Hindawi
Scientific Programming
Volume 2023, Article ID 8552624, 15 pages
https://doi.org/10.1155/2023/8552624

Research Article

Architecture of Deep Convolutional Encoder-Decoder Networks for Building Footprint Semantic Segmentation

Abderrahim Norelyaqine,[1] **Rida Azmi**,[2] **and Abderrahim Saadane**[3]

[1]*Department of Mineral Engineering, Mohammedia School of Engineers, Rabat 10090, Morocco*
[2]*Center of Urban Systems– CUS, Mohammed VI Polytechnic University (UM6P), Benguerir 43150, Morocco*
[3]*Department of Geology, Faculty of Sciences of Rabat, Rabat 10090, Morocco*

Correspondence should be addressed to Abderrahim Norelyaqine; norelyaqine.abdou@gmail.com

Received 6 May 2022; Revised 19 July 2022; Accepted 2 September 2022; Published 25 April 2023

Academic Editor: Tongguang Ni

Building extraction from high-resolution aerial images is critical in geospatial applications such as telecommunications, dynamic urban monitoring, updating geographic databases, urban planning, disaster monitoring, and navigation. Automatic building extraction is a massive task because buildings in various places have varied spectral and geometric qualities. As a result, traditional image processing approaches are insufficient for autonomous building extraction from high-resolution aerial imaging applications. Automatic object extraction from high-resolution images has been achieved using semantic segmentation and deep learning models, which have become increasingly important in recent years. In this study, the U-Net model was used for building extraction, initially designed for biomedical image analysis. The encoder part of the U-Net model has been improved with ResNet50, VGG19, VGG16, DenseNet169, and Xception. However, three other models have been implemented to test the performance of the model studied: PSPNet, FPN, and LinkNet. The performance analysis through the intersection of union method has shown that U-Net with the VGG16 encoder presents the best results compared to the other models with a high IoU score of 83.06%. This research aims to examine the effectiveness of these four approaches for extracting buildings from high-resolution aerial data.

1. Introduction

Collecting urban geographic information and updating data timely are crucial and vital challenges for better management of cities in the fast urbanization and building of megacities. The accuracy of information extraction may be considerably improved by using high-resolution remote sensing images. Experts and scholars from all over the world have focused on remote sensing data classification methods in recent decades, ranging from supervised and unsupervised classifications based on traditional statistical analysis [1]. Among these, the pixel-based statistical classification approach has emerged as the most popular and well developed, with promising results in particular domains [2]. On the other hand, traditional pixel-based classification algorithms primarily use spectral data and have limited effectiveness in categorizing high-resolution multispectral urban images with similar spectra into separate categories [3]. To develop more accurate categorization maps, geographic information such as geometric and spatial characteristics and textural information must be used.

In recent years, object-oriented classification algorithms have attracted researchers' interest [3, 4]. It has been demonstrated to have the ability to overcome the suffering from some forms with per-pixel analysis, such as the omission of geometric and contextual information. The fundamental concept is to divide the image into objects with specific meanings and then categorize them using the items spectral, form, and textural properties. This technique considers additional discriminative features and conforms to human visual interpretation patterns, resulting in a new way of thinking about data extraction [5]. While

several studies have demonstrated the benefits of object-based classification over pixel-based classification, there has been less focus on its possible drawbacks. However, the object-based technique has its own set of constraints. Image segmentation errors include both over-segmentation and under-segmentation. These segmentation issues can affect the categorizing process in two ways: (1) Poorly segmented image objects with over-segmentation or under-segmentation errors produce image objects that span multiple classes, introducing classification errors because all pixels in each mixed image object must be assigned to the same class; (2) features extracted from poorly segmented image objects with over-segmentation or under-segmentation errors do not represent the properties of real objects on the Earth's surface (e.g., shape and surface area), so they may not be useful and may even reduce the accuracy of the classification.

Image segmentation is an essential and vital phase in (GEographic) Object-Based Image Analysis (GEOBIA or OBIA). The quality of image segmentation significantly influences the final feature extraction and classification in OBIA. In traditional segmentation methods, images are usually divided into several disjoint regions based on grayscale, color, texture, and shape. Typical segmentation methods include pixel-based statistical classification, segmentation methods based on thresholds, edges, regions, and graph theory, and object-based image segmentation.

The basic idea of the threshold-based segmentation method is to calculate the grayscale threshold based on the grayscale features of the image and compare the grayscale value of each pixel of the image with the threshold to obtain its category. For example, Li et al. [6] used the wavelet transform and adaptive global threshold method to extract the labelling information of building groups according to the distribution and texture characteristics of building groups to achieve segmentation; Wu et al. [7] proposed a method based on the line intercept histogram. Multi-threshold segmentation methods and edge-based segmentation methods [8] mainly perform edge detection based on the sudden change of image edge grayscale, color, texture, and other features. Differential operators such as Prewitt [9] perform edge detection on the image, identify the edge information of the image, and complete the segmentation. The basic idea of the segmentation method based on graph theory is to associate the image segmentation problem with the minimum segmentation problem of the graph and finally realize the segmentation effect. For example, Felzenszwalb et al. [10] introduced an image segmentation method based on graph representation, proposed a variable component model algorithm based on the greedy clustering algorithm, and established a segmentation algorithm based on graph theory. However, due to the rich spectral information contained in remote sensing images, traditional feature extraction methods still have significant limitations for demanding remote sensing image segmentation application scenarios, and their classification accuracy cannot meet the actual needs for dealing with huge image data and serious image interference. Therefore, traditional classifiers are unsuitable for complex image classification and, more precisely, building extraction.

Urban system studies are promising for using particular resolution and very high spatial satellite image data. Thus, for Earth monitoring, the development of various sensors has substantially expanded the availability of high-resolution remote sensing images since the launch of the first satellite, giving accurate terrestrial scene interpretation and an enormous potential for meaningful. Identifying rooftops is one of the most challenging satellite image analyses, but essential tasks for object extraction. Many remote sensing applications, such as disaster monitoring, geographic databases, urban planning, etc., can benefit from this data. However, with high spatial and spectral quality RS data, manually distinguishing buildings from other objects and delineating their outlines are time consuming and costly. As a result, there have been several attempts to develop automated building extraction technologies.

Some algorithms for building detection based on high-resolution satellite and aerial data use specific building appearance criteria, such as uniform spectral reflectance values. The fundamental issue with these techniques is that the building is confused with other objects having similar spectral reflectance. Many methods for building extraction use multispectral images that provide for a scene set criteria height information, like relatively homogenous structures following a given pattern. However, these techniques are severely constrained since the established criteria only work for specific types of buildings and do not apply to regions with complex and varied structures. Different data sources might provide each other with complementing information.

In recent years, deep learning has shown significant promise for meeting the challenging demands of remote sensing image processing. Deep learning has shown to be a very effective collection of technologies in recent years, sometimes even surpassing human abilities to perform highly computational jobs. The RS community's interest in deep learning approaches is expanding rapidly, and several architectures have been developed in recent years to handle RS difficulties, frequently with excellent results. Deep learning is an emerging machine learning algorithm that has attracted extensive attention from researchers because of its remarkable effect on image feature learning. Compared with traditional image classification methods, it does not require artificial feature description and extraction of target images but learns features from training samples autonomously through neural networks and extracts higher-dimensional and abstract features, and these features related to the classifier are closely related and solves the difficult problem of manual feature extraction and classifier selection. It is an end-to-end model. The essential advantage of the deep learning-based image classification method compared to the traditional image classification method is that it can automatically learn more abstract data features through the deep architecture without designing specific artificial features for specific image data or classification methods, significantly improving the performance of image classification. Deep Learning (DL) outperforms its predecessors significantly; it is based on a traditional neural network. Furthermore, in order to construct multi-layer learning models, DL uses both transformations and graph technologies at the same time.

The latest DL algorithms have achieved excellent results in various applications, including natural language processing (NLP), visual data processing, and audio and voice processing. Convolutional neural networks (CNNs) with more hidden layers have a more complicated network structure and can learn and express features more effectively than classic machine learning approaches [11, 12]. In remote sensing, the use of CNN has become crucial with the appearance of multispectral data at a very high spatial resolution. However, Figure 1 shows the number of publications in the last six years that use CNN and different techniques to classify high-resolution satellite data. This exponential number of publications shows the importance of the deep learning approach in automatic object recognition from high and very high spatial and spectral resolution images.

