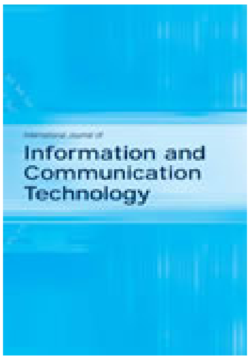
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



Information and Communication Technology

International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Improved DeepLabv3+ connected augmented reality technology for building target extraction in urban environmental design

Jie Chen, Qian Wu

DOI: 10.1504/IJICT.2024.10062708

Article History:

Received:	15 April 2023
Last revised:	07 May 2023
Accepted:	15 November 2023
Published online:	03 May 2024

Improved DeepLabv3+ connected augmented reality technology for building target extraction in urban environmental design

Jie Chen and Qian Wu*

Sichuan Technology and Business University,
Chengdu 610000, Sichuan, China
Email: 13880218789@163.com
Email: 13693407018@163.com
*Corresponding author

Abstract: Aiming at the problem of inaccurate segmentation of building edges in remote sensing images, the imprecise segmentation of building edges in remote sensing images by deep learning models is an important research direction for remote sensing intelligence applications. This paper proposes a lightweight remote sensing image building extraction method based on DeepLabv3. The skeleton network uses DeepLabv3 to connect the IEU-Net structure. Secondly, in order to solve the problem of limited feature richness of the model, the morphological construction index MBI is introduced to participate in the classification process of the model together with the RGB band of the remote sensing image. Finally, in the model prediction, corresponding to IELoss, a strategy of ignoring edge prediction is adopted to obtain the best building extraction results. Our proposed method can effectively overcome the problem of insufficient edge pixel features of samples, suppress the influence of road and building shadows on the results, and improve the extraction accuracy of houses and buildings in remote sensing images.

Keywords: building extraction; boundary perception; DeepLabv3+.

Reference to this paper should be made as follows: Chen, J. and Wu, Q. (2024) ‘Improved DeepLabv3+ connected augmented reality technology for building target extraction in urban environmental design’, *Int. J. Information and Communication Technology*, Vol. 24, No. 5, pp.54–73.

Biographical notes: Jie Chen graduated from Southwest Minzu University with a Master’s degree. Currently teaches at the School of Art, Sichuan Technology and Business University, and mainly researching environmental design, smart cities, low-carbon cities, and other fields.

Qian Wu graduated from Southwest Minzu University with a Master’s degree. Currently teaches at the School of Art, Sichuan Technology and Business University, and mainly researching visual communication design, visual cultural communication, UI design, interface design, and other fields.

1 Introduction

In recent years, China's aerospace surveying and mapping technology has made great progress. With the launch of 'high score', 'sky painting' and 'resources' series satellites, China has initially possessed the surveying and mapping ability with aerospace remote sensing as the main means (Yuan et al., 2021) the development of remote sensing technology has improved the spatial resolution of remote sensing images! Because the features in the image are more obvious, the information and noise increase accordingly. How to accurately extract buildings from high-resolution remote sensing images has become a research hotspot (Zhang et al., 2021).

Building extraction can be regarded as a specific image segmentation task – segmenting buildings from the surrounding background. The algorithms are mainly divided into four categories: threshold based, edge based, region based and classification based methods. The threshold based method assigns pixels with different values to different parts through manually or automatically selected thresholds (Chen et al., 2021), but can not distinguish between different regions with similar gray values; The edge based method uses Gaussian (Ji and Wei, 2019) and Sobel (Liu et al., 2021) edge detection filters to detect the mutation between adjacent pixels and generate boundaries for segmentation; Traditional building extraction from remote sensing images is generally based on implicit features such as brightness, contrast, entropy and morphological features, and the final results are obtained by artificial threshold or machine learning algorithm. The differential morphological profile model uses the image structure information, context and spectral information constructed by a variety of morphological operations to extract buildings.

With the development of deep learning, convolutional neural network, which can automatically learn and extract deep-seated features, is more and more widely used in target detection (Li et al., 2020; Liu et al., 2019), image classification (Wu et al., 2021), significant target detection, and has also made some progress in remote sensing image processing. Maggiori et al. (Xie et al., 2020) and Yuan et al. (Lv et al., 2020) improved pixel level semantic segmentation for buildings based on full convolutional network (FCN) framework, but the segmentation result is rough, and the segmentation area is uneven and incomplete. Liu et al. (2021) offer a new deep-learning structure named MultiRes-UNet network, which is an improved version of the original UNet network; Furthermore, the semantic edge information is combined with semantic polygons to solve the problem of irregular semantic polygons and enhance the boundaries of semantic polygons. Experimental results show that after adding semantic edges, the proposed network is successful in constructing object extraction from aerial images. Wu et al. (Zhao et al., 2020) proposed the multi constraint full convolution network (MC-FCN) for building extraction, and marked the missed detection and false detection through different colours, but there are false detection and missed detection for the complex background. Ji (He and Jiang, 2021) proposed scale invariant remote sensing image building extraction network (sU-Net) to promote building extraction to a new automation level. However, due to the complexity of remote sensing imaging mechanism, building itself and background environment, there are still problems of fuzzy boundary and incomplete extraction area. Zhang et al. (Huang et al., 2016) constructed a sparse constrained semantic segmentation model, which improved the extraction speed, but failed to identify small buildings, misdetect ground objects, and failed to segment the edges of some buildings. To sum up, there are still some problems in building extraction from remote sensing images, such as missed detection of small targets, blurred segmentation boundary and incomplete region. Improved snake model focuses on ranging from colour aerial images and optical images Extract buildings from LiDAR data. Omube model combines building segmentation based on total neighbourhood change with object-oriented analysis. According to the different extraction complexity of different buildings in the segmented image, a hierarchical building extraction strategy with multi feature fusion is adopted to obtain the final building extraction result. Morphological building index MBI extracts buildings by establishing the relationship between implicit features of buildings and morphological operators. Badrinarayanan et al. Proposed SEGNet network (Zhang et al., 2020b), which is an end-to-end network architecture based on pixels and an optimisation of FCN. It follows the idea of FCN for image semantic segmentation, and the network integrates Due to the characteristics of coding decoding structure and hopping network, the model can obtain more accurate output characteristic map, and can also obtain more accurate classification results when the training samples are limited. Zhang et al. (Shao et al., 2020) proposed a high-resolution remote sensing image building extraction algorithm based on sparse constraint SEGNet. The regular term and dropout are added to the SEGNet model. The algorithm introduces pyramid pooling module and Lorentz function sparse constraint factor to construct a new semantic segmentation model. This method reduces the over fitting phenomenon of the model and can extract richer semantic features (Weiyang and Liu, 2020; Cao et al., 2021; Lee and Nishikawa, 2020; Luo et al., 2020). However, small buildings are not recognised, ground features are incorrectly detected, and the edges of some buildings are not well divided.

