IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 1, 2022, accepted March 17, 2022, date of publication April 4, 2022, date of current version April 13, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3164401

IRU-Net: An Efficient End-to-End Network for Automatic Building Extraction From Remote Sensing Images

MD. ABDUL ALIM SHEIKH[1], TANMOY MAITY[2], (Senior Member, IEEE), AND ALOK KOLE[3], (Senior Member, IEEE)

[1]Department of Electronics and Communication Engineering, Aliah University, Kolkata 700160, India
[2]Department of Mining Machinery Engineering, Indian Institute of Technology (ISM), Dhanbad, Dhanbad 826004, India
[3]Department of Electrical Engineering, RCC Institute of Information Technology, Kolkata 700015, India

Corresponding author: Md. Abdul Alim Sheikh (alim.sheikh16@gmail.com)

ABSTRACT Automatic extraction of buildings from High-Resolution Remote Sensing (RS) Imagery is of great practical interest for numerous applications; including urban planning, change detection, disaster management, estimation of human population, and many other geospatial related applications. This paper proposes a novel efficient Improved ResU-Net architecture called IRU-Net, integrating spatial pyramid pooling module with an encoder-decoder structure, in combination with Atrous convolutions, modified residual connections, and a new skip connection between the encoder-decoder features for automatic extraction of buildings from RS images. Moreover, a new dual loss function called binary cross-entropy-dice-loss (BCEDL) is opted that make cross-entropy (CE) and dice loss (DL) and consider both local information and global information to decrease the class imbalance influence and improve the building extraction results. The proposed model is examined to demonstrate its generalization on two publicly available datasets; the Aerial Images for Roof Segmentation (AIRS) Dataset and the Massachusetts buildings dataset. The proposed IRU-Net achieved an average F-1 accuracy of 92.34% for the Massachusetts dataset and 95.65% for the AIRS dataset. When compared to other state-of-the-art deep learning-based models such as SegNet, U-Net, E-Net, ERFNet and SRI-Net, the overall accuracy improvements of our IRU-Net model is 9.0% (0.9725 vs. 0.8842), 5.2% (0.9725 vs. 0.9218), 3.0% (0.9725 vs. 0.9428), 1.4% (0.9725 vs. 0.9588) and 0.93% (0.9725 vs. 0.9635), for AIRS dataset and 11.6%, 5.9%, 3.1%, 2.7% and 1.4%, for Massachusetts building dataset. These results prove the superiority of the proposed model for building extraction from high-resolution RS images.

INDEX TERMS Building extraction, deep learning, encoder-decoder network, atrous spatial pyramid pooling, remote sensing imagery, cross-entropy and dice loss.

I. INTRODUCTION

With the rapid development of different sensors, the availability of high-resolution RS imagery is significantly increased [1]. These valuable data provide a huge potential for meaningful and accurate terrestrial object interpretation. Among which, building is one of the most important types of terrestrial objects and automatic detection of buildings plays a vital role in a wide range of RS applications, such as urban planning and reconstruction, change detection, Disaster management, estimation of human population, 3D city modelling, real-estate management, illegal building survey, geographic information systems, etc. [2]–[4]. Although building detection can be achieved manually by human experts, but it is very time-consuming, labour-intensive and expensive to extract buildings from RS images. As a result, the traditional image processing-based method, which is over-dependent upon manual extraction of features, cannot solve the problems of large-scale dataset interpretation and does not fulfil the requirements of nowadays practical applications [5]. Therefore, there are strong efforts to develop automatic, accurate, and computationally fast methodologies to extract buildings [6]. However, extracting buildings accurately and efficiently from RS imagery is still a challenging task with

The associate editor coordinating the review of this manuscript and approving it for publication was Oguzhan Urhan.

several difficulties. The diverse characteristics of building including colour, shape, size, material, and interference of building shadows and trees increase difficulty and challenge for accurate and reliable building extraction [7], [8]. On the other hand, many objects e.g., roads, parking lots in high-resolution RS images are highly similar to buildings in appearance due to low inter-class variance and high intra-class variance [9]. Therefore, automatic extraction of buildings accurately and efficiently from RS imagery remains a challenge that attracts huge research interests [10]. Over the past few years, a variety of methods have been proposed by researchers to extract buildings from high-resolution RS images. They can be divided into two groups: classical image processing-based and Deep Learning-based methods. In classical image processing-based algorithms, usually features are extracted manually and they need prior knowledge which is leading to time-consuming, labour-intensive, and limits their accuracy [8], [9].

Classical building extraction techniques exploit the characteristics of the texture, spectrum, geometry, edge, and shadow [11]–[18] as feature descriptors for buildings extraction from RS images. Since these features vary under different illumination conditions, sensor types and building architectures, traditional methods can resolve only particular issues with specific data. Therefore, fusing data sources, such as multi-spectral images with either stereo Digital Surface Models (DSMs) or light detection and ranging (LiDAR) DSM or synthetic aperture radar (SAR) [19] were reported in [20]–[23] to distinguish non-building areas that are highly similar to buildings to increase the accuracy of building extraction. More recently, several machine learning techniques have been introduced for pixel-wise classification such as Support Vector Machines [24], K-Means [25], Adaptive Boosting [26], Random Forests [27], and Conditional Random Fields (CRFs) [28]. Some algorithms utilized specific criteria of building appearances like the uniform spectral reflectance values using morphological building/shadow index and mutual information [29]–[31]. These approaches rely heavily on manually extracted features which usually change with the application area and fail to detect buildings with other objects having a similar appearance like roads in RS images.

Based on recent advances, Deep Learning (DL) [32] is proving to be a very successful set of tools for several image understanding tasks, segmentation, classification tasks and other applications including RS image analysis [19]. Convolutional Neural Networks (CNNs) [33], [34] are one of the most successful DL architectures. Mnih [35] first introduced the CNNs for building extraction which set remarkable progress in computer vision and photogrammetry research. Recently, CNNs have made great achievement in a wide variety of image segmentation task and is proving prominent models in remote sensing applications [33]. Various CNN architectures for automatic building extraction have been adopted in the literature. In the early phases, the patch-based CNN models such as GoogleNet [36], Visual Geometry Group (VGG) [37], AlexNet [38], Deep Residual Network (ResNet) [39], and DenseNet [40] have outperformed traditional machine learning methods on segmentation and classification applications. Some researchers also utilized patch-based CNN methods to segment buildings in RS images and achieved improved performance [41]. However, the patch-based method needs overlapping patches to predict each pixel, which causing redundant computations. However, because of the inability to preserve the spatial information of contextual features and consistency, patch-based CNN models are not the optimal solution for building segmentation [42].

Whereas early works mainly use patched-based CNNs, Fully Convolutional Network (FCNs) [43] based approaches [44]–[46] are often used for building extraction from RS images and achieve reliable results with high accuracy. In FCN, fully connected layers are replaced by up-sampling layers so that the output preserves spatial information of contextual features [43]. Zou *et al.* [45] proposed a Hierarchically Fused FCN (HF-FCN) which approached a strategy by hierarchically fusing the information from the multi-scale receptive fields of the network built on the basis of VGG-16 architecture for robust building extraction. However, HF-FCN is a FC network applied to every pixel individually and it significantly enlarges the number of parameters in the neural network. The employment of pooling layers causes the loss of detailed information also.

Over the past few years, many FCN-based variants have been proposed to achieve more accurate segmentation results. The SegNet [47] and U-Net [48] are two classic models with symmetric encoder-decoder structures, which were both regarded as effective architectures due to their capabilities of recovering semantic details. Wang *et al.* [8] propose a novel network ENRU-Net, composed of a U-shape encoder-decoder structure and an improved non-local block named asymmetric pyramid non-local block (APNB) for accurate building extraction from high- resolution aerial imagery. Li *et al.* [10] proposed a U-Net-based semantic segmentation method for building footprints extraction from high-resolution multispectral satellite images using the SpaceNet building dataset and Multi-Source GIS Data. Abdollahi *et al.* [49] integrated semantic edge information and segmentation information for building extraction from aerial images using U-Net. However, lack of global information limits the performance for building extraction task proposed in [8] and [49]. The model fails to detect small size buildings due the scale invariance of buildings under many complex scenarios. Bittner *et al.* [50] proposed a methodology using FCN architecture to automatically generates a binary building mask out of a Digital Surface Model (DSM). In this study, the FCNs were trained on set of patches which needs overlapping to predict each pixel, which caused redundant computations. Moreover, the model is not optimal because of the inability to preserve the spatial information of contextual features and consistency.

However, the FCNs and other encoder-decoder architectures, like the Seg-Net [47] or DenseNet [40], only apply some layers to generate final output neglecting the fine details. Nonetheless, the classical U-Net applications has two main limitations: 1) the parameters on both sides of the bottleneck layers are updated before the intermediate layers. This makes intermediate layers less powerful in terms of semantic representations [51]; 2) the sparsity applied in the intermediate features limits the generalization performance.

Liu *et al.* [52] incorporated the spatial pyramid pooling module into the encoder-decoder architecture for building extraction. The main drawback of this model is additional computational cost and memory consumption. Ji *et al.* [53] proposed a scale-robust FCN using ASPP structures for building extraction from aerial and satellite imagery. In another work, Ji *et al*. [54] proposed Siamese U-Net to improve segmentation performance by multi-scale input. But the deep symmetry architecture means it needs heavy-weight decoder which leads to high memory consumption and low inference speed. Li *et al*. [55] proposed encoder-decoder architecture and employed Dense-block [40] as their core module. But the Dense-block makes networks need large memory consumption and computational cost. For the problem of multi-scale building extraction, [56]–[58] integrated hierarchical results extracted from multiple models, based on feature pyramid network or a design-specific CNN models for accurately buildings extraction. Despite the improvement of segmentation accuracy, the challenges of building extraction still exist. First, the engagement of pooling layers causes the loss of detailed information, and coarse upsampling layers without the detailed information, would reduce the recognition accuracy of small buildings, especially the contours. Second, the coarse upsampling layers, and the orthodox structures of FCNs, leads to numerous misclassifications when extracting buildings from RS images.

To achieve further improvement in accuracy, Zhang and Wang [59] proposed a method called JointNet, which is a novel neural network for extraction of both roads and buildings built on the integration of dense connectivity and atrous convolution, which employs the propagation efficiency of the dense connectivity pattern and the large receptive field of atrous convolution layer. The main drawbacks of the model are small training epochs and needs long training time. Zhang *et al.* [60] proposed a novel fully convolutional network, called the Web-Net, which uses the Ultra-Hierarchical Sampling (UHS) block to absorb and fuse the inter-level feature maps to propagate the feature maps among different levels to perform the building extraction on high-resolution remote sensing images. The proposed Web-Net performed not well in the extraction of buildings that were mixed with vegetation or shadows. Liu *et al.* [61] proposed a novel FCN-based network named SRI-Net in which spatial residual inception (SRI) module was proposed to capture and aggregate multi-scale contexts for a better semantic representation by successively fusing multi-level features. This network requires high computational cost, large memory consumption and too much time to train. Abdollahi, Pradhan and Alamri [62] 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. It is not useful specifically in RS which also needs large memory allocation for high-resolution data with constraint computational resources.

For further improving accuracy and to preserve the structure consistency, DeepLab family [7], [63]–[65], some FCN-based models [28], [33] utilize postprocessing and additional context module, such as CRFs, dense-CRF [63], exponential linear units and ASPP [64], [65]. Although these networks significantly improved segmentation performance, they generally need high computational cost, large memory consumption and too much time to train and are difficult to apply for the application of DL in RS [9]. For example, Shrestha and Vanneschi [66] proposed a building extraction method using conditional random fields (CRFs) and exponential linear units. Alshehhi *et al.* [33] used a patch-based CNN architecture and proposed a post-processing method integrating low-level features of adjacent regions. Though, the improvement of results is obtained by post-processing methods but within a specific range, and the quality of results strongly depends on the initial segmentation accuracy [38].

