu, 2015), U-Net, SegNet (Badrinarayanan et al., 2017) and DeepLab v3(Chen et al., 2017). Xu et al. (2021) improved U-net combining with attention mechanism, and developed a loss function. They achieved higher F1-score of more than 10.78% compare to U-net. Liu et al. (2021) demonstrated the effectiveness of pyramid scene parsing and a residual connection in enabling U-net to catch global context information by the large-scale experimental. The DeepLab v3 (Chen et al., 2017) architecture proposed by Chen et al. (2017) combined a Atrous Spatial Pyramid Pooling (ASPP) and an Encoder-Decoder, which achieved higher performance than that of ResNet-38 (He et al., 2016), and PSPNet (Zhao, 2016). In addition, Yang et al. (2020) utilized the DeepLabv3 architecture and the dilated convolution to extract building footprints, which proved that combining the DeepLabv3 and dilated convolution is more effective than using DeepLabv3 only. Zhang et al. (2020) used deep separable convolution to improve U-net, which proved the efficiency of XU-Net. Zhou et al. (2020) introduced encoder and decoder sub-networks and connected them via a series of nested, dense skip connection, which showed high performance as shown in their experiment. This network solved the disadvantage of the skip - connection, and used deep supervision to help network choose path by itself. Other modifications includes combining U-Net and Mask R-CNN as an ensemble model(Vuola et al., 2019) and involving a pre-trained VGG-encoder in U-Net (Shvets, 2018) developed r to improve the U-Net.

However, all methods mentioned above cannot fully overcome detailed information loss during edge and corner detection, which are important for building footprint extraction in terms of extraction accuracy. To address the issue, Salient object

*Corresponding author: Zhimeng He, 20191235004@nuist.edu.cn

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLIII-B1-2022-39-2022 | © Author(s) 2022. CC BY 4.0 License.

detection is proposed which focuses on edge preserving. At present, binary label is more fashion in works compared with class activation map, and more benefit for segmentation application. There are mainstream researches like center-surround difference (L. et al., 1998) and (T. et al., 2007), uniqueness prior (K et al., 2013) and (P et al., 2013) and backgrounds prior (Y. et al., 2012) and (W. et al., 2014). However, they haven't achieved considering context information and clear edge, therefore the accuracies are expected to be improved continued. Wang et al. (2016) proposed a method based on Fast R-CNN framework (Girshick, 2015) to generate a saliency map both clear and context information. The imagery is segmented into super-pixel regions and edge regions to achieve state-of-the-art performance.

This paper presents a new method called enhanced U-Net (eU-Net), which is based on U-Net. For this method, the dilated convolution (Yu and Koltun, 2015), the dropout module (Hinton et al., 2013), the jump connection (Ronneberger. O, 2015), the short-cut connection (He et al., 2016) and a pre-module were applied to preserve detailed information of building footprints. The rest of the paper is organized as follows. Section 2 describes the pre-module and the eU-Net architecture. Section 3 presents the results of two experiments and demonstrates the performance of eU-Net. Section 4 concludes the paper.

2. METHOD

2.1 The pre-modules

To preserve more detail information, a pre-module was inserted before passing the images to the deep learning network. The pre-module converted the original images to 6-band tensors via Principal Component Analysis (PCA) (Ke and Sukthankar, 2004) (only the first component was kept), Canny edge detection operator (Canny and John, 1986) and the Red Green Index. The Canny edge detection was an edge detection operator that uses a multi-stage algorithm to detect a wide range of edges in images. The Red Green Index is like the Normalized Difference Vegetation Index (NDVI), which is an index to detect live green plant canopies in multispectral remote sensing data.

The pre-modules were used for (1) emphasizing edges and (2) enhancing inter-band links. Deep neural networks can be seen as multi-layer matrix multiplication, which can only find multiplicative relationships between bands. In this study, the red and green bands were chosen to calculate Red Green Index because the buildings in the image were more obvious (Figure 1).

