IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 16, 2019, accepted August 31, 2019, date of publication September 11, 2019, date of current version September 20, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2940527

Automatic Building Extraction on High-Resolution Remote Sensing Imagery Using Deep Convolutional Encoder-Decoder With Spatial Pyramid Pooling

YAOHUI LIU[1,2], LUTZ GROSS[2], ZHIQIANG LI[3], XIAOLI LI[3], XIWEI FAN[1], AND WENHUA QI[1]
[1] Institute of Geology, China Earthquake Administration, Beijing 100029, China
[2] School of Earth and Environmental Sciences, The University of Queensland, Brisbane, QLD 4072, Australia
[3] China Earthquake Networks Center, Beijing 100045, China

Corresponding author: Zhiqiang Li (lzhq9028@163.com)

This work was supported in part by the Key Special Fund for the Study on Rapid Assessment of Multi-source Earthquake Loss under Grant 201308018-5, in part by the National Key Research and Development Program of China under Grant 2018YFC1504403 and Grant 2018YFC1504503, in part by the China Earthquake Administration Special Project Surplus Fund (High Resolution Rapid Post-Earthquake Assessment Techniques), and in part by the Fundamental Research Funds for the Institute of Geology, China Earthquake Administration, under Grant IGCEA1610.

ABSTRACT Automatic extraction of buildings from remote sensing imagery plays a significant role in many applications, such as urban planning and monitoring changes to land cover. Various building segmentation methods have been proposed for visible remote sensing images, especially state-of-the-art methods based on convolutional neural networks (CNNs). However, high-accuracy building segmentation from high-resolution remote sensing imagery is still a challenging task due to the potentially complex texture of buildings in general and image background. Repeated pooling and striding operations used in CNNs reduce feature resolution causing a loss of detailed information. To address this issue, we propose a light-weight deep learning model integrating spatial pyramid pooling with an encoder-decoder structure. The proposed model takes advantage of a spatial pyramid pooling module to capture and aggregate multi-scale contextual information and of the ability of encoder-decoder networks to restore losses of information. The proposed model is evaluated on two publicly available datasets; the Massachusetts roads and buildings dataset and the INRIA Aerial Image Labeling Dataset. The experimental results on these datasets show qualitative and quantitative improvement against established image segmentation models, including SegNet, FCN, U-Net, Tiramisu, and FRRN. For instance, compared to the standard U-Net, the overall accuracy gain is 1.0% (0.913 vs. 0.904) and 3.6% (0.909 vs. 0.877) with a maximal increase of 3.6% in model-training time on these two datasets. These results demonstrate that the proposed model has the potential to deliver automatic building segmentation from high-resolution remote sensing images at an accuracy that makes it a useful tool for practical application scenarios.

INDEX TERMS Deep learning, high-resolution remote sensing imagery, building extraction, fully convolutional networks, encoder-decoder.

I. INTRODUCTION

Automatic extraction of buildings from remote sensing imagery is of great significance for many applications, including urban planning, navigation, and disaster management [1]–[6]. Recent years have witnessed a massive improvement in the capabilities of remote sensing techniques, which has led to a dramatic increase in the availability and accessibility of high-resolution remote sensing images [7]–[9]. The availability of high-quality data for spatially large areas, it is possible to perform accurate image segmentation targeting the extraction of buildings. However, the diverse characteristics of buildings including color, shape, material, size; and the interference of building shadows

The associate editor coordinating the review of this manuscript and approving it for publication was Mohammad Shorif Uddin.

and vegetation; mean that the development of accurate and reliable building extraction methods is still a challenging task [10].

Over the past few decades, various approaches for feature extraction from images have been developed. Spatial and textural features of an image are extracted through mathematical descriptors, such as Haar spaces [11], Scale-invariant Feature Transform (SIFT) [12], Local Binary Patterns (LBP) [13], Grey Level Co-occurrence Matrix (GLCM) [14], and Histogram of Oriented Gradients (HOG) [15]. More recently pixel-by-pixel predictions were introduced on the basis of extracted features through classifiers such as Support Vector Machines (SVM) [16], Adaptive Boosting (AdaBoost) [17], Random Forests [18], K-Means [19], and Conditional Random Fields (CRF) [20]. However, these methods rely heavily on manual feature design and implementations, which generally change with the application area. As a consequence, they can easily introduce bias and poor generalization and are time-consuming and labor-intensive.

In recent years, alongside advancements in computational capabilities and the availability of large volumes of data, the use of deep learning technology [21], especially convolutional neural networks (CNNs), has emerged as a powerful tool in many domains, particularly computer vision [22]. The extraction of buildings from images is a problem of semantic segmentation for which CNNs are particularly suitable as they automatically learn semantical information from input images and derive classifications through sequential convolutional and fully connected layers. In the early stages, patch-based CNN approaches, including Visual Geometry Group (VGG) [23], Deep Residual Network (ResNet) [24], and DenseNet [25], have outperformed traditional machine learning methods on classification tasks. Mnih [26] and Mnih and Hinton [27] proposed an automatic CNN method for extracting roads and buildings. As part of their investigation, the authors established a corresponding large-scale dataset, namely the Massachusetts roads and buildings dataset as reference for further developments. Saito *et al.* [28], [29] proposed a CNN framework to extract roads and buildings without pre-processing steps. However, because of the inefficiency of the sliding window used in their approach, the frame of the patch-based CNN is not optimal for addressing the building segmentation task [30]. Fully convolutional networks (FCNs) overcome this shortcoming substantially by replacing fully connected layers with up-sampling layers so that the output preserves spatial information of contextual features [31]. Many networks such as SegNet [32], and DeconvNet [22] have extended this approach.