Although the abovementioned models have achieved progress in tackling the issue of building extraction, they revealed several limitations. Two aspects in building extraction still exist. The first one is the high intraclass variance of buildings and the low interclass difference between buildings and other nonbuilding objects. The other one is the scale invariance of buildings under many complex scenarios. Most of these structures revealed poor success in building extraction purposes in heterogeneous areas such as vegetation covers, shadows, and parking lots where these obstacles encompass buildings.

More recently, following the great success of Generative Adversarial Networks (GANs), Luc *et al.* [67] proposed to train an adversarial network to inspires the segmentation network to generate label maps that cannot be distinguished from the reference map. In this way, the joint distribution of at each pixel location of all label variables can be measured all together, and can enforce forms of high-order consistency that cannot be enforced by pixelwise classification. Abdollahi *et. al* [4] proposed an end-to-end convolutional neural network called GAN for building footprint extraction from high resolution aerial images utilizing SegNet model with Bi-directional Convolutional LSTM (BConvLSTM). However, traditional adversarial networks are known hard to train and face the danger of model collapse. This can lead to an optimization problem for segmentation network.

To trade-off between efficiency and accuracy, a variety of FCN-based architectures have been designed, including ESFNet [1], ARC-Net [9], ENet [68], ERFNet [69], EDANet [70], MobileNet family [71], ShuffleNet family [72], EU-Net [73], DE-Net [74], DR-Net [75], DeepReID [76] and learning to rank [77]. All of these recent

networks are aiming to compromise performance and efficiency. So, there is still room for further improvement. CNN-based building extraction algorithms have mainly been encoder-decoder-based, which loses the spatial details in the encoder stage and recovers by fusing shallow feature maps during the decoder stage. However, it causes imprecise localization on building boundaries since the coarse features propagated from shallow layers and small size buildings may be unrecognized. Moreover, the extracted features are always partly restricted by the local respective field, and large-scale buildings with low texture are always discontinuous and perforated when extracted.

To better balance the accuracy and efficiency, this paper proposes an improved novel dense deep learning-based convolutional network called IRU-Net which adopts U-Net structure, residual learning, atrous convolutions and spatial pyramid pooling (SPP) to extract buildings from two public building datasets, the Massachusetts Building Dataset [35] and the Aerial Images for Roof Segmentation (AIRS) dataset [78]. As far as we know, the proposed technique has not been used in the literature and this is for the first time this kind of approach has been proposed for the given task. The atrous spatial pyramid pooling (ASPP) [79] module is added as a bridge between the encoder and the decoder path to extract features at multiple spatial scales and comprise some more spatial details and at the same time up-samples the feature maps to learn global contextual information. Moreover, the common skip connection used in U-Net is replaced with a new path utilizing a chain of convolutional operations along with the skip connections to pass the features from the encoder to the decoder to lessen the semantic gaps between the encode–decoder features [49]. In addition, a new residual unit is used in the encoder and decoder path by incorporating convolution kernel with a size of 1×1, a step size of 1, and a batch normalization layer as the identity mapping function to overcome the problem of dimensionality change of the input image during convolution in the residual unit. Also, a new objective loss function based on binary cross-entropy and dice loss (BCEDL) [55] was opted to combine local information (CE) and global information (DL)and reduce the influence of class imbalance, and thereby increase the building segmentation results. The key contributions of this study are summarized as follows:

1) We design a novel efficient network, called IRU-Net, which could efficiently capture multi-scale features and effectively utilize the detailed context information of buildings at various scales.
2) To mitigate the semantic gaps between encoder and decoder features, the common skip connection used in the U-Net is replaced with a new path utilizing 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.
3) a new dual loss function called binary cross-entropy-dice-loss (BCEDL) is opted that make cross-entropy (CE) and dice loss (DL) and consider both local information (CE) and global information (DL) to decrease the class imbalance influence and improve the building extraction results.
4) The proposed method is evaluated qualitatively and quantitatively on two public building labeling datasets, the Massachusetts Dataset [35] and the AIRS Dataset [78], and demonstrates the excellent performance of the proposed model. Compared with established models such as SegNet [47], Residual U-Net (ResU-Net) [80], E-Net [68], ERFNet [69], and SRI-Net [61], the proposed IRU-Net could achieve higher accuracies and F-1 scores on both the two datasets for the problem of building extraction.

The remainder of this paper is organized as follows. The second section gives overall methodology of the proposed IRU-Net architecture. Sections III illustrates the test dataset, experimental settings, evaluation metrics, experiential outcomes of the proposed model and detailed comparison, respectively. Lastly, Section IV describes the significant findings and conclusion of this study.

II. PROPOSED METHEDOLOGY

The overall methodology of the proposed IRU-Net based method for building extraction from high-resolution RS imagery is illustrated in Figure 1. At the first step, two different building datasets called Massachusetts Building dataset [35] and AIRS dataset [78] are used to prepare the training, validation and test images for training the IRU-Net model and evaluate the performance of the proposed method. Then, the architecture of the proposed IRU-Net approach along with the new BCEDL [81] function is designed.

A. PROPOSED IRU-NET ARCHITECTURE

We propose an efficient Improved ResU-Net architecture called IRU-Net, integrating ASPP module with an encoder-decoder structure, in combination with modified residual units, skip connection and Atrous convolutions for building segmentation in high-resolution RS images. In this work, we utilize a 7-level architecture of deep ResU-Net for building extraction, as shown in Figure 3. The network comprises of three parts: encoding path, ASPP as bridge connector and decoding path. A brief explanation of each of the parts is given in the following subsections.

1) ENCODER AND DECODER

The encoder extracts spatial features from the training data [52]. It consists of a single STEM block and three encoder blocks. A single STEM block, which differs from the convolution blocks only for lacking the initial Batch Normalization (BN) and Rectified Linear Unit (ReLU) operations, processes the initial input. Each encoder block is built with residual units which consist of two 3×3 convolution blocks and an identity mapping. Each convolution block includes a BN layer, a ReLU activation layer and a convolutional layer.

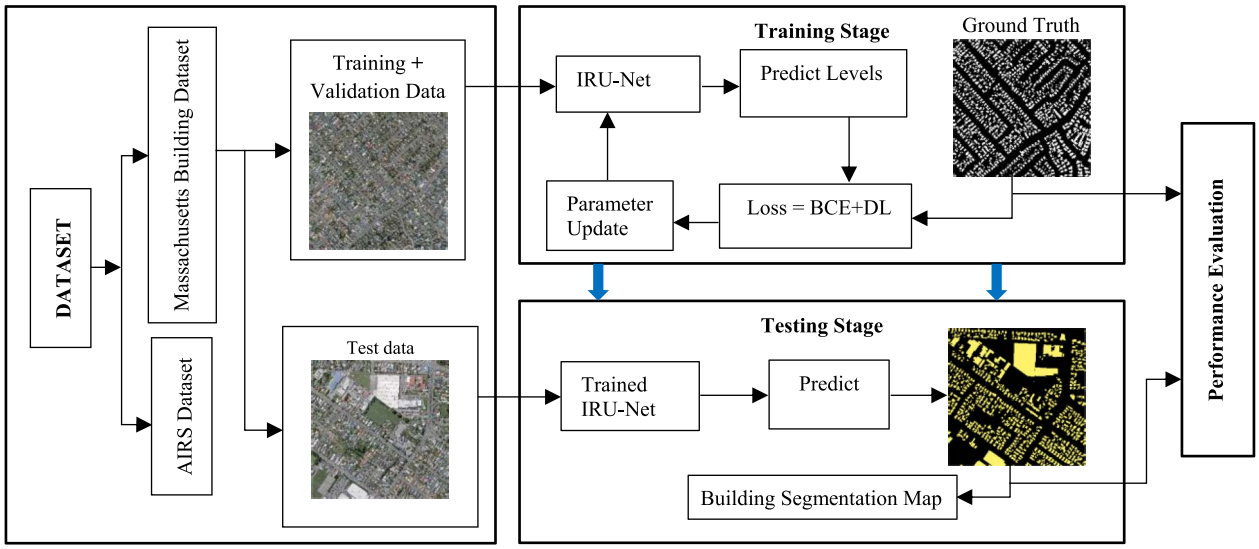

FIGURE 1. The overall framework of the proposed IRU-Net network for building extraction.

The identity mapping connects input and output of the unit. A stride of 2 is applied to the first convolution block to reduce the feature map by half. At the same time, the output of each encoder block is fused with the corresponding decoder block layer via a skip connection as shown in Figure 3, which makes full use of the semantic information and improves the segmentation accuracy.

This creates an information propagation path that allows signals to spread more easily between low-level and advanced features; this not only facilitates backpropagation during training but also improves model segmentation accuracy. The ASPP module is integrated as a bridge between the encoding and the decoding path. The decoding or expanding path restores the feature map to a pixel-wise categorization, i.e., semantic segmentation [52]. It comprises of three residual units each of which is preceded by an up-sampling of feature maps from a lower level and concatenation with the feature maps from the corresponding encoding path. The output of the last decoder block is passed through ASPP, and finally, a 1×1 convolution and a sigmoid activation layer are added on top of the IRU-Net to project the multi-channel feature maps into the desired segmentation.

2) RESIDUAL UNIT

The residual block propagates information over layers, allowing to build a deeper neural network that could solve the training degradation problem in each of the encoders while at the same time reducing the computational cost [69], [80]. The residual unit consists of two parts: the identity mapping part and the residual part. Identity mapping mainly integrates the input with the output generated by the residual part, which enables the fusion of subsequent features. Each residual unit is generally composed of multiple convolution layers, BN, and ReLU activation function. The deep ResU-net consists a series of stacked sequence residual units. Each residual unit

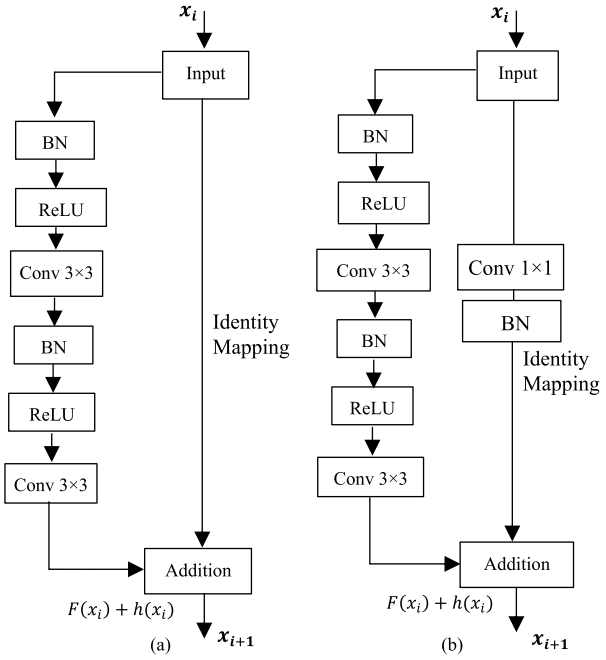

FIGURE 2. Building blocks of neural networks. A residual block with identity mapping used in the proposed ResU-net.

can be defined as a general form:

$$\left.\begin{aligned} y_i &= h(x_i) + F(x_i, W_i) \\ x_{i+1} &= f\,(y_i) \end{aligned}\right\} \tag{1}$$

whereby x_i and x_{i+1} denote to the input and output of the i-th residual unit; $f\,(y_i)$ and $F\,(\cdot)$ are the activation and the residual functions, respectively; and $h\,(\cdot)$ is the identity mapping function $h\,(x_i) = x_i$. Figure 2(a) shows a pre-activated residual unit. A linear projection W_i is used to

