Machine Learning with Applications 6 (2021) 100194

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

Integrating semantic edges and segmentation information for building extraction from aerial images using UNet

Arnick Abdollahi, Biswajeet Pradhan*

The Centre for Advanced Modelling and Geospatial Information Systems (CAMGIS), Faculty of Engineering and IT, University of Technology Sydney, Sydney 2007, NSW, Australia

ARTICLE INFO

Keywords:
AIRS
Buildings mapping
Deep learning
MultiRes-UNet
UNet architecture
Remote sensing

ABSTRACT

Understanding urban dynamics, such as estimating population, urban development, and several other uses, necessitates up-to-date large-scale building maps. Since aerial imagery provides enough textural and structural details, it has been utilized as a critical data source for building detection. However, accurate mapping of building objects from aerial imagery is a challenging task. This problem is attributed due to presence of vegetation and shadows in images that present similar spectral values and transparency as a building class. To deal with the issues mentioned above, we offer a new deep-learning structure named MultiRes-UNet network, which is an improved version of the original UNet network. In the proposed network, we utilized the MultiRes block to assimilate the features learned from the data at various scales and comprise some more spatial details. Also, we suggest the incorporation of several convolutional operations along with the skip connections to mitigate the differences between the encode–decoder features. Furthermore, we integrated semantic edge information with semantic polygons to solve the issue of irregular semantic polygons and enhance the boundary of semantic polygons. We tested our network on aerial images for roof segmentation dataset, and the experimental results exhibited that the proposed network can improve the quantitative results of Intersection Over Union to 0.78% after adding semantic edges. We also used state-of-the-art comparative models such as UNet, DeeplabV3, ResNet, and FractalNet networks to show the competency of the introduced network, and the results prove the success of the introduced network for building object extraction from aerial imagery.

1. Introduction

Building objects are one of the significant terrestrial features because they play an essential task in many applications, such as geographic information systems, real-estate management, population estimation, urban planning, and other geospatial related applications (Vakalopoulou, Karantzalos, Komodakis, & Paragios, 2015). An enormous amount of remote sensing data is being accumulated each day with the rapid advancement of sensor technologies. Therefore, extracting building objects by leveraging the fast-updated and affordable remote sensing imagery has been a significant practical interest since high-resolution remote sensing data became more available and cost effective (Sumer & Turker, 2013). Manual delineation of building objects from images consumes considerable effort and time. Remote sensing technologies and novel data science provide possibilities for automatic building detection to contribute to urban dynamic mapping and lessen extensively manual works (Abdollahi, Pradhan, Gite and Alamri, 2020; Ji, Wei, & Lu, 2018). However, the automatic detection of building objects from remote sensing imagery has been a challenge because of the heterogeneity and complicated appearance of these objects in mixed backgrounds. Designing features that can best present a building object is traditionally the principal method for extracting building objects from remote sensing images. The most common utilized factors, such as semantic and height (Zhong, Xu, Yang, & Hu, 2015), shape (Dunaeva & Kornilov, 2017), shadow (Chen, Shang, & Wu, 2014), edge (Li & Wu, 0000), texture (Zhang, 1999), spectrum (Zhong, Huang, & Xie, 2008), and color (Sirmacek & Unsalan, 2008), can alter under various conditions of building architecture, scale, atmospheric circumstances, surroundings, light, and sensor quality. The practical feature design is far from a common procedure for automatic building extraction, given that it solves only particular issues with particular data.

In recent years, convolutional neural network (CNN) has been popularly used in various remote sensing field (Abdollahi, Pradhan, & Alamri, 2021; Hong, Noh, & Han, 2015). CNN generally maps the main

The code (and data) in this article has been certified as Reproducible by Code Ocean: (https://codeocean.com/). More information on the Reproducibility Badge Initiative is available at https://www.elsevier.com/physical-sciences-and-engineering/computer-science/journals.

* Corresponding author.
E-mail addresses: arnick.abdollahi@uts.edu.au (A. Abdollahi), Biswajeet.Pradhan@uts.edu.au (B. Pradhan).

https://doi.org/10.1016/j.mlwa.2021.100194
Received 12 April 2021; Received in revised form 15 October 2021; Accepted 16 October 2021
Available online 24 October 2021

input to successive vectors (a regression issue) or to specified multiple or binary labels (a classification issue) by automatically learning multilevel representations (Maggiori, Tarabalka, Charpiat, & Alliez, 2016). The CNN gradually replaces the conservative typical feature handcrafting in classification or detection applications using the powerful "representation learning" capability. Notably, the CNN application on building extraction significantly simplifies the feature design and has illustrated promising outcomes (Yuan, 2017). The commonly utilized CNN architectures contain ResNet (He, Zhang, Ren, & Sun, 2016), AlexNet (Krizhevsky, Sutskever, & Hinton, 2017), GoogleNet (Szegedy et al., 2015), and VGGNet (Simonyan & Zisserman, 2014), which have been successfully implemented in image segmentation and classification. A single class label is typically the output of these architectures in image classification and CNN has been contributed extensively to image semantic segmentation. In 2015, Long, Shelhamer, and Darrell (2015) developed a pixel-to-pixel fully convolutional network (FCN) by extending the main CNN structure to enable dense prediction. In a typical FCN structure, the level of convolutions was utilized to downsample feature maps; then, the low-resolution features were upsampled to the original input by transposed convolutions (Zeiler, Krishnan, Taylor, & Fergus, 2010). Since the development of FCN, different types of FCN structures, such as UNet (Ronneberger, Fischer, & Brox, 2015), DeconvNet (Noh, Hong, & Han, 2015), and SegNet (Badrinarayanan, Kendall, & Cipolla, 2017), have been suggested. The latter approaches mainly leveraged FCN-based structures for semantic segmentation of remote sensing imagery because the earlier approaches that used non-FCN-based structures are computationally and memory extensive (Volpi & Tuia, 2016).

FCN-based approaches are exclusively used for building detection from remote sensing data. Wu et al. (2018) performed end-to-end building segmentation from aerial imagery using multi-constraint FCN architecture. Abdollahi, Pradhan and Alamri (2020) applied a new FCN architecture called Seg-UNet, which is a mixture of SegNet and UNet structures, to extract building objects from a Massachusetts building dataset. In Yuan (2017), a simple architecture of FCN model that combines several layers of activation into pixel level prediction was proposed. In addition, the signed distance function of building borders, which has an enhanced representation power, was introduced for presenting output. Maggiori et al. (2016) reduced the tradeoff between identification and accurate localization by applying an end-to-end FCN structure for the dense and pixel-wise classification of Massachusetts building imagery. Yang et al. (2018, 2018) applied dense-attention structure for building detection from Postdam building dataset. The suggested model includes DenseNets and spatial attention fusion mechanism that can efficiently obtain high-level feature information to overcome noises and strengthen feature distribution. Xu, Wu, Xie, and Chen (2018) used deep residual networks to detect building objects from Potsdam and Vaihingen datasets. To optimize the produced classification map via the prosed model and remove salt-and-pepper noises, a guided filter was utilized in the post-processing stage. In Yang, Wu et al. (2018), Yang, Yuan et al. (2018), various types of FCN structures, such as conditional random field as recurrent neural network, branch-out FCN, and SegNet, were proposed for building detection from aerial imagery with a 1 m spatial resolution. Additional near infra-red information and signed-distance labels were fused into the building detection architecture to advance the results. Another work (Chen et al., 2018) implemented various state-of-the-art deep FCN frameworks, such as pyramid scene parsing network, feature pyramid network (FPN), and FPN with multi-scale feature fusion for building roof detection from large-scale benchmark aerial imagery.