Classification accuracy of FCN-based methods is improving. The challenge for semantic segmentation is to obtain more precise boundaries of objects and to address misclassification of small objects. Two aspects need to be addressed: Firstly, pooling layers or convolution striding used between convolution layers can augment the receptive field, and at the same time down-sample resolution of feature maps causing a loss of spatial information. Secondly, objects of the same category can exist at multiple shape spatial scales, and consequently, small objects are difficult to classify correctly [33]. Therefore, simply employing up-sampling operations such as deconvolution or bilinear interpolation after the feature extractor parts of a network cannot guarantee reliable results at high accuracy on high-resolution remote sensing data. Many network structures have been proposed to handle these problems; among them, the pyramid pooling module (PSPNet) [33] and the encoder-decoder structure (U-Net) [34] which are well established and have been shown to perform well.

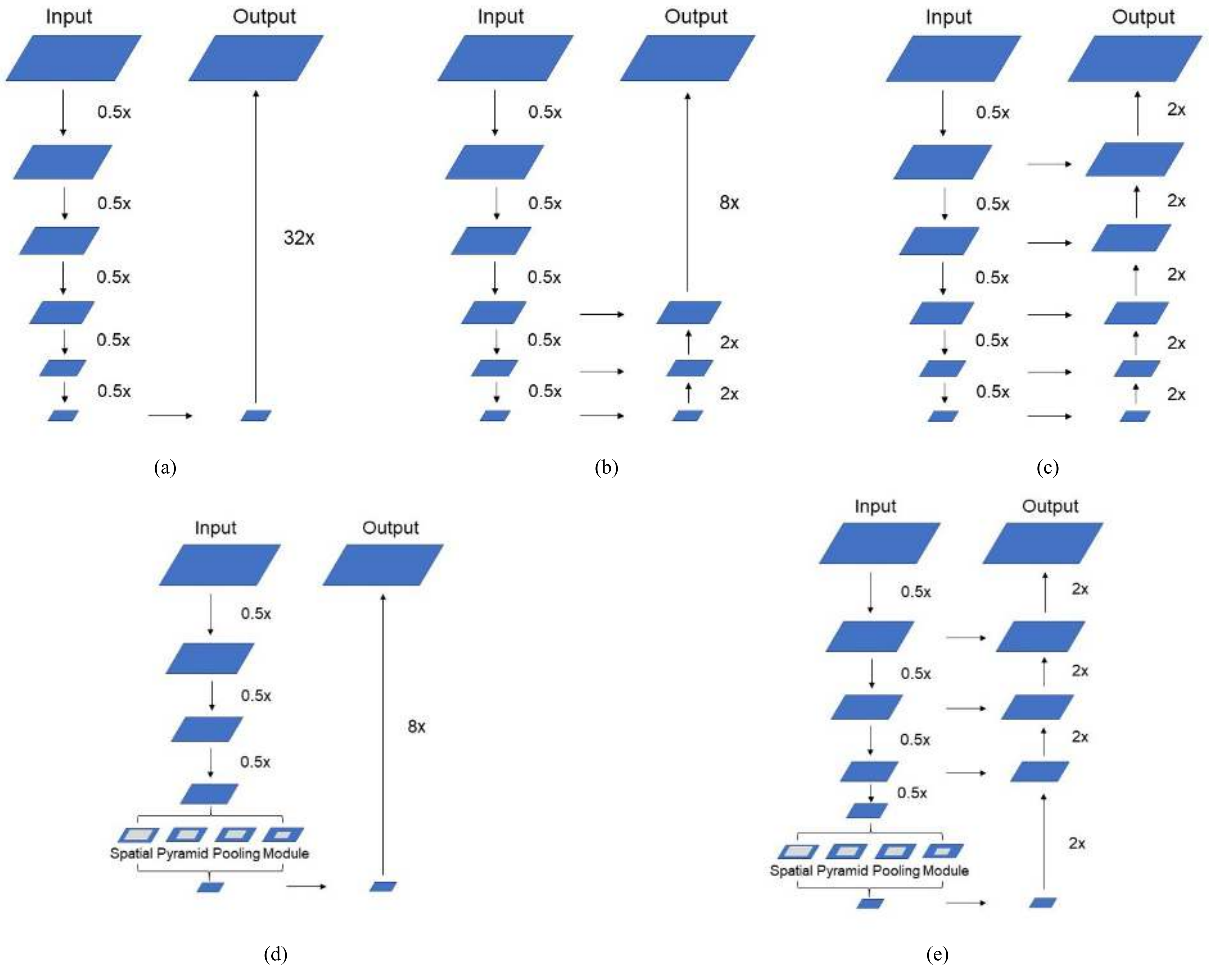
The spatial pyramid pooling structure of the PSPNet model and the atrous spatial pyramid pooling structure, an improved version as part of Deeplab [35], aim to handle the problem of segmenting objects at different spatial scales. The approach is well-known for achieving robust and efficient performance for dense semantic labeling. The network structure consists of several branches of dilated convolution operations to enlarge the receptive field. Spatial pyramid pooling shows better performance at pixel-level prediction tasks such as scene parsing and semantic segmentation. However, the PSPNet model only utilizes FCN based on ResNet as the backbone and lacks up-sampling capabilities [10].