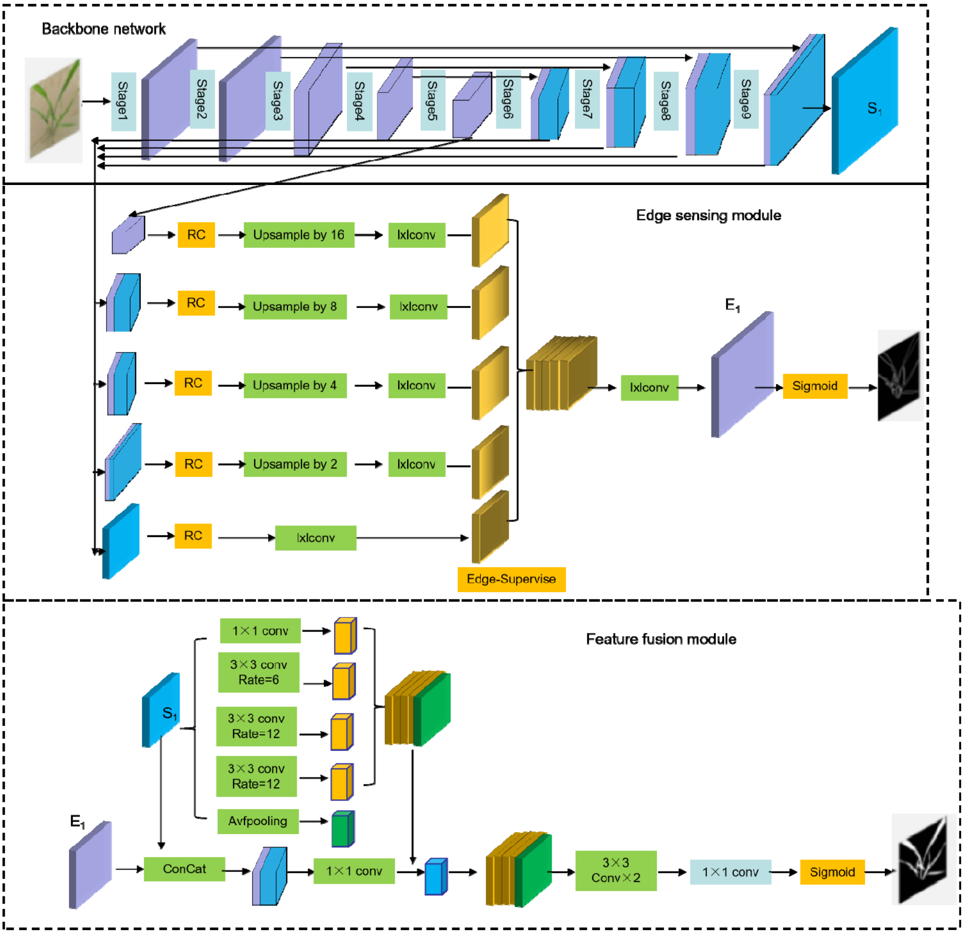
This paper proposes a lightweight remote sensing image building extraction method based on DeepLabv3. Automatic detection and extraction of buildings from remote sensing images is of great significance in the fields of smart city construction, land use investigation, disaster emergency assessment, and military target reconnaissance. Aiming at the problem of inaccurate segmentation of building edges in remote sensing images, the imprecise segmentation of building edges in remote sensing images by deep learning models is an important research direction for remote sensing intelligence applications. The skeleton network uses DeepLabv3 to connect the IEU-Net structure. Our proposed method can effectively overcome the problem of insufficient edge pixel features of samples, suppress the influence of road and building shadows on the results, and improve the extraction accuracy of houses and buildings in remote sensing images.

2 Method and principle

2.1 Model network architecture

The semantic segmentation network U-Net is in a ‘U’ shape as a whole. The feature map is dimensionally spliced by jumping connection, which can retain more location and feature information. The segmentation performance is better than other network structures on small sample datasets, and is suitable for the semantic segmentation task of small-size objects (Xie et al., 2020; Zhang et al., 2020a; Pasquali et al., 2019; Gurieva and Ilyina, 2020; Karimi et al., 2019). Therefore, based on the U-Net network model, this paper builds a crop seedling plant segmentation network that can jointly learn the regional semantics and edge features of the target. The overall structure of the network is shown in Figure 1, which mainly includes three parts: backbone network, edge perception module and feature fusion module (FFM). Among them, backbone network is mainly used to extract the semantic segmentation features of crop plant region. The edge sensing module (EAM) extracts the edge features of crop plants by introducing side output and using the edge label of the image to supervise it in depth. The FFM is mainly used to fuse the semantic segmentation features and edge features of plant regions to improve the accuracy of network detection.

Figure 1 Overall network architecture (see online version for colours)

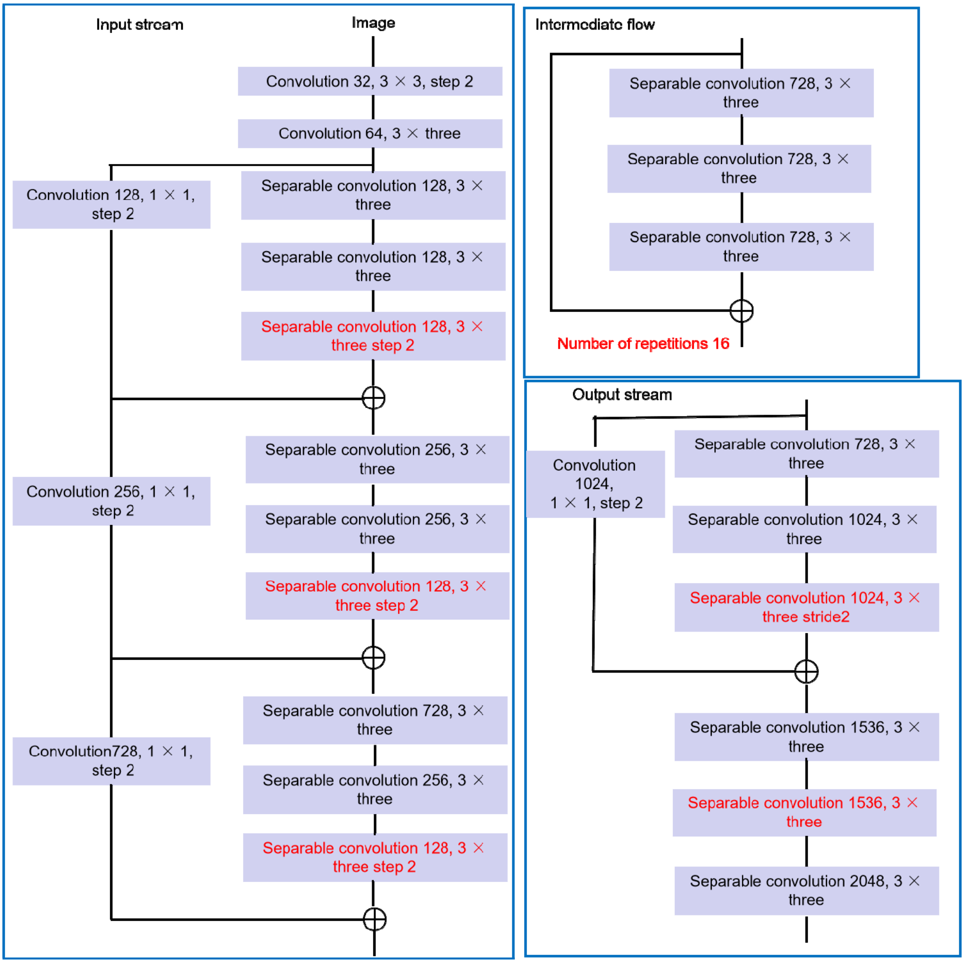


2.2 *DeepLabv3+ model*

DeepLabv3+ introduces the encoder-decoder structure, which is mainly} divided into encoder and decoder. In this structure, the resolution of the extracted features of the encoder can be controlled arbitrarily the backbone network improves the original exception ‘and applies deep separable convolution to ASPP and decoder modules.