Although the abovementioned frameworks have attained progress in tackling the issue of building detection, they revealed several limitations. Most of these frameworks unveiled poor success in building detection purposes in heterogeneous areas such as shadows, vegetation covers, and parking lots where these obstacles enclose buildings. Thus, we proposed a new deep learning framework named MultiRes-UNet, an improved version of the UNet network, which we believe will improve the results of other deep learning structures in the building detection domain. We presumed that the UNet network may be lacking in specific criteria and then suggested some modifications to it to identify possible improvement scopes. We added MultiRes block to the model to assimilate the features learned from the data at various scales and comprise some more spatial details. Moreover, we replaced the common skip connection used in the UNet with a new shortcut path named Res path. In Res path, we utilized a chain of convolutional operations to pass the features from the encoder to the decoder instead of merging the feature maps from the encoder part with those from the decoder part in a straight-forward manner. The semantic gaps between encoder and decoder features were expected to decrease using these extra non-linear operations. We tested our network based on aerial images for roof segmentation (AIRS) dataset, which included over 220,000 buildings and presented a broad coverage of aerial images with a spatial resolution of 7.5 cm. We integrated building semantic edges with semantic polygons to detect buildings accurately. Specifically, we used semantic edges to: (i) realize the distinction between adjacent buildings, make semantic polygons more appropriate for actual building forms, (ii) solve the issue of irregular semantic polygons, and (iii) enhance the boundary of semantic polygons. The rest of this manuscript is presented as follows. The second section gives an overview of the suggested MultiRes-UNet framework. Sections 3 and 4 depict the experiential outcomes and detailed comparison, respectively. Lastly, Section 5 describes the significant findings of this study.

2. Methodology

The structure of the presented MultiRes-UNet network is detailed in this section. The initial part explains the framework of MultiRes block and then discusses the Res path, which is a new shortcut path for passing the encoder feature maps to the decoder part. Ultimately, the architecture of the presented MultiRes-UNet network is elucidated.

2.1. MultiRes block

In most cases of remote sensing images originating from different modalities, building objects are of various scales and are irregular. Thus, a deep learning framework could serve well to analyze these objects across multiple context sizes. A sequence of two convolutions with 3×3 kernel size was utilized after every pooling operation and unsampled convolution in the original UNet (Ronneberger et al., 2015) structure. This series of two 3×3 convolutions matched a 5×5 convolution (Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016). Thus, the incorporation of 3×3 and 7×7 convolutions in parallel to the 5×5 convolution (Fig. 1(e)) is the easiest way to augment the UNet structure with multi-resolution analysis following the method of inception framework. Thus, the architecture of UNet will be facilitated to adapt the features learned from the images of various context sizes by the replacement of convolution operations with inception-like blocks. Using strided convolution operations (Wang et al., 2018) is another feasible way. However, the introduction of extra convolutional operations in parallel extremely proliferates the memory requirement despite enhancing performance. Thus, we used a sequence of lightweight and smaller convolutions (3×3) to factorize the more expensive and larger of 5×5 and 7×7 convolutions which are shown in Fig. 1(f). To extract the spatial features from various scales, we obtained the three convolutions' outputs and concatenated them together. We performed these steps because the outputs of the 5×5 and 7×7 convolutional layers can be approximated by the second and third 3×3 convolutional operations, respectively. This modification is still considerably memory demanding, although it hugely decreases the memory requirement. The reason is that if two convolutions are presented in a sequence in a deep neural structure, the number of filters in the initial convolution holds a quadratic influence over the memory (Szegedy et al., 2015).

To hamper the initial layers' memory requirement from propagating to the deeper section of the framework, we constantly proliferated the filters for the three successive convolutional layers (from 1 to 3) rather than using the equal number of filters in those layers. In addition, to determine extra spatial information, we added a residual connection for the introduction of 1×1 convolutions (Drozdzal, Vorontsov, Chartrand, Kadoury, & Pal, 2016). Fig. 1(g) depicts this arrangement, which is called a "MultiRes block".

2.2. Res path

The introduction of skip connection between the encoder layers corresponding to the decoder layers is an original contribution of the UNet network (Ronneberger et al., 2015). This condition can preserve the disintegrated spatial features that are lost during the pooling operation. However, the skip connection has defects, as described in the following. A feasible semantic gap exists between two collections of features being concatenated because the initial layers in the encoder part of the UNet model compute the low-level features, whereas the deep layers in the decoder part compute the notable higher-level features. Before the initial pooling layer, the encoder was fused with the decoder after the last unsampled layer using the first skip connection. Hence, the concatenation of these inconsistent collections of features can possibly negatively influence the prediction procedure because they can cause inconsistencies during the learning process. As we moved toward the subsequent skip connections, the amount of inconsistency was expected to gently reduce because the encoder features were not only concatenated with the features from the decoder part of the newer layers but also moved with further processing.

Thus, we suggested accommodating several convolutional operations along the skip connections to mitigate the difference between the encode–decoder features. In addition, we introduced a residual connection rather than utilize the normal convolutional operation because this process yields ample deep structures and eases the learning process (Szegedy, Ioffe, Vanhoucke, & Alemi, 2017). We initially passed the features through a sequence of convolutions and merged them with the decoder features instead of simply merging the encoder and decoder features. The semantic gaps between encoder and decoder features were expected to decrease using these extra non-linear operations. Fig. 2 depicts the suggested shortcut called "Res path". In particular, the 1×1 filters accompanied the residual connections, and the 3×3 filters were utilized in the convolutions.

2.3. Architecture of MultiRes-UNet

We used the suggested MultiRes block instead of the series of two convolutions in the proposed MultiRes-UNet structure. To control the number of filters of the convolutions inside the MultiRes blocks, we allocated a W parameter for every block. The value of W is computed as follows:

$$W = \alpha \times U \tag{1}$$

where the number of filters in the corresponding layer of the UNet network is defined as U and a scaler coefficient is defined as α. The parameter W preserves an analogous connection between the suggested MultiRes-UNet network and the main UNet network. After every pooling or transposing of layers, the value of W became double, similar to the original UNet network. To maintain the number of parameters in our proposed network to a level lesser than that of the UNet, we assigned $\alpha = 1.67$. We assigned the number of filters in our proposed network as $U = [32, 64, 128, 256, 512]$. We also allocated filters of $[\frac{W}{6}]$, $[\frac{W}{3}]$, and $[\frac{W}{2}]$ to the three succeeding convolutions, respectively. This finding is due to the following. Instead of maintaining the number of filters the same, the number of filters in the succeeding convolutions within a MultiRes block must be expanded, as we pointed out in Section 2.1.