The encoder-decoder structure proposed by U-Net is widely used in the field of semantic segmentation. It introduces a bottom-up/top-down architecture with skip connections that combine both lower and higher layers to generate the final result. Although it achieves better performance, U-Net has no extraction of multi-scale contextual features capability [6].

Another way to improve the performance of building extraction is post-processing. For example, Shrestha and Vanneschi [36] proposed a building extraction method using conditional random fields (CRFs) and exponential linear units. Alshehhi *et al.* [37] used a patch-based CNN architecture and proposed a post-processing method integrating low-level features of adjacent regions. However, post-processing methods are only able to improve results within a specific range, and the quality of results strongly depends on the accuracy of the initial segmentation [38].

Motivated by the analysis above, we propose a novel deep CNN to specifically improve the classification accuracy for building segmentation in high-resolution remote sensing images. To this end, a U-shape structure of an encoder and decoder path is applied as the backbone. In addition, the spatial pyramid pooling module is integrated as a bridge between the encoding and the decoding path. We refer to the proposed model as the USPP model. This approach enables extraction of features at multiple spatial scales and at the same time up-samples the feature maps to learn global contextual information. The main contributions of this study are summarized as follows:

(1) By combining the U-shape encoder-decoder structure and the spatial pyramid pooling module, the proposed USPP model can capture multi-scale features and effectively



(a)

(b)

(c)

(d)

(e)

FIGURE 1. Typical convolutional neural network structures in semantic segmentation. (a) The FCN-32s, (b) The FCN-8s (with skip connection), (c) Encoder-Decoder, (d) Spatial Pyramid Pooling Module, and (e) USPP.

restore detailed context information of buildings at all scales of the footprint.

(2) The qualitative and quantitative experiments for two public building labeling datasets, the Massachusetts Dataset [26] and the INRIA Aerial Image Labeling Dataset [39], demonstrate the excellent performance of the USPP model. Compared with established models such as SegNet, FCNs, U-Net, Tiramisu [40], and FRRN [41], higher accuracies and F1 scores are achieved for the problem of building extraction at small extra training costs.

The subsequent sections in this paper are organized as follows: The details of the USPP model are proposed in Section 2. In Section 3, datasets and experimental settings are described. Section 4 provides the experimental results of the USPP model and quantitative and qualitative comparisons with established models. A discussion and some conclusions are presented in Sections 5 and 6, respectively.

II. METHODOLOGY

In this paper, a dense semantic labeling system for automatic building segmentation from high-resolution remote sensing imagery is proposed. It combines a U-shaped encoder-decoder structure with spatial pyramid pooling. In the following this approach, which we call USPP model, is explained in more details.

A. U-SHAPE STRUCTURE

U-Net is a typical fully convolutional network which was originally designed for biomedical image segmentation [34]. The main structure of U-Net is similar to the letter U. It has an encoder that extracts spatial features from the training data, and a decoder that constructs the segmentation map from the encoded features. It also uses skip-connections to preserve features. Furthermore, to enable a more efficient operation of the network structure, fully connected layers are deprecated in the network, significantly reducing the number of parameters that need to be trained. Due to its excellent performance, U-Net has been one of the most popular architecture for semantic segmentation. However, it suffers from loss of global contextual information during the encoder phase reducing spatial resolution of the resulting feature maps. Moreover, this information cannot be recovered in the decoder phase. Figure 1 (a)-(c) illustrates details of typical FCN structures for semantic segmentation.

TABLE 1. Configurations of the encoder block in USPP (symbols h and w represent the height and width of the input layer, respectively).

Layer	Type	Kernel Size	Scale	Connect to
Conv_1	(h, w, k)	(3, 3)	-	Input
BN_1	(h, w, k)	-	-	Conv_1
ReLU_1	(h, w, k)	-	-	BN_1
Conv_2	(h, w, k)	(3, 3)	-	ReLU_1
BN_2	(h, w, k)	-	-	Conv_2
ReLU_2	(h, w, k)	-	-	BN_2
Maxpool_1	(h/2, w/2, k)	-	(2, 2)	ReLU_2

B. SPATIAL PYRAMID POOLING MODULE

The spatial pyramid pooling module draws from PSPNet [33]. Pyramid pooling aims to overcome the limitation of the fixed-size requirement for the CNN input image. It is applied to extract features in multiple scales and synthesize global information [42]. Details of spatial pyramid pooling module structure are illustrated in Figure 2. Its integration into FCN-8s is shown in Figure 1 (d). The pyramid pooling module comprises four steps; pyramid pooling, convolution, up-sampling, and concatenation operations, see Zhao *et al.* [33] for details. Through pyramid pooling, spatial features on four different spatial scales can be identified. In order to enhance the nonlinear learning ability of the multiscale features, 1 × 1 convolution is added to maintain the size of features and to reduce the number of each features channels by an N-th of the number of channels of the input feature map; N is the number of pyramid pooling scales, typically chosen to be four following the works by Zhao *et al.* [33] and Yu *et al.* [43]. The convoluted feature maps are further interpolated using bilinear filtering to match the size of the input feature map. The input feature map is finally concatenated with four up-sampled feature maps so that global context features can be maintained with multi-scale features. For the pooling operation, we adopt adaptive average pooling as illustrated in Figure 2. Four levels with bin sizes of 1 × 1, 2 × 2, 3 × 3, and 6 × 6 are used in the spatial pyramid pooling module. Note that the number and size of pyramid levels can be modified. They are related to the size of the feature map fed into the pyramid pooling layer [33].