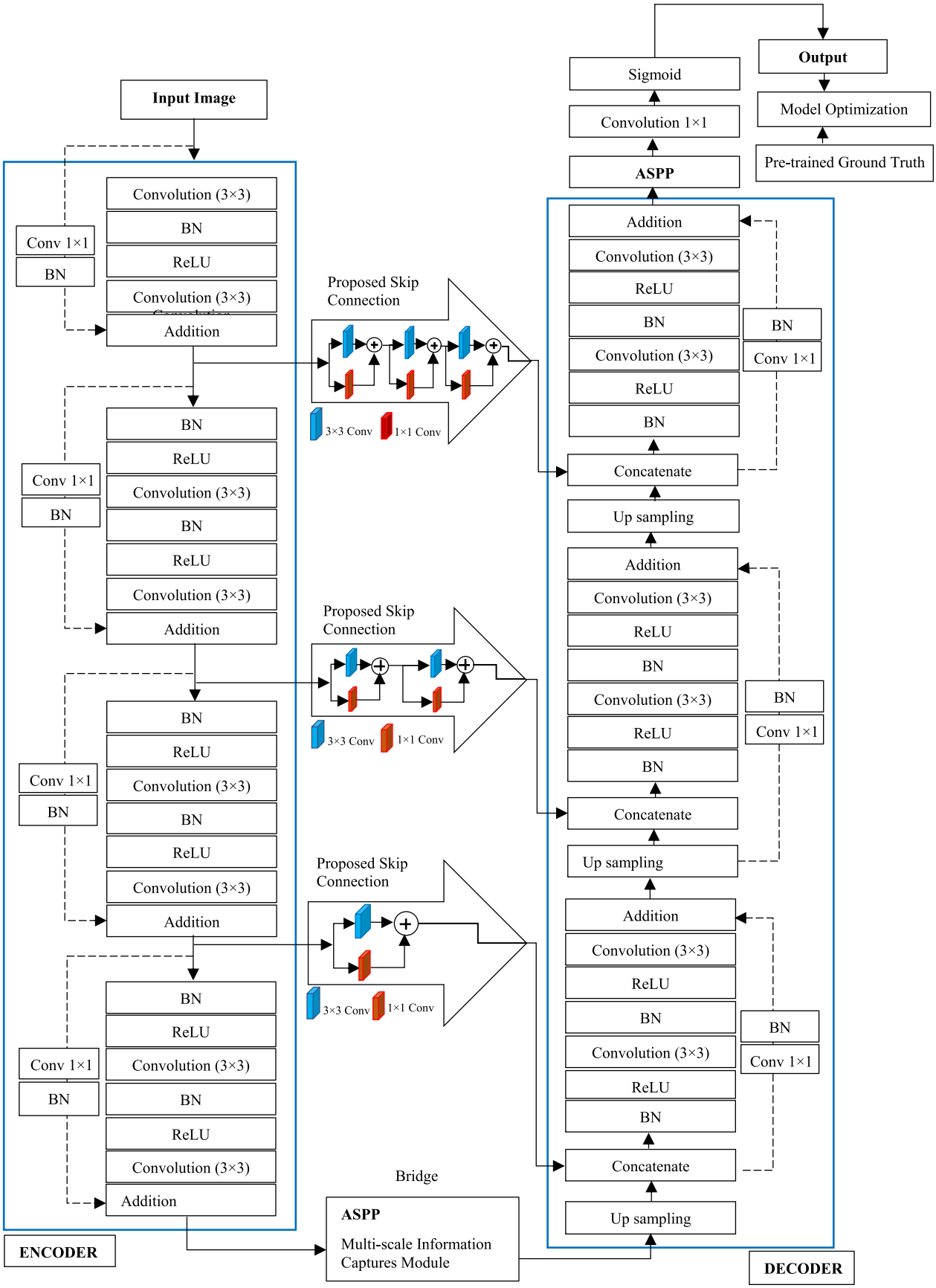

FIGURE 3. Proposed IRU-Net structure.

maintain the dimension of the input and output of the convolutional layers.

The identity mapping integrates the input and output of the residual neural network unit. The dimension of the input image also needs to be changed during identity mapping as the dimensionality of the corresponding input image changes during convolution, The residual unit as proposed in [11] has this kind of problem. To overcome this, a convolution kernel with a size of 1×1, a step size of 1, and a BN layer was used as the identity mapping function to build our deep residual U-Net. Figure 2 shows the difference between the residual network unit and the proposed residual network unit. Figure 2(a) is the structure of the residual network unit and, Figure 2(b) is the structure of the proposed residual network unit.

3) SKIP CONNECTION BETWEEN ENCODER AND DECODER PATH

The skip connection between the encoder and decoder layers was introduced in the U-Net network [48]. However, in the skip connection presented in [48], a possible semantic gap exists between two sets of features being fused because the initial layers in the encoder path of the U-Net model compute the low-level features, whereas the deep layers in the decoder path compute the prominent higher-level features. After the initial addition layer, the encoder was fused with the decoder after the last upsampled layer using the first skip connection. Hence, the concatenation of these different collections of features can perhaps negatively affect the prediction process because they can cause conflicts during the learning process. The volume of gap is expected to slightly reduce as we moved to the subsequent skip connections, because the encoder features were not only fused with the features from the decoder path of the newer layers but also moved with additional processing.

Hence, a new skip connection is proposed consisting several convolutional operations to lessen the difference between the encoder and decoder feature maps. 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 [36]. At first, the features are passed through a chain of convolutions and fused them with the decoder features instead of simply integration the encoder and decoder features. The semantic gaps between encoder and decoder features are expected to lessen using this chain of operations. Figure 4 illustrates the proposed skip connection. Precisely, the residual connections are accompanied by the 1×1 filters, and the 3×3 filters were utilized in the convolutions.

It is anticipated that, the strength of the semantic gaps between the encoder features maps and decoder ones are decreased as we passed through the internal shortcut paths. The number of convolutional blocks utilized along the three skip connections are also gradually decreased to 3, 2 and 1. Furthermore, filters of 64, 128, and 256 are used in the three skip connection blocks to consider the number of feature maps in encoder–decoder.

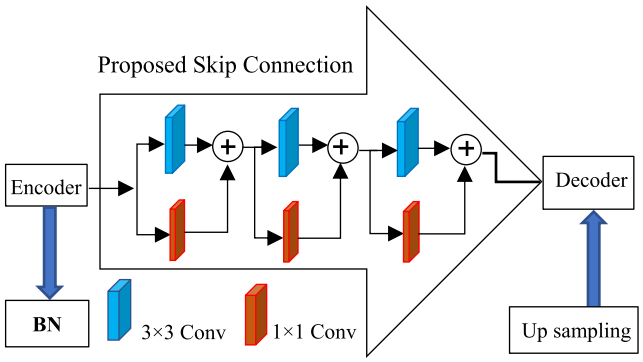

FIGURE 4. Proposed skip connection. A series of convolutions are used to pass the features from the encoder to decoder part instead of directly merging the features maps.

4) ATROUS SPATIAL PYRAMIDAL POOLING

The proposed architecture consists of a spatial pyramid pooling module called ASPP [52] to capture and aggregate multi-scale contextual information and helps in propagating fine detailed information from earlier layers to higher levels, and at the same time up-samples the feature maps to learn global contextual information in order to produce more accurate classification [52], [79].

The ASPP, the middle part of the proposed IRU-Net, acts as a bridge connecting the encoding and decoding paths in our architecture. Figure 5 presents the detailed structure of the ASPP module in the IRU-Net. In ASPP, the contextual information is captured at multiple scales [79] and many parallel atrous convolutions [7] with different rates in the input feature map are fused at the end [35]. Comparing with the standard convolutional layer, the atrous convolutions can effectively increase the receptive field of the network without extra down-samplings. In this work, the ASPP module is employed with a convolution of size 1×1 and three branches of atrous convolution with rate 6, 12, 18 as a connector in block after encoder to effectively capture multi-scale contextual information. Global average pooling is considered to acquire the feature map at the image level, permitting these two results to be combined and convoluted. The ASPP model yields promising results on segmentation of building by giving useful multi-scale information.

B. NETWORK TRAINING LOSS FUNCTION

Loss functions set the rules to evaluate the distance between network prediction and ground truth [82]. For our case, building extraction can be seen as a binary segmentation problem where the binary cross-entropy (BCE) loss function is most commonly used, as given by Equation (2). However, building segmentation from RS images have an imbalance problem between building pixels and background pixels, where the BCE loss function is prone to get stuck in local minima, and the network tends to predict the background for

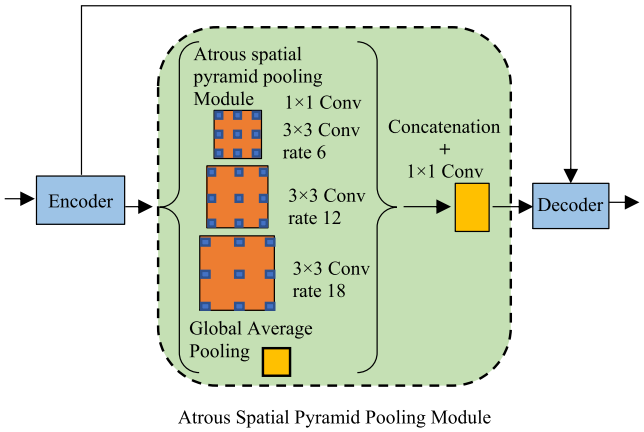

FIGURE 5. The proposed IRU-Net with ASPP integration: The ASPP module is inserted between the encoder-decoder network. The feature map generated by the encoder is processed by ASPP, and then, the output is fed into the decoder path.

a good loss value and fails to learn representative features of the minor class [81]. Therefore, the foreground area is usually partly identified or even missed. One way to tackle this problem is to assign a prior weight on each class when computing loss—a bigger weight for buildings in this case—which introduces additional hyper-parameters that need careful tuning. Another way is to choose a less biased function, such as dice and BCE loss [82].

In this study, since we have the same issue of imbalance classes such as building pixels (foreground) and non-building pixels (background) we opted a new dual objective loss function (BCEDL) [81] which combines both Binary Cross-Entropy loss function (BCE) (Equation (2)) and Dice coefficient (DL) (Equation (3)) to (i) integrating local information and global information, (ii) reduce the influence of class imbalance, and (iii) improve the building segmentation results.

The binary cross entropy (BCE) function is given as follows:

$$BCE\ (p_i, g_i) = -\sum_{i=1}^{N} (g_i \log (p_i) + (1 - g_i) \log (1 - p_i)) \quad (2)$$

where N is the pixel number, p_i and g_i represents the value of *i-th* pixel in the model prediction results and the ground truth value, respectively.

The Dice loss (DL) between two binary classes is defined as

$$DL = \frac{2 \sum_i^N p_i g_i}{\sum_i^N p_i^2 + \sum_i^N g_i^2} \quad (3)$$

where $g_i \in G$ is the ground truth pixels, $p_i \in P$ is the predicted binary pixels and N defines as total pixels.

Equation (4) defines the new loss function (L) which integrates the global information and local information to extract buildings more accurately.

$$LOSS_{BCE+DL} = BCE\ (p_i, g_i) + DL\ (p_i, g_i) \quad (4)$$

The proposed loss function integrates dice loss and BCE loss by addition to combine the advantages of both functions. BCE loss function well evaluates the misclassification and is easy to calculate the gradient mathematically despite the abovementioned flaw. The dice loss function, as given by Equation (3), is built on a dice coefficient that evaluates the overlap between the prediction and the ground truth whose values are ranging from 0 and 1. The more they match, the nearer dice coefficient is to one, pushing the dice loss to zero.

III. EXPERIMENTAL RESULTS AND DISCUSSION

In order to measure the effectiveness of IRU-Net for building extraction from high-resolution aerial imagery, we conduct numerous experiments on two public datasets: the Massachusetts Buildings Dataset [35] and the AIRS dataset [78]. The performance and efficiency of IRU-Net is also compared with some state-of-the-art models in semantic segmentation, including SegNet [47], ResU-Net [80], E-Net [68], ERFNet [69] and SRI-Net [61].

A. DATASETS

To verify the effectiveness of IRU-Net for building segmentation from high-resolution RS imagery, we conducted experiments on two public dataset building datasets: the AIRS Dataset [78] and the Massachusetts Building Dataset [35]. This section introduces the information of the data set used to train the proposed IRU-Net. It should be noted that both datasets are publicly available.