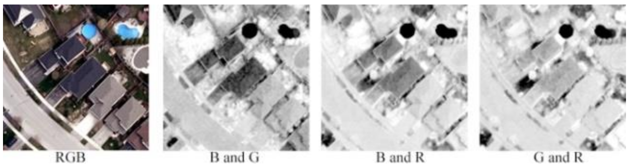

Figure 1. Ratio results of different bands

The inter-band ratio, Red Green Index, can be calculated by the formula 1.

$$g_{rg}(i,j) = 0.5 + 0.5 * (\frac{b_g(i,j)-b_r(i,j)}{b_g(i,j)+b_r(i,j)}) \qquad (1)$$

where (i,j) represented the coordinates of the pixel points on the image, and $\mathrm{g}_{\mathrm{rg}}(\mathrm{i},\mathrm{j})$ was the pixel value at point (i,j) on the image. The $\mathrm{b}_{\mathrm{g}}(\mathrm{i},\mathrm{j})$ was the pixel value at point (i, j) on the image of green band. The $\mathrm{b}_{\mathrm{r}}(\mathrm{i},\mathrm{j})$ was the pixel value at point (i, j) on the image of red band.

2.2 eU-Net network architecture

U-Net was an effective network because of its jump connections, symmetric structure, and multi-scale feature extraction capability. The jump connection was useful to prevent the gradient explosion and the Vanishing Gradients; the symmetrical structure of the encoding and decoding not only eased the jump connection but it also facilitated end-to-end classification (Hao et al., 2020); the multi-scale feature extraction acquired both detailed information and high-level semantic information which made the network acquire more information from the images. However, the results of U-Net were commonly inaccurate with respect to building boundary detection.

Figure 3 shows the architecture of eU-Net, the Dropout module was employed between the encoder and decoder to avoid over-fitting (Hinton et al., 2013). Dropout not only improved the robustness of the network, but also reduced the training time to 1-p of the original time (p is the probability of stopping units.). The Dropout strategy was frozen at the testing phase, which ensured a stable performance.

Dilated convolution was used to increase the field size of the module. The increase in the size of the kernel was determined by the dilation rate. In this study, dilated convolution was applied after the pre-module. Feature maps, after the dilated convolutions with the dilation rates of 2, 3 and 5, respectively, were connected to three encoding modules (except the first module) via a jump connection.
The encoding and decoding units were aligned with the Y-residual unit (Figure 2) of the short-cut connections to increase the robustness of the model. The receptive field with different sizes was helpful for analyzing the targets with different size, while concatenating the receptive field by increasing the size of the kernel will cause an increase in the number of parameters. Therefore, in the eU-Net, two 3x3 kernels are applied instead of one 5x5 kernel which resulted in the same receptive field but with fewer parameters (Szegedy et al., 2015). The concatenate algorithm is applied to merge the output from the 1x1 and the 3x3 convolutions layers.