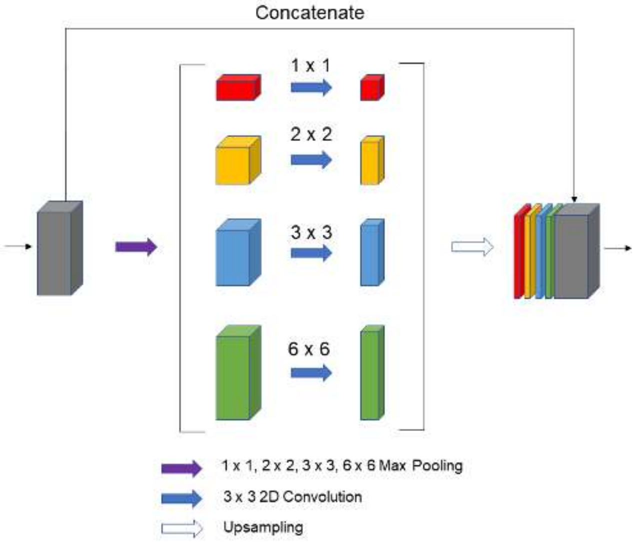


FIGURE 2. An illustration of the spatial pyramid pooling module structure with four pooling scales in semantic segmentation.

C. THE USPP MODEL

Inspired by VGG [23], U-Net [34], and PSP-Net [33], an end-to-end symmetric training structure is proposed to predict pixel-level results and generate final segmentation maps. Figure 3 and Figure 1 (e) present the detailed structure of the proposed USPP model. It contains four encoder blocks, one spatial pyramid pooling module, and four decoder blocks. For the encoder phase, the VGG-11 architecture is used as the backbone. As presented in Table 1, each encoder block contains two successive 3 × 3 convolutional layers followed by a 2 × 2 MaxPooling layer [44] to down-sample the input images. Each convolutional layer is then followed by a Batch-Normalization (BN) layer [45] and a nonlinear activation function of the rectified linear unit (ReLU) [46]. Batch-Normalization is introduced to ease training and enable concatenation of feature maps from different layers. ReLU is a widely-applied activation function in CNNs and is defined in Equation (1) for input z.

$$ReLU\ (z) = max(0, z) \tag{1}$$

The ReLU function helps to reduce the number of calculations and avoid overfitting during the training phase. The four encoder blocks use 64, 128, 256, and 512 kernels.

One crucial modification of USPP in comparison to the encoder-decoder structure is that the bottom layer in the encoder phase is replaced with the spatial pyramid pooling module as a connector between the encoding and the decoding path. This enables extraction of features at multiple scales and up-sampling of feature maps and learning of global context information.

Table 2 shows details of the decoder block. The decoder block uses a transposed convolutional layer instead of the pooling layer. Additionally, each up-sampled map is concatenated with the corresponding feature map from the encoding path through skip connections. Finally, there is an output

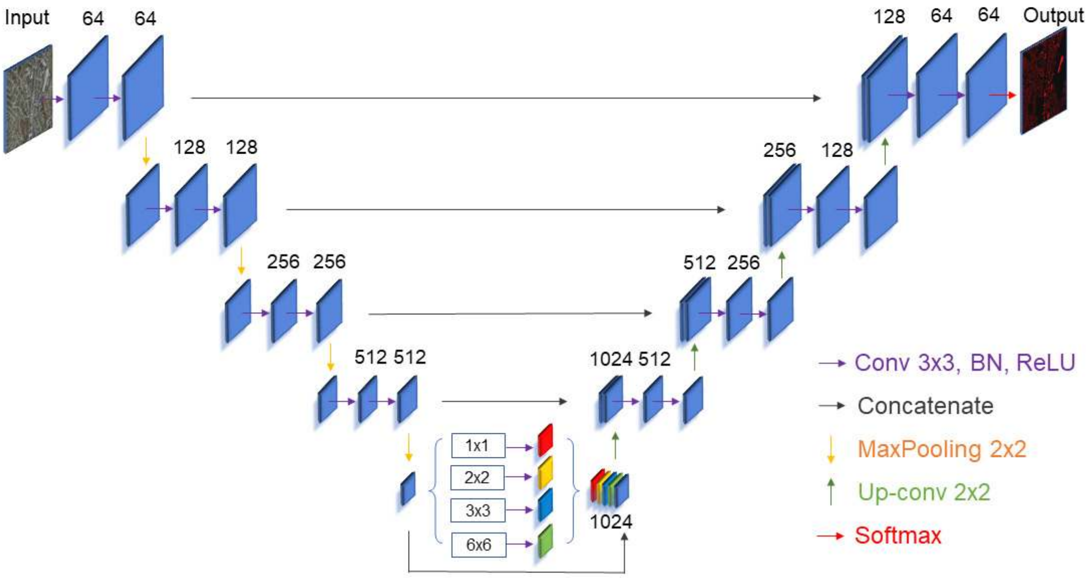


FIGURE 3. The architecture of USPP with the fusion of spatial pyramid pooling and encoder-decoder structure.