High-resolution remote sensing images have rich spatial information but contain fewer bands. In order to extract abstract features with sufficient discriminative power and robustness, in recent years, people mainly automatically extract deep-level features from image data through learning methods. CNN is commonly used in remote sensing image classification and can be divided into patch-based CNN and fully convolutional neural network (FCN) [13]. Patch-based CNN can effectively learn the spatial-spectral joint features of the pixels to be classified and their neighborhoods and has been widely used in the field of hyperspectral classification [14]. However, the network has a large number of repeated computations, which limits its application in large-scale high-resolution remote sensing imagery tasks. The trained FCNN can classify all the input image pixels through one forward pass, which is more efficient than the patch-based CNN [15]. Therefore, FCNN is widely used in large-scale high-resolution remote sensing image building extraction tasks [16]. The learning of image features by CNN is realized by optimizing each layer of convolution kernels in the network. The static structure of the network determines the mode of feature learning, and the data determines the specific feature extraction results, thus showing certain robustness. The feature fusion methods used in Residual Networks (ResNet) and DenseNets (DenseNet), that is, feature map addition and feature map connection, have a profound impact on CNN optimization research.

Building extraction from remote sensing images and comparing the performance of different models of the semantic segmentation network is our primary motivation for this paper. However, this paper allowed us to:

(i) show the importance of the deep learning model in the classification of satellite images with very high spatial resolution;

(ii) minimize subjectivity in urbanized areas with the most important step in the classification process being segmentation;

(iii) compare the four improved DL architecture (U-Net, LinkNet, FPN, and PSPNet) with five different initialized and pre-trained encoders (VGG16, VGG19, ResNet50, DenseNet169, and Xception);

(iv) improve overall classification accuracy in the Massachusetts aerial image dataset.

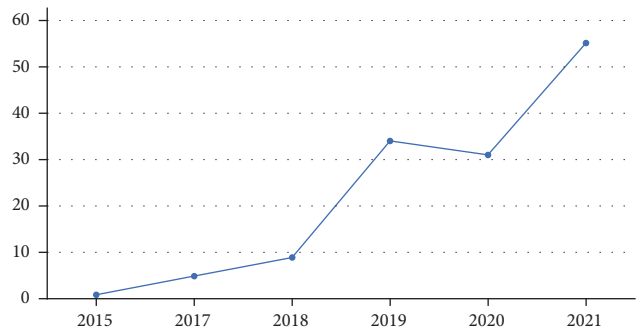


Figure 1: Number of published papers using CNN in the last six years according to the Scopus database.

2. Related Works

2.1. Semantic Segmentation. The semantic segmentation of remote sensing images aims to assign a land cover label to each pixel in the image, which can be understood as a pixel-level classification problem. Fully convolutional neural networks (FCN) were suggested by the authors in [17] to overcome the limitations of convolutional neural networks applied to the field of semantic segmentation. FCN has usually adopted an encoder-decoder system, with the encoder being a subsampling network, which is mainly used to learn multilevel semantic features. The decoder is generally defined as an oversampling network and is used primarily to map the semantic features learned by the encoder to the pixel space of the original resolution for pixel-level classification. Currently, in the field of remote sensing, researchers have made many improvements to FCN based on the characteristics of remote sensing images. For example, considering the rich and diverse categories of remote sensing objects and complex boundaries, Long et al. [18] improved the decoder by designing deconvolution and jump connections, improving the extraction effect of the edge details of the remote sensing object. To solve the problem of fuzzy edge details of objects extracted, the FCN method was proposed [19] by reducing the expansion factor of hole convolution to aggregate local features. Aiming at the multiscale problem of ground objects in complex remote sensing scenes, Hamaguchi et al. [20] proposed using a closed convolution neural network to complete information diffusion between feature maps at different levels to achieve multiscale feature fusion. As discussed elsewhere [21], based on the idea of clustered convolution design, an efficient spatial pyramid network with holes is proposed to complete the multiscale information extraction of remote sensing features. Moreover, considering the problem that FCN cannot adaptively take the long-range dependencies between different objects because of the fixed receptive

field, the researchers used recurrent neural networks, self-attention mechanisms, and other methods to model the long-range contexts of remote sensing objects further to improve the semantics of segmentation accuracy [22].

2.2. Building Extraction. Several authors have utilized deep learning models to extract urban features from image data with very high spatial resolution, and with the progress of convolution, the degree of feature abstraction continues to increase, and the receptive field also increases, which inevitably leads to the loss of spatial details. Most of the FCNs used for building extraction use an encoder-decoder structure, which has the characteristics of level-by-level decoding and can recover spatial information. U-Net effectively recovers spatial information by fusing the feature maps of the encoded segment and the corresponding decoded segment and shows excellent potential in the task of building extraction [23]. In addition, buildings in high-resolution images have multi-scale characteristics, and the characteristics of vertical imaging in remote sensing images make their semantic features quite complex. There are many ground objects with similar colors and textures to building roofs.

The U-Net family [24] suggested two innovative classifiers for multi-object segmentation to extract roads and buildings. The multi-level context gating U-Net (MCG-U-Net) and the bi-directional ConvLSTM U-Net model are the two models discussed. The proposed methods generate detailed segmentation maps that preserve boundary information even in complex backgrounds by combining tightly-coupled convolutions, bidirectional ConvLSTM, and squeeze-and-excitation modules. The researchers also devised an essential efficient loss function known as boundary-aware loss (BAL), which allowed a network to focus on complex semantic segmentation regions such as overlapping areas, tiny objects, complex objects, and object boundaries while still delivering high-quality segmentation results. To employ building features from high-resolution aerial images, researchers [25] developed a unique deep neural network called the Seg-U-Net approach, which is a blend of Segnet and U-Net algorithms. They utilized the Massachusetts building dataset for their analysis. Consequently, the accuracy of the contributions increased to 92.73 percent. The authors in [26] established a unique multi-task loss to solve the difficulty of retaining semantic segmentation borders in high-resolution satellite images. The loss is based on differing output representations of the segmentation mask, according to the researchers, and biases the network to focus more on pixels near borders. The authors demonstrate that the technique outperforms state-of-the-art methods by 9.8% on the Intersection over Union (IoU) measure without extra post-processing steps using the Inria aerial image labeling dataset. The U-Net model with ResNet50 as an encoder was used in [27] to increase and improve the accuracy to extract buildings from the Massachusetts dataset.

3. Methodology

3.1. Sematic Segmentation with Fully Convolutional Network. To improve pixel-level image segmentation by traditional CNNs, the authors in [18] proposed a fully convolutional neural network, which achieves high accuracy in image-level classification and regression tasks, usually by connecting multiple fully connected layers after multiple convolutional layers. The N-dimensional feature vector is used to predict the probability value of each N category, and then the category of the input image is obtained. The difference between the above task and the extraction of the building in remote sensing images is that each pixel in the input image is classified to obtain a pixel-level classification result. Although CNN can define sliding windows centered on individual pixels and model the window features to obtain semantic segmentation results at the pixel level, the time complexity increases significantly due to the large amount of duplicate information generated by the overlapping areas between adjacent windows. In addition, the choice of window size will also be a challenge: A too-small window will lose target contextual information and reduce accuracy; a too-large window will increase the computational and memory load.

To address these issues, FCN has outperformed CNN. FCN uses deconvolution to up-sample high-dimensional feature maps to obtain prediction results similar to the input image, rather than utilizing fully connected layers to create feature vectors to forecast probabilities after multilayer convolution and pooling, as shown in Figure 2. This network topology prevents the propagation process from losing spatial information from the input image, allowing each pixel in the image to be predicted. Furthermore, the FCN does not have to perform a window-by-window calculation on the picture, dramatically improving computational efficiency.

Although FCN enhancement can reach the same segmentation result as the input image size, the predicted image is frequently too smooth, resulting in more severe information loss. The fundamental reason for this is that the input image is clustered many times, allowing neurons at the tail end to receive more information, resulting in a broader perceptual field. However, the image loses information, as a result making the edge contours extracted less desirable. Consequently, FCN integrates the low-dimensional feature map into the feature pyramid with the output after deconvolution to overcome the problem mentioned above and increase the accuracy of extracting detailed information. Consequently, U-Net [23] extends this idea of merging low dimensional features with high-dimensional features.