On the other hand, we introduced a new shortcut path (Res path) for combining the encoder and decoder features and replaced it with the common skip connections used in the original UNet model. In the suggested Res path, we implemented several convolutional layers on the feature maps disseminating from the contracting part (encoder) to the expansive part (decoder). As pointed out in Section 2.2, we assumed that as we passed through the internal shortcut paths, the intensity of the semantic gaps between the encoder features maps and decoder ones were reduced. Thus, the number of convolutional blocks utilized along the four Res paths also steadily decreased to 4, 3, 2 and 1. Moreover, in the four Res paths blocks, we utilized filters of 32, 64, 128, and 256 to consider the number of feature maps in encoder–decoder. The Rectified Linear Unit activation function (ReLU) (Abdollahi & Pradhan, 2021a, 2021b) was used to activate the entire convolutional layers utilized in the suggested network except for the output layer. In addition, all of them were batch normalized (Ioffe & Szegedy, 2015). Table 1 and Fig. 3 depict the architectural details and a diagram of the suggested MultiRes-UNet network, respectively. We utilized the binary cross-entropy (BCE) (Ibtehaz & Rahman, 2020) function as the loss function of the MultiRes-UNet model to train the model as follows:

$$BCE(X, Y, \hat{Y}) = \sum_{px \in X} -(y_{px} \log(\hat{y}_{px}) + (1 - y_{px}) \log(1 - \hat{y}_{px})) \tag{2}$$

where X, Y, and $\hat{Y}$ are the input image, corresponding ground truth image, and predicted segmentation map, respectively. Meanwhile, for a pixel px, y_{px} is the ground truth value, and $\hat{y}_{px}$ is the model prediction. The loss function J for a batch including n imagery can be defined as follows:

$$J = \frac{1}{n} \sum_{i=1}^{n} BCE(X_i, Y_i, \hat{Y}_i) \tag{3}$$

3. Results

3.1. Dataset preparation

We used AIRS (Chen et al., 2018) dataset, which includes 1047 aerial images with the original spatial dimension of $10\,000 \times 10\,000$ and spatial resolution of 7.5 cm. Given computational restraints, we cut the original images into the size of 1536×1536. Consequently, we utilized 1250 images in our experiment. We divided the dataset into 1225 images for training and validation and 25 images for the testing set. We just selected 25 images with different backgrounds and complexity to test the effectiveness of the proposed model for building extraction from aerial images. Fig. 4 exhibits several examples of the used imagery.

3.2. Evaluation metrics

The presented technique's efficiency was evaluated using three measurement metrics: Matthews Correlation Coefficient (MCC), Intersection Over Union (IOU), and F1 (Abdollahi & Pradhan, 2021a, 2021b). F1 is a term that refers to a mixture of precision and recall metrics. The MCC gives a value between -1 and $+1$ and is defined as a correlation coefficient between recognized binary classifications and the predicted ones. The IOU factor is calculated by dividing the total number of mutual pixels between the real and classified masks by the total number of present pixels in both masks.

3.3. Experimental setting

We utilized data augmentation to increase the size of our dataset and avoid over-fitting. We used the rotation technique of 90°, 180°, and 270° chased by flipping vertically and horizontally to expand the dataset. The MultiRes-UNet model's training to optimize the loss function was implemented utilizing the widely used Adam optimizer with a learning rate of $1e{-}4$. To avoid overfitting, a dropout probability

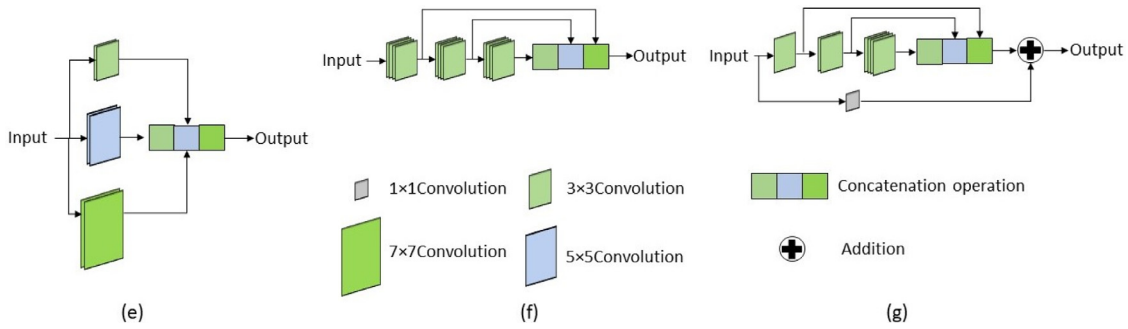

Fig. 1. Design of the proposed MultiRes block operation. (e) Simple inception block that enabled us to adapt spatial features from various scales by utilizing the 3 × 3, 5 × 5, and 7 × 7 convolutions in parallel and merging the produced feature maps. (f) Sequence of lightweight and smaller 3 × 3 convolutions used to factorize the more expensive and larger 5 × 5 and 7 × 7 convolutions. (g) Arrangement of MultiRes block; we added a 1 × 1 filter to maintain dimensions along with a residual connection and gradually incremented the number of filters in the series of three layers.

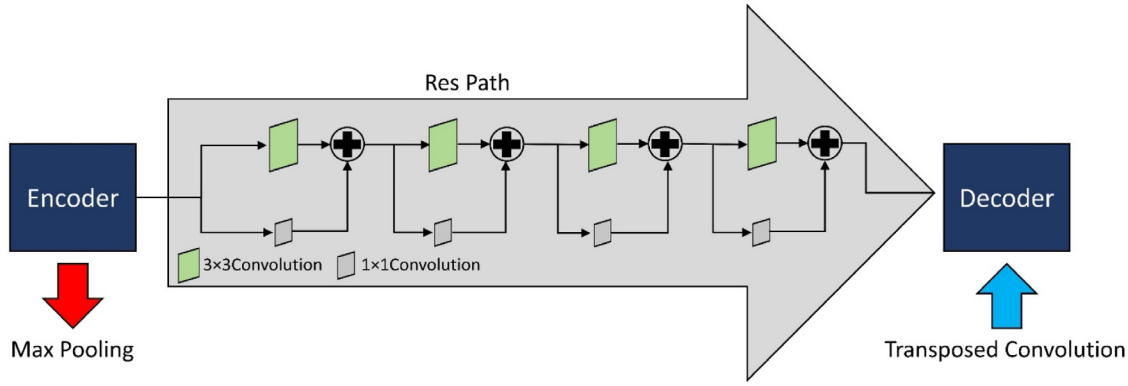

Fig. 2. Suggested Res path structure. We utilized a chain of convolutions to pass the features from the encoder to decoder part instead of directly merging the features maps from the encoder part with those from the decoder part. In addition, to ease the learning process, we suggested residual connections.

Table 1
Architectural details of the presented MultiRes-UNet network.