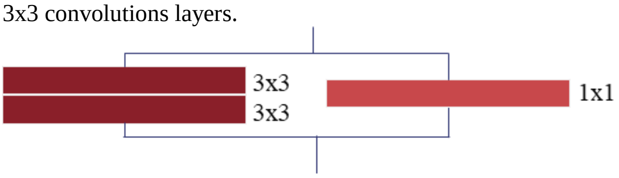

Figure 2. The Y-residual unit

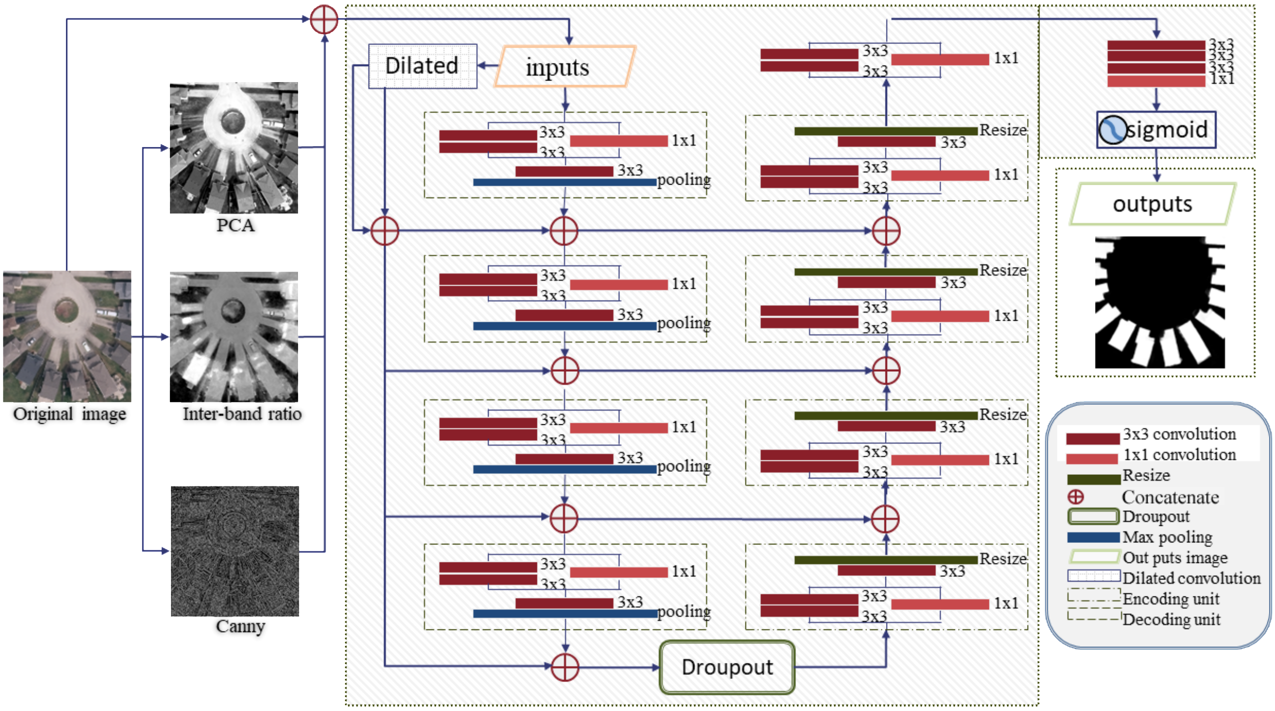

Figure 3. Architecture of eU-Net

2.3 Evaluation metrics

In this study, the metrics include Overall Accuracy (OA), Recall, Precision, F1-score, and Intersection over Union (IoU) were applied to evaluate the performance of building footprint extraction models. All metrics were calculated as follows:

$$\text{Overall} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}} \tag{2}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

$$\text{F1} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

$$\text{IoU} = \frac{\text{area(C)} \cap \text{area(G)}}{\text{area(C)} \cup \text{area(G)}} \tag{6}$$

Where the True Positive (TP) was the number of correctly predicted pixels indicating the positive classes; the False Positive (FP) was the number of predicted pixels wrongly indicating the negative classes; the False Negative (FN) was the number of predicted pixels wrongly indicating the positive classes; the True Negative (TN) was the number of predicted pixels correctly indicating the negative classes.

2.4 Implementation Details

In this experiment, a Rectified Linear Unit (ReLU) was selected as the activation function except for the output layer, which suppressed negative values and prevented the resulting gradient exploring and vanishing. Batch size, loss function, learning rate and epochs have been set to 4, binary cross-entropy loss, 1e-4 and 100 respectively, which were same as the experimental setting of He et al. (2021). All algorithms were trained and tested on a GeForce RTX 2060 with CUDA 10.2. The experiments were performed using the Keras library with the TensorFlow framework.

3. RESULTS AND DISCUSSION

3.1 Experiment Data

In this study, the eU-net model and other models were tested on the Waterloo Building Dataset, which was released by He et al. (2021) on Harvard Dataverse. The Waterloo Building Dataset (Figure 4) provided very-high-spatial-resolution aerial ortho-imagery consisting of red, green and blue bands. It covers the Kitchener-Waterloo area in Ontario, Canada, contains 117,000 manually labelled buildings, and extends over an area of 205.8 km^2. At a spatial resolution of 12 cm, it is the highest resolution publicly available building footprint extraction dataset in North America. This dataset covered variety types of building and other object like roads, trees, bare land et al. This dataset contains 69, 792 patches, including 42, 147, 20, 768 and 6, 877 pairs of images and masks for training, validation and testing.