3.2. Model Used

3.2.1. U-Net Architecture. Figure 3 depicts the U-Net structure, consisting of two feature encoding and decoding steps. The raw input is convolved and subsampled layer by layer in the feature encoding step to obtain high-level semantic features with lower spatial resolution. In the decoding step, the underlying features are increased by a factor of 2 layer by layer through the upward convolution operation, concatenated with the same layer features in the encoding step, and returned to

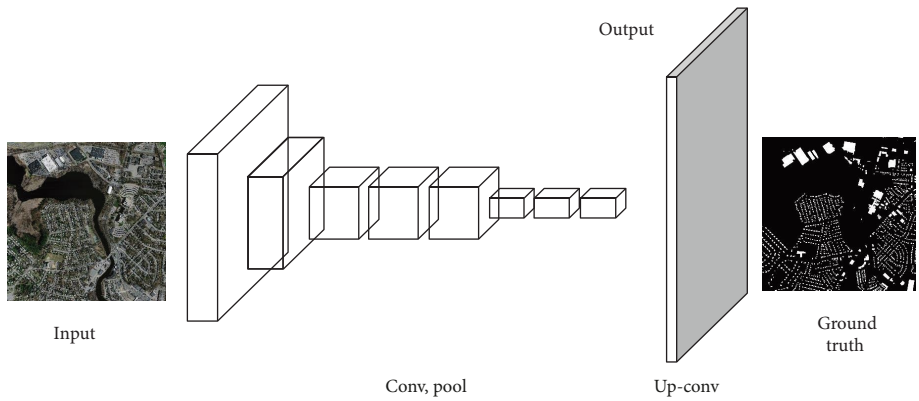


Figure 2: Architecture of fully convolutional neural networks.

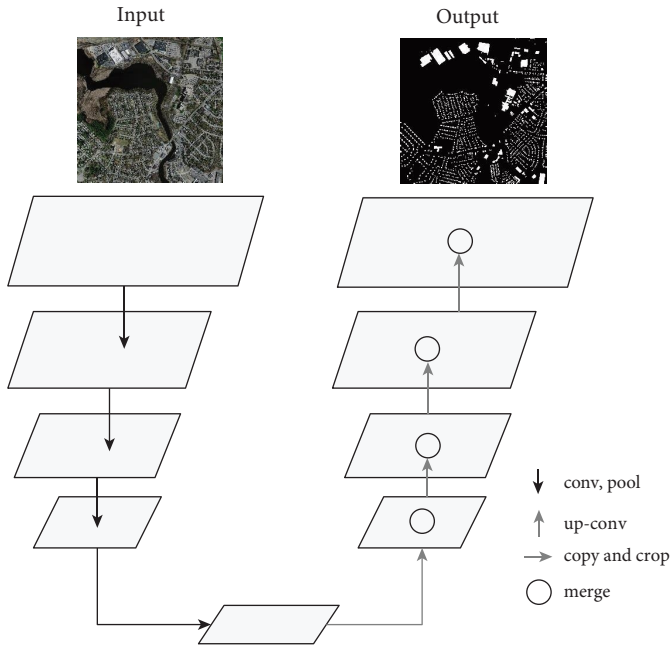


Figure 3: Architecture overview of U-Net.

the original image scale. At the original scale, the difference between the current model predictions and the ground truth reference is used to form the network parameters via back-propagation. U-Net only performs image pixel class classification in the last layer. Although U-Net uses some information from the previous layers in the encoding step, its ability to generalize to multi-scale information is limited.

3.2.2. Pyramid Scene Parsing Network (PSPNet). Multiscale information is also essential for enhancing the accuracy of semantic segmentation. The multiscale receptive field can learn information from objects of different sizes combined with the image scale context. For example, global scene classification can provide category distribution information for semantic image segmentation, and the pyramid clustering module obtains category distribution information by using clustering layers with larger convolution kernels. A spatial pyramid scene parsing network (PSPNet) [28] was proposed to acquire information about the overall scene. As shown in Figure 4, to extract the features from the input image, the convolutional neural network (CNN) model is used and the feature map is sent to the pyramid clustering model. In addition, to extract multiscale information from the images, the model integrates four parallel clustered features of different scales and transforms any size feature map into a fixed-length feature vector. To capture global features, 1x1 convolutions is used to reduce the number of channels to 1/4 of the original size after each clustering operation at different scales. Before ungrouping, the feature maps are restored to their original size using bilinear interpolation, and then connected to the feature maps before pooling. Finally, a convolutional layer generates the final prediction result. The spatial pyramid pooling model leverages distinct spatial information and combines global and local information to get a global understanding of the scene.

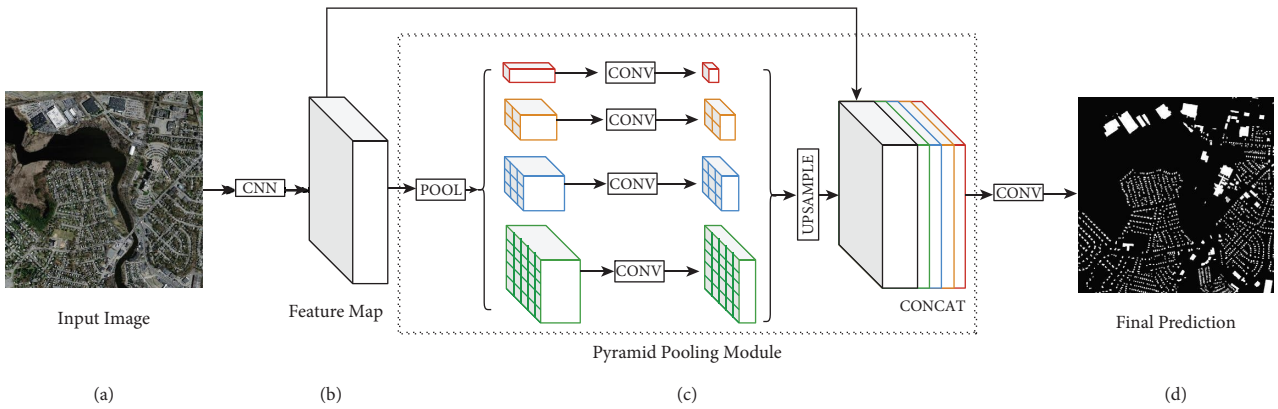


Figure 4: Architecture overview of PSPNet.

3.2.3. LinkNet. LinkNet, a real-time semantic segmentation network, was suggested by [29]. DeconvNet and SegNet employ clustering indices to recover spatial information lost during subsampling, whereas LinkNet sends spatial information directly from the encoder to the matching decoder, conserving as much of the image's spatial information as feasible. As shown in Figure 5, this method directly connects the shallow feature map in the encoder module to the decoder module of the corresponding size, that is, the output of each encoder module is used as the input of the corresponding decoder module, which not only uses the accurate position information of the shallow layer but also avoids adding redundant parameters and computations, resulting in improved computational speed while ensuring accuracy.

3.2.4. Feature Pyramid Network (FPN). Convolution and pooling operations are performed on the original image in convolutional neural networks to create feature maps of various layers and sizes. The network surface layer is more interested in detailed information, but the deep layer is more interested in semantic information, which might assist us in precisely detecting the target. As a result, the typical convolutional neural network makes predictions based on the feature maps of the final convolutional layer. The FPN is an end-to-end network in which feature maps are created through a succession of convolutional processes, predictions are formed at each step, and feature maps are utilized for each prediction layer identified at the appropriate resolution [30]. This guarantees that each layer has sufficient resolution and solid semantic characteristics. By weighing the outcomes of each prediction step, the FPN gets the final loss function. The principle is to accumulate surface and deep features, as the surface features provide more accurate location information. In contrast, the deep network's location information is inaccurate due to multiple subsampling and oversampling operations, and their combined use builds a deeper FPN (Figure 6) that integrates multiple layers of feature information and produces various features.

3.3. Backbone of Network. This study adopts models with end-to-end fully convolutional neural network structures, consisting of a decoder and an encoder. The encoder learns the target features hierarchically to gradually reduce the spatial resolution and gradually increase the receptive field. Among the features learned by the encoder, shallow features have more spatial information, including edge, contour, and location information, while deep features have more semantic category information. The decoder restores the spatial resolution of the features learned by the encoder and produces the prediction results with a similar spatial resolution as the input image. Considering in remote sensing that the scales of buildings images are quite different and there are both large buildings and small residential buildings in the same image, the spatial information lost in the encoding process should be compensated in the network design process, and the features of different scales should be integrated for decoding.