MultiRes-UNet							
MultiRes blocks	Layer	Size of filter	#Filters	Path	Layer	Size of filter	#Filters
Block 1	Convolution	3 × 3	8		Convolution	3 × 3	64
	Convolution	3 × 3	17		Convolution	1 × 1	64
Block 9	Convolution	3 × 3	26		Convolution	3 × 3	64
	Convolution	1 × 1	51	Res Path 1	Convolution	1 × 1	64
Block 2	Convolution	3 × 3	17		Convolution	3 × 3	64
	Convolution	3 × 3	35		Convolution	1 × 1	64
Block 8	Convolution	3 × 3	53		Convolution	3 × 3	64
	Convolution	1 × 1	105		Convolution	1 × 1	64
Block 3	Convolution	3 × 3	35		Convolution	3 × 3	128
	Convolution	3 × 3	71		Convolution	1 × 1	128
Block 7	Convolution	3 × 3	106	Res Path 2	Convolution	3 × 3	128
	Convolution	1 × 1	212		Convolution	1 × 1	128
Block 4	Convolution	3 × 3	71		Convolution	3 × 3	128
	Convolution	3 × 3	142		Convolution	1 × 1	128
Block 6	Convolution	3 × 3	213		Convolution	3 × 3	256
	Convolution	1 × 1	426	Res Path 3	Convolution	1 × 1	256
Block 5	Convolution	3 × 3	142		Convolution	3 × 3	256
	Convolution	3 × 3	284		Convolution	1 × 1	256
	Convolution	3 × 3	427	Res Path 4	Convolution	3 × 3	512
	Convolution	1 × 1	853		Convolution	1 × 1	512

of 0.5 (Srivastava, Hinton, Krizhevsky, Krizhevsky, & Salakhutdinov, 2014) was used during model training. The suggested network was trained with batch size 1, and sigmoid function was used at the last convolutional operation to generate the probability of 0 or 1 (Shi, Liu, & Li, 2018). The trained network was then implemented to the test data for building extraction. To evaluate the performance, the segmentation maps were compared against the ground truth images. In the current work, the whole process of training and testing the presented network for building detection was executed under TensorFlow backend and Keras framework with a memory of 24 GB, a GPU Nvidia Quadro RTX 600 and a computation capacity of 7.5.

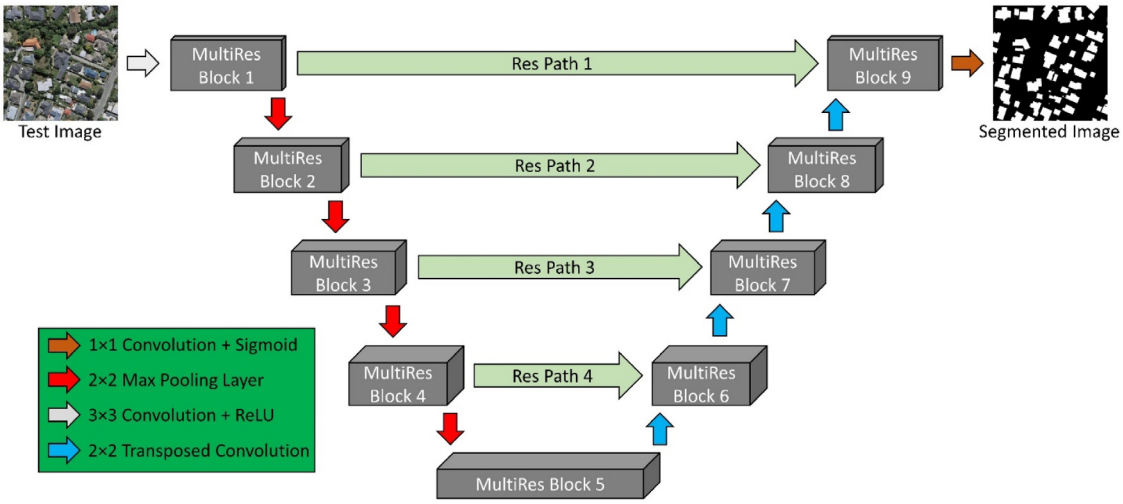

Fig. 3. Architecture of the presented MultiRes-Unet network. We proposed MultiRes block instead of using a series of two convolutional blocks in the original UNet network. Moreover, we replaced the common skip connections with the suggested Res path.

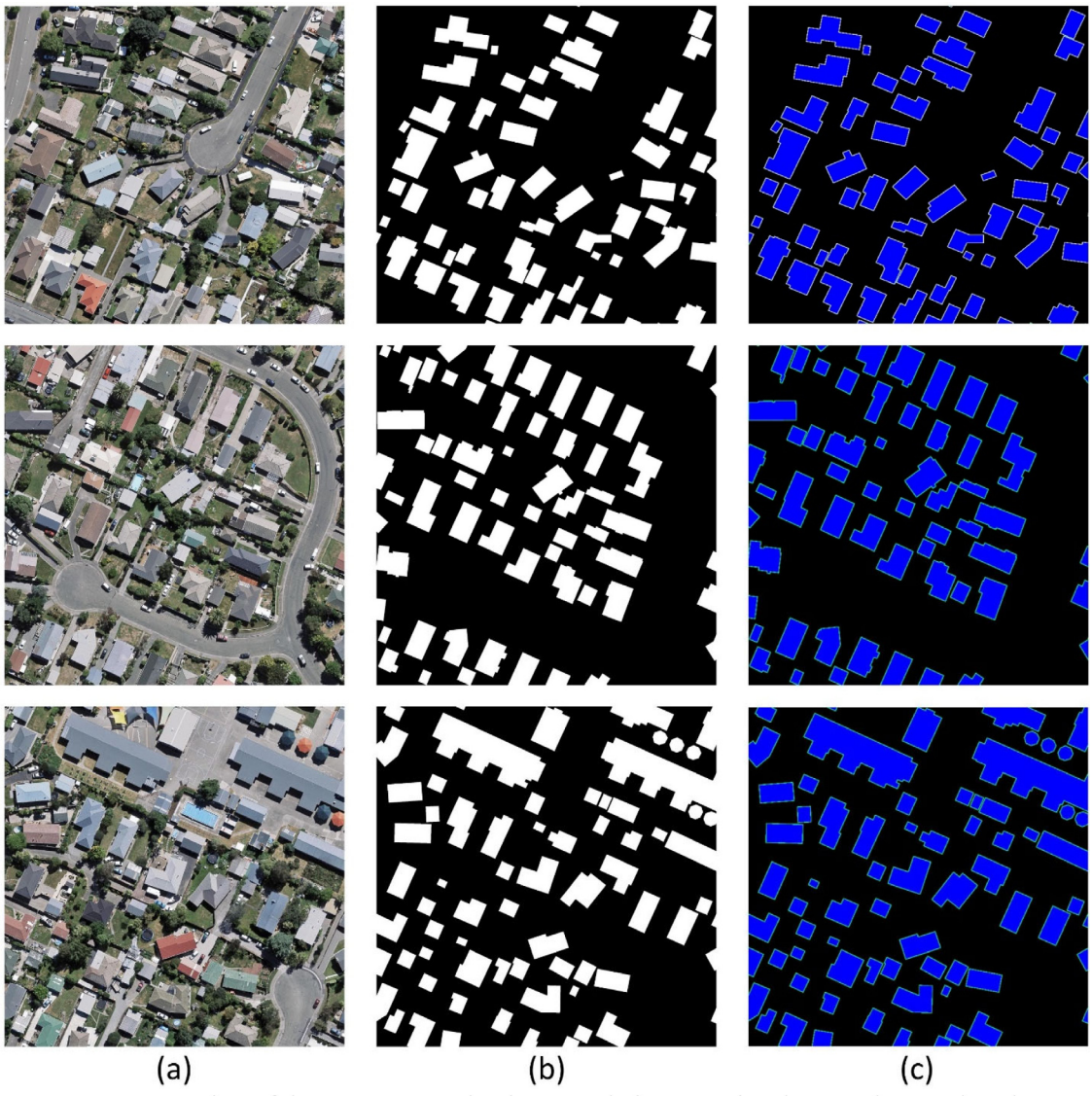

Fig. 4. Demonstration of three representative images, their semantic edges, and ground truth maps for AIRS dataset. (a) exhibit the main RGB images; (b) is corresponding segmentation ground truth maps, and (c) is superposition between semantic edges and segmentation ground truth maps.