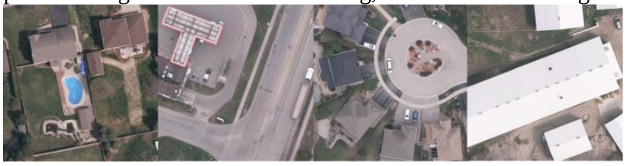

Figure 4. The imagery diagram

3.2 Ablation study

In this section, the contribution of the pre-modules to the performance of the method using the Waterloo Building Dataset (He et al., 2021) supports, the mosaic results of eU-Net shown in Figure 6. The following five experiments were conducted: (1) an experiment without the pre-module (baseline), (2) an experiment without the PCA, but with the edge detection and inter-band ratio operation, (3) an experiment without custom ratios, but with the PCA and inter-band ratio operations, (4) an experiment without edge detection, but with the inter-band ratio operations and PCA, and (5) an experiment with the pre-module. Table 2 presents the results and Figure 5 shows the training accuracy and loss.

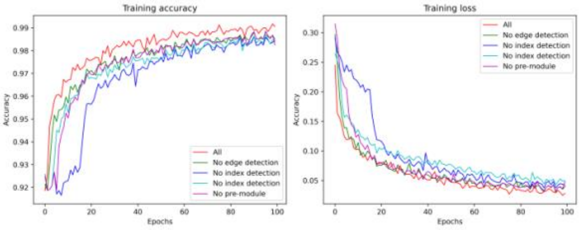

Figure 5. Training accuracy and loss

	OA (%)	**IoU (%)**	**mIoU (%)**	**Precision (%)**	**Recall (%)**	**F_1 -score (%)**	**FPS (s)**
No pre-modules (1)	97.3	80.3	87.9	90.6	87.6	89.1	12.3
No Canny (2)	97.1	85.3	90.9	**91.6**	92.5	92.0	9.5
No PCA (3)	97.3	85.6	91.2	91.4	93.1	92.2	7.8
No Inter-band (4)	96.6	82.1	89.0	85.5	95.3	90.2	8.7
eU-Net (5)	**97.8**	**87.9**	**92.6**	87.5	**93.1**	**93.5**	8.5

Table 2. Results of ablation experiments

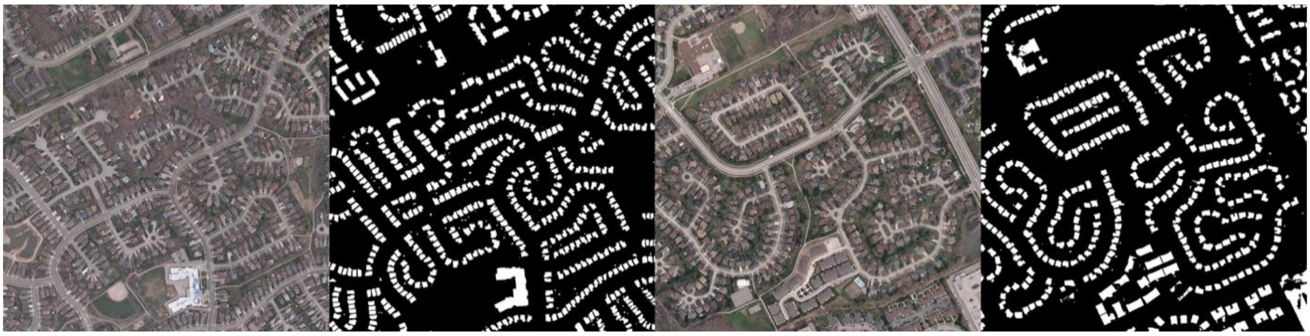

Figure 6. Examples of eU-Net study mosaic results

As shown ablation study results in Table 2 and Figure 7, the pre-module yielded a significant improvement in all the metrics compared to its baseline. The F_1-score increased from 89.1% to 93.5 %, which indicated that the pre-module can effectively promote deep learning networks to improve their accuracy. Apart from the baseline, the lowest F_1-score of 90.2% occur in Experiment (4), which removed the inter-band ratio compared to Experiment (5). This removal resulted in the decrease of F_1-score by 3.4%, which proved that the inter-band module is the most effectively module of the three modules for preprocessing. Experiment (3) showed the PCA components have a negligible contribution to the performance improvement in terms of F_1-score, which decreased by 1% when PCA components were removed from the pre-modules.