Verify the importance of the depth of the decoder and encoder layers and improve the proposed networks. This paper uses VGG16, VGG19, ResNet50, Densenet169, and Xception as pre-trained encoders on a large ImageNet dataset [31]. The addition of the encoder-decoder module aims mainly at improving the detailed information of the segmentation by restoring the original pixel information.

3.3.1. VGG as Backbone. VGG is a 16–19 layer deep convolutional network used by the Visual Geometry Group (VGG) at the University of Oxford in the 2014 ILSVRC (ImageNet) competition based on the AlexNet network. The model achieves a success rate of 92.5% in the top 5 of the validation set [32]. It inputs a color image of size 224∗224 px and classifies it into one of 1000 classes. Then, it returns a vector of size 1000, which contains the probabilities of belonging to each class. The automatic feature extraction exploits only the convolutional part of a pre-trained network. It uses it as a feature extractor of the images to feed the classifier. Using a multi-scale learning strategy to increase the amount of data, the model shows that the deeper the network, the better the results.

3.3.2. ResNet as Backbone. In [33], a ResNet to solve the problem of degraded deep network learning is proposed. ResNet adds constant mapping using a shortcut structure, which maps features X at the lower level directly to the

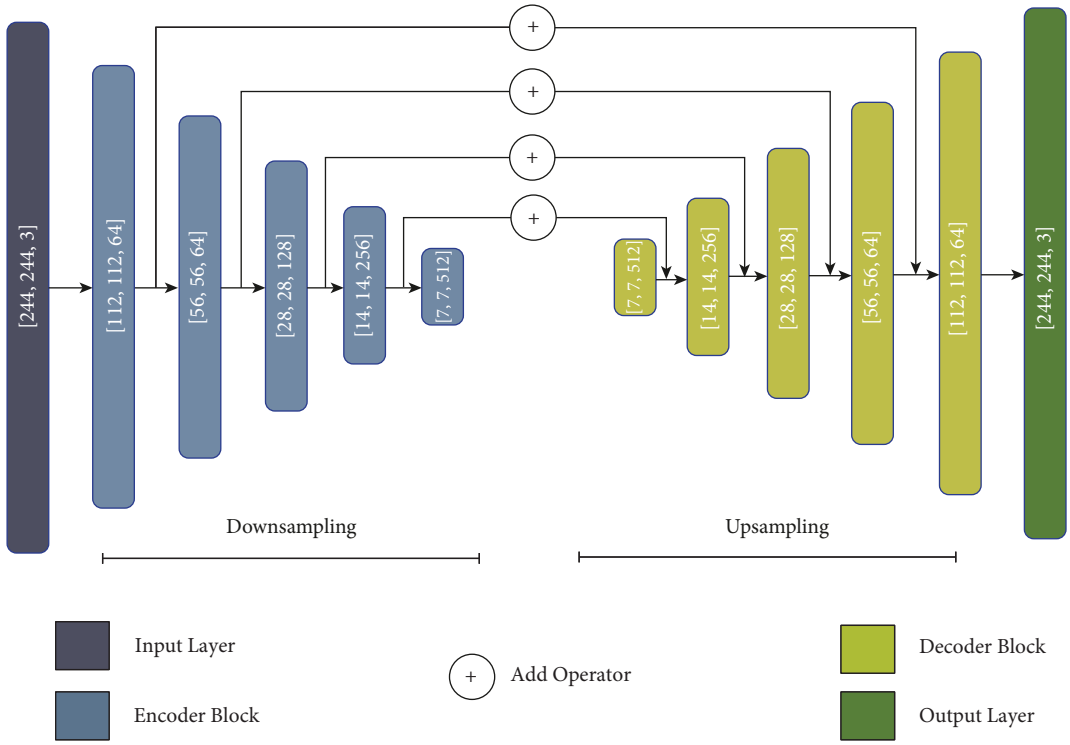


Figure 5: An illustration of the LinkNet architecture.

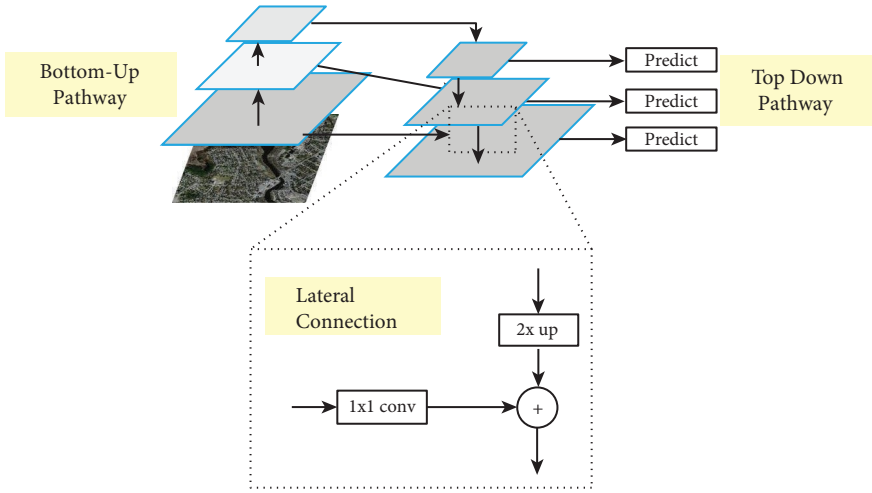


Figure 6: An illustration of the FPN architecture.

network at the higher level. Assuming that the input to a neural network segment is X and the desired output is $H(X)$, the shortcut converts the original learning target $H(X)$ to $H(X)-X$ so that the whole network needs to learn a portion of the difference between the output and the input, simplifying the target and the difficulty of learning the network.

3.3.3. DenseNet as Backbone. Based on the ResNet network, Huang et al. [34] proposed a DenseNet model that connects each layer of the network to all previous layers in a feed-forward way while designing each layer to be particularly narrow and learning very few feature maps to reduce redundancy, which achieves accuracy comparable to ResNet on ImageNet but requires much fewer parameters.

3.3.4. Xception as Backbone. With a separable depthwise convolution, Xception replaces the inception modules [35] and adds residual links. This type of approach considerably, without changing the number of parameters, reduces the use of resources during the matrix calculation.

Usually, the encoder structures in segmentation tasks are similar, mainly derived from the network structures used for classification tasks. This has the advantage that the weighting

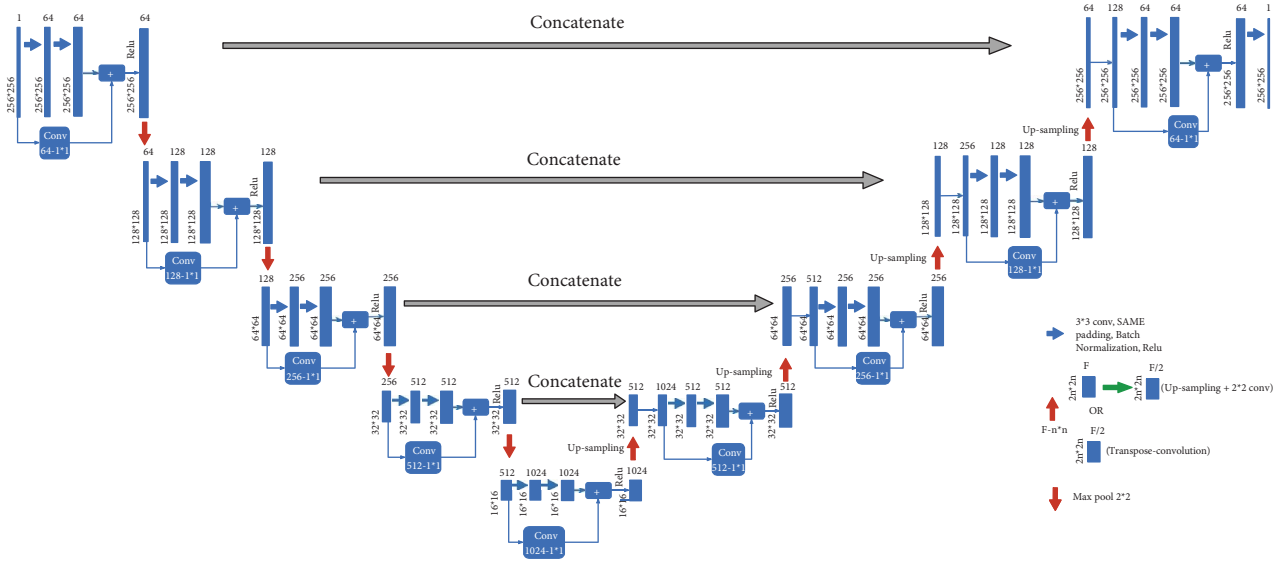


Figure 7: An illustration of the U-Net model with ResNet backbone.

parameters of the classification network trained in the large database can be borrowed to achieve better results through transfer learning. Therefore, the decoder difference largely determines the effect of a segmentation network based on the encoder-decoder structure.