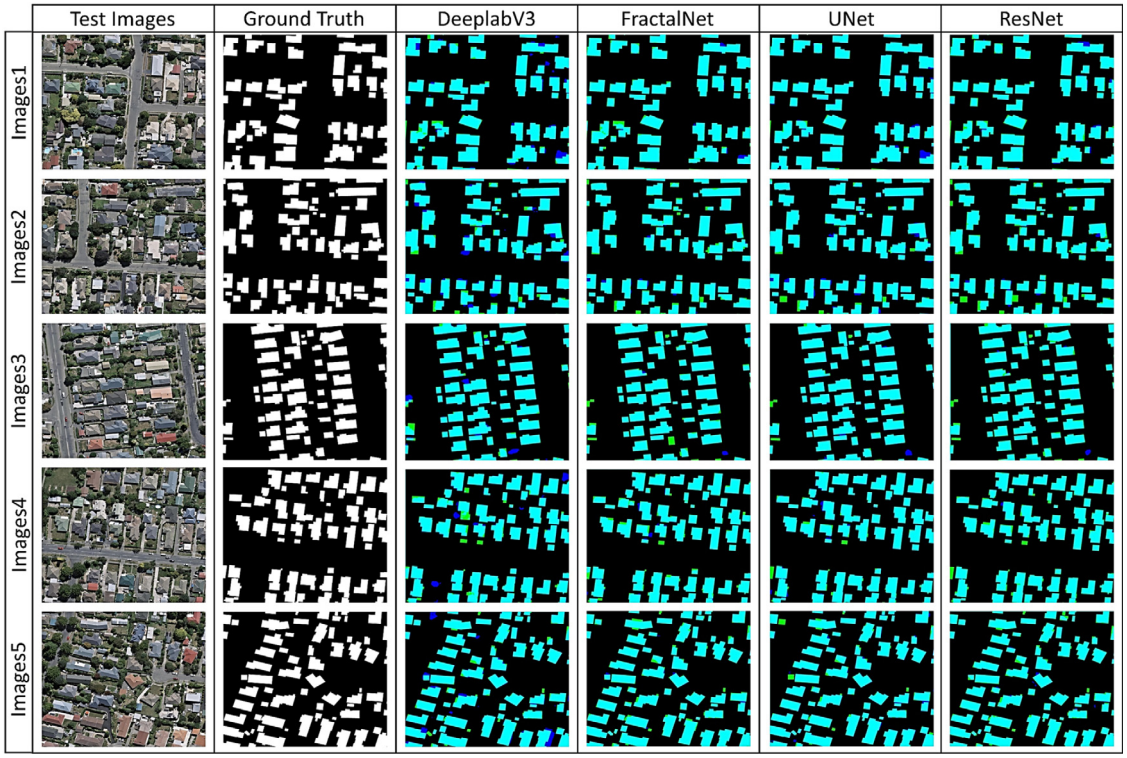

Fig. 5. Visualization outcomes of comparative networks, such as DeeplabV3, FractalNet, UNet and ResNet. FPs, FNs, and TPs are exhibited in blue, yellow, and white colors, respectively. . (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.4. Results of experiments

We compared the presented MultiRes-UNet network to other state-of-the-art deep learning techniques such as UNet network (Ronneberger et al., 2015), DeeplabV3 architecture (Chen, Papandreou, Kokkinos, Murphy, & Yuille, 2017), ResNet framework (He et al., 2016), and FractalNet network (Larsson, Maire, & Shakhnarovich, 2016) to validate its results. Fig. 5 exhibits the visual results attained by other comparative techniques. There are six columns and five rows in this figure. In the first and second columns, the RGB and reference images are exhibited, respectively. The third, fourth, fifth, and sixth columns display the outcomes attained by DeeplabV3, FractalNet, UNet, and ResNet architectures, respectively. Also, Fig. 6 shows the qualitative results obtained by the proposed MultiRes-UNet model without and with augmentation techniques.

Figs. 5 and 6 show that the proposed MultiRes-UNet model and other comparative techniques can obtain accurate building segmentation maps in general. However, the proposed MultiRes-UNet network can produce more accurate building maps than the others. The network predicted less FPs (exhibited in blue color) and less FNs (exhibited in yellow color), which led to the preservation of the boundary information of building objects and generation of high-resolution building segmentation maps. Such a finding may be due to the use of MultiRes block and Res path in the network. In Res path, the semantic gaps between encoder and decoder features are expected to lessen because a chain of convolutional operations was used to pass the features from the encoder to the decoder instead of directly merging the features maps from encoder part with those from decoder part. MultiRes block was also used to reconcile spatial features from various context sizes. Thus, the performance of the MultiRes-UNet network for building detection was improved using these modifications.

Furthermore, we computed the performance assessment metrics to determine the efficacy of the introduced network for building object detection. Table 2 exhibits the correctness of every particularized judgment factor for building extraction. The bold and underlined values depict the best and second-best values, respectively. As the table illustrates, ResNet method achieved the least amount of IOU with 88.84% for building detection. DeeplabV3 and FractalNet methods ranked the fourth and third methods in building detection, with IOU of 89.48% and 90.91%, respectively. The UNet network was ranked the second-best method for building extraction with achieving the IOU accuracy of 92.40%. The UNet model can improve the values of IOU to 1.49% compared with the FractalNet, which is the third-best method. By contrast, the proposed MultiRes-UNet network can achieve higher accuracy of IOU than all the other methods. The network achieved 93.14% accuracy for IOU, which is 0.74% and 2.23% higher than that of the second- (UNet) and third-best (FractalNet) methods, respectively. In addition, we compared the quantitative results achieved by the proposed model without and with data augmentation. The proposed model could improve the IOU accuracy to 3.1% compared to the MultiRes-UNet model without data augmentation. In summary, the results proved the capability of the MultiRes-Unet model for building semantic segmentation from aerial imagery. Fig. 7 also demonstrates the proposed model's performance accuracy on training and validation datasets over 100 epochs. Depending on the reduction in model loss and increase in model accuracy over time, the model has learnt efficient features to classify the images and extract building areas. In fact, the training and validation accuracy/loss are close together in the learning curve and the model reduced over-fitting.

4. Discussion

For accurate detection of the building objects from aerial imagery, we incorporated semantic edges with the semantic polygons. Fig. 8 indicates the visualization results attained by the MultiRes-UNet network before and after adding the semantic edges to the semantic polygons. The figure is presented in four columns and three rows. In the first and second columns, the original RGB imagery and corresponding label maps are shown, respectively, while the results of the building detection before and after integration of the semantic edges are demonstrated in the third and fourth columns, respectively. As the

Table 2
Accuracy assessment factors for computing the quantitative results attained by the MultiRes-UNet and other comparative networks. The underlined and bold values show the second-best and best values, respectively.