3.3 Performance of eU-Net and Comparative Study

As shown in Table 3, a comparative study was conducted between the eU-Net and the methods used in He et al. (2021) using the Waterloo Building Dataset. For a fair comparison, we used the same experiment settings as those in He et al. (2021). Specifically, the batch size, loss function, learning rate and epochs were set to 4, binary cross-entropy loss, 1e^{-4} and 100, respectively. eU-Net achieved higher values in terms of IoU, mIoU, Recall and F_1-score produced results like all other methods. The result demonstrated that eU-Net performs well compared to state-of-the-art DL-based methods in building footprint extraction. There are the number of parameters for eU-Net (ours) and Mask R-CNN (He et al., 2021), which are 61,422,472 and 64,266,590.

	OA (%)	IoU (%)	mIoU (%)	Precision (%)	Recall (%)	F1-score (%)	FPS (s)
FCN-8s (Wu, 2015)	77.1	25.0	50.1	26.1	85.5	40.0	19.6
U-Net (Ronneberger, 2015)	86.7	37.3	61.4	39.2	88.4	54.3	14.9
DeepLab v3+ (Chen et al., 2017)	97.3	72.7	84.9	88.6	80.3	84.2	17.6
Fast SCNN (Poudel et al., 2019)	77.3	23.0	49.3	24.8	76.1	37.4	**24.0**
HRNet v2 (Sun et al., 2019)	**97.8**	76.6	87.1	**92.5**	81.7	86.8	18.2
Mask R-CNN (ESRI)	96.6	64.6	80.5	89.2	70.1	78.5	-
Mask R-CNN (He et al., 2021)	95.3	59.4	77.2	71.7	77.5	74.5	-
eU-Net (ours)	97.8	**87.9**	**92.6**	87.5	**93.1**	**93.5**	8.5

Table 3. Results of comparison experiments

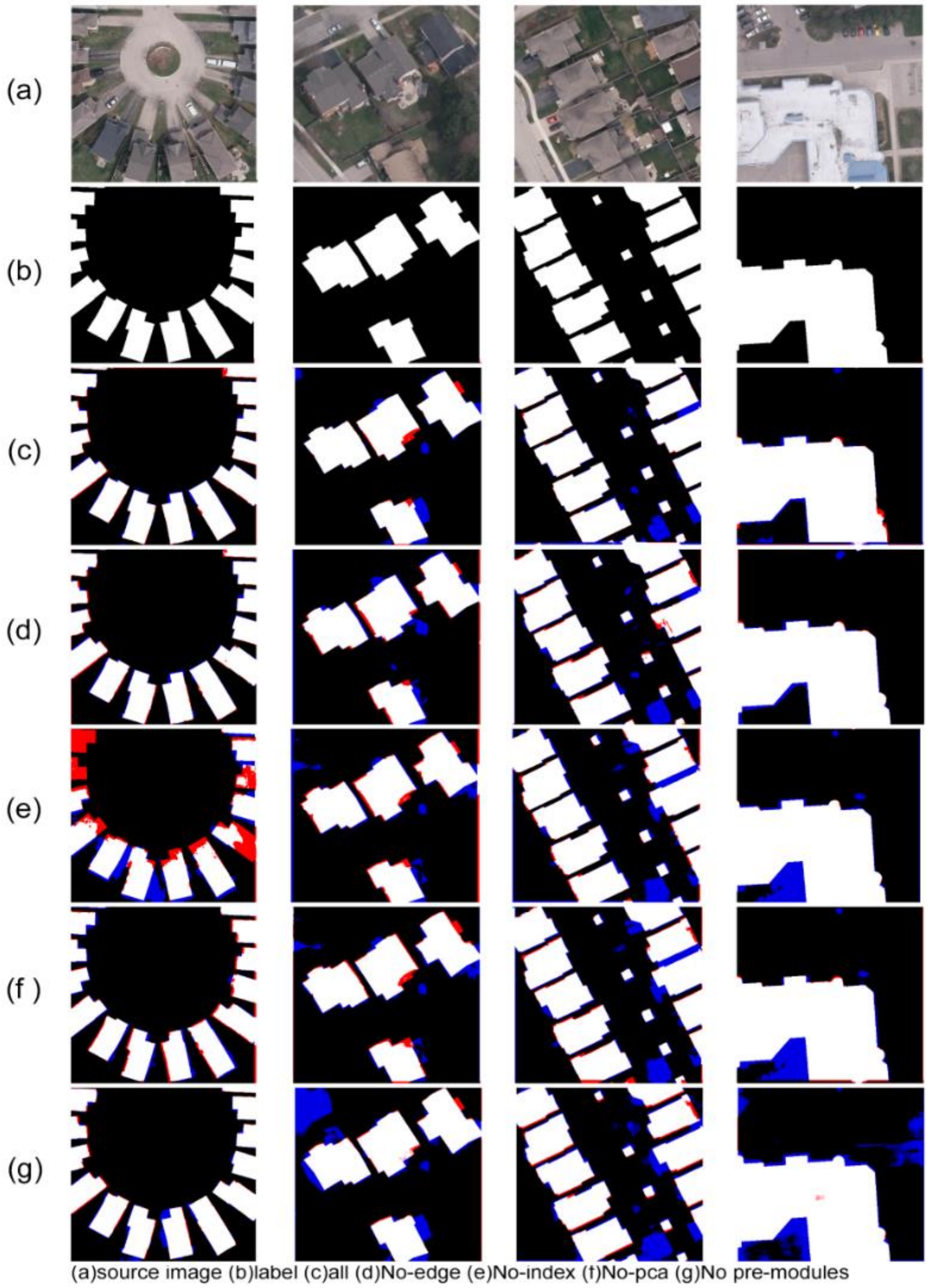
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(a)source image (b)label (c)all (d)No-edge (e)No-index (f)No-pca (g)No pre-modules