An example of the Res-U-Net (U-Net model with ResNet backbone) is shown in Figure 7.

3.4. Dice Loss Function. The cross-entropy loss function (equation (1)) is often used in binary image segmentation problems. The improvement of cross-entropy is that it is easy to calculate the gradient, but when used in the building extraction problem, it will focus more on identifying the categories with high proportion due to the imbalance of samples, making it difficult to extract categories with few samples. After statistics, the ratio of building to non-building pixels in the Massachusetts dataset is about 1 : 10. To solve this problem, this study chooses the dice loss function to complement the cross-entropy loss function to reduce in building extraction the impact of sample imbalance, which is defined as equation (2),

$$L_1 = -\sum_{n=1}^{N} \left(y_n' \log y_n + (1 - y_n') \log (1 - y_n)\right), \tag{1}$$

where y_n' represents the true label class, building pixels are 1, non-building are 0, $y_n \in [0, 1]$ represents the predicted class probability, N is the total number of pixels in a sample, and n is one of the pixels,

$$L_2 = 1 - \frac{2\sum_{n=1}^{N} p_n \times t_n}{\sum_{n=1}^{N} p_n + \sum_{n=1}^{N} t_n}, \tag{2}$$

$$L_3 = L_1 + L_2, \tag{3}$$

where p_n and t_n represent the predicted category and the true label category of the pixel, respectively, and the rest of the parameters are defined in the same way as in formula (1).

According to equations (1) and (2), when there are too many non-building pixels, the cross-entropy function will make the network tend to reinforce the learning of non-building and increase the predicted category probability of non-building pixels to reduce the loss. In contrast, the dice loss function only focuses on the correct classification of the building pixels. Therefore, in this study, the dice loss function $L1$ (equation (1)) and the cross-entropy loss function $L2$ (equation (2)) are added to obtain a composite loss function $L3$ (equation (3)) that combines dice and cross-entropy, which improves the performance of the network classification ability when buildings have few pixels.

4. Experiments and Analysis

4.1. Data Description. The Massachusetts dataset, created by Mnih [36], was captured in Massachusetts, USA, and contains labels for buildings and roads, which were used in this experiment only for building extraction. The dataset contains 137 training images, 10 test images, and 4 validation images, with 3 bands of red, green, and blue, all 1500 pixels in length and width, and a spatial resolution of 1 m, covering an area of approximately 340 km^2. As mentioned in Figure 8, an original RGB image with its validation mask whose objects (buildings in this case) are shown in binary.

4.2. Data Augmentation. In general, the larger the amount of data, the more easily the model can learn representative features. Due to the high cost of acquiring new data, there are various data enhancement techniques to increase the amount of data, such as zooming in, zooming out, rotating, flipping, color changes, etc. In this experiment, zooming, rotating, and horizontal and vertical flips were used to enhance the data. Figure 9 depicts the results: Figures 9(a) and 9(b) show the original raw image and the modified image, respectively.

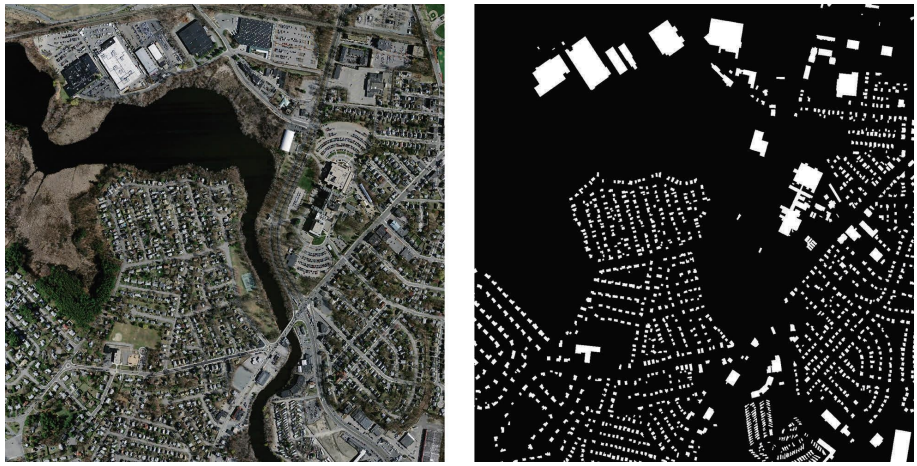


Figure 8: An example of the original and ground truth mask of testing Massachusetts buildings' dataset.

4.3. Implementation Details. The internal parameters of the neural network can be obtained by iterating the optimization algorithm, while some hyperparameters need to be set artificially to guide the model during learning, such as learning rate, optimization function, weight decay parameter, etc.

The optimization problem is one of the most important research directions in computational mathematics. In the field of deep learning, the choice of the optimization algorithm is also the top priority of a model. The Adam optimization [37] function was used in this paper; it is one of the most popular optimizers in deep learning. It is suitable for many types of problems, including models with sparse or noisy gradients. Its ease of fine-tuning makes it possible to achieve good results quickly. The Adam optimizer combines the advantages of AdaGrad and RMSProp. Adam uses the same learning rate for each parameter and adapts independently as learning progresses.

The learning rate is considered as one of the essential hyperparameters for optimizing deep neural networks; by acting on its convergence, it sets the conditions of its operation before the learning process. Indeed, a too high learning rate leads to essential weight updates, and the convergence becomes unstable. On the other hand, for a low learning rate, the convergence is slowed down with a possibility of falling into local minima. The popular approach used in deep learning to have the optimal learning rate is to start learning with a high value to accelerate the gradient descent and reduce it later to improve the accuracy [38]. Practically, this involves initializing an $\alpha 0$ to a high value at the beginning and then decreasing it by a constant multiplicative factor during the learning phase until the validation error reaches a stable value or when the learning error does not decrease anymore [39].

The initial learning rate was 0.0001; it could be formulated as in equation (4),

$$lr = 0.0001 \times \left(1 - \frac{iter}{\max iter}\right)^{0.9}. \tag{4}$$

In our experiments, the training and testing process for building detection was implemented in the PyTorch framework using the Nvidia TesUla K80 graphics card. The batch size was 16 with 100 epochs.

4.4. Evaluation Metrics. To validate the semantic segmentation performance of the proposed method, in this paper, we use four indicators (precision, recall, F1 score, and IoU (intersection over union)) to evaluate the performance of different methods on the dataset. The IoU indicator is, often referred to as the intersection over union ratio, also known as the Jaccard index, and it is a statistic for determining how accurate an object detector is on a given dataset, which is often used not only in semantic segmentation evaluation but is frequently used in object detection problems, such as remote sensing images. As the name suggests, IoU is the ratio of intersection and union between the target and the prediction (equation (5)),

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{A \cap B|}{|A| + |B| - |A \cap B|}, \tag{5}$$