1) AIRS DATASET

This dataset is proposed by Chen *et al.* [78]. The AIRS dataset includes 1047 aerial images with the original spatial dimension of 10000 × 10000 and spatial resolution of 7.5 cm. The dataset covers a surface area of about 450 km2 in Christchurch, New Zealand and the whole aerial image and the corresponding ground truth are provided. Given computational restraints, we cut the original images into the size of 1536 × 1536. Due to the limitation in GPU memory, we randomly cropped all images into grid patches with a size 256 × 256 pixels. Subsequently, 1250 images are utilized in our experiment. We divided the dataset into a training set, a validation set, and a test set, consisting of 1125 images, 100 images, and 25 images, respectively. Figure 6 shows examples of input images and its corresponding label. The white color represents the buildings and the black color presents the background.

2) MASSACHUSETTS BUILDING DATASET

Massachusetts buildings dataset is proposed by Mnih [35]. It consists of 151 aerial RGB images of the Boston area with a spatial resolution of 1 m/pixel. Each image has a size of 1500 × 1500 pixels. There are 137 images in the training set, 10 images in the test set, and 4 images for validation with no overlapping areas. Similarly, we cropped the training and validation images with a size 256 × 256 pixels. After scanning, 3530 good quality images and corresponding labels

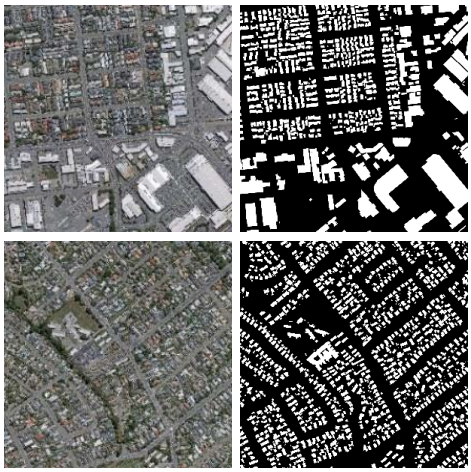

FIGURE 6. Examples of AIRS dataset images and its corresponding label.

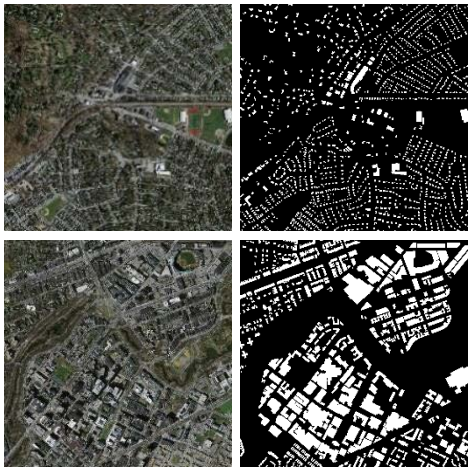

FIGURE 7. Examples of massachusetts building dataset images and its corresponding label.

are selected in which 3356 images are used for training, 144 images for validation and remaining 30 are used for testing, respectively. Also, data augmentation is used such as flipping images horizontally and vertically to expand the dataset. Figure 7 shows examples of input images and its corresponding label.

B. EXPERIMENTAL SETTINGS

The proposed IRU-Net building extraction model is implemented on the deep learning framework named PyTorch. The experiments were conducted on computer servers with two NVIDIA GeForce GTX 1080 Ti with a memory of 11GB. In addition, for easy network training, we randomly cropped all images in a size of 256 × 256 pixels for model training and validation. For training phase, the proposed network is optimized with ADAM optimizer [83] with an initial learning rate of 0.0005, weight decay of 0.00002 and momentum 0.9. Models has been trained with 300 epochs for both the AIRS dataset and the Massachusetts Building Dataset, respectively with mini-batch size set to 16.

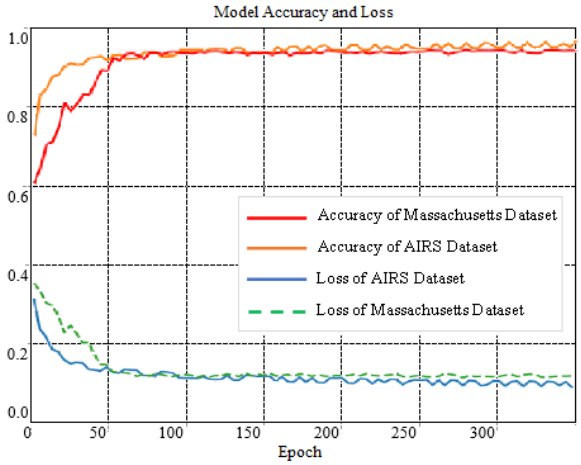

FIGURE 8. The accuracy and loss of the proposed model for training the massachusetts building dataset and AIRS dataset.

We use poly learning rate strategy for converging quickly that is computed as Equation (5) to adjust the learning rate:

$$lr = lr_{init}\left(1 - \frac{iter}{iter_{max}}\right)^{power} \tag{5}$$

where the initial learning rate is 0.0005 with power of 0.9. Figure 8 displays the dynamic accuracies and losses of the Massachusetts and AIRS datasets with increasing epochs. It is evident that the loss gradually decreases while the accuracy increases and stays at a high and stable level.

C. EVALUATION METRICS

In order to measure the effectiveness of IRU-Net for building extraction, the 'Overall Accuracy' (*OA*), 'Precision', 'Recall', 'F1-score', 'Intersection-over-Union (*IoU*)' are used as quality metrics [29], [81]. The *recall* value indicates the percentage of the ground truth road pixels detected. The *Precision* indicates the percentage of the correctly classified road pixels among all predicted pixels of the classifier. Finally, IoU is the number of pixels that are common between the predicted and ground truth divided by the total number of available pixels over both masks. The *F1-Score* indicates the harmonic average of *Precision* and *Recall.* The values of these metrics are in the range of 0 to 1, and higher values indicate better classification performance. The metrics are calculated as follows:

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

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

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

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{9}$$

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

where TP denotes the true positive; *TN* is the number of true negatives; FP denotes the false positive, and FN denotes the false negative.

D. COMPARISON OF BUILDING SEGMENTATION ALGORITHMS

We compared the suggested IRU-Net model with other state-of-the-art approaches to verify the performance of the proposed IRU-Net model in building segmentation. We selected CNN-based approaches such as the SegNet [47], ResU-Net [80], E-Net [68], ERFNet [69] and SRI-Net [61], for comparison, as the proposed model is a pixel-wise segmentation method. All experiments are evaluated based on five widely-used metrics: Overall Accuracy (OA), Recall, Precision, Intersection over Union (IoU) and F1-score (F1) [29], [81].

1) QUANTITATIVE AND QUALITATIVE RESULTS ON THE AIRS DATASET

For evaluating the effectiveness of the proposed IRU-Net model for building extraction, the qualitative segmentation results are presented for all six models on the AIRS dataset in Figure 11. There are five columns and 8 rows in this figure. The yellow, blue, red and black pixels of the maps denote the predictions of TP, FP, FN, and TN, respectively. In the first and second rows, the RGB and corresponding reference images are displayed. The third, fourth, fifth, sixth, seventh and eight rows display the outputs attained by SegNet [47], ResU-Net [80], E-Net [68], ERFNet [69] and SRI-Net [61], and proposed IRU-Net model, respectively. SegNet and ResU-Net return more FP (blue) and FN (red) than the other methods. SegNet returns slightly more false positives (blue) compared to other models. By contrast, the proposed IRU-Net model shows significantly less false positives (blue) and false negatives (red) than the other models, while maintaining high completeness in building segmentation on the AIRS dataset. Therefore, the segmentation map obtained by the proposed model is smoother than that of the other models, with higher accuracy and fewer FPs.

The quantitative comparison of the different networks for the five test images is presented in the first five rows in Table 1 and the average performance is shown in the last row of the Table 1. The proposed IRU-Net delivers improvements on all evaluation metrics over the other models except for precision. The IRU-Net model achieved best result among all models on OA metric with an improvement of 0.93% (0.9725 vs. 0.9635) over the next best model SRI-Net. As for Precision, the ERFNet model holds the highest values and gains 0.33% over IRU-Net (0.9668 vs. 0.9636). For Recall, the E-Net, SRI-Net, and IRU-Net methods scored significantly better performance over the other three methods while IRU-Net achieves the best value being 1.91% (0.9571 vs. 0.9388) ahead of the SRI-Net method. Similarly, IRU-Net achieves the best F1-score being 1.1% (0.9565 vs. 0.9459) ahead of the SRI-Net method where SRI-Net and ERFNet achieves the best model amongst the others. For the IoU metric, IRU-Net has scored the best value 1.2% ahead of ERFNet (0.9227 vs. 0.9108) and even 1% ahead of SRI-Net (0.9227 vs. 0.9133). Compared to the ResU-Net, IRU-Net yields a higher F1-score by 2.9% (0.9565 vs.0.9292) and OA by 5.2% (0.9725vs.0.9218). The overall average quantitative comparison of the different networks is depicted graphically in Figure 9.

TABLE 1. Quantitative comparison with state-of-the art models and the proposed IRU-Net on the AIRS dataset.

		SegNet	ResU-Net	E-Net	ERF Net	SRI-Net	Proposed Method
Image 1	Recall	0.8023	0.9113	0.9329	0.9331	0.9345	**0.9544**
	Precision	0.8858	0.9176	0.9406	0.9629	0.9525	**0.9639**
	F1-Score	0.8452	0.9282	0.9451	0.9421	0.9428	**0.9567**
	IoU	0.7868	0.8956	0.9034	0.9047	0.9091	**0.9195**
	OA	0.8772	0.9291	0.9455	0.9528	0.9663	**0.9718**
Image2	Recall	0.7982	0.9115	0.9396	0.9389	0.9415	**0.9601**
	Precision	0.8808	0.9266	0.9413	**0.9675**	0.9531	0.9609
	F1-Score	0.8472	0.9304	0.9464	0.9493	0.9443	**0.9538**
	IoU	0.7871	0.8928	0.9087	0.9119	0.9129	**0.9246**
	OA	0.8891	0.9147	0.9426	0.9587	0.9615	**0.9705**
Image3	Recall	0.8152	0.9245	0.9439	0.9431	0.9433	**0.9602**
	Precision	0.8895	0.9197	0.9418	**0.9727**	0.9542	0.9692
	F1-Score	0.8545	0.9312	0.9439	0.9462	0.9485	**0.9569**
	IoU	0.7863	0.8923	0.9031	0.9149	0.9123	**0.9255**
	OA	0.8948	0.9234	0.9391	0.9611	0.9623	**0.9725**
Image4	Recall	0.8177	0.9195	0.9379	0.9374	0.9385	**0.9548**
	Precision	0.8803	0.9167	0.9451	**0.9658**	0.9547	0.9621
	F1-Score	0.8565	0.9319	0.9399	0.9439	0.9462	**0.9586**
	IoU	0.7898	0.8932	0.9079	0.9103	0.9135	**0.9227**
	OA	0.8897	0.9197	0.9443	0.9601	0.9625	**0.9758**
Image5	Recall	0.8171	0.9255	0.9371	0.9361	0.9362	**0.9559**
	Precision	0.8818	0.9179	0.9406	**0.9651**	0.9591	0.9618
	F1-Score	0.8481	0.9245	0.9474	0.9438	0.9479	**0.9565**
	IoU	0.7768	0.8937	0.9063	0.9123	0.9185	**0.9214**
	OA	0.8701	0.9223	0.9427	0.9615	0.9649	**0.9718**
Average	Recall	0.8101	0.9185	0.9383	0.9377	0.9388	**0.9571**
	Precision	0.8836	0.9197	0.9419	**0.9668**	0.9547	0.9636
	F1-Score	0.8503	0.9292	0.9445	0.9451	0.9459	**0.9565**
	IoU	0.7854	0.8935	0.9059	0.9108	0.9133	**0.9227**
	OA	0.8842	0.9218	0.9428	0.9588	0.9635	**0.9725**

2) COMPARISON ON THE MASSACHUSETTS BUILDING DATASET

Building extraction results from the different models on a sample in the Massachusetts Building dataset are presented in Figure 12 for a qualitative comparison. It is clear that SegNet yields more FNs while U-Net yields more FPs in comparison to the other models. Overall, SRI-Net and IRU-Net predict buildings reasonably fine.