TABLE 2. Configurations of the decoder block in USPP (symbols h, w, and d represent the height, width, and depth of the input layer, respectively).

Layer	Type	Kernel Size	Scale	Connect to
Deconv_1'	(2 × h', 2 × w', d)	-	(2, 2)	Input'
Skip_1'	(2 × h', 2 × w', d + k)	-	-	Deconv_1' & BN_2
Conv_1'	(2 × h', 2 × w', k')	(3, 3)	-	Skip_1'
BN_1'	(2 × h', 2 × w', k')	-	-	Conv_1'
ReLU_1'	(2 × h', 2 × w', k')	-	-	BN_1'
Conv_2'	(2 × h', 2 × w', k')	(3, 3)	-	ReLU_1'
BN_2'	(2 × h', 2 × w', k')	-	-	Conv_2'
ReLU_2'	(2 × h', 2 × w', k')	-	-	BN_2'

layer after the encoder path which performs a pixel-wise classification. This output layer is a 1 × 1 convolutional layer with a sigmoid activation function.

D. MODEL TRAINING AND TESTING

The following training and testing procedure is applied: First, the images, including samples and labels, are separated into training and testing datasets. Data augmentation, such as flipping and random cropping, are employed to increase the complexity within the data set and to reduce over-fitting during training [47]. Then, the USPP model is trained using the training data set; the training procedure is based on the gradient descent algorithm [48] that utilizes updated parameters calculated by the loss function and backpropagation to improve the performance of the network [49]. Instead of the simple mean square error (MSE), binary cross-entropy [50] is chosen to calculate the loss between every prediction and relative ground truth. Finally, testing data are fed into the trained model with the optimal model parameters, and the results are evaluated and compared on some evaluation metrics. An overview of the proposed building extraction system is illustrated in Figure 4.

III. EXPERIMENTAL DATASETS AND EVALUATION

To verify the effectiveness of USPP as a tool for building segmentation from high-resolution remote sensing imagery, extensive experiments have been conducted on two classical datasets; the Massachusetts and INRIA datasets. Furthermore, the performance of USPP is compared to similar architecture with alternative sub-components and with other CNN architectures. All experiments were evaluated based on five major metrics, including Overall Accuracy, Precision, Recall, F1-score, and Mean IoU.

A. DATASETS

The first dataset investigated is the Massachusetts buildings dataset assembled by Mnih [26]. The dataset consists of 137 training, 4 validation, and 10 testing images, covering a surface of 2.25 square kilometers of urban and suburban areas of Boston (MA) in the United States. The size of each image

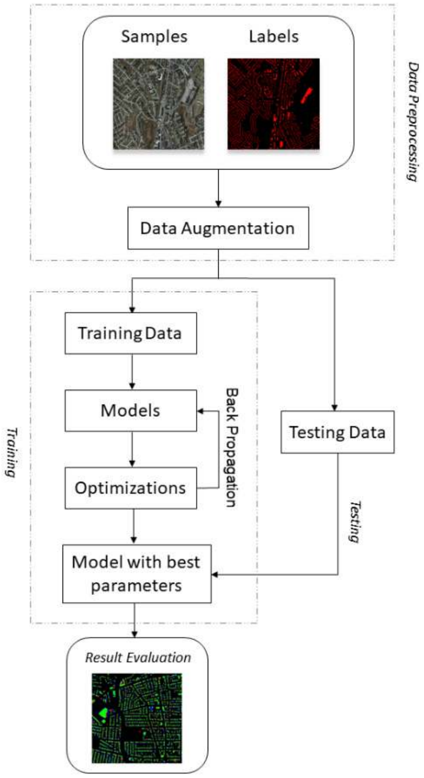


FIGURE 4. The schematic workflow of this study.

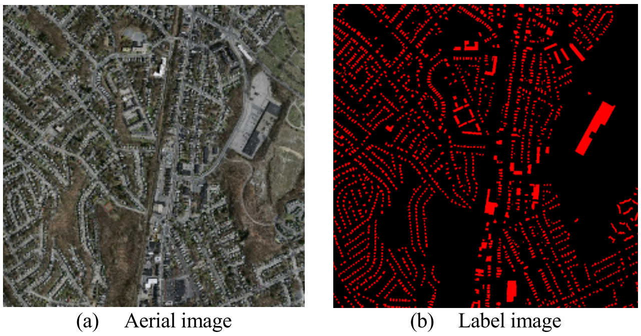

(a) Aerial image (b) Label image

FIGURE 5. Image and label example from the Massachusetts Dataset. (a) Aerial image; (b) Label image; red for buildings and black for the background.

is 1500 × 1500 pixels with the spatial resolution of 1 meter per pixel and is composed of red, green, and blue (RGB) channels. An example of an input image and its building labeling are shown in Figure 5.

The second dataset is the INRIA Aerial Image Labeling Dataset [39], comprising of 360 ortho-rectified aerial RGB images. This dataset covers different cities including Austin, Chicago, Kitsap, Western/Eastern Tyrol, Vienna, Bellingham, and San Francisco. Figure 6 presents an example image and its building labeling. The spatial resolution of images is 0.3 m with an image size of 5000 × 5000 pixels and spatial coverage of 1500 × 1500 m^2. The images comprise an overall area of 810 square kilometers. Only two semantic classes - buildings and non-buildings - were considered as the ground truth.