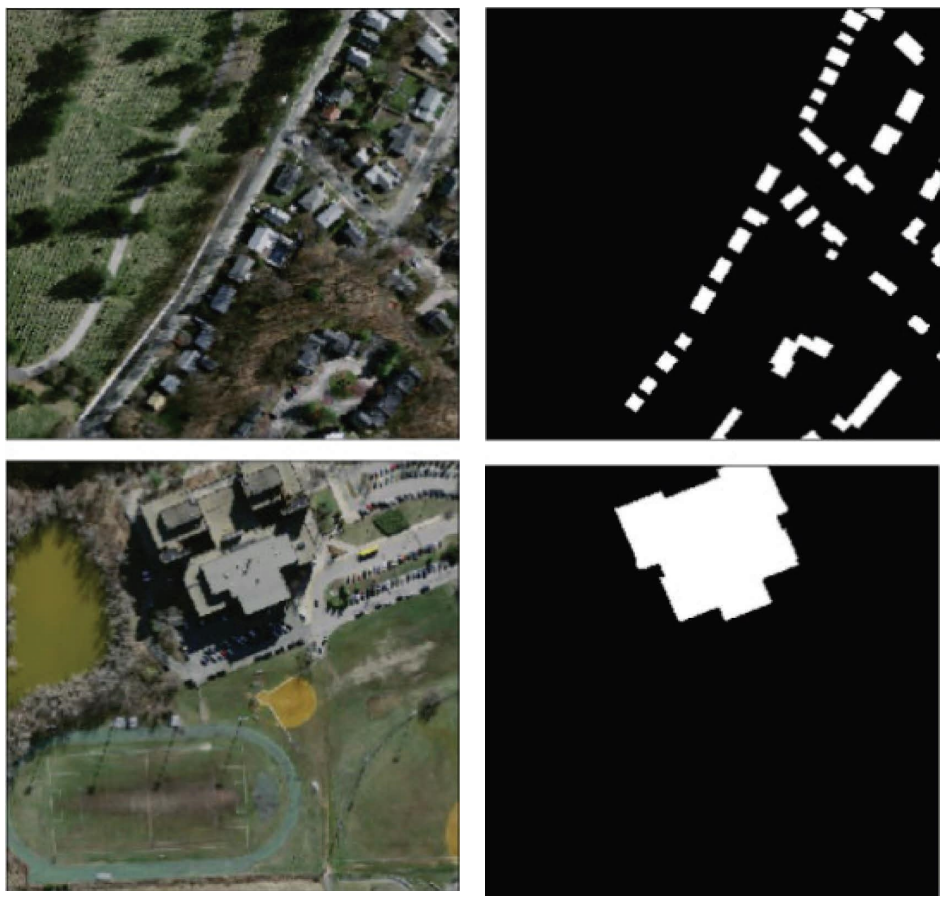
where A represents the buildings and characteristics predicted by different methods and B represents the map of the actual characteristics of the building.

The precision is expressed as the ratio of the number of correctly predicted positive samples to the number of all predicted positive samples,

$$Precision = \frac{TP}{TP + FP}. \tag{6}$$

The recall is expressed as the ratio of the number of correctly predicted positive samples to the number of all positive samples in the test set,

$$Recall = \frac{TP}{TP + FN}. \tag{7}$$



(a) (b)

Figure 9: Examples of data augmentation pre-processing. (a) Transform image and (b) transform ground truth mask.

F1score is the geometric mean between precision and recall, also known as the harmonic mean, and is an index to measure the precision of the binary classification model,

$$F1\ score = \frac{2 * Precision * Recall}{Precision + Recall}, \tag{8}$$

where TP (true positive) refers to the number of correctly classified positive samples, FP (false positive) refers to the number of negative samples mislabeled as positive samples), TN (True Negative) refers to the number of correctly classified negative samples), and FN (False Negative) refers to the number of positive samples incorrectly marked as negative samples).

5. Results

The proposed networks are built with deeper encoding and decoding layers to achieve better building segmentation results. The FCN method uses a simple convolutional coding layer. Due to its low encoding and decoding layers, it cannot fully extract the variable features from the building features, which leads to poor building feature extraction results.

This research undertakes five experiments based on VGG16, VGG19, ResNet50, DenseNet169, and Xception as a backbone for comparative analysis to demonstrate the relevance of the depth of encoder and decoder layers in building the network for each model. In this work, the efficiency of several deep learning-based models (U-Net, FPN, PSPNet, and LinkNet) in extracting buildings from high-resolution aerial images was evaluated, and in this case, IoU technique, Fscore, precision, and recall were implemented and used. Each of these models may offer several distinct advantages over the others. For example, the U-Net with VGG16 is a shallower model than others and has a basic network topology. On the other hand, PSPNET architecture considers the image's global context when predicting local level predictions, resulting in improved performance on benchmark datasets such as cityscapes and PASCAL VOC 2012.

Because the buildings in each region have various distribution characteristics, and it is difficult for each approach to reach a full optimum impact in different locations, the results of FPN in the test set are not as excellent as the other techniques, as shown in Table 1, but U-Net with VGG16 surpasses all other methods in IoU at the same time. Second, we have LinkNet, which is always using the VGG16 decoder, followed by VGG19, which outperforms U-Net in recall and has an interesting Fscore that is very near top model. Based on this first IoU comparison, we may infer that the U-Net and LinkNet classifiers perfectly match the VGG decoder.

Table 1: Comparison metrics of the models tested in this study.

Model	Encoder	IOU	*F*1-score	Precision	Recall
PSPNet	VGG16	0.7784	0.8743	0.8405	0.911
	VGG19	0.7772	0.8735	0.8426	0.9068
	ResNet50	0.7176	0.8452	0.8261	0.8653
	DenseNet169	0.6669	0.7946	0.7762	0.8139
	Xception	0.6528	0.7844	0.7566	0.8144
LinkNet	VGG16	0.8179	0.8992	0.8585	0.9439
	VGG19	0.8193	0.9001	0.8568	**0.9481**
	ResNet50	0.82	0.9003	0.8671	0.9363
	DenseNet169	**0.8238**	**0.9027**	**0.8808**	0.9256
	Xception	0.8127	0.8959	0.8715	0.9216
FPN	VGG16	0.6952	0.7681	0.7403	0.7981
	VGG19	0.6812	0.7528	0.7252	0.7826
	ResNet50	0.6656	0.7481	0.7255	0.7721
	DenseNet169	0.6632	0.7476	0.7288	0.7673
	Xception	0.6471	0.7336	0.7120	0.7567
U-Net	VGG16	**0.8302**	**0.9067**	**0.8846**	0.9298
	VGG19	0.8296	0.9064	0.8827	0.9314
	ResNet50	0.8233	0.9023	0.8741	0.9324
	DenseNet169	0.826	0.9041	0.8788	0.931
	Xception	0.821	0.901	0.8782	0.925

Table 2: Comparison table between three other models using IoU metric.

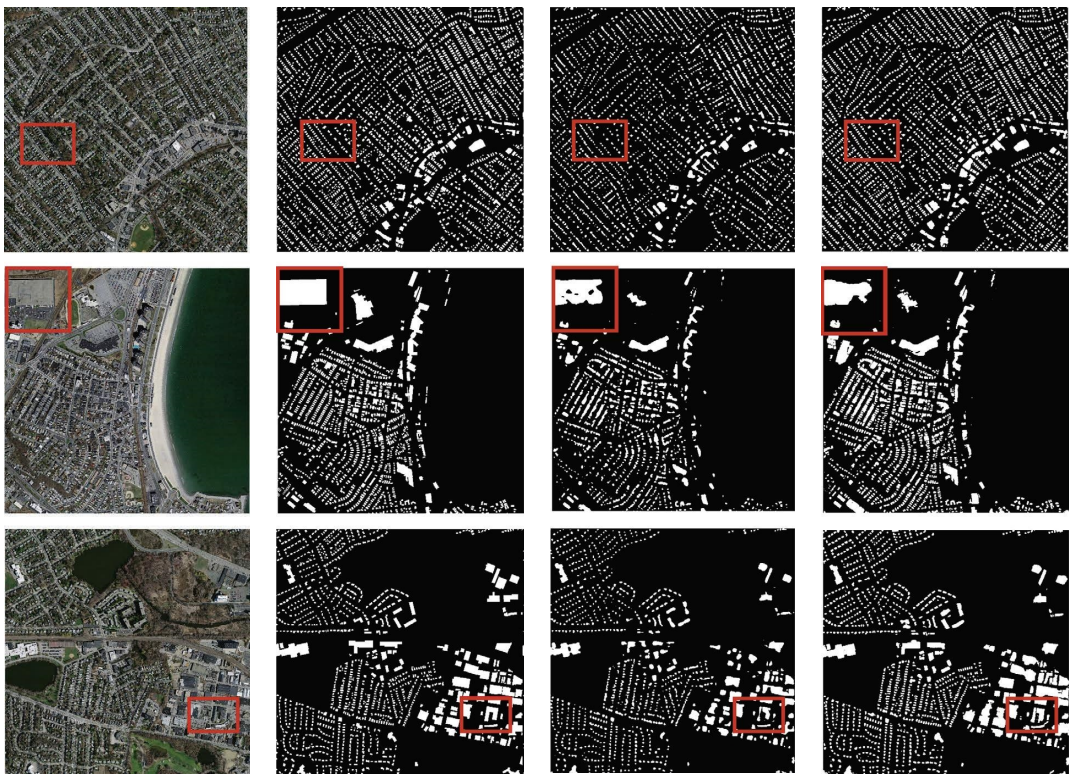
Model	IoU
MHA-Net [41]	0.7446
ABNet [40]	0.8165
U-Net + ResNet50 [42]	0.8263
U-Net + VGG16	**0.8302**

Second best Fscore is for LinkNet with the DenseNet169 encoder, followed by ResNet50.