Encoder part: the serial at route revolution is used in the trunk DCNN. Decoder part: one part is the feature of DCNN output after four times of sampling, and the other part is the result of DCNN output after parallel hole convolution. In order to prevent the high-level features obtained by encoder from being weakened, 1 × 1 convolution is used to reduce the dimension of low-level features. After the two features are fused, 3 × 3 convolution is used to further fuse the features, and bilinear interpolation method is used for 4 times up sampling to obtain the segmentation prediction result of the same size as the original image.

Figure 2 Network structure in DeepLabv3+ (see online version for colours)


Input stream
Image
Convolution 32, 3 × 3, step 2
Convolution 64, 3 × three
Convolution 128, 1 × 1, step 2
Separable convolution 128, 3 × three
Separable convolution 128, 3 × three
Separable convolution 128, 3 × three step 2
Separable convolution 256, 3 × three
Convolution 256, 1 × 1, step 2
Separable convolution 256, 3 × three
Separable convolution 128, 3 × three step 2
Separable convolution 728, 3 × three
Convolution728, 1 × 1, step 2
Separable convolution 256, 3 × three
Separable convolution 128, 3 × three step 2
Intermediate flow
Separable convolution 728, 3 × three
Separable convolution 728, 3 × three
Separable convolution 728, 3 × three
Number of repetitions 16
Output stream
Separable convolution 728, 3 × three
Convolution 1024, 1 × 1, step 2
Separable convolution 1024, 3 × three
Separable convolution 1024, 3 × three stride2
Separable convolution 1536, 3 × three
Separable convolution 1536, 3 × three
Separable convolution 2048, 3 × three

DeepLabv3+ improves the model in the semantic segmentation task. Compared with the before the improvement, the input flow of DeepLabv3+ remains unchanged, but there are more intermediate flows. All maximal pooling is replaced by depth separable convolution. After each 3 × 3 depth convolution, batch normal (BN) and rectified linear units are added (Weiyang and Liu, 2020; Yi et al., 2019; Hu et al., 2017). Structure is shown in Figure 2.

2.3 IEU net network structure

With the proposal of end-to-end full convolution neural network, Han et al. (2020) designed U-Net network with better segmentation effect. The name of U-Net network comes from the 'U' shape of the whole network (Buzzi et al., 2021; Hu et al., 2021; Ji et al., 2019; Schuegraf and Bittner, 2019). On the left is the contraction path used to extract high-dimensional feature information, and on the right is an expansion path, which is used to locate accurately. Aiming at the promotion of U-Net network, this paper proposes an IEU-Net model for house and building extraction from high-resolution remote sensing images. The structure is shown in Figure 2. The white box represents the early extracted features of the shrinkage path, which contains abstract but rich spatial information. The yellow box represents the result of up sampling convolution. Its features are extracted through the whole architecture, including detailed features with low spatial information. D in Figure 3 represents the dropout processing result. Compared with the U-Net model, the most important structure of IEU-Net is that its loss function ignores the edge cross entropy function IELoss. IELoss is based on the cross entropy function CELoss (Bittner et al., 2017), a loss function designed to solve the problem of insufficient edge pixel features of the sample image. In the second channel, the value is 1 at the location of houses and buildings and 0 at the location of non houses and buildings. The tensor composed of the pixel category probability value obtained by the forward propagation of the sample is the value in the range of layer *m* [0, 1], which is also simply expressed as p_c^i $\{c = 1, 2, \ldots, M\}$. Then shrink p_c^i and y_c^i. The process of the difference is to update and improve the model parameters and obtain the optimal solution of the parameters. The deep learning model generally uses the loss function to quantify the calculation p_c^i and y_c^i. The difference between is loss. The specific formula of classification cross entropy function CELoss is as follows:

$$CELoss = \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{M} y_c^i \log\left(p_c^i\right) \tag{1}$$

As shown in Figure 3, if the sample image edge region is ignored when calculating the loss value, the calculation region is no longer the whole image region a, but the middle region a, which means that the over fitting phenomenon caused by insufficient image edge pixel features is avoided, and the classification accuracy on the test set can be improved to a certain extent. The loss function at this time is called ignore edge cross entropy function and is recorded as IELoss. The specific formula is shown in equation (2). In IEU net network, ignore edge cross entropy function IELoss is the loss function of the model.

$$IELoss = \frac{1}{r \times N}\sum_{i-1}^{r\times N}\sum_{c=1}^{M} y_c^i \log\left(p_c^i\right) \tag{2}$$

where *r* represents the proportion of the number of selected regional pixels in the total image pixels, $r = \alpha / A$.

Figure 3 IEU-Net model structure (see online version for colours)

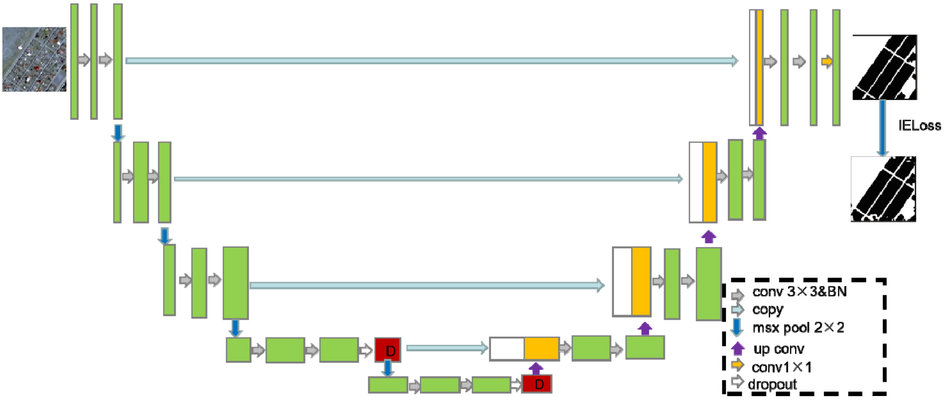
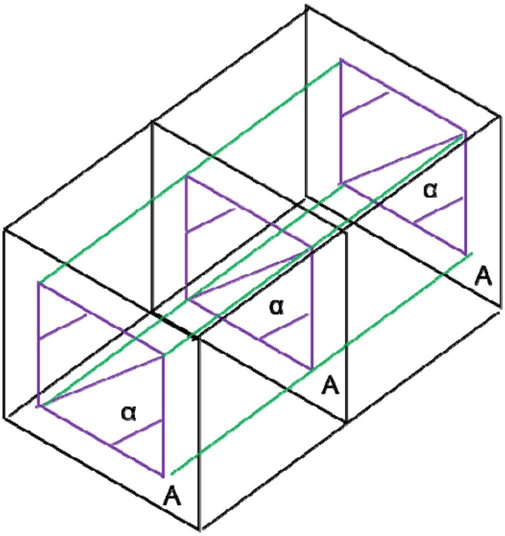


Figure 4 The calculation area for IELoss (see online version for colours)



Add BN method: for deep neural networks such as IEU net, as long as the data distribution of the first few layers of the network changes slightly, the later layers will be accumulated and amplified. Once the input data distribution of a certain layer of the network changes greatly, IEU net model needs to learn new data distribution and update parameters, which will greatly reduce the training speed of the network. In order to solve this situation, a batch normalisation BN (batch normalisation) processing method is added after each convolution. BN is a strategy proposed by Ioffe and Szegedy to do normalisation processing before entering the next layer of the network (Sun e al., 2018, 2019; Maltezos et al., 2019; Deng et al., 2021). The normalisation layer here is a network layer that can be learned and has parameters, as shown in equation (3). This paper is written in 3 steps each time. BN layer is added after convolution operation.