Figure 7. Examples of ablation study results

4. CRITICAL ASSESSMENT

The manual search for inter-band ratio relationships and edge enhancement generally has limited effectiveness. When changing targets, specific pre-processing needs to be adapted. Edge enhancement by manually selected features will significantly improve extraction accuracy, but in forest extraction it tends to cause a pepper noise. In future research, the pre-processing module of this paper is replaced by a deep learning network to do feature enhancement. Combining two deep learning networks, ensuring them carry special functions that one for feature enhancement as pre-processing and another one for extraction. Deep learning networks should be modularization, and then become more controllable and make the functions of each module recognizable.

5. CONCLUSION

The primary goal of this research was to extract the building footprints with precise edges. For this purpose, a new deep learning network named eU-Net was developed, which consisted of two components: the pre-modules and the deep learning neural network. The pre-modules were constructed using PCA, edge detection and the inter-band ratio results. The processed 6-band image set was transferred to the designed FCN. And the FCN employed dilated convolutions, jump connections, Y-residual units and U-type architecture. The comparative study demonstrated that method used in this study exhibited high performance in building footprint extraction compared to other commonly used and state-of-the-art methods

REFERENCES

Badrinarayanan, V., Kendall, A., Cipolla, R.J.I.T.o.P.A., Intelligence, M., 2017. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. 1-1.

Canny, John, 1986. A computational approach for edge detection. PAMI-8, 679-698.

Chen, L.C., Papandreou, G., Schroff, F., Adam, H., 2017. Rethinking Atrous Convolution for Semantic Image Segmentation. arXiv.1706.05587.

Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.J.S., Cham, 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. DOI: 10.1007/978-3-030-01234-2_49.

Cortes, C., Vapnik, V.J.M.L., 1995. Support-Vector Networks. 20, 273-297.

Girshick, R., 2015. Fast R-CNN. IEEE International Conference on Computer Vision (ICCV), pp. 1440-1448.

He, K., Gkioxari, G., Dollár, P., Girshick, R.J.I.T.o.P.A., Intelligence, M., 2017. Mask R-CNN. 2017 IEEE International Conference on Computer Vision (ICCV). DOI: 10.1109/TPAMI.2018.2844175.

He, K., Zhang, X., Ren, S., Sun, J.J.I., 2016. Deep Residual Learning for Image Recognition. DOI: 10.1109/CVPR.2016.90

Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R., 2013. Improving neural networks by preventing co-adaptation of feature detectors.