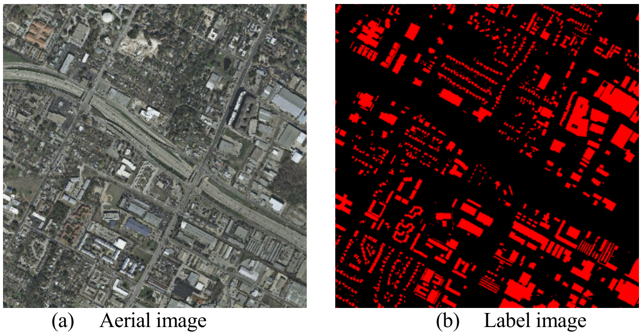

(a) Aerial image (b) Label image

FIGURE 6. Image and label example from the INRIA Dataset. (a) Aerial image; (b) Label image; red for buildings and black for the background.

Following previous researches [39, 51], we choose the first five images of each city from the training set for validation.

B. DATA AUGMENTATION

Deep convolutional neural networks require large amounts of training data, which are not always available during the learning phase. Data augmentation is essential to teach the network the desired invariance and robustness properties and to avoid overfitting when only a few training samples are available [34]. The training patch was processed by a sliding window of 256 × 256 pixels to generate data for model training and cross-validation. Horizontal or vertical flipping was applied randomly with a probability of 0.5. In addition, windows were rotated by 90, 180, and 270 degrees. The pixel values of each image were scaled to the interval [0, 1] by dividing by 255. No application-specific post-processing was performed. Since the final layer was activated by a Sigmoid function, it generated outputs in the range [0, 1]. The final segmentation map of the input images was produced by applying a threshold of 0.5.

C. EXPERIMENTAL SETTINGS

The implementation of USPP is based on the deep learning library PyTorch [52]. All experiments were carried out on computer servers with an Intel®Xeon®CPU E5-2630 v4 (2.20GHz), 64GB of memory (RAM) and two NVIDIA GeForce GTX 1080 Ti (11GB). Parallelization [53] was implemented at the PyTorch level to make use of the available GPU performance and to accelerate calculations. During the training phase, the Adam stochastic optimizer [54] with an initial learning rate of 0.0001 was used. All models were trained for 300 epochs with a mini-batch size of 16. Each batch contained cropped images that were randomly selected from the training patches. The changing accuracies and losses of the Massachusetts and INRIA datasets with increasing epochs are shown in Figure 7. It is evident that the error gradually decreases while the accuracy increases and stays at a high and stable level.

D. EVALUATION METRICS

To evaluate the quantitative performance of different CNN methods, the ‘Overall Accuracy’ (OA), ‘Precision’,

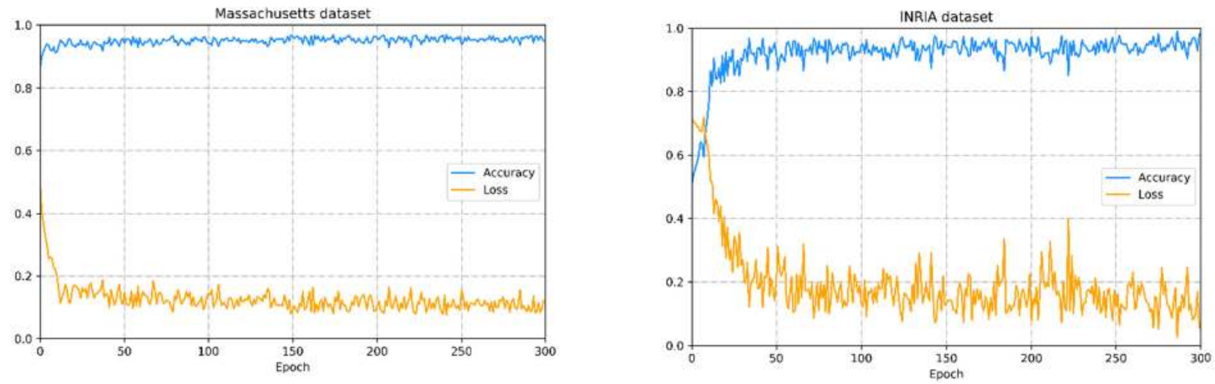


FIGURE 7. **The accuracy and loss of the proposed model for training the datasets. Left: Massachusetts dataset. Right: INRIA dataset.**

TABLE 3. **Quantitative result of different methods including SegNet, FCN, U-Net, Tiramisu, FRRN, and USPP on the Massachusetts testing dataset. The highest values for the different metrics are highlighted in bold.**

Methods	OA	Precision	Recall	F1-score	Mean *IoU*
SegNet	0.862	0.882	0.822	0.851	0.740
FCN	0.894	0.895	0.867	0.881	0.787
U-Net	0.904	0.899	0.869	0.891	0.803
Tiramisu	0.882	0.903	0.837	0.869	0.768
FRRN	0.869	**0.928**	0.796	0.857	0.749
USPP	**0.913**	0.908	**0.892**	**0.900**	**0.818**