We further conducted a quantitative comparison with different models on the Massachusetts Building dataset. The results of the quantitative comparison are summarized in Table 2. In contrast to the AIRS dataset where ERFNet performed the second best, SRI-Net shows the best performance amongst the established methods. SRI-Net has the highest recall of 0.9177 with an improvement of 0.39% (0.9177 vs. 0.9141) while the proposed IRU-Net performs best except for

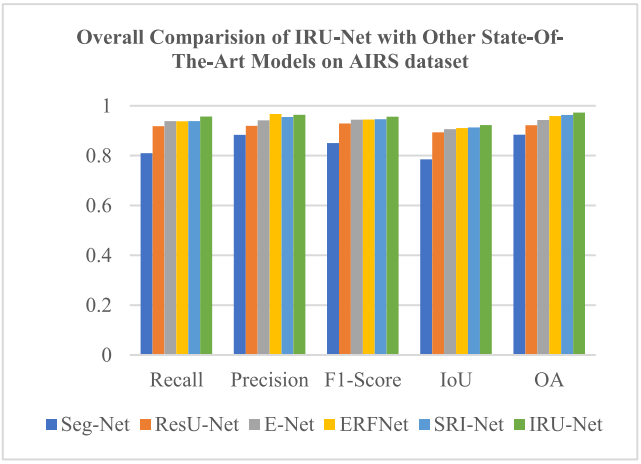

FIGURE 9. Quantitative comparison with state-of-the art models and the proposed IRU-Net on the AIRS dataset.

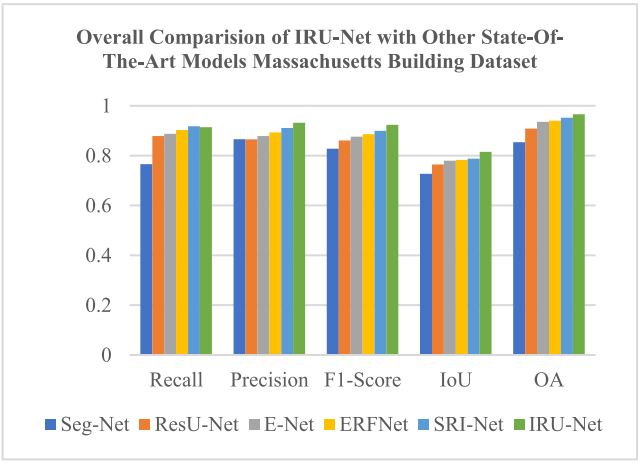

FIGURE 10. Quantitative comparison with state-of-the art models and the proposed IRU-Net on the massachusetts building dataset.

TABLE 2. Quantitative comparison with state-of-the art models and the proposed IRU-Net on the massachusetts building dataset.

		Seg Net	ResU-Net	E-Net	ERF Net	SRI-Net	Proposed Method
Image 1	Recall	0.7631	0.8673	0.8883	0.8995	**0.9168**	0.9118
	Precision	0.8616	0.8668	0.8845	0.8949	0.9171	**0.9384**
	F1-Score	0.8192	0.8672	0.8821	0.9025	0.9092	**0.9231**
	IoU	0.7315	0.7609	0.7734	0.7834	0.7991	**0.8178**
	OA	0.8502	0.9111	0.9352	0.9412	0.9507	**0.9642**
Image2	Recall	0.7731	0.8785	0.8923	0.9147	0.9238	**0.9242**
	Precision	0.8732	0.8595	0.8845	0.9056	0.9115	**0.9391**
	F1-Score	0.8282	0.8572	0.8771	0.8831	0.9085	**0.9253**
	IoU	0.7215	0.7636	0.7885	0.7841	0.7941	**0.8191**
	OA	0.8519	0.9026	0.9405	0.9417	0.9514	**0.9684**
Image3	Recall	0.7682	0.8785	0.8875	0.8985	**0.9195**	0.9189
	Precision	0.8691	0.8697	0.8691	0.8891	0.9161	**0.9304**
	F1-Score	0.8327	0.8585	0.8791	0.8955	0.9085	**0.9358**
	IoU	0.7401	0.7721	0.7861	0.7851	0.7813	**0.8101**
	OA	0.8595	0.9015	0.9345	0.9429	0.9525	**0.9673**
Image4	Recall	0.7629	0.8898	0.8788	0.8999	**0.9115**	0.9032
	Precision	0.8661	0.8701	0.8799	0.8908	0.9021	**0.9104**
	F1-Score	0.8298	0.8628	0.8774	0.8789	0.8885	**0.9052**
	IoU	0.7305	0.7705	0.7902	0.7884	0.7854	**0.8112**
	OA	0.8554	0.9105	0.9326	0.9431	0.9536	**0.9645**
Image5	Recall	0.7621	0.8801	0.8908	0.8995	**0.9168**	0.9122
	Precision	0.8604	0.8595	0.8745	0.8849	0.9091	**0.9397**
	F1-Score	0.8292	0.8572	0.8651	0.8709	0.8837	**0.9275**
	IoU	0.7135	0.7556	0.7604	0.7734	0.7791	**0.8178**
	OA	0.8512	0.9161	0.9372	0.9318	0.9522	**0.9662**
Average	Recall	0.7659	0.8788	0.8875	0.9024	**0.9177**	0.9141
	Precision	0.8661	0.8651	0.8785	0.8931	0.9112	**0.9316**
	F1-Score	0.8278	0.8606	0.8762	0.8862	0.8997	**0.9234**
	IoU	0.7274	0.7645	0.7797	0.7829	0.7878	**0.8152**
	OA	0.8536	0.9084	0.936	0.9401	0.9521	**0.9661**

the Recall score. In the testing case, IRU-net achieves the best values on OA metric with an improvement of 1.45% compared to SRI-Net (0.9661 vs 0.9521). For precision, IRU-Net achieves an improvement over the next best model SRI-Net of 2.19% (0.9316 vs. 0.9112). For Recall, IRU-Net achieves an improvement of 3.9% over ResU-Net (0.9141 vs. 0.8788). As for F1-score and IoU, IRU-Net obtains the highest value over the other models and outperforms SRI-Net by 2.6% (0.9234 vs 0.8997) and 3.4% (0.8152 vs 0.7878) and outperforms ERFNet by 4% (0.9234 vs 0.8862) and 3.9% (0.8152 vs 0.7829) which are considered to be state-of the-art networks for segmentation. Compared to the ResU-Net, IRU-Net yields a higher F1-score by 6.8% (0.9234 vs.0.8606) and a higher IoU by 6.2% (0.8152 vs. 0.7645). For Overall Accuracy, IRU-Net holds the highest values with a gain of 6.0% compared to ResU-Net (0.9661 vs. 0.9084). The overall average quantitative comparison of the different networks is depicted graphically in Figure 10.

E. EFFECT OF ASPP

The ASPP model has shown promising results on building segmentation tasks by providing useful multi-scale information which improve the accuracy of buildings segmentation in different sizes, especially medium-sized to over-sized buildings [73]. One crucial innovation of the IRU-Net is that it employs the ASPP module as a bridge between the encoder and the decoder. To test the performance, we conducted a comparison experiment with and without the ASPP module of IRU-Net on the AIRS dataset. As presented in Table 3, the model with ASPP shows an obvious improvement over the model without ASPP across all evaluation metrics. The comparison result shows the improvement of efficiency and applicability of the ASPP module as a bridge connector between encoder-decoder path of the proposed IRU-Net model for building extraction from high-resolution RS images [33]. The overall effect of ASPP on AIRS dataset is depicted graphically in Fig. 13.

As it can be seen from Figure 13, the proposed IRU-Net with ASPP achieves the best values on OA metric with an improvement of 0.915% compared to IRU-Net without ASPP (0.9725 vs 0.9636). IRU-Net with ASPP could accomplish higher average accuracy over the IRU-Net without ASPP and outperforms by 1.306% (0.9571 vs 0.9501) for Recall, 1.017% (0.9636 vs 0.9543) for precision, 1.181% (0.9565 vs

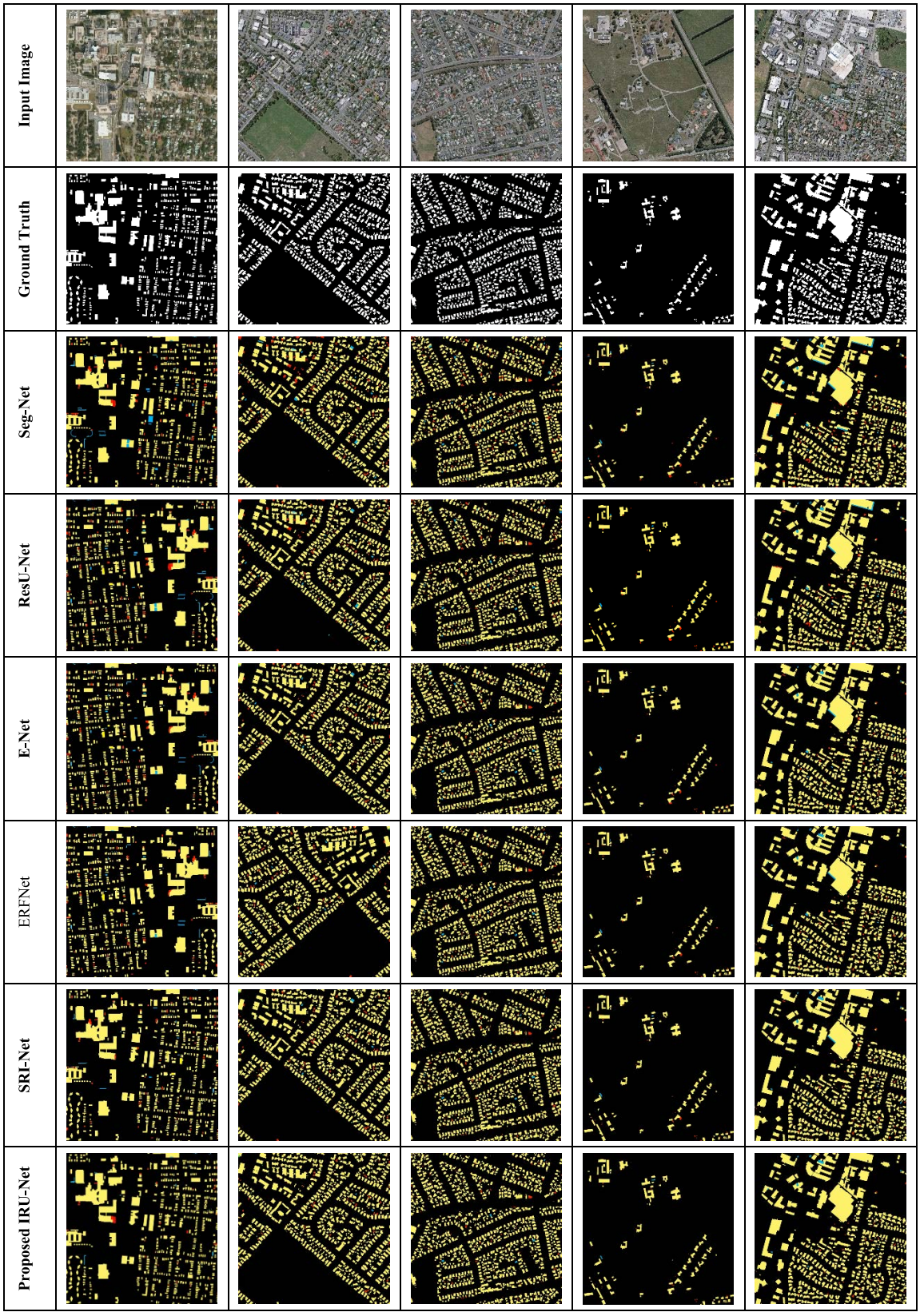

FIGURE 11. Building extraction results from AIRS dataset achieved by proposed model and other state-of-the-art methods. First and second rows show the original images and corresponding ground truth. The third, fourth, fifth, sixth, seventh and eighth rows are the results achieved by the SegNet [47], ResU-Net [72], E-Net [62], ERFNet [63] and SRI-Net [59], and the proposed IRU-Net architecture, respectively. The black (background), yellow, blue and red colours represent TNs, TPs, FPs, and FNs, respectively.