Table 2 compares the test data to numerous novel segmentation approaches reported on the aerial image dataset in the previous two years. When compared to the AttentionBuildNet (ABNet) model provided in [40], our findings suggest that the U-Net approach with VGG increases the IoU by 1.73 percent. A second comparison is with the MHA-NET models provided by [41], in which our model has a UoI of 11.5 percent, a significant improvement over the MHA-NET models. When compared to the Res-U-Net approach described in [42], VGG U-Net improves IoU by 0.52 percent [42].

Figure 10 depicts the experimental visualization, where Figure 10(a) is the original image of the data set in the Massachusetts region and Figure 10(b) is the original image labelling. Figures 10(c)–10(e) illustrate the best possible outcomes of each model, PspNet, LinkNet, and U-Net (e). In comparison to the segmentation findings, the prediction results show that the U-Net technique in this study can better differentiate the borders between buildings, create fewer misclassified pixels with less loss of edge information, and capture crisper features. It can produce a more precise and realistic extraction. A well-defined segmentation enables for better categorizing of the sought objects. The findings of the proposed model were fascinating for both high and low density areas.

There are several types of buildings (large shopping mall, residential, industrial, etc.), and as shown in Figure 11, U-Net can better distinguish the large buildings with a high accuracy compared to PspNet, which finds much difficult to extract them better. This is due to the structure of the building, which has a remarkable similarity to the parking lots. However, with a simple building structure, as shown in the first row of Figure 11, most models are also closer to the ground truth. Solar shadows on the building itself can influence the building extraction too much. The results of building segmentation in a large area can reflect the degree of model training, as shown in Figure 12. It can be seen from the segmentation results that neither of the two-deep neural networks (PspNet and LinkNet) can fully achieve accurate building segmentation, which indicates that there is still a gap between the trained model and the actual segmentation model. On the other hand, we can see that Figure 12 confirms even more that U-Net with vgg16 distinguishes buildings better and can adapt to different types of high resolution remote sensing images. Furthermore, we can see that U-Net with vgg16 represents less False Negative than the other models and less False Positive due to the shadow caused by the buildings, which represents a significant challenge to increase the accuracy.



Test Image (a) | Ground truth (b) | PSPNet-VGG16 (c) | LinkNet-VGG16 (d)

U-Net-VGG16 (e)

Figure 10: Examples of the extracted results on the Massachusetts building dataset. (a) Test image, (b) ground truth, (c) PSPNet-VGG16, (d) LinkNet-VGG16, and (e) U-Net-VGG16.

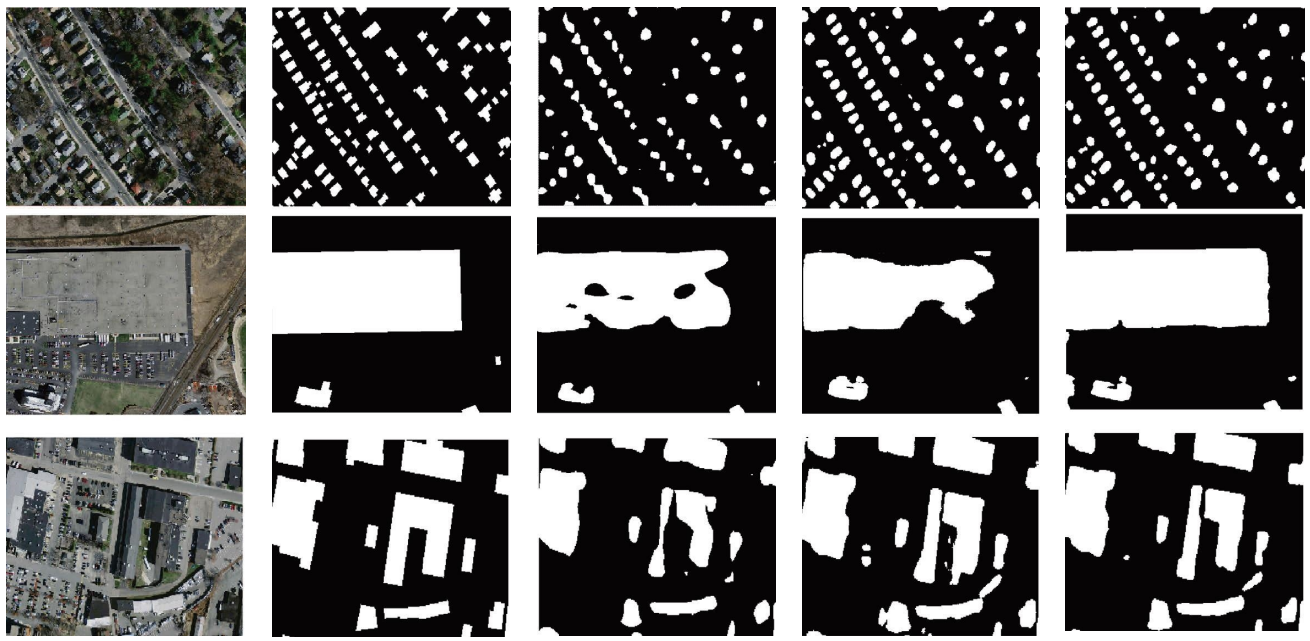


Figure 11: Performance comparison of building with different textures. (a) Test image, (b) ground truth, (c) PSPNet-VGG16, (d) LinkNet-VGG16, and (e) U-Net-VGG16.

Figure 12: Visual comparisons. White = TP (the predicted result is a building, and the ground truth is also a building), black = TN (the ground truth is a non-building, and the predicted result is a non-building), yellow = FN (the ground truth is a building, and the predicted result is a non-building), and red = FP (the ground truth is a non-building, but the predicted result is a building). (a) Test image, (b) ground truth, (c) PSPNet-VGG16, (d) LinkNet-VGG16, and (e) U-Net-VGG16.

6. Conclusions

Building segmentation from remote sensing images must be accurate and automated for applications such as urban planning and catastrophe management. The current state of development of key deep learning approaches for image categorization and building instance extraction from high-resolution remote sensing images is discussed. Furthermore, this paper mainly focuses on the four most advanced auto-encoder methods U-Net, PSPNet, LinkNet, and FPN, with an improvement of the models using VGG, ResNet, DenseNet, and Xception as backbone. The feature similarity of different pixel types was weakened to effectively separate pixels from urban and complex background areas. Considering that existing classical image classification methods based on deep learning have many limitations, such as generating blurred edges and losing detailed information. Since environmental information and building information are easily confused, which leads to mediocre extraction results, a new loss function is proposed, which allows the model to update the parameters faster and more stably. Training and testing are performed on the Massachusetts aerial image dataset with a coverage of 340 km^2. The results show that the U-net model with VGG16 as backbone achieves the best result with 83.06%, where it outperforms all presented models.

In addition, the presence of solar shadows, occlusions, and differences in the characteristics of the building itself will have some impact on the integrity of the building extraction. It is not exhaustive to consider only the color or brightness characteristics of the pixel itself and its local area. In future work, it is necessary to study the shadows and occlusions of buildings in the image to improve the building extraction effect.

Data Availability

The Mssachusetts Building data used to support the findings of this study are available at https://www.cs.toronto.edu/~vmnih/data/.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] R. Azmi, O. B. Alami, A. E. Saadane, I. Kacimi, and T. Chafiq, "A modified and enhanced normalized built-up index using multispectral and thermal bands," *Indian Journal of Science and Technology*, vol. 8, no. 1, pp. 1–11, 2015.