'Recall', 'F1-score', and mean of Intersection-over-Union ('Mean*IoU*') are used as quality metrics. 'Overall Accuracy' is defined as the number of correctly classified pixels divided by the total number of test pixels. 'Precision' is the percentage of correctly classified positive pixels amongst all pixels predicted as positive. 'Recall' is the percentage of correctly classified positive pixels among all true positive pixels. 'F1-score' is a combination of precision and recall. 'Mean*IoU*' is applied to characterize the accuracy at the segment level [55]. The values of these metrics are in the range of 0 to 1, and higher values indicate better classification performance. The five metrics can be calculated as follows:

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

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

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

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{5}$$

$$Mean\ IoU = \frac{TP}{TP + FP + FN} \tag{6}$$

where *TP* is the number of true positives, *TN* is the number of true negatives, *FP* is the number of false positives, and *FN* is the number of false negatives.

IV. RESULTS

A. COMPARISON ON THE MASSACHUSETTS DATASET

USPP is compared to the established networks SegNet, FCN, U-Net, Tiramisu, and FRRN for semantic segmentation. Figure 8 shows the qualitative segmentation results of the CNN models on the Massachusetts dataset. SegNet and FRRN return more false positives and false negatives than the other methods. U-Net returns slightly more false positives but less false negatives compared to FCN. By contrast, the proposed USPP model shows significantly less false positives and false negatives than the other methods, while maintaining high completeness in building segmentation.

The quantitative comparison of the different networks on the whole testing dataset is presented in Table 3. It demonstrates that the proposed USPP delivers improvements on all performance indicators over the other models except for 'Precision'. In the testing case, the USPP model is the best among all models on Overall Accuracy score with a gain of 1.0% (0.913 vs. 0.904) over the next best model U-Net. As for Precision, the FRRN model holds the highest values and gains 2.2% over USPP (0.928 vs. 0.908). USPP still performs better than U-Net by 1.0% (0.908 vs. 0.899) over the entire testing dataset. For Recall, the U-Net, FCN, and USPP method shows significantly better performance over the other three methods while USPP achieves the best value being 2.6% ahead of the U-Net method (0.892 vs. 0.869). USPP achieves the best F1-score where U-Net is again the best model amongst the others. For Mean *IoU*, USPP (0.818)