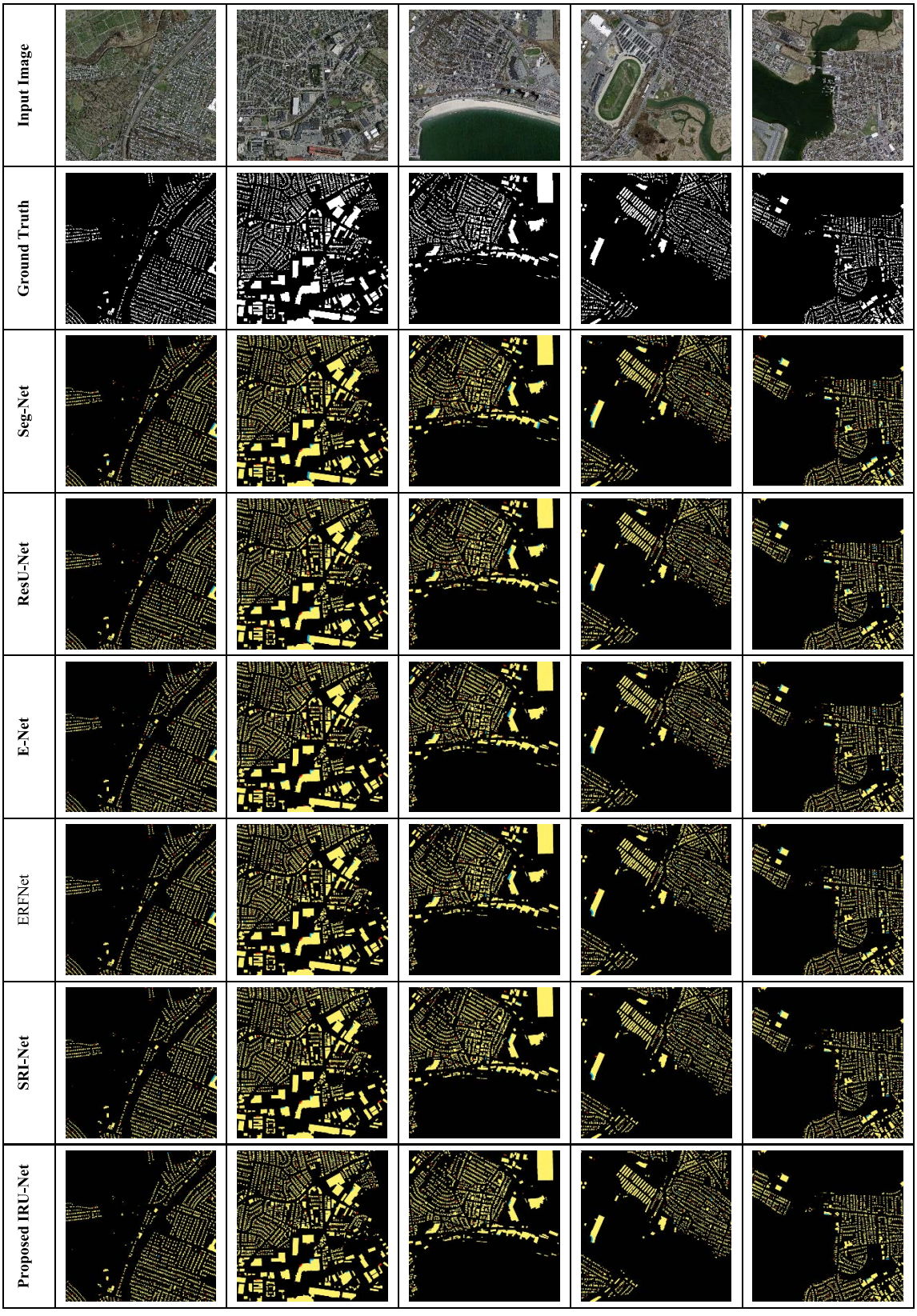

FIGURE 12. Building extraction results from massachusetts building dataset achieved by proposed model and other state-of-the-art methods. First and second rows show the original images and corresponding ground truth. The third, fourth, fifth, sixth, seventh and eighth rows are the results achieved by the SegNet [47], ResU-Net [72], E-Net [62], ERFNet [63] and SRI-Net [59], and the proposed IRU-Net architecture, respectively. The black (background), yellow, blue and red colours represent TNs, TPs, FPs, and FNs, respectively.

TABLE 3. Comparison of the IRU-Net with or without the ASPP module on the AIRS dataset.

		Image1	Image2	Image3	Image4	Image5	Average
IRU-Net Without ASPP	Recall	0.9492	0.9493	0.9513	0.9489	0.9476	0.9501
	Precision	0.9519	0.9549	0.9663	0.9543	0.9543	0.9543
	F1-Score	0.9489	0.9428	0.9488	0.9521	0.9541	0.9451
	IoU	0.9045	0.9054	0.9075	0.9040	0.9040	0.9040
	OA	0.9589	0.9619	0.9624	0.9603	0.9661	0.9636
IRU-Net With ASPP	Recall	0.9544	0.9601	0.9602	0.9548	0.9559	**0.9571**
	Precision	0.9639	0.9609	0.9692	0.9621	0.9618	**0.9636**
	F1-Score	0.9567	0.9538	0.9569	0.9586	0.9565	**0.9565**
	IoU	0.9195	0.9246	0.9255	0.9227	0.9214	**0.9227**
	OA	0.9718	0.9705	0.9725	0.9758	0.9718	**0.9725**

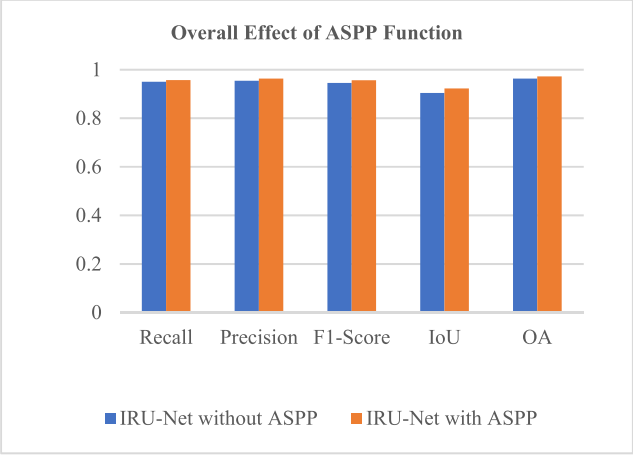

FIGURE 13. The comparative quantitative evaluation measured in terms of Recall, Precision, F1-score, IoU and OA with effect of ASPP.

0.9451) for F-1 score, 1.311% (0.9227 vs 0.9040) for IoU and 1.275 (0.9725 vs. 0.9601) for OA respectively for AIRS dataset.

F. EFFECT OF LOSS FUNCTION

Moreover, we assessed the accuracy measurements of IRU-Net with BCE loss function, IRU-Net with DL loss function, and IRU-Net with BCEDL loss function for Massachusetts building and AIRS datasets to review the fitness of the proposed model for building extraction. Table 4 and Table 5 illustrate the accuracy of each defined metric for the Massachusetts building and AIRS datasets, respectively. As it can be seen from both Tables, the proposed IRU-Net+BCEDL network could accomplish higher average accuracy than IRU-Net+BCE, IRU-Net+DL, for Recall, precision, F-1 score, IoU and OA with 91.41%, 93.16%, 92.34%, 81.52% and 96.61%, respectively for Massachusetts building dataset; and 95.71%, 96.36%, 95.65%, 92.27% and 97.25%, respectively for AIRS dataset. It is clear that the proposed IRU-Net with BCEDL loss function network achieves good results for the building extraction from both

TABLE 4. Comparing IRU-Net model with BCE, DL and BCEDL loss functions for building extraction form massachusetts building dataset.

		Image1	Image2	Image3	Image4	Image5	Average
IRU-Net + BCE	Recall	0.8987	0.9116	0.9088	0.8928	0.9054	0.9035
	Precision	0.9234	0.9264	0.9214	0.9015	0.9308	0.9207
	F1-Score	0.9102	0.9122	0.8611	0.8985	0.9142	0.8992
	IoU	0.8075	0.8102	0.8003	0.8016	0.8018	0.8043
	OA	0.9545	0.9518	0.9536	0.9552	0.9548	0.9539
IRU-Net + DL	Recall	0.9012	0.9168	0.9115	0.8975	0.9085	0.9071
	Precision	0.9284	0.9302	0.9256	0.9059	0.9326	0.9245
	F1-Score	0.9154	0.9175	0.8655	0.9005	0.9196	0.9037
	IoU	0.8112	0.8122	0.8042	0.8072	0.8058	0.8081
	OA	0.9598	0.9565	0.9588	0.9585	0.9578	0.9583
IRU-Net + BCEDL	Recall	0.9118	0.9242	0.9189	0.9032	0.9122	**0.9141**
	Precision	0.9384	0.9391	0.9304	0.9104	0.9397	**0.9316**
	F1-Score	0.9231	0.9253	0.9358	0.9052	0.9275	**0.9234**
	IoU	0.8178	0.8191	0.8101	0.8112	0.8178	**0.8152**
	OA	0.9642	0.9684	0.9673	0.9645	0.9662	**0.9661**

TABLE 5. Comparing IRU-Net model with BCE, DL and BCEDL loss functions for building extraction on AIRS dataset.

		Image1	Image2	Image3	Image4	Image5	Average
IRU-Net + BCE	Recall	0.9428	0.9489	0.9442	0.9445	0.9428	0.9446
	Precision	0.9539	0.9513	0.9572	0.9528	0.9536	0.9538
	F1-Score	0.9419	0.9428	0.9485	0.9496	0.9431	0.9452
	IoU	0.9024	0.9118	0.9124	0.9135	0.9128	0.9106
	OA	0.9579	0.9575	0.9604	0.9605	0.9644	0.9601
IRU-Net + DL	Recall	0.9475	0.9539	0.9486	0.9502	0.9493	0.9499
	Precision	0.9561	0.9553	0.9601	0.9608	0.9566	0.9578
	F1-Score	0.9449	0.9468	0.9475	0.9505	0.9481	0.9476
	IoU	0.9058	0.9155	0.9154	0.9161	0.9188	0.9143
	OA	0.9609	0.9612	0.9654	0.9665	0.9662	0.9640
IRU-Net + BCEDL	Recall	0.9544	0.9601	0.9602	0.9548	0.9559	**0.9571**
	Precision	0.9639	0.9609	0.9692	0.9621	0.9618	**0.9636**
	F1-Score	0.9567	0.9538	0.9569	0.9586	0.9565	**0.9565**
	IoU	0.9195	0.9246	0.9255	0.9227	0.9214	**0.9227**
	OA	0.9718	0.9705	0.9725	0.9758	0.9718	**0.9725**

datasets and determines that the segmented building sections are close to ground truth, verifying the effectiveness of our approach in building extraction. Furthermore, the proposed IRU-Net with BCEDL loss function network could achieve higher efficiency on the segmentation results than the other comparative approaches. The overall effect of loss function on AIRS Dataset and Massachusetts building dataset is depicted graphically in Figure 14 and Figure 15.

Compared to the IRU-Net+BCE, IRU-Net+BCEDL yields higher OA by 1.262% (0.9661 vs.0.9539), F1-score by 2.62% (0.9234 vs.0.8992), recall 1.159 (0.9141 vs. 0.9035), precision by 1.17% (0.9316 vs. 0.9207) and a higher IoU by 1.337% (0.8151 vs. 0.8043) for Massachusetts building dataset; and for AIRS dataset 1.25% (0.9571 vs 0.9446) for Recall, 0.98% (0.9636 vs 0.9538) for precision, 1.13% (0.9565 vs 0.9452) for F-1 score, 1.21% (0.9227 vs 0.9106) for IoU and 1.24% (0.9725 vs 0.9601) for OA respectively.