		Image1	Image2	Image3	Image4	Image5	Average
DeeplabV3	F1	0.9462	0.9359	0.9516	0.9432	0.9455	0.9445
	MCC	0.9233	0.9075	0.9327	0.9177	0.9188	0.9200
	IOU	0.8978	0.8794	0.9076	0.8925	0.8966	0.8948
FractalNet	F1	0.9541	0.9472	0.9519	0.9529	0.9561	0.9524
	MCC	0.9363	0.9263	0.9353	0.9334	0.9368	0.9336
	IOU	0.9121	0.8997	0.9081	0.9099	0.9158	0.9091
UNet	F1	0.9631	0.9509	0.9651	0.9635	0.9597	<u>0.9605</u>
	MCC	0.9475	0.9294	0.9516	0.9471	0.9403	<u>0.9432</u>
	IOU	0.9288	0.9064	0.9326	0.9296	0.9224	<u>0.9240</u>
ResNet	F1	0.9441	0.9294	0.9412	0.9443	0.9456	0.9409
	MCC	0.9232	0.9038	0.9221	0.9232	0.9230	0.9191
	IOU	0.8940	0.8680	0.8889	0.8944	0.8969	0.8884
MultiRes-UNet without augmentation	F1	0.9501	0.9222	0.9550	0.9583	0.9518	0.9475
	MCC	0.9302	0.8913	0.9387	0.9402	0.9297	0.9260
	IOU	0.9049	0.8556	0.9138	0.9198	0.9080	0.9004
MultiRes-UNet with augmentation	F1	0.9707	0.9541	0.9681	0.9660	0.9636	**0.9645**
	MCC	0.9585	0.9343	0.9558	0.9509	0.9460	**0.9491**
	IOU	0.9430	0.9122	0.9381	0.9342	0.9297	**0.9314**

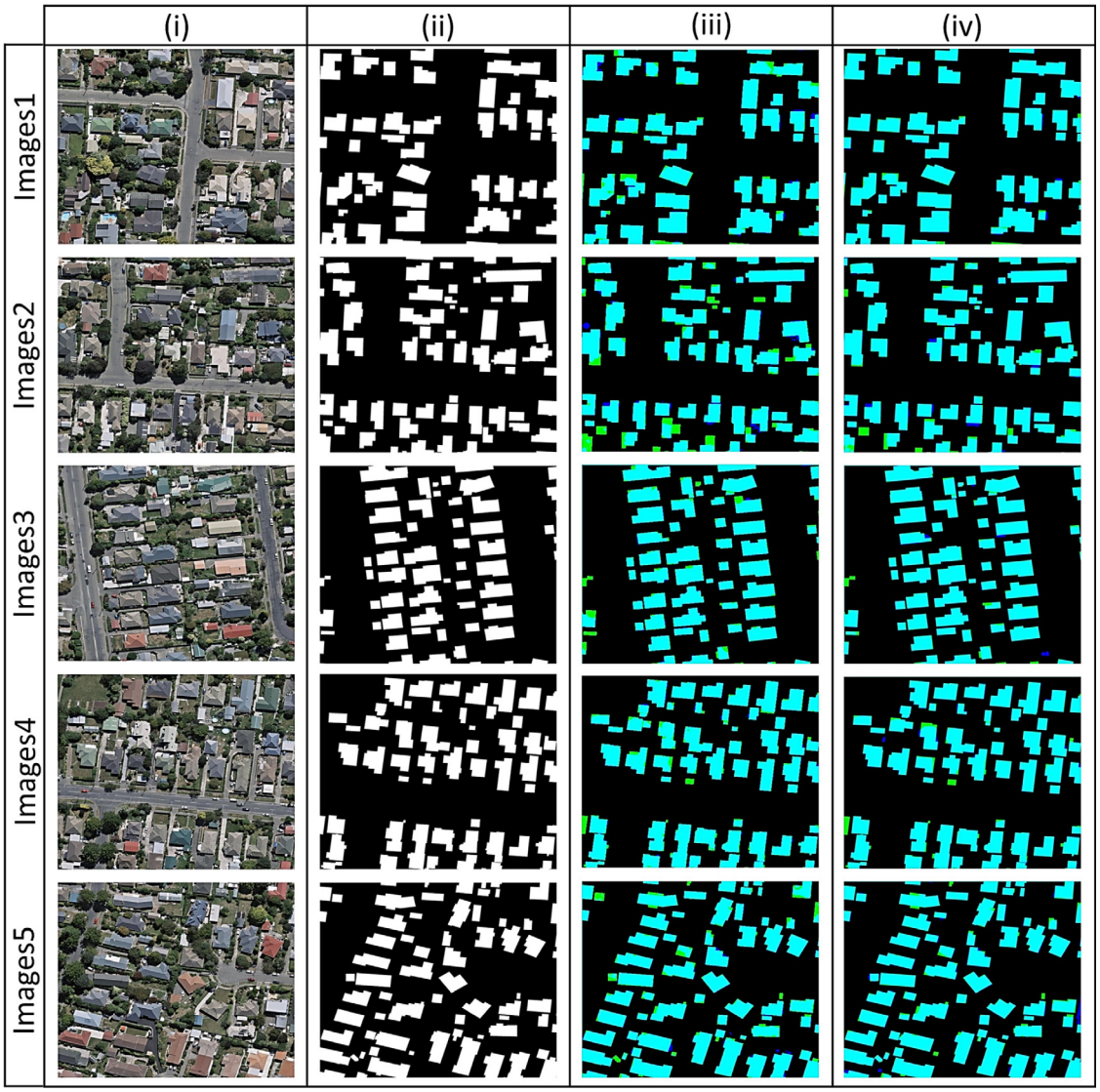

Fig. 6. Visualization outcomes of the proposed model. (i) original images, (ii) ground truth images, (iii) results of MultiRes-UNet without augmentation, and (iv) results of MultiRes-UNet with augmentation. FPs, FNs, and TPs are exhibited in blue, yellow, and white colors, respectively. . (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

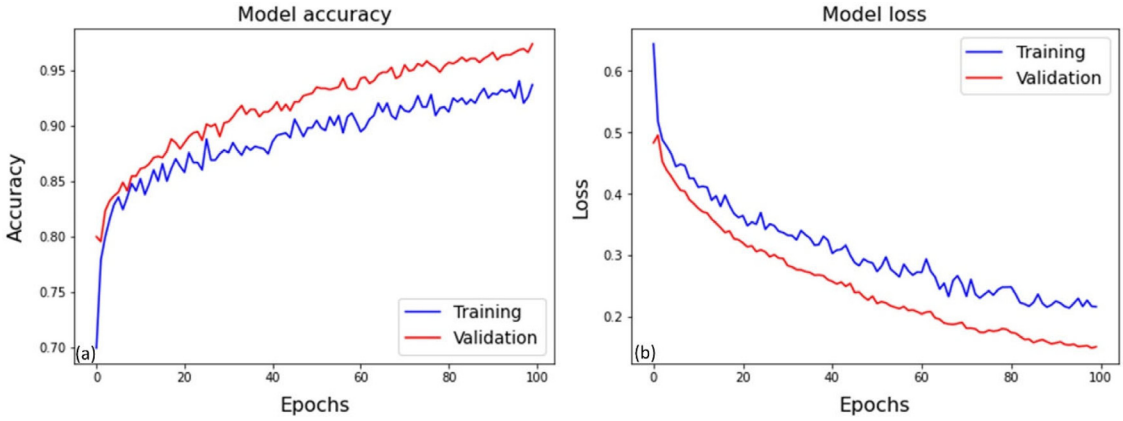

Fig. 7. (a) Model accuracy, and (b) model loss of the proposed MultiRes-UNet Network.