[2] R. C. Weih and N. D. Riggan, "OBJECT-BASED classification vs. pixel-based classification: comparitive importance of multi-resolution imagery," *Environmental Science, Mathematics*.

[3] M. D. Hossain and D. Chen, "Segmentation for Object-Based Image Analysis (OBIA): a review of algorithms and challenges from remote sensing perspective," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 150, pp. 115–134, 2019/04/01/2019.

[4] R. Azmi, H. Amar, and A. Norelyaqine, "Generate knowledge base from very high spatial resolution satellite image using robust classification rules and genetic programming," in *Proceedings of the 2020 IEEE International conference of Moroccan Geomatics (Morgeo)*, pp. 1–6, Casablanca, Morocco, May 2020.

[5] A. Rida, A. Hicham, and N. Abderrahim, "Optimization of object-based image analysis with genetic programming to generate explicit knowledge from WorldView-2 data for urban mapping," in *Geospatial Intelligence*, pp. 157–169, Springer, Berlin, Germany, 2022.

[6] L. L. C. W. M. J. T. C. S. f. A. M. Shuli, "Segmentation of Remote Sensing Images Based on Adaptive Global Threshold and Fused Markers," 2013.

[7] S. Wu, Y. Wu, and J. Zhou, "Multi-level thresholding for remote sensing image of urban area based on line intercept histogram," *CAAI Transactions on Intelligent Systems*, vol. 10, 2018.

[8] J. Wang, X. Yang, X. Qin, X. Ye, Q. J. I. G. Qin, and R. S. Letters, "An efficient approach for automatic rectangular building extraction from very high resolution optical satellite imagery," vol. 12, no. 3, pp. 487–491, 2014.

[9] L. Yang, X. Wu, D. Zhao, H. Li, and J. Zhai, "An improved Prewitt algorithm for edge detection based on noised image,"vol. 3, pp. 1197–1200, in *Proceedings of the 2011 4th International congress on image and signal processing*, vol. 3, pp. 1197–1200, IEEE, Shanghai, China, October 2011.

[10] P. F. Felzenszwalb and D. P. J. I. Huttenlocher, "Efficient graph-based image segmentation," *International Journal of Computer Vision*, vol. 59, no. 2, pp. 167–181, 2004.

[11] N. Y. Q. Abderrahim, S. Abderrahim, and A. Rida, "Road segmentation using u-net architecture," in *Proceedings of the 2020 IEEE International conference of Moroccan Geomatics (Morgeo)*, pp. 1–4, IEEE, Casablanca, Morocco, May 2020.

[12] N. Abderrahim, A. Saadane, and A. Rida, "Deep convolution neural network for automated method of road extraction on aerial imagery," in *Geospatial Intelligence*, pp. 31–40, Springer, Berlin, Germany, 2022.

[13] Q. Yuan, H. Shen, T. Li et al., "Deep learning in environmental remote sensing: achievements and challenges," *Remote Sensing of Environment*, vol. 241, Article ID 111716, 2020.

[14] F. Fan, W. Shuangting, Z. Jin, W. Chunyang, and O. Progress, "Hyperspectral images classification based on multi-feature fusion and hybrid convolutional neural networks," *Laser & Optoelectronics Progress*, vol. 58, no. 8, Article ID 0810010, 2021.

[15] E. Maggiori, Y. Tarabalka, G. Charpiat, P. J. I. T. Alliez, and R. sensing, "Convolutional neural networks for large-scale remote-sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 2, pp. 645–657, 2017.

[16] Z. Zhang and Y. J. R. S. Wang, "JointNet: a common neural network for road and building extraction," *Remote Sensing*, vol. 11, no. 6, p. 696, 2019.

[17] Y. Lihua, W. Lei, Z. Wenwen, L. Yonggang, and W. J. A. G. E. C. S. Zengkai, "Deep metric learning method for high resolution remote sensing image scene classification," vol. 48, no. 6, p. 698, 2019.

[18] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, Boston, MA, USA, June 2015.

[19] Y. Liu, D. Minh Nguyen, N. Deligiannis, W. Ding, and A. J. R. S. Munteanu, "Hourglass-shapenetwork based semantic segmentation for high resolution aerial imagery," *Remote Sensing*, vol. 9, no. 6, p. 522, 2017.

[20] R. Hamaguchi, A. Fujita, K. Nemoto, T. Imaizumi, and S. Hikosaka, "Effective use of dilated convolutions for segmenting small object instances in remote sensing imagery," in *Proceedings of the 2018 IEEE winter conference on applications of computer vision (WACV)*, pp. 1442–1450, IEEE, Lake Tahoe, Nevada, USA, March 2018.

[21] H. Wang, Y. Wang, Q. Zhang, S. Xiang, and C. J. R. S. Pan, "Gated convolutional neural network for semantic segmentation in high-resolution images," *Remote Sensing*, vol. 9, no. 5, p. 446, 2017.

[22] R. Shang, J. Zhang, L. Jiao, Y. Li, N. Marturi, and R. J. R. S. Stolkin, "Multi-scale adaptive feature fusion network for semantic segmentation in remote sensing images," *Remote Sensing*, vol. 12, no. 5, p. 872, 2020.

[23] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," in *Proceedings of the International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, Munich, Germany, October 2015.

[24] A. Abdollahi, B. Pradhan, N. Shukla, S. Chakraborty, and A. Alamri, "Multi-object segmentation in complex urban scenes from high-resolution remote sensing data," *Remote Sensing*, vol. 13, no. 18, p. 3710, 2021.

[25] A. Abdollahi, B. Pradhan, and A. M. Alamri, "An ensemble architecture of deep convolutional Segnet and Unet networks for building semantic segmentation from high-resolution aerial images," *Geocarto International*, vol. 37, no. 12, pp. 3355–3370, 2020.

[26] B. Bischke, P. Helber, J. Folz, D. Borth, and A. Dengel, "Multi-task learning for segmentation of building footprints with deep neural networks," in *Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP)*, pp. 1480–1484, Taipei, Taiwan, September 2019.

[27] K. Bittner, F. Adam, S. Cui, M. Körner, and P. Reinartz, "Building footprint extraction from VHR remote sensing images combined with normalized DSMs using fused fully

convolutional networks," *Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 8, pp. 2615–2629, 2018.

[28] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2881–2890, Honolulu, HI, USA, July 2017.

[29] A. Chaurasia and E. Culurciello, "Linknet: exploiting encoder representations for efficient semantic segmentation," in *Proceedings of the 2017 IEEE Visual Communications and Image Processing (VCIP)*, pp. 1–4, IEEE, Petersburg, FL, USA, December 2017.

[30] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, Honolulu, HI, USA, July 2017.

[31] O. Russakovsky, J. Deng, H. Su et al., "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[32] K. Simonyan and A. J. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, https://arxiv.org/abs/1409.1556.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, Las Vegas, Nevada, USA, June 2016.

[34] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, Honolulu, HI, USA, July 2017.

[35] F. Chollet, "Xception: deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251–1258, Honolulu, HI, USA, July 2017.

[36] V. Mnih, *Machine Learning for Aerial Image Labeling*, University of Toronto, Toronto, Ontario, Canada, 2013.

[37] D. P. Kingma and J. J. Ba, "Adam: A Method for Stochastic Optimization," 2014, https://arxiv.org/abs/1412.6980.

[38] Y. Bengio, "Practical recommendations for gradient-based training of deep architectures," in *Neural Networks: Tricks of the Trade*, pp. 437–478, Springer, Berlin, Germany, 2012.

[39] L. Bottou, "Stochastic gradient descent tricks," in *Neural Networks: Tricks of the Trade*, pp. 421–436, Springer, Berlin, Germany, 2012.

[40] P. Das and S. Chand, "AttentionBuildNet for building extraction from aerial imagery," in *Proceedings of the 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, pp. 576–580, IEEE, Greater Noida, India, November 2021.

[41] J. Cai, Y. J. I. J. T. A. E. O. Chen, and R. Sensing, "MHA-net: Multipath Hybrid Attention Network for Building Footprint Extraction from High-Resolution Remote Sensing Imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, 2021.

[42] A. Norelyaqine and A. Saadane, "Deep learning for building extraction from high-resolution remote sensing images," in *Proceedings of the International Conference on Advanced Technologies for Humanity*, pp. 116–128, Springer, Rabat, Morocco, November 2021.