As it can be seen from both Table 4 and V, the proposed IRU-Net+BCEDL outperforms IRU-Net+DL by 0.752%

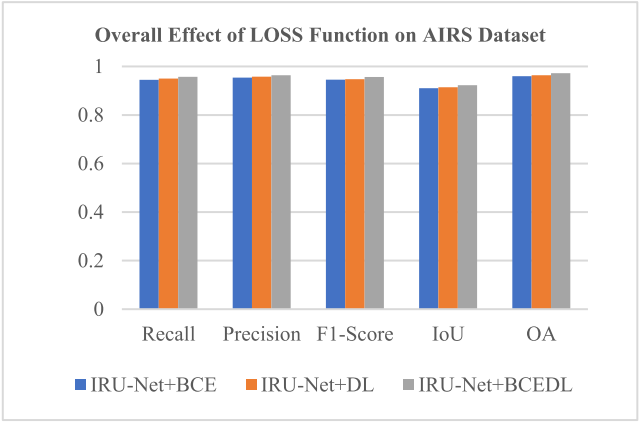

FIGURE 14. The comparative quantitative evaluation measured in terms of Recall, Precision, F1-score, IoU and OA with effect of loss functions on AIRS dataset.

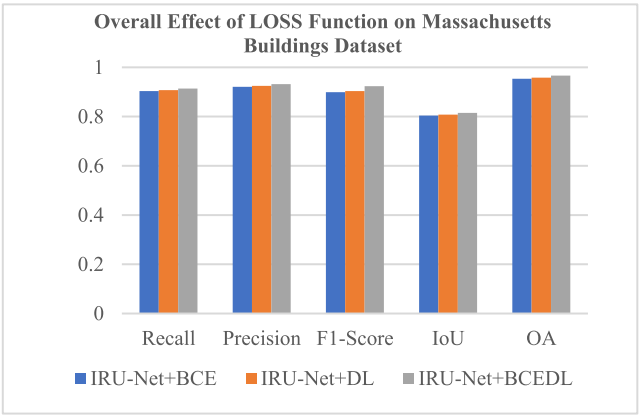

FIGURE 15. The comparative quantitative evaluation measured in terms of Recall, Precision, F1-score, IoU and OA with effect of loss functions on massachusetts building dataset.

(0.9571 vs 0.9499) for Recall, 0.601% (0.9636 vs 0.9578) for precision, 0.930% (0.9565 vs 0.9476) for F-1 score, 0.910% (0.9227 vs 0.9143) for IoU and 0.874% (0.9725 vs 0.9640) respectively for AIRS dataset. Compared to the IRU-Net+DL, IRU-Net+BCEDL yields higher OA by 0.807% (0.9661 vs.0.9583), F1-score by 2.133% (0.9234 vs. 0.9037), recall by 0.675 (0.9141 vs. 0.9071), precision by 0.762% (0.9316 vs. 0.9245) and a higher IoU by 1.337% (0.8152 vs. 0.8081) for Massachusetts building dataset.

G. MODEL EFFICIENCY

To address the model efficiency, we further compared the computational cost of different models in terms of floating-point of operations (FLOPs) and the number of trainable parameters [1]. In deep learning, the complexity of networks could be measured by these two metrics. Higher FLOPs and more trainable parameters correspond to greater complexity of a model. The amount of computation consumption of all models is calculated on a 256 × 256 RS image. As shown in Table 6, we compared our proposed

TABLE 6. Comparison of flops and trainable parameters between IRU-Net and other state-of-the-art models.

Model	FLOPs (G)	Parameters (M)
Seg-Net	80.323	39.44
U-Net	16.591	32.08
E-Net	12.215	5.38
ERFNet	14.674	12.06
SRI-Net	73.494	26.78
Proposed IRU-Net	11.113	6.01

IRU-Net model with other state-of-the-art networks (i.e., Seg-Net, U-Net, E-Net, ERFNet and SRI-Net). All the comparison results are based on test set with the same training environment and configuration.

As shown in Table 6, it could be seen that U-Net has the smallest number of trainable parameters and FLOPs because of its simple structure than Seg-Net. Seg-Net still has so many parameters of 39.44M though it has much lighter decoder than other encoder-decoder architectures. Our IRU-Net achieves 11.113G FLOPs and 6.01 M parameters. The proposed IRU-Net has lowest number of FLOPs. The proposed IRU-Net has relatively lowest FLOPs, while the number of required training parameters is lowest compared with other models except E-net. However, due to the extra parameters in the upsampling path, IRU-Net has more parameters than E-Net and second-lowest in trainable parameters among all models.

The results in Table 6 indicate that our IRU-Net model obtain much better efficiency than other established state-of-the-art models.

H. LIMITATIONS

Despite the improvements in semantic segmentation of buildings from RS images the proposed IRU-Net, some issues remain to be considered. With the rapid development of RS technology, the availability of high-resolution RS imagery with abundant features and spectral information is significantly increased [34]. Extraction of building from RS images plays a vital role in a wide range of RS applications but highly challenging task, due to the diverse characteristics of building and poses a major challenge for computer vision and image processing researchers. The proposed IRU-Net model is able to helps to improve the accuracy of semantic segmentation.

However, this model may fail to generalize to areas with complex and heterogeneous buildings because the datasets used in this research do not cover images from different sensors, such as hyperspectral images, DSMs or Light Detection and Ranging (LiDAR) DSM and SAR images. Spectral information is not enough since roads and building roofs can have similar texture. Moreover, in observing only two-dimensional images, we lose the third dimension—height. As a result, accuracy and robustness of the extraction results could be improved by integration of different data sources. Therefore, fusing data sources, such as multi-spectral images

with either stereo DSMs or LiDAR DSM rather than the use of only a single data source can be used for solving these problems and as a result, improve image interpretation. However, these data provide information complementary to the data in the visual spectrum, and therefore there is the potential that training models with these additional data may lead to better segmentation results.

IV. CONCLUSION

Accurate and automatic building segmentation from RS imagery is essential for application areas such as urban planning and disaster management. In this paper, we proposed a CNN framework, named IRU-Net, to perform building segmentation on high-resolution RS images. The significant contribution of this work is the analysis of the advantages of existing FCN-based models and the development of a novel model signifying that the two powerful tools the encoder-decoder and spatial pyramid pooling module need to be fused to improve building segmentation task. Moreover, a new skip connection is utilized to mitigate the semantic gaps between the encoder and decoder features. Also, we executed a new loss function termed BCEDL to reduce the problem of class imbalance in our datasets and improved the result of building segmentation.

Experiments were conducted on two public building datasets: the Massachusetts and AIRS datasets. The results show that the proposed IRU-Net model achieves high accuracy on these two datasets. The qualitative and quantitative comparison with the state-of-the-art models SegNet, ResU-Net, E-Net, ERFNet and SRI-net have demonstrated that IRU-Net outperforms these models. Compared with the ResU-Net, IRU-Net gains 5.97% (0.9661 vs. 0.9084) and 5.2% (0.9725vs.0.9218) improvements in Overall Accuracy for Massachusetts datasets and AIRS dataset with the small increase of 3.6% and 2.1% in model-training time on the Massachusetts and the AIRS dataset respectively.

CONFLICT OF INTEREST

All authors have no conflict of interest.

REFERENCES

[1] J. Lin, W. Jing, H. Song, and G. Chen, "ESFNet: Efficient network for building extraction from high-resolution aerial images," *IEEE Access*, vol. 7, pp. 54285–54294, 2019, doi: 10.1109/ACCESS.2019.2912822.

[2] Y. Liu, Z. Li, B. Wei, X. Li, and B. Fu, "Seismic vulnerability assessment at urban scale using data mining and GIScience technology: Application to Urumqi (China)," *Geomatics, Natural Hazards Risk*, vol. 10, no. 1, pp. 958–985, Jan. 2019.

[3] T. Panboonyuen, K. Jitkajornwanich, S. Lawawirojwong, P. Srestasathiern, and P. Vateekul, "Semantic segmentation on remotely sensed images using an enhanced global convolutional network with channel attention and domain specific transfer learning," *Remote Sens.*, vol. 11, no. 1, p. 83, 2019.

[4] A. Abdollahi, B. Pradhan, S. Gite, and A. Alamri, "Building footprint extraction from high resolution aerial images using generative adversarial network (GAN) architecture," *IEEE Access*, vol. 8, pp. 209517–209527, 2020, doi: 10.1109/ACCESS.2020.3038225.

[5] 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 J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 8, pp. 2615–2629, Aug. 2018.

[6] Q. Zhu, C. Liao, H. Hu, X. Mei, and H. Li, "MAP-Net: Multiple attending path neural network for building footprint extraction from remote sensed imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 6169–6181, Jul. 2021, doi: 10.1109/TGRS.2020.3026051.

[7] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[8] S. Wang, X. Hou, and X. Zhao, "Automatic building extraction from high-resolution aerial imagery via fully convolutional encoder-decoder network with non-local block," *IEEE Access*, vol. 8, pp. 7313–7322, 2020, doi: 10.1109/ACCESS.2020.2964043.

[9] Y. Liu, J. Zhou, W. Qi, X. Li, L. Gross, Q. Shao, Z. Zhao, L. Ni, X. Fan, and Z. Li, "ARC-Net: An efficient network for building extraction from high-resolution aerial images," *IEEE Access*, vol. 8, pp. 154997–155010, 2020, doi: 10.1109/ACCESS.2020.3015701.

[10] W. Li, C. He, J. Fang, J. Zheng, H. Fu, and L. Yu, "Semantic segmentation-based building footprint extraction using very high-resolution satellite images and multi-source GIS data," *Remote Sens.*, vol. 11, no. 4, p. 403, Feb. 2019.

[11] J. Du, D. Chen, R. Wang, J. Peethambaran, P. T. Mathiopoulos, L. Xie, and T. Yun, "A novel framework for 2.5-D building contouring from large-scale residential scenes," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 4121–4145, Jun. 2019.

[12] Z. Li, W. Shi, Q. Wang, and Z. Miao, "Extracting man-made objects from high spatial resolution remote sensing images via fast level set evolutions," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 2, pp. 883–899, Feb. 2014.

[13] N. L. Gavankar and S. K. Ghosh, "Automatic building footprint extraction from high-resolution satellite image using mathematical morphology," *Eur. J. Remote Sens.*, vol. 51, no. 1, pp. 182–193, 2018.

[14] J.-P. Burochin, B. Vallet, M. Brédif, C. Mallet, T. Brosset, and N. Paparoditis, "Detecting blind building façades from highly overlapping wide angle aerial imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 96, pp. 193–209, Oct. 2014.

[15] M. Cote and P. Saeedi, "Automatic rooftop extraction in nadir aerial imagery of suburban regions using corners and variational level set evolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 1, pp. 313–328, Jan. 2012.

[16] G. Zhou and X. Zhou, "Seamless fusion of LiDAR and aerial imagery for building extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 11, pp. 7393–7407, Nov. 2014.

[17] M. A. A. Sheikh and S. Mukhopadhyay, "Noise tolerant classification of aerial images into manmade structures and natural-scene images based on statistical dispersion measures," in *Proc. Annu. IEEE India Conf. (INDICON)*, Dec. 2012, pp. 653–658, doi: 10.1109/INDCON.2012.6420699.

[18] M. A. A. Sheikh, "A novel self-assessed approach for classification of manmade objects and natural scene images from aerial images," in *Proc. Annu. IEEE India Conf.*, Dec. 2011, pp. 1–7, doi: 10.1109/INDCON.2011.6139328.

[19] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS J. Photogramm. Remote Sens.*, vol. 152, pp. 166–177, Jun. 2019.

[20] M. Awrangjeb, C. Zhang, and C. S. Fraser, "Automatic extraction of building roofs using LIDAR data and multispectral imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 83, pp. 1–18, Sep. 2013.

[21] S. Du, Y. Zhang, Z. Zou, S. Xu, X. He, and S. Chen, "Automatic building extraction from LiDAR data fusion of point and grid-based features," *ISPRS J. Photogramm. Remote Sens.*, vol. 130, pp. 294–307, Aug. 2017.