$$y^{(k)} = \gamma^{(k)} \overline{x}^{(k)} + \beta^{(k)} \tag{3}$$

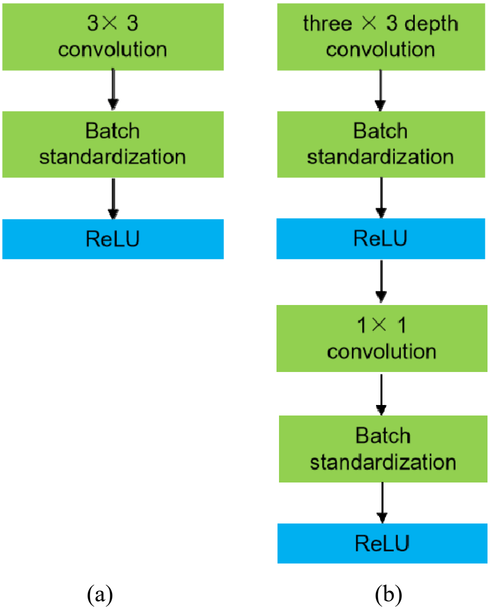
where $y^{(k)}$ is the batch normalisation result of layer K, $\overline{x}^{(k)}$ is the normalised result of standard deviation, $\gamma^{(k)}$, $\beta^{(k)}$ is the learning parameter.

2.4 Deep separable convolution

The core content of the backbone network in this paper is deep separable convolution. Yes In the remote sensing image processing with a large amount of noise and information, compared with the conventional convolution operation, the depth separable convolution has the advantages of less parameters and short training time. When the accuracy remains unchanged, it can better quickly extract the characteristic information of buildings from the remote sensing image.

The size of convolution kernel represents the size of receptive field. The larger the convolution kernel, the larger the receptive field. If the convolution kernel is too large, the amount of calculation will increase. DW is a convolution kernel corresponding to an input channel, and each input channel is spatially convoluted independently. PW is used to combine the output of DW, that is, each channel is convoluted separately without changing the number of channels. The convolution operation is shown in Figure 5.

Figure 5 (a) Standard convolution and (b) depth separable convoluton (see online version for colours)



In the deep separable convolution in this paper, DW is used to convolute different input channels respectively, and then PW is used to combine the above outputs. The low-level feature information contains more edge information, which is beneficial to improve the training accuracy. The high-level feature information extracts more complex features, and the decoding end restores the target detail information and spatial dimension through bilinear interpolation upsampling.

2.5 Accuracy evaluation index

Confusion matrix is usually used to determine the accuracy of feature extraction in the field of semantic segmentation. The definition of confusion matrix is shown in Table 1. TP is the correctly detected building feature, TN is the correctly detected non-building feature, FP is the incorrectly detected non-building feature, and FN is the incorrectly detected non building feature.

Table 1 Confusion matrix

Confusion matrix		*True value*	
		Building	*Non-building*
Estimate	Building	TP	FP
	Non-building	FN	TN

Semantic segmentation is regarded as a multi classification problem. The confusion matrix can be used to compare the classification results of the predicted output with the marked real value at the pixel level, and evaluate the predicted output results of each pixel. The extraction accuracy of buildings can be determined by accuracy, accuracy, recall and average intersection and union ratio.

Accuracy refers to the proportion of correctly predicted pixels in the total pixels, which is defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

Precision refers to the proportion of true positive examples in the predicted positive examples, which is defined as:

$$Precision = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

Recall refers to how many positive cases in the sample are correctly predicted, which is defined as:

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

Average intersection union ratio () is a commonly used evaluation index for semantic segmentation. It calculates the intersection and union ratio of two sets of real value and predicted value, which is defined as:

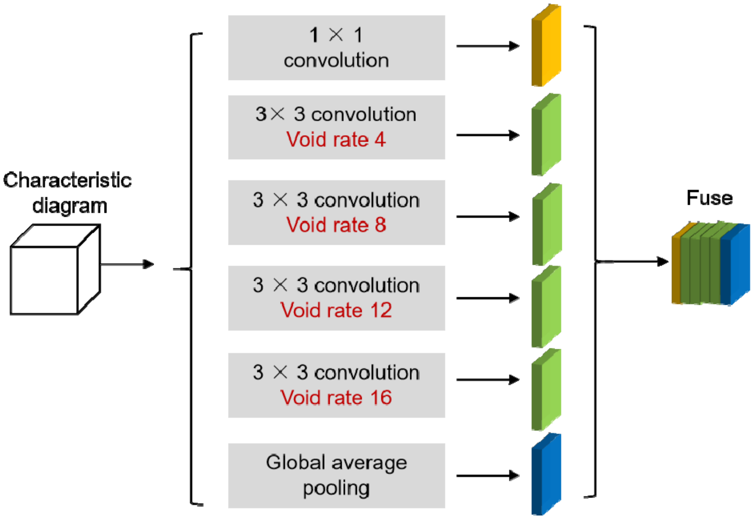
$$MIOU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{TP}{TP + FN + FP} \tag{7}$$

where *k* represents the number of categories. The average intersection union ratio reflects the coincidence degree between the predicted graph and the real graph. The closer the ratio is to 1, the higher the coincidence degree and the higher the quality of semantic segmentation.

2.6 Pyramid pooling of empty space

ASPP is formed by the fusion of atrus revolution and spatial pyramid pooling (SPP) (Zhu et al., 2021; Shrestha and Vanneschi, 2018; Zheng et al., 2018). It can effectively extract multi-scale semantic features in remote sensing images, so it is widely used in building extraction of remote sensing images. The hole rate combination of ASPP module hole convolution in DeepLabv3+ is 6, 12 and 18. With the feature extraction by the backbone network, the resolution of the feature map will gradually decrease. On the contrary, the smaller target can be segmented by the convolution kernel with smaller hole rate, and the smaller hole rate can make feature extraction more effective. The feature map obtained through the backbone network mobilenetv2 is input into the ASPP module in this paper. After 1× 1 convolution operation, 3× 3 convolution operation with different hole rates and the final pooling operation, the segmented targets of different sizes are convoluted to extract the feature map in turn, and the output 6 feature maps are fused to obtain the feature map generated by ASPP in this paper. The improved ASPP structure in this paper is shown in Figure 6.

Figure 6 Atrous spatial pyramid pooling structure of in this paper (see online version for colours)



3 Experiment and analysis

The public dataset Rmote-A building dataset is used for training and testing. The dataset is a high-resolution remote sensing image dataset published in 2019 (Protopapadakis et al., 2021), which is suitable for building extraction. The training dataset includes 4736 remote sensing images, RGB images of buildings and 4736 binary label images; The test set contains 2416 remote sensing images, building images and 2416 binary label images. The experimental results were evaluated by intersection over union (IOU), precision (P) and recall (R). The intersection and union ratio refers to the intersection of building pixels and real positive pixels detected by the algorithm and the ratio between their union, which is defined as follows:

$$IOU = \frac{TP}{TP + FN + FP} \tag{8}$$

where *TP* represents the correctly detected building features. *FP* represents a non-building feature incorrectly detected as a building feature. *FN* represents a building feature incorrectly detected as a non-building feature. Accuracy refers to the percentage of real pixels in building pixels detected by the algorithm, which is defined as follows:

$$P = \frac{TP}{TP + FP} \tag{9}$$

Recall rate refers to the percentage of building pixels detected by the algorithm in the positive pixels of the real label, which is defined as follows:

$$R = \frac{TP}{TP + FN} \tag{10}$$

In order to verify the effectiveness of the algorithm in this paper, firstly, the overall performance of Banet is analysed, and then compared with U-Net, SEGNet, MC-FCN, Su net and LSPNet algorithms. The experiment was carried out under windows10 64 bit system (Feng et al., 2020).