Iglovikov, V., Shvets, A., 2018. TernausNet: U-Net with VGG11 Encoder Pre-Trained on ImageNet for Image Segmentation. DOI: 10.48550/arXiv.1801.05746

K. Shi, K.W., J. Lu, and L. Lin, 2013. Pisa: pixelwise image saliency by aggregating complementary appearance contrast measures with spatial priors in CVPR. DOI: 10.1109/CVPR.2013.275

Ke, N.Y., Sukthankar, R., 2004. PCA-SIFT: a more distinctive representation for local image descriptors, IEEE Computer Society Conference on Computer Vision & Pattern Recognition. DOI：10.1109/CVPR.2004.1315206

L. Itti, C.K., and E. Niebur, 1998. A model of saliency-based visual attention for rapid scene analysis. IEEE TPAMI. DOI：10.1109/34.730558

Liu, W., Xu, J., Guo, Z., Li, E., Li, X., Zhang, L., Liu, W., 2021. Building Footprint Extraction from Unmanned Aerial Vehicle Images Via PRU-Net: Application to Change Detection. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 14, 2236-2248.

P. Jiang, H.L., J. Yu, and J. Peng, 2013. Salient region detection by ufo: Uniqueness, focusness and objectness in ICCV. DOI: 10.1109/ICCV.2013.248

Poudel, R., Liwicki, S., Cipolla, R., 2019. Fast-SCNN: Fast Semantic Segmentation Network. arXiv:1902.04502.

Ronneberger. O, F.P., Brox. T, 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. Springer International Publishing. DOI: 10.1007/978-3-319-24574-4_28.

Sun, K., Zhao, Y., Jiang, B., Cheng, T., Wang, J., 2019. High-Resolution Representations for Labeling Pixels and Regions. DOI: 10.48550/arXiv.1904.04514

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2015. Rethinking the Inception Architecture for Computer Vision. DOI: 10.48550/arXiv.1512.00567

T. Liu, J.S., N.-N. Zheng, X. Tang, and H.-Y. Shum, 2007. Learning to detect a salient object. CVPR. DOI: 10.1109/TPAMI.2010.70

Vuola, A.O., Akram, S.U., Kannala, J.J.I., 2019. Mask-RCNN and U-net Ensembled for Nuclei Segmentation. DOI: 10.1109/ISBI.2019.8759574

W. Zhu, S.L., Y. Wei, and J. Sun, 2014. Saliency optimization from robust background detection in CVPR. DOI: 10.1109/CVPR.2014.360.

Wang, S., Hou, X., Zhao, X., 2020. Automatic Building Extraction from High-Resolution Aerial Imagery via Fully Convolutional Encoder-Decoder Network with Non-Local Block. IEEE Access 8, 7313-7322.

Wang, X., Ma, H., Chen, X., You, S.J.I.T.I.P., 2016. Edge Preserving and Multi-Scale Contextual Neural Network for Salient Object Detection. PP, 1-1.

Wei, Y., Liu, X., 2020. The Application of Deep Convolution Neural Network to Building Extraction in Remote Sensing Images. World Scientific Research Journal 6, 136-144.

Wu, X., 2015. Fully Convolutional Networks for Semantic Segmentation. Computer Science. DOI: 10.1109/CVPR.2015.7298965

Xu, L., Liu, Y., Yang, P., Chen, H., Zhang, H., Wang, D., Zhang, X., 2021. HA U-Net: Improved Model for Building Extraction from High Resolution Remote Sensing Imagery. Ieee Access 9, 101972-101984.

Y. Wei, F.W., W. Zhu, and J. Sun, 2012. Geodesic saliency using background priors. in ECCV. DOI: 10.1007/978-3-642-33712-3_3.

Yu, F., Koltun, V., 2015. Multi-Scale Context Aggregation by Dilated Convolutions. arXiv:1511.07122v3.

Zhang, X., Zheng, Y., Liu, W., Peng, Y., Wang, Z., 2020. An improved architecture for urban building extraction based on depthwise separable convolution. Journal of Intelligent & Fuzzy Systems 38, 5821-5829.

Zhao, H.S., J.， Qi, X.， Wang, X.， Jia, J., 2016. Pyramid Scene Parsing Network. IEEE Computer Society. DOI: 10.1109/CVPR.2017.660.

Zhou, Z., Siddiquee, M., Tajbakhsh, N., Liang, J.J.I.T.o.M.I., 2020. UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation. 39, 1856-1867.