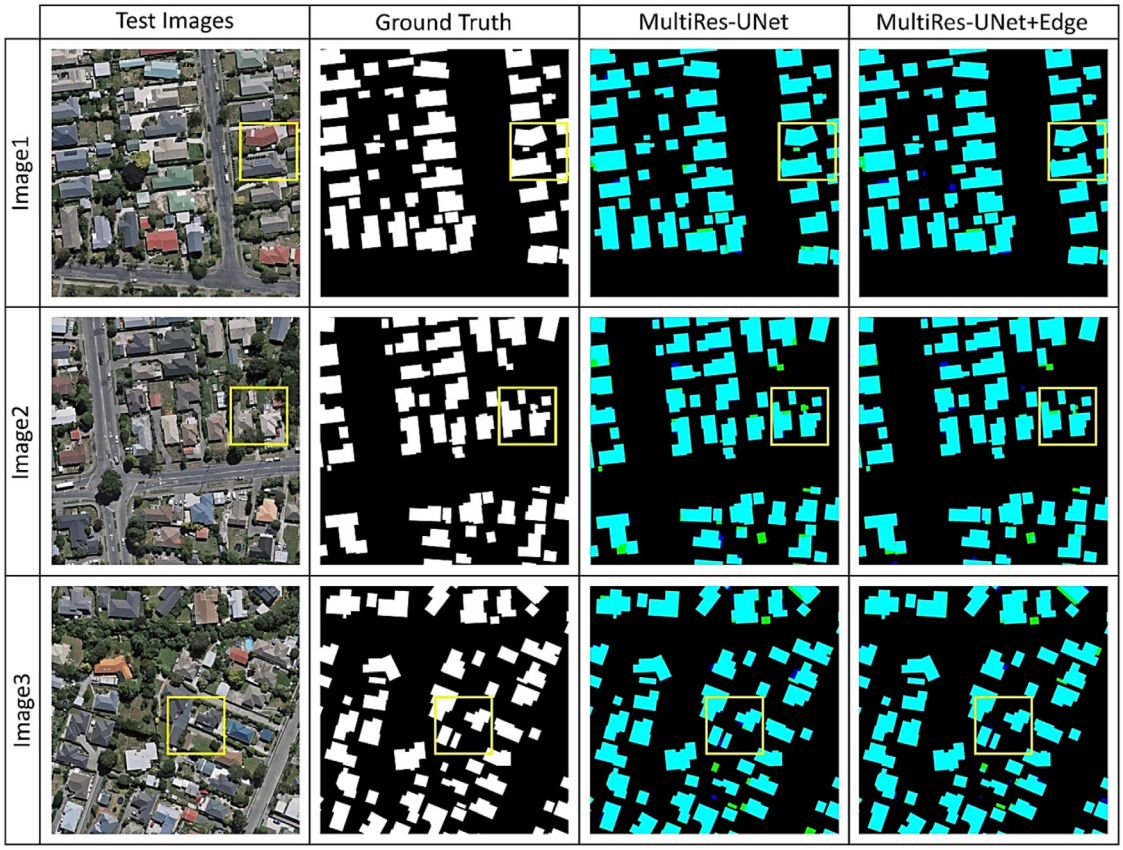

Fig. 8. Visualization outcomes of the proposed MultiRes-UNet network before and after integration of semantic edge information. The FPs, FNs, and TPs are exhibited in blue, yellow, and white colors, respectively. The yellow boxes show the FP and FN prediction. . (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

figure shows, the proposed network could improve the quantitative results after integrating semantic edges and generate high-resolution building segmentation maps. This is because the proposed model could reduce the number of FP pixels and detect the boundary of buildings more accurately after using building edges. Thus, by incorporating semantic edges with semantic buildings, the proposed technique's performance in building detection was increased. In fact, we solved the issue of incompleteness and distinctive of semantics edges and realized complete buildings semantic extraction using semantic polygons. Furthermore, we recognized the distinction between adjacent buildings, adapted semantic polygons to the real building's shape, solved irregular semantic polygons issues, and strengthened the semantic polygon boundaries using semantic edges, which leads to improving the boundary of semantic polygons and creating high-resolution building segmentation maps. In addition, we assessed the performance metrics to realize the effect of adding semantic edges for the proposed model in building detection. Table 3 indicates the qualitative results attained by the model before and after the integration of semantic edges. As the table depicts, the proposed MultiRes-UNet model can achieve 96.56%, 95.16%, and 93.35% accuracy for the F1, MCC, and IOU, respectively, before adding the semantic edges information. In contrast, the model can obtain 96.98%, 95.73%, and 94.13% accuracy for the F1, MCC, and IOU metrics, respectively, after using the edge information. The proposed network can enhance the results of F1, MCC, and IOU to 0.42%, 0.57%, and 0.78%, respectively, which confirmed the influence

Table 3
Accuracy assessment factors for assessing the quantitative results attained by the MultiRes-UNet network before and after integrating the semantic edge information.

		Image1	Image2	Image3	Average
MultiRes-UNet	F1	0.9715	0.9600	0.9654	0.9656
	MCC	0.9592	0.9447	0.9509	0.9516
	IOU	0.9445	0.9230	0.9330	0.9335
MultiRes-UNet + Edge Information	F1	0.9733	0.9659	0.9702	**0.9698**
	MCC	0.9617	0.9525	0.9576	**0.9573**
	IOU	0.9480	0.9339	0.9420	**0.9413**

*Bold values show the best values.

of edge information in identifying the difference between adjacent buildings and improving the boundary of semantic polygons.

5. Conclusion

In this research, we executed a new deep learning structure named MultiRes-UNet model, which is a modified version of the original UNet network, to detect buildings from aerial imagery. In the proposed model, we used MutiRes block to adapt spatial features from various scales and used Res path with a collection of convolutions for passing the encoder features to the decoder section rather than combining the features from the encoder with those from the decoder straightly. We trained our model based on the AIRS dataset containing over 220,000 buildings with a spatial resolution of 7.5 cm and a broad coverage of aerial images. Moreover, semantic edge information was integrated with the semantic building to make semantic polygons more proper for real buildings form and improve the accuracy of buildings boundaries. After integration, the quantitative results demonstrated that the proposed network can enhance the results of IOU to 0.78%, which confirmed the influence of edge information in recognizing the difference between adjacent buildings and improving the boundary of semantic polygons. In addition, we used state-of-the-art comparative models to show the competency of the introduced network in building semantic segmentation. The experiential consequences declared the success of the advised network for building object extraction from aerial imagery.

CRediT authorship contribution statement

Arnick Abdollahi: Conceptualization, Methodology, Data curation, Writing – original draft. **Biswajeet Pradhan:** Conceptualization, Writing – review & editing, Supervised the project including funding.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

The Centre for Advanced Modelling and Geospatial Information Systems, Faculty of Engineering and IT, University of Technology Sydney, Australia, funded this research.