3.1 Overall performance analysis

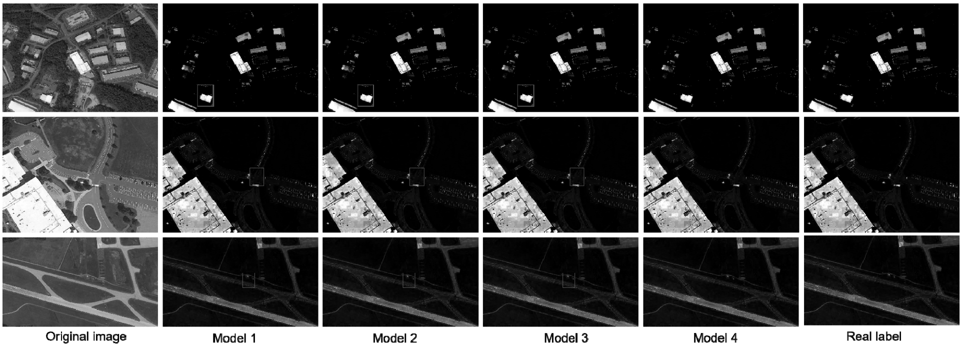
In order to verify the effectiveness of the interactive aggregation module, feature enhancement network and structural similarity loss function in the boundary perception model. Model 1 does not add interactive aggregation module and feature enhancement network, and is trained by the combination of binary cross entropy loss function and structural similarity loss function. Model 2 adds an interactive aggregation module on the basis of model 1, and the loss function is the same as model 1. In model 3, a feature enhancement network is added on the basis of model 2, and the binary cross entropy loss function is used for training. Model 4 is a model proposed by the author, which adopts the combination of binary cross entropy loss function and structural similarity loss function. All experiments were tested on Rmote-A dataset, and the performance of different network models was compared through quantitative calculation of objective evaluation indexes. The results are shown in Table 2, and the visual effect comparison is shown in Figure 7.

Table 2 Comparison of objective indexes of experimental results of different network models

Model	*Network structure*			*Loss function*		*Evaluating indicator*	
	Baseline	*IAM*	*FEN*	*LBCE*	*LSSIM*	*IoU*	*P*
1	√			√	√	0.814	0.847
2	√	√		√	√	0.831	0.857
3	√	√	√	√		0.837	0.860
4	√	√	√	√	√	0.850	0.867

It can be seen from Table 1 that the intersection and union of model 2 is 1.7% higher than that of model 1, and the accuracy is 1%. Compared with model 2, the objective evaluation indexes of test images in model 4, i.e., intersection ratio and accuracy, are improved by 1.8% and 1.1% respectively. Therefore, after adding the FFM and feature enhancement network, the method proposed by the author can improve the objective evaluation index of the test image. Compared with model 3, the intersection union ratio and accuracy of model 4 are improved by 1.2% and 0.9% respectively. The effectiveness of the structural similarity loss function is verified based on the use of binary cross entropy loss. As can be seen from Figure 7, there are many missed and false detections in model 1. Compared with model 1, the false detection of model 2 is significantly reduced, but there are still many missed detection and incomplete internal areas of buildings. The boundary of model 3 is very fuzzy and can not accurately segment the outline of the building. In contrast, model 4 achieves better extraction effect, with clearer boundary, clearer outline and more complete internal area of the building. Through the comparison of objective indicators and subjective visual effects, the effectiveness of the network structure and loss function designed by the author is verified.

Figure 7 Comparison of visual effects of different network models



3.2 Model training

When DeepLabv3+ is used for training in this paper, the training parameters are shown in Table 3, in which the basic learning rate is 1 × 10^{-4}, the learning rate attenuation factor is 0.1, and the number of batches is set to 20, indicating the number of learning images sent into the deep learning network in a batch. The input sample image size is 256 × 256 pixels, the convolution layer uses relu as the activation function, and the total number of global training is set to 307,000 steps. At the end of each round of training, the loss and accuracy will be calculated once in the verification set, and the model will be saved once. Finally, the model with the best performance in the verification set will be selected.

Table 3 Training parameters

Parameter	*Parameter value*
Basic learning rate	0.0001
Learning rate decay factor	0.1
Number of batches	20
Sample size	256 × 256

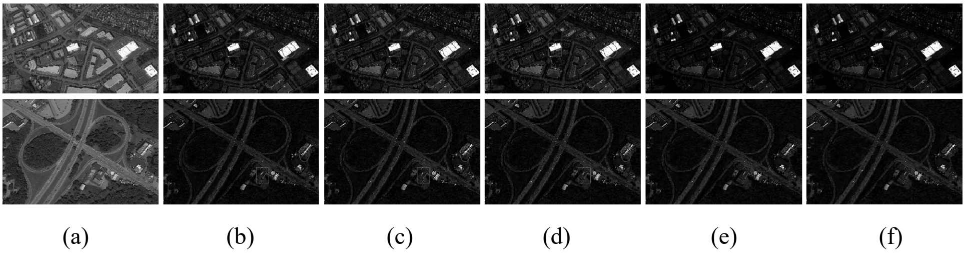
3.3 Comparative analysis of the method in this paper and the classical method

In order to verify the superiority of the method in this paper, the objective indexes are compared with the current mainstream building extraction methods such as U-Net, SEGNet, MC-FCN. The experiment is also trained on Rmote-A dataset. In the training process, relu is used as the activation function and Adam algorithm is used for network optimisation. The initial learning rate is 0.000 1, and all parameters are initialised with normal distribution. The Rmote-A test set images are tested, and the objective evaluation index results are shown in Table 4. Two images with different scales and complex background are intercepted to compare the visual effects of the extraction results, as shown in Figure 8.

Table 4 Comparison of objective indexes between BANet and classical methods

Experiment	*Method*	*IoU*	*P*	*R*
1	U-Net	0.723	0.744	0.766
2	SEGNet	0.730	0.751	0.773
3	MC-FCN	0.742	0.763	0.785
4	SU-Net	0.777	0.798	0.82
5	LSPNet	0.753	0.774	0.796
6	BANet (method in text)	0.850	0.871	0.893

Figure 8 Visual effect comparison between the method in the text and the classical method, (a) the original image (b) U-Net (c) Segnet (d) SU-Net (e) method of this article (f) true label



It can be seen from the figure that there are a large number of false detections in the U-Net extraction results shown in column (b) of Figure 8. The error detection of SEGNet extraction results shown in column (c) of Figure 8 is reduced, but there are problems of fuzzy boundary and incomplete region. It can be seen from the image in column (d) of Figure 8 that the accuracy of Su net extraction results is improved, but there are still cases where the boundary is not clear enough and there is false detection. It can be seen from the image in column (e) of Figure 8 that the image boundary extracted by the method in this paper is clearer and the building area is more complete. According to the visual effect, compared with other methods, there are false detection and the extraction area is not fine. The method proposed by the author improves the visual effect and the evaluation index.