[22] S. A. N. Gilani, M. Awrangjeb, and G. Lu, "An automatic building extraction and regularisation technique using LiDAR point cloud data and orthoimage," *Remote Sens.*, vol. 8, no. 3, p. 258, Mar. 2016.

[23] G. Sohn and I. Dowman, "Data fusion of high-resolution satellite imagery and LiDAR data for automatic building extraction," *ISPRS J. Photogramm. Remote Sens.*, vol. 62, no. 1, pp. 43–63, 2007.

[24] J. Inglada, "Automatic recognition of man-made objects in high resolution optical remote sensing images by SVM classification of geometric image features," *ISPRS J. Photogramm. Remote Sens.*, vol. 62, no. 3, pp. 236–248, Aug. 2007.

[25] T. Celik, "Unsupervised change detection in satellite images using principal component analysis and k-means clustering," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 4, pp. 772–776, Oct. 2009.

[26] Ö. Aytekin, U. Zöngür, and U. Halici, "Texture-based airport runway detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 3, pp. 471–475, May 2013.

[27] Y. Dong, B. Du, and L. Zhang, "Target detection based on random forest metric learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 4, pp. 1830–1838, Apr. 2015.

[28] E. Li, J. Femiani, S. Xu, X. Zhang, and P. Wonka, "Robust rooftop extraction from visible band images using higher order CRF," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4483–4495, Aug. 2015.

[29] M. A. A. Sheikh, A. Kole, and T. Maity, "A multi-level approach for change detection of buildings using satellite imagery," *Int. J. Artif. Intell. Tools*, vol. 27, no. 8, Dec. 2018, Art. no. 1850031, doi: 10.1142/S0218213018500318.

[30] X. Huang and L. Zhang, "Morphological building/shadow index for building extraction from high-resolution imagery over urban areas," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 5, no. 1, pp. 161–172, Feb. 2012.

[31] A. O. Ok, "Automated detection of buildings from single VHR multispectral images using shadow information and graph cuts," *ISPRS J. Photogramm. Remote Sens.*, vol. 86, pp. 21–40, Dec. 2013.

[32] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[33] R. Alshehhi, P. R. Marpu, W. L. Woon, and M. D. Mura, "Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 130, pp. 139–149, Aug. 2017.

[34] J. Hui, M. Du, X. Ye, Q. Qin, and J. Sui, "Effective building extraction from high-resolution remote sensing images with multitask driven deep neural network," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 5, pp. 786–790, May 2018.

[35] V. Mnih, "Machine learning for aerial image labeling," Ph.D. dissertation, Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, 2013.

[36] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.

[37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[38] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 2, pp. 84–90, Jun. 2012.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[40] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[41] S. Saito, T. Yamashita, and Y. Aoki, "Multiple object extraction from aerial imagery with convolutional neural networks," *J. Imag. Sci. Technol.*, vol. 60, no. 1, pp. 104021–104029, Jan. 2016, doi: 10.2352/J.ImagingSci.Technol.2016.60.1.010402.

[42] X. Wei, K. Fu, X. Gao, M. Yan, X. Sun, K. Chen, and H. Sun, "Semantic pixel labelling in remote sensing images using a deep convolutional encoder-decoder model," *Remote Sens. Lett.*, vol. 9, no. 3, pp. 199–208, 2018.

[43] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[44] Z. Zhong, J. Li, W. Cui, and H. Jiang, "Fully convolutional networks for building and road extraction: Preliminary results," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2016, pp. 1591–1594.

[45] T. Zuo, J. Feng, and X. Chen, "HF-FCN: Hierarchically fused fully convolutional network for robust building extraction," in *Proc. Asian Conf. Comput. Vis.*, 2016, pp. 291–302.

[46] Y. Xu, L. Wu, Z. Xie, and Z. Chen, "Building extraction in very high resolution remote sensing imagery using deep learning and guided filters," *Remote Sens.*, vol. 10, no. 1, p. 144, Jan. 2018.

[47] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Jan. 2017.

[48] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Springer, 2015, pp. 234–241.

[49] A. Abdollahi and B. Pradhan, "Integrating semantic edges and segmentation information for building extraction from aerial images using UNet," *Mach. Learn. Appl.*, vol. 6, Dec. 2021, Art. no. 100194.

[50] K. Bittner, S. Cui, and P. Reinartz, "Building extraction from remote sensing data using fully convolutional networks," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci. Arch.*, vol. 42, no. W1, pp. 481–486, 2017, doi: 10.5194/isprs-archives-XLII-1-W1-481-2017.

[51] G. Wu, X. Shao, Z. Guo, Q. Chen, W. Yuan, X. Shi, Y. Xu, and R. Shibasaki, "Automatic building segmentation of aerial imagery using multi-constraint fully convolutional networks," *Remote Sens.*, vol. 10, no. 3, p. 407, 2018.

[52] Y. Liu, L. Gross, Z. Li, X. Li, X. Fan, and W. Qi, "Automatic building extraction on high-resolution remote sensing imagery using deep convolutional encoder-decoder with spatial pyramid pooling," *IEEE Access*, vol. 7, pp. 128774–128786, 2019, doi: 10.1109/ACCESS.2019.2940527.

[53] S. Ji, S. Wei, and M. Lu, "A scale robust convolutional neural network for automatic building extraction from aerial and satellite imagery," *Int. J. Remote Sens.*, vol. 40, no. 9, pp. 3308–3322, May 2019.

[54] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2019.

[55] L. Li, J. Liang, M. Weng, and H. Zhu, "A multiple-feature reuse network to extract buildings from remote sensing imagery," *Remote Sens.*, vol. 10, no. 9, p. 1350, Sep. 2018.

[56] K. Fu, Z. Chang, Y. Zhang, G. Xu, K. Zhang, and X. Sun, "Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 161, pp. 294–308, Mar. 2020.

[57] G. Sun, H. Huang, A. Zhang, F. Li, H. Zhao, and H. Fu, "Fusion of multiscale convolutional neural networks for building extraction in very high-resolution images," *Remote Sens.*, vol. 11, no. 3, p. 227, 2019.

[58] S. Wei, S. Ji, and M. Lu, "Toward automatic building footprint delineation from aerial images using CNN and regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 2178–2189, Mar. 2020, doi: 10.1109/TGRS.2019.2954461.

[59] Z. Zhang and Y. Wang, "JointNet: A common neural network for road and building extraction," *Remote Sens.*, vol. 11, no. 6, p. 696, 2019.

[60] Y. Zhang, W. Gong, J. Sun, and W. Li, "Web-Net: A novel nest networks with ultra-hierarchical sampling for building extraction from aerial imageries," *Remote Sens.*, vol. 11, no. 16, p. 1897, Aug. 2019, doi: 10.3390/rs11161897.

[61] P. Liu, X. Liu, M. Liu, Q. Shi, J. Yang, X. Xu, and Y. Zhang, "Building footprint extraction from high-resolution images via spatial residual inception convolutional neural network," *Remote Sens.*, vol. 11, no. 7, p. 830, Apr. 2019.

[62] 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 Int.*, pp. 1–16, Dec. 2020, doi: 10.1080/10106049.2020.1856199.

[63] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," 2014, *arXiv:1412.7062*.

[64] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[65] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," 2018, *arXiv:1802.02611*.

[66] S. Shrestha and L. Vanneschi, "Improved fully convolutional network with conditional random fields for building extraction," *Remote Sens.*, vol. 10, no. 7, p. 1135, Jul. 2018.

[67] P. Luc, C. Couprie, S. Chintala, and J. Verbeek, "Semantic segmentation using adversarial networks," in *Proc. NIPS Workshop Adversarial Training*, 2016, pp. 1–12.

[68] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "ENet: A deep neural network architecture for real-time semantic segmentation," 2016, *arXiv:1606.02147*.

[69] E. Romera, J. M. Alvarez, L. M. Bergasa, and R. Arroyo, "ERFNet: Efficient residual factorized ConvNet for real-time semantic segmentation," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 263–272, Jan. 2018.

[70] S.-Y. Lo, H.-M. Hang, S.-W. Chan, and J.-J. Lin, "Efficient dense modules of asymmetric convolution for real-time semantic segmentation," in *Proc. ACM Multimedia Asia*, Dec. 2019, pp. 1–6.

[71] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[72] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6848–6856.
[73] W. Kang, Y. Xiang, F. Wang, and H. You, "EU-Net: An efficient fully convolutional network for building extraction from optical remote sensing images," *Remote Sens.*, vol. 11, no. 23, p. 2813, Nov. 2019.
[74] H. Liu, J. Luo, B. Huang, X. Hu, Y. Sun, Y. Yang, N. Xu, and N. Zhou, "DE-Net: Deep encoding network for building extraction from high-resolution remote sensing imagery," *Remote Sens.*, vol. 11, no. 20, p. 2380, Oct. 2019, doi: 10.3390/rs11202380.
[75] M. Chen, J. Wu, L. Liu, W. Zhao, F. Tian, Q. Shen, B. Zhao, and R. Du, "DR-Net: An improved network for building extraction from high resolution remote sensing image," *Remote Sens.*, vol. 13, no. 2, p. 294, Jan. 2021, doi: 10.3390/rs13020294.
[76] S. U. Khan, T. Hussain, A. Ullah, and S. W. Baik, "Deep-ReID: Deep features and autoencoder assisted image patching strategy for person re-identification in smart cities surveillance," *Multimedia Tools Appl.*, pp. 1–22, Jan. 2021, doi: 10.1007/s11042-020-10145-8.
[77] S. U. Khan, I. U. Haq, N. Khan, K. Muhammad, M. Hijji, and S. W. Baik, "Learning to rank: An intelligent system for person reidentification," *Int. J. Intell. Syst.*, pp. 1–26, Jan. 2022, doi: 10.1002/int.22820.
[78] Q. Chen, L. Wang, Y. Wu, G. Wu, Z. Guo, and S. L. Waslander, "Temporary removal: Aerial imagery for roof segmentation: A large-scale dataset towards automatic mapping of buildings," *ISPRS J. Photogramm. Remote Sens.*, vol. 147, pp. 42–55, Jan. 2019.
[79] B. Yu, L. Yang, and F. Chen, "Semantic segmentation for high spatial resolution remote sensing images based on convolution neural network and pyramid pooling module," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 9, pp. 3252–3261, Sep. 2018.
[80] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-Net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, May 2018.
[81] A. Abdollahi, B. Pradhan, and A. Alamri, "VNet: An end-to-end fully convolutional neural network for road extraction from high-resolution remote sensing data," *IEEE Access*, vol. 8, pp. 179424–179436, 2020.
[82] S. A. Taghanaki, Y. F. Zheng, S. K. Zhou, B. Georgescu, P. Sharma, D. Xu, D. Comaniciu, and G. Hamarneh, "Combo loss: Handling input and output imbalance in multi-organ segmentation," *Comput. Med. Imag. Graph.*, vol. 75, pp. 24–33, Oct. 2019.
[83] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

MD. ABDUL ALIM SHEIKH received the Master of Technology degree from Jadavpur University, Kolkata, India, in 2008. He is currently pursuing the Ph.D. degree with the Indian Institute of Technology (Indian School of Mines), India. He is an Assistant Professor at the Electronics and Communication Engineering Department, Aliah University, Kolkata. His research interests include image processing and computer vision, remote sensing image analysis, and machine learning. He is a life-time member of IEI, India.

TANMOY MAITY (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from Bengal Engineering and Science University, Shibpur, India. He is currently an Associate Professor of mining machinery engineering at the Indian Institute of Technology (Indian School of Mines), India. His research interests include instrumentation/power electronics engineering and signal processing. He is a member of the Institution of Engineers (IEI).

ALOK KOLE (Senior Member, IEEE) received the Ph.D. degree in artificial intelligence and intelligent control. He is currently a Professor with the Department of Electrical Engineering, RCCIIT, Kolkata, India. His current research interests include image processing, artificial intelligence, machine learning, control, computer vision, pattern recognition, and remote sensing image analysis. He is a member of the Institution of Engineers (IEI).