References

Abdollahi, A., & Pradhan, B. (2021a). Integrated technique of segmentation and classification methods with connected components analysis for road extraction from orthophoto images. *Expert Systems with Applications*, Article 114908. http://dx.doi.org/10.1016/j.eswa.2021.114908.

Abdollahi, A., & Pradhan, B. (2021b). Urban vegetation mapping from aerial imagery using explainable AI (XAI). *Sensors*, *21*, 4738.

Abdollahi, A., Pradhan, B., & Alamri, A. M. (2020). An ensemble architecture of deep convolutional segnet and unet networks for building semantic segmentation from high-resolution aerial images. *Geocarto International*, 1–13.

Abdollahi, A., Pradhan, B., & Alamri, A. (2021). RoadVecNet: a new approach for simultaneous road network segmentation and vectorization from aerial and google earth imagery in a complex urban set-up. *GIScience & Remote Sensing*, 1–24. http://dx.doi.org/10.1080/15481603.2021.1972713.

Abdollahi, A., Pradhan, B., Gite, S., & Alamri, A. (2020). Building footprint extraction from high resolution aerial images using generative adversarial network (GAN) architecture. *IEEE Access*, Article 209517-209527. http://dx.doi.org/10.1109/ACCESS.2020.3038225.

Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). Segnet: A deep convolutional encoder–decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis Machine Intelligence*, *39*, 2481–2495.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis Machine Intelligence*, *40*, 834–848.

Chen, D., Shang, S., & Wu, C. (2014). Shadow-based building detection and segmentation in high-resolution remote sensing image. *Journal of Multimedia*, *9*, 181–188.

Chen, Q., Wang, L., Wu, Y., Wu, G., Guo, Z., & Waslander, S. L. (2018). Aerial imagery for roof segmentation: A large-scale dataset towards automatic mapping of buildings. *ISPRS Journal of Photogrammetry and Remote Sensing*, *147*, 42–55.

Drozdzal, M., Vorontsov, E., Chartrand, G., Kadoury, S., & Pal, C. (2016). The importance of skip connections in biomedical image segmentation. In *Deep learning and data labeling for medical applications* (pp. 179–187). http://dx.doi.org/10.1007/978-3-319-46976-8_19.

Dunaeva, A. V. e., & Kornilov, F. A. (2017). Specific shape building detection from aerial imagery in infrared range. *Vychislitelnaya Matematika i Informatik*, *6*, 84–100.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Hong, S., Noh, H., & Han, B. (2015). Decoupled deep neural network for semi-supervised semantic segmentation. In *Advances in neural information processing systems* (pp. 1495–1503). Available from: https://arxiv.org/abs/1506.04924.

Ibtehaz, N., & Rahman, M. S. (2020). MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation. *Neural Networks*, *121*, 74–87.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. (pp. 448–456). Available from: https://arxiv.org/abs/1502.03167.

Ji, S., Wei, S., & Lu, M. (2018). Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Transactions on Geoscience Remote Sensing*, *57*, 574–586.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, *60*, 84–90.

Larsson, G., Maire, M., & Shakhnarovich, G. (2016). Fractalnet: Ultra-deep neural networks without residuals. (pp. 1–11). Available from: https://arxiv.org/abs/1605.07648.

Li, Y., & Wu, H. Adaptive building edge detection by combining LiDAR data and aerial images. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 37*, 197-202.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition, Vol. 39* (pp. 3431–3440).

Maggiori, E., Tarabalka, Y., Charpiat, G., & Alliez, P. (2016). Convolutional neural networks for large-scale remote-sensing image classification. *IEEE Transactions on Geoscience Remote Sensing*, *55*, 645–657.

Noh, H., Hong, S., & Han, B. (2015). Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision* (pp. 1520–1528).

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 234–241). 210.1007/1978-1003-1319-24574-24574_24528.

Shi, Q., Liu, X., & Li, X. (2018). Road detection from remote sensing images by generative adversarial networks. *IEEE Access*, *6*, 25486–25494.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. https://arxiv.org/abs/1409.1556.

Sirmacek, B., & Unsalan, C. (2008). 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). http://dx.doi.org/10.1109/ISCIS.2008.4717854.

Srivastava, N., Hinton, G., Krizhevsky, A., Krizhevsky, A., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*, 1929–1958.

Sumer, E., & Turker, M. (2013). An adaptive fuzzy-genetic algorithm approach for building detection using high-resolution satellite images. *Computers, Environment Urban Systems*, *39*, 48–62.

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence, Vol. 31*. (1). https://ojs.aaai.org/index.php/AAAI/article/view/11231.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1–9).

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818–2826).

Vakalopoulou, M., Karantzalos, K., Komodakis, N., & Paragios, N. (2015). 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).

Volpi, M., & Tuia, D. (2016). Dense semantic labeling of subdecimeter resolution images with convolutional neural networks. *IEEE Transactions on Geoscience Remote Sensing*, *55*, 881–893.

Wang, P., Chen, P., Yuan, Y., Liu, D., Huang, Z., Hou, X., et al. (2018). Understanding convolution for semantic segmentation. In *2018 IEEE winter conference on applications of computer vision (WACV)* (pp. 1451–1460).

Wu, G., Shao, X., Guo, Z., Chen, Q., Yuan, W., Shi, X., et al. (2018). Automatic building segmentation of aerial imagery using multi-constraint fully convolutional networks. *Remote Sensing*, *10*, 407.

Xu, Y., Wu, L., Xie, Z., & Chen, Z. (2018). Building extraction in very high resolution remote sensing imagery using deep learning and guided filters. *Remote Sensing*, *10*, 144.

Yang, H., Wu, P., Yao, X., Wu, Y., Wang, B., & Xu, Y. (2018). Building extraction in very high resolution imagery by dense-attention networks. *Remote Sensing*, *10*, 1768.

Yang, H. L., Yuan, J., Lunga, D., Laverdiere, M., Rose, A., & Bhaduri, B. (2018). Building extraction at scale using convolutional neural network: Mapping of the united states. *IEEE Journal of Selected Topics in Applied Earth Observations Remote Sensing*, *11*, 2600–2614.

Yuan, J. (2017). Learning building extraction in aerial scenes with convolutional networks. *IEEE Transactions on Pattern Analysis Machine Intelligence*, *40*, 2793–2798.

Zeiler, M. D., Krishnan, D., Taylor, G. W., & Fergus, R. (2010). Deconvolutional networks. In *2010 IEEE computer society conference on computer vision and pattern recognition* (pp. 2528–2535).

Zhang, Y. (1999). Optimisation of building detection in satellite images by combining multispectral classification and texture filtering. *ISPRS Journal of Photogrammetry Remote Sensing*, *54*, 50–60.

Zhong, S., Huang, J., & Xie, W. (2008). A new method of building detection from a single aerial photograph. In *2008 9th international conference on signal processing* (pp. 1219–1222).

Zhong, C., Xu, Q., Yang, F., & Hu, L. (2015). Building change detection for high-resolution remotely sensed images based on a semantic dependency. In *2015 IEEE international geoscience and remote sensing symposium (IGARSS)* (pp. 3345–3348).