3.4 Results performance evaluation index

The performance of building extraction of IEU net model is described quantitatively (Hu and Guo, 2019). In this paper, the confusion matrix is used to evaluate the accuracy. The confusion matrix is an M-matrix composed of the number of pixels classified into a certain category and the number of pixels whose truth test is this category × M-Size matrix, where m is the total number of categories, and the evaluation factors include overall accuracy OA, kappa coefficient, etc. The calculation method of each evaluation factor is shown in equation (10).

$$\begin{cases} OA = \sum_{\mathrm{i=1}}^{M} \frac{x_{ii}}{N} \\ Kappa = \frac{N\sum_{\mathrm{i=1}}^{M} \frac{x_{ii}}{N} - \sum_{\mathrm{i=1}}^{M} \left(x_{i+}x_{+i}\right)}{N^2 - \sum_{i=1}^{M} \left(x_{i+}x_{+i}\right)} \end{cases} \tag{11}$$

where X_{ii} refers to the diagonal elements of the confusion matrix, X_{i+}, and N refers to the total number of pixels. The overall accuracy OA represents the probability that the classified result is consistent with the actual type of the region corresponding to the reference data for each random sample. Kappa coefficient represents the proportion of error reduction between classification and completely random classification.

3.5 Experimental results and analysis

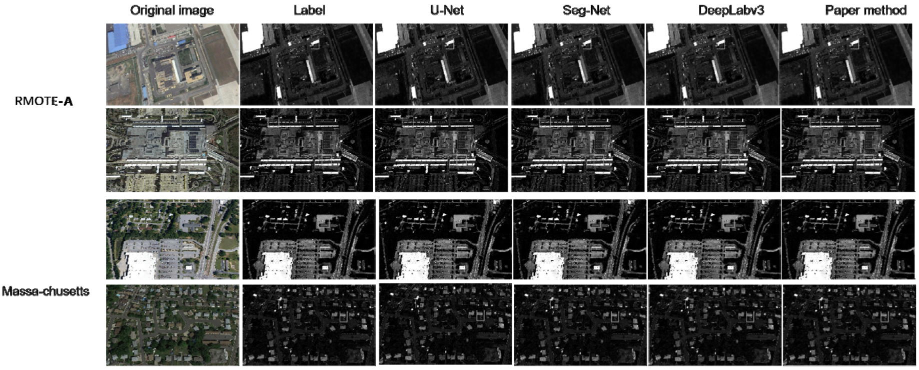
Table 6 shows the evaluation index values of different models in the two datasets, and the evaluation indexes are mainly intersection union ratio and F1 score. Compared with other classical models, the intersection and merging ratio of experimental results in the two datasets is higher, and the accuracy of building extraction from remote sensing images is further improved. Compared with DeepLabv3+ model, the intersection and merger ratio of this method in Rmote-A dataset is increased by 2.71%, F1 score is increased by 2.13%, intersection and merger ratio in Massachusetts dataset is increased by 2.04%, F1 score is increased by 2.31%, and the evaluation index value of U-Net and SEGNet is low. In general, the method proposed in this paper is improved compared with other models, and has high effectiveness for building extraction. Because the model in this paper uses lightweight network, it is the backbone network of DeepLabv3+ model, this method has less backbone network parameters, so the training time is shorter, which can effectively improve the training speed of the model.

The building extraction results are shown in Figure 9. In the prediction results, images are randomly selected as the comparative analysis of the experimental results. In the Rmote-A dataset, the extraction results of U-Net and SEGNet are similar, and the overall effect is poor. The extraction of small buildings sometimes fails or the extraction area is very small. However, the extraction accuracy of small buildings still needs to be improved. The boundary information extraction of small buildings is not perfect. When extracting buildings with complex boundaries, the boundary detail information extraction is not fine enough. When extracting large buildings, some holes or extraction blur occasionally appear.

Table 5 Building extraction evaluation results

Dataset	*Model*	*Consolidation ratio/%*	*F1 score/%*	*Training time/h*
RMOTE-A	U-Net	76.17	86.27	10.53
	SEGNet	76.04	86.14	14.52
	PSPNet	77.12	87.22	10.11
	DeepLabv3+	78.55	88.65	9.54
	Method of this article	71.26	81.36	5.15
Massachusetts	U-Net	70.14	80.24	12.64
	SEGNet	70.76	80.86	16.63
	PSPNet	71.04	81.14	12.22
	DeepLabv3+	73.47	83.57	11.65
	Method of this article	75.51	85.61	7.26

Figure 9 RMOTE-A and Massa-chusetts building extraction results (see online version for colours)



In this paper, the image is visually interpreted in ArcGIS software (Shrestha and Vanneschi, 2018), the elements of the obtained vector file are transformed into grid to obtain the label image, that is, the real surface image recording the location of houses and buildings in the remote sensing image. As shown in Figure 10, the loss values of the training set and the verification set gradually decrease with the number of iterations until they become stable, and the accuracy rate gradually increases until they become stable.

In order to verify the effectiveness of ignoring the edge cross entropy function IELoss and ignoring the edge prediction in IEUNet model in solving the insufficient characteristics of edge pixels compared with the cross entropy function CELoss and ordinary prediction As shown in Figure 11, taking *r* value of 0.5 as an example, the overall accuracy OA using IELoss and ignoring edge prediction is 5.03% higher and kappa value is 0.165 higher than that using CELoss and ordinary prediction, which fully illustrates the effectiveness of IELoss and ignore edge prediction. Figure 11 shows the extraction results of houses and buildings corresponding to different r values. We can intuitively see that the results using IELoss are more accurate. In order to exclude the influence of ignoring edge prediction, IELoss is verified separately. In this paper, the value of r is 1.0, that is, the experimental result model using CELoss as the loss function performs r=0.5 ignoring edge prediction. As shown in Table 2, using IELoss results in a 2.74% higher overall accuracy OA and a 0.093 higher Kappa value than CELoss, proving that IELoss plays a role in the accuracy improvement in Figure 11.

Figure 10 Accuracy and loss of training set and validations (see online version for colours)

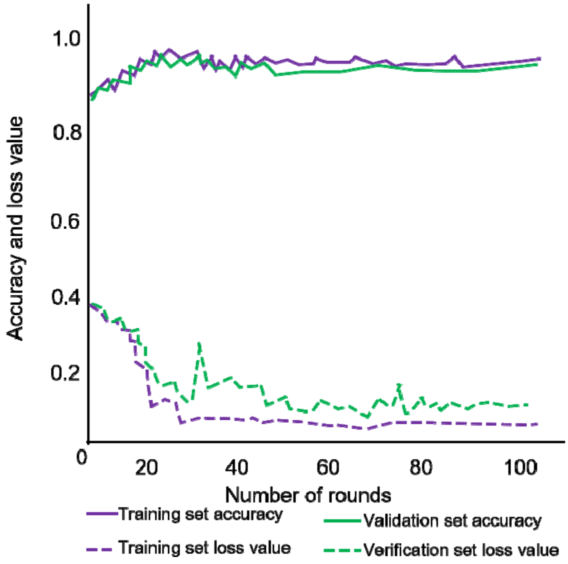
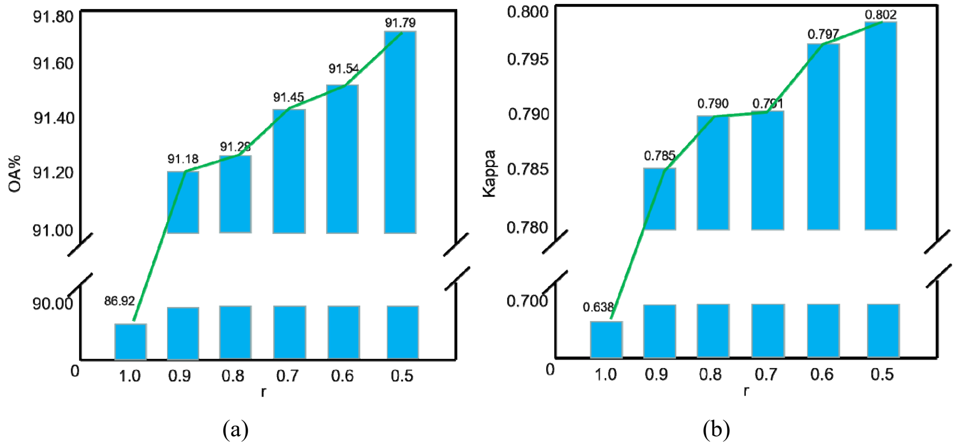
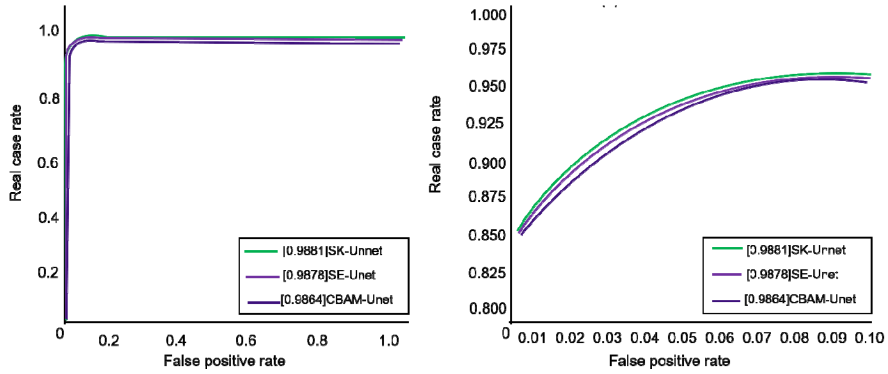


Figure 11 OA and Kappa of prediction results corresponding to different values, (a) distribution of change in OA (b) distribution of change in Kappa (see online version for colours)



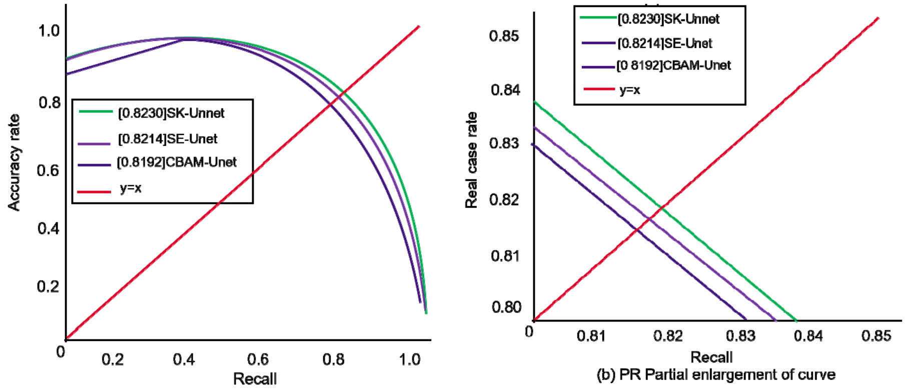
In order to further verify the extraction accuracy of the added model for buildings, training and testing are carried out respectively. The verification accuracy is shown in Table 3. It can be seen that compared with the original U-Net model, the model module in this paper improves IOU by 1.02% and 0.63%, recall by 1.76% and 1.22%, and F1 score by 0.58% and 0.38%. After the SK module is added, the extraction accuracy is improved the highest, which is increased by 1.06% on IOU, 3.60% on recall and 0.6% on F1 score as show on Figures 12 and 13.

Figure 12 Overall and partial ROC curves of the model, (a) ROC overall of curve (b) partial enlargement of curve (see online version for colours)



(a) (b)

Figure 13 Overall and partial diagram of PR curve of model, (a) PR overall curve (b) PR enlargement of curve (see online version for colours)



(a) (b)

4 Conclusions

Aiming at the problems of blurred boundary, inaccurate extraction results, missed detection of buildings, and false detection of ground objects in building extraction, a boundary-aware-based building extraction network is proposed that does not rely on any priori conditions. The skeleton network uses DeepLabv3 to connect the IEU-Net structure. It is used to learn the features of missed targets and improve the accuracy of the prediction results. The feature refinement network is used to further refine the boundaries and regions of the enhanced results. Finally, in order to accelerate the convergence of the model and learn the boundary information better, two methods are given. On the basis of building extraction, related industry applications such as building change detection can be carried out to facilitate the effective management of land resources. In addition, the network can also be extended to similar image detection and binary segmentation, such as other feature extraction. The addition of MBI data can overcome the influence of road and building shadows to a certain extent, and extract the edge information of buildings more accurately.

References

Bittner, K., Cui, S. and Reinartz, P. (2017) 'Building extraction from remote sensing data using fully convolutional networks', *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences-ISPRS Archives*, Vol. 42, No. W1, pp.481–486.

Buzzi, S., Grossi, E., Lops, M., et al. (2021) 'Radar target detection aided by reconfigurable intelligent surfaces', *IEEE Signal Processing Letters*, Vol. 7, No. 28, pp.1315–1319.

Cao, R., Wang, Y., Zhao, B. et al. (2021) 'Ship target imaging in airborne SAR system based on automatic image segmentation and ISAR technique', *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol. 2, No. 14, pp.1985–2000.

Chen, M., Wu, J., Liu, L., Zhao, W., Tian, F., Shen, Q., Zhao, B. and Du, R. (2021) DR-Net: an improved network for building extraction from high resolution remote sensing image', *Remote Sensing*, Vol. 13, No. 2, p.329.

Deng, W., Shi, Q. and Li, J. (2021) 'Attention gate based encoder-decoder network for automatical building extraction', *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol. 25, No. 99, p.1.

Feng, W., Sui, H., Hua, L. et al. (2020) 'Building extraction from VHR remote sensing imagery by combining an improved deep convolutional encoder-decoder architecture and historical land use vector map', *International Journal of Remote Sensing*, Vol. 41, No. 17, pp.6595–6617.

Gurieva, V.A. and Ilyina, A. (2020) 'Prospects for the development of building ceramics, based on stale slags from the non-ferrous metallurgy of Orenburg', *Solid State Phenomena*, Vol. 299, No. 3, pp.210–215.

Han, J., Moradi, S., Faramarzi, I. et al. (2020) 'Infrared small target detection based on the weighted strengthened local contrast measure', *IEEE Geoscience and Remote Sensing Letters*, Vol. 18, No. 9, pp.1670–1674.

He, S. and Jiang, W. (2021) 'Boundary-assisted learning for building extraction from optical remote sensing imagery', *Remote Sensing*, Vol. 13, No. 4, p.760.

Hu, P., Wu, F., Peng, J. et al. (2017) 'Automatic abdominal multi-organ segmentation using deep convolutional neural network and time-implicit level sets', *International Journal of Computer Assisted Radiology and Surgery*, Vol. 9, No. 12, pp.399–411.

Hu, Q., Zhen. L., Yao, M. et al. (2021) 'Automated building extraction using satellite remote sensing imagery', *Automation in Construction*, Vol. 123, No. 4, p.103509.

Hu, Y. and Guo, F. (2019) 'Building extraction using mask scoring R-CNN network', *Computer Science and Application Engineering*, Vol. 131, No. 3, p.1427.

Huang, X., Yuan, W., Li, J. et al. (2016) 'A new building extraction postprocessing framework for high-spatial-resolution remote-sensing imagery', *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol. 10, No. 2, pp.654–668.

Ji, S. and Wei, S. (2019) 'Building extraction via convolutional neural networks from an open remote sensing building dataset', *Acta Geodaetica et Cartographica Sinica*, Vol. 48, No. 4, p.448.

Ji, S., Wei, S. and Lu, M. (2019) 'A scale robust convolutional neural network for automatic building extraction from aerial and satellite imagery', *International Journal of Remote Sensing*, Vol. 40, No. 9, p.779.

Karimi, D., Zeng, Q., Mathur, P. et al. (2019) 'Accurate and robust deep learning-based segmentation of the prostate clinical target volume in ultrasound images', *Medical Image Analysis*, Vol. 5, No. 57, pp.186–196.

Lee, J. and Nishikawa, R.M. (2020) 'Cross-organ, cross-modality transfer learning: feasibility study for segmentation and classification', *IEEE Access*, Vol. 12, No. 8, pp.210194–210205.

Li, D., Shen, X., Yu, Y., Guan, H., Li, J., Zhang, G. and Li, D. (2020) 'Building extraction from airborne multi-spectral LiDAR point clouds based on graph geometric moments convolutional neural networks', *Remote Sensing*, Vol. 12, No. 19, p.3186, https://doi.org/10.3390/rs12193186.

Liu, Y., Geng, L., Zhang, W. et al. (2021) 'Survey of video based small target detection', *Journal of Image and Graphics*, Vol. 9, No. 4, pp.122–134.

Liu, Y., Gross, L., Li, Z. et al. (2019) 'Automatic building extraction on high-resolution remote sensing imagery using deep convolutional encoder-decoder with spatial pyramid pooling', *IEEE Access*, Vol. 7, No. 1, pp.128774–128786.

Luo, X.J., Oyedele, L.O., Ajayi, A.O., Akinade, O.O., Owolabi, H.A. and Ahmed, A. (2020) 'Feature extraction and genetic algorithm enhanced adaptive deep neural network for energy consumption prediction in buildings', *Renewable and Sustainable Energy Reviews*, Vol. 131, No. 4, p.1078.

Lv, B., Peng, L., Wu, T. and Chen, R. (2020) 'Research on urban building extraction method based on deep learning convolutional neural Network', *IOP Conference Series Earth and Environmental Science*, Vol. 502, No. 1, p.671.

Maltezos, E., Doulamis, A., Doulamis, N. and Ioannidis, C. (2019) 'Building extraction from LiDAR data applying deep convolutional neural networks', *IEEE Geoscience and Remote Sensing Letters*, Vol. 16, No. 1, p.1249.

Pasquali, G., Iannelli, G.C. and Dell'Acquam, F. (2019) 'Building footprint extraction from multispectral, spaceborne earth observation datasets using a structurally optimized U-Net convolutional neural network', *Remote Sensing*, Vol. 11, No. 23, p.326.

Protopapadakis, E., Doulamis, A., Doulamis, N. et al. (2021) 'Stacked autoencoders driven by semi-supervised learning for building extraction from near infrared remote sensing imagery', *Remote Sensing*, Vol. 13, No. 3, p.371.

Schuegraf, P. and Bittner, K. (2019) 'Automatic building footprint extraction from multi-resolution remote sensing images using a hybrid FCN', *ISPRS International Journal of Geo-Information*, Vol. 8, No. 4, p.323.

Shao, Z., Tang, P., Wang, Z., Saleem, N., Yam, S. and Sommai, C. (2020) 'BRRNet: a fully convolutional neural network for automatic building extraction from high-resolution remote sensing images', *Remote Sensing*, Vol. 12, No. 6, p.1309.

Shrestha, S. and Vanneschi, L. (2018) 'Improved fully convolutional network with conditional random fields for building extraction', *Remote Sensing*, Vol. 10, No. 7, p.471.

Shrestha, S. and Vanneschi, L. (2018) 'Improved fully convolutional network with conditional random fields for building extraction', *Remote Sensing*, Vol. 10, No. 7, p.1135.

Sun, G., Huang, H., Zhang, A., Li, F., Zhao, H. and Fu, H. (2019) 'Fusion of multiscale convolutional neural networks for building extraction in very high-resolution images', *Remote Sensing*, Vol. 11, No. 3, p.1031.

Sun, Y., Zhang, X., Zhao, X. and Xin, Q. (2018) 'Extracting building boundaries from high resolution optical images and LiDAR data by integrating the convolutional neural network and the active contour model', *Remote Sensing*, Vol. 10, No. 9, p.778.

Weiyang and Liu, X. (2020) 'The application of deep convolution neural network to building extraction in remote sensing images', *World Scientific Research Journal*, Vol. 6, No. 3, pp.136–144.

Wu, J., Chen, P., Liu, Y. et al. (2021) 'High-resolution remote sensing information extraction method for earthquake-damaged buildings', *Geography and Geographic Information Science*, Vol. 2021, No. 2013-3, pp.35–38.

Xie, Y., Zhu, J., Cao, Y., Feng, D., Hu, M., Li, W., Zhang, Y. and Fu, L. (2020) 'Refined extraction of building outlines from high-resolution remote sensing imagery based on a multifeature convolutional neural network and morphological filtering', *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol. 13, DOI: 10.1109/JSTARS.2020.2991391.

Yi, Y., Zhang, Z., Zhang, W., Zhang, C., Li, W. and Zhao, T. (2019) 'Semantic segmentation of urban buildings from VHR remote sensing imagery using a deep convolutional neural network', *Remote Sensing*, Vol. 11, No. 15, p.244.

Yuan, Q., Shafri, H.Z.M., Alias, A.H. and bin Hashim, S.J. (2021) 'Multiscale semantic feature optimization and fusion network for building extraction using high-resolution aerial images and LiDAR data', *Remote Sensing*, Vol. 13, No. 13, p.1035.

Zhang, X., Zheng, Y., Liu, W., Peng, Y. and Wang, Z. (2020a) 'An improved architecture for urban building extraction based on depthwise separable convolution', *Journal of Intelligent & Fuzzy Systems*, Vol. 38, No. 5, p.102.

Zhang, Y., Li, W., Gong, W., Wang, Z. and Sun, J. (2020b) 'An improved boundary-aware perceptual loss for building extraction from VHR images', *Remote Sensing*, Vol. 12, No. 7, p.11.

Zhang, Y., Liu, C. and Tang, R. (2021) 'Building semantic information extraction based on full convolution neural network and parameter transfer', *IOP Conference Series: Earth and Environmental Science*, Vol. 783, No. 1.

Zhao, B., Wang, C., Fu, Q. et al. (2020) 'A novel pattern for infrared small target detection with generative adversarial network', *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 59, No. 5, pp.4481–4492.

Zheng, X., Lei, Q., Yao, R. et al. (2018) 'Image segmentation based on adaptive K-means algorithm', *EURASIP Journal on Image and Video Processing*, Vol. 2018, No. 1, pp.1–10.

Zhu, Y., Liang, Z., Yan, J. et al. (2021) 'E-D-Net: automatic building extraction from high-resolution aerial images with boundary information', *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol. 14, No. 99, p.1.