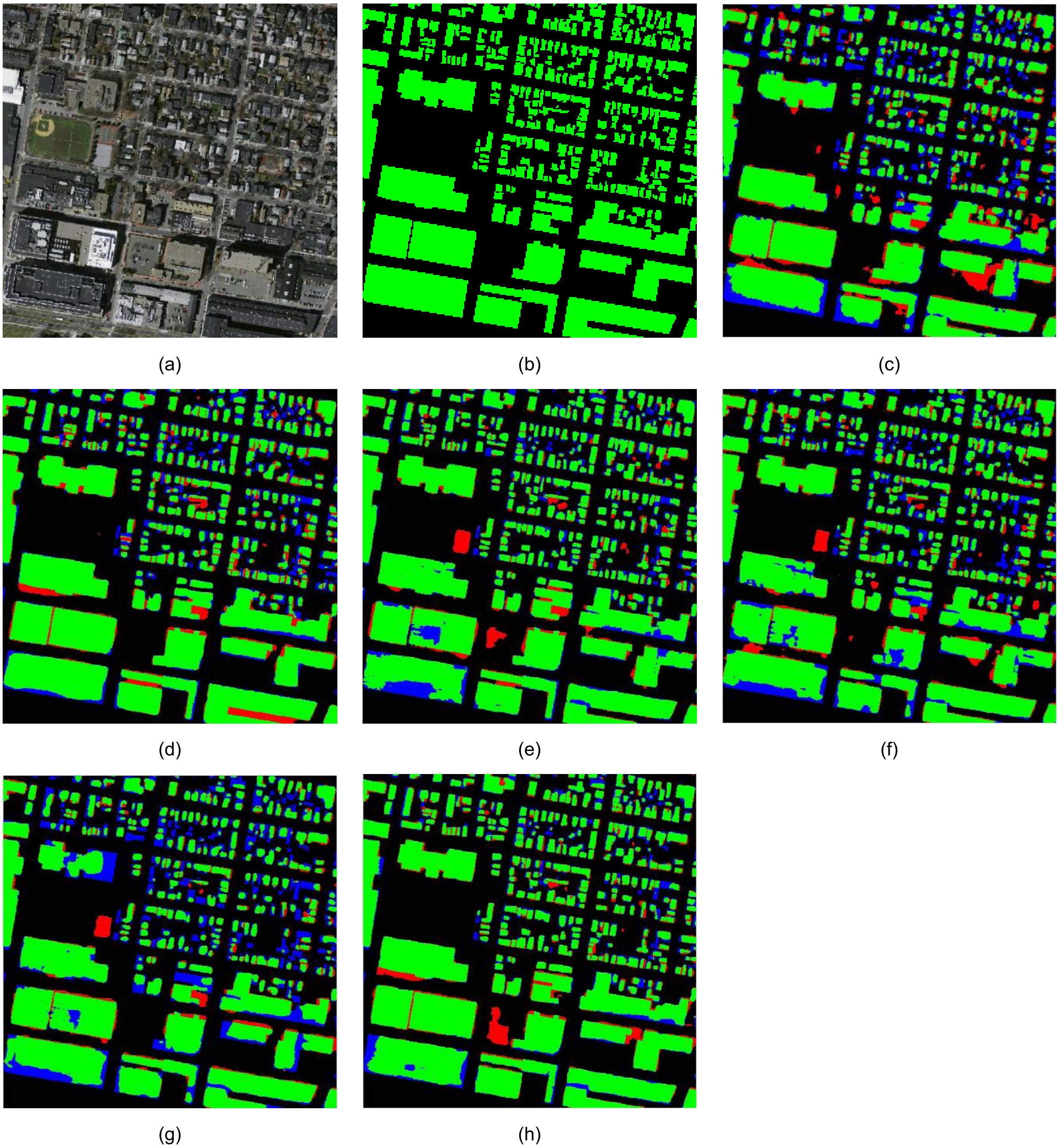


FIGURE 8. Segmentation results of different methods on the Massachusetts dataset. (a) Original input image. (b) Target map (ground truth). (c) The output of SegNet. (d) The output of FCN. (e) The output of U-Net. (f) The output of Tiramisu. (g) The output of FRRN. (h) The output of USPP. The green, red, blue, and black pixels of the maps represent the predictions of true positive, false positive, false negative and true negative, respectively.

has made the highest score 1.9% ahead of U-Net (0.803), and 9.2% ahead of FRNN (0.749) respectively.

B. COMPARISON ON THE INRIA DATASET

Building extraction results from the different models on a sample in the INRIA dataset are presented in Figure 9 for a qualitative comparison. It is clear that SegNet returns more false negatives while U-Net returns more false positives in comparison to the other methods. Overall U-Net and USPP predict building outlines reasonably well.

We further conducted a quantitative comparison with different models on the INRIA dataset. The results of the quantitative comparison are summarized in Table 4. In contrast to the Massachusetts dataset where U-Net performed the best, FCN shows the best performance amongst the established methods. However, the proposed USPP model performs best except for the Precision score, where FCN obtains the highest value 1.7% ahead of USPP (0.918 vs. 0.903). For Overall Accuracy, USPP holds the highest values with a gain of 3.6% compared to U-Net (0.909 vs. 0.877). For Recall,

TABLE 4. Quantitative result of different methods including SegNet, FCN, U-Net, Tiramisu, FRRN, and USPP on the INRIA testing dataset. The highest values for the different metrics are highlighted in bold.

Methods	OA	Precision	Recall	F1-score	Mean *IoU*
SegNet	0.865	0.867	0.767	0.849	0.737
FCN	0.889	**0.918**	0.831	0.872	0.773
U-Net	0.877	0.885	0.837	0.860	0.755
Tiramisu	0.869	0.911	0.801	0.853	0.743
FRRN	0.875	0.915	0.808	0.858	0.752
USPP	**0.909**	0.903	**0.882**	**0.893**	**0.806**

USPP achieves an improvement of 5.4% over U-Net (0.882 vs. 0.837). As for F1-score and Mean *IoU*, USPP obtains the highest value over the other models and outperforms Tiramisu and FRRN which are considered to be state-of-the-art networks for segmentation. Compared to the classic U-Net, USPP yields a higher F1-score by 3.8% (0.893 vs. 0.860) and a higher Mean *IoU* by 6.8% (0.806 vs. 0.755).

C. COMPARISON ON THE INDEPENDENT BUILDINGS

For a more detailed analysis of the segmentation results, several independent buildings were randomly selected from the two datasets. Figure 10 presents the results of these samples generated by the SegNet (without skip connection), U-Net (without spatial pyramid pooling), and the proposed USPP. In general, USPP performs better than the other two models. For the SegNet model shown in the second row, major parts of buildings are not accurately extracted (see columns a, c, d, e, and f). As depicted in the third row, U-Net is able to present contours of buildings more accurately (in columns c, g, and h), but it still produces large numbers of false negatives (blue in columns a, e, and f). In the fourth row, USPP shows the best performance in building extraction and noise reduction compared with the other two methods. However, all methods failed to segment the entire outline of buildings which are partially obscured, for instance by trees (see column a).

D. COMPUTATIONAL EFFICIENCY

Computational cost is also a significant efficiency indicator in deep learning. It represents the complexity of the in-depth learning model [6] where the costs for training and testing quantify the differences in complexity between CNN models. Considering the relatively close performance, a quantitative comparison of the computing times for the different deep-learning methods was conducted and is presented in Tables 5 and 6.

For the model-training time, SegNet, FCN, U-Net, and USPP require about 750 min on the Massachusetts dataset while Tiramisu and FRRN require about 30% more time (∼1050 min). For the INRIA dataset, the training times are higher, and again Tiramisu and FRRN require significantly higher training time (18% more, ∼900 vs. ∼1000 min). This can be attributed to the fact that both have a more complex structure and include additional layers. For both datasets, FCN and SegNet model take the least training time, but USPP requires only a slightly higher training time; e.g., relative to the U-Net model, there is an increase of only 3.6% and 1.0% on the Massachusetts dataset and INRIA dataset respectively.

TABLE 5. Comparison of model-training time in minutes of SegNet, FCN, U-Net, Tiramisu, FRRN, and USPP on the two datasets for 300 epochs. For each dataset, the minimum is highlighted in bold.

	Massachusetts dataset	INRIA dataset
SegNet	749	**891**
FCN	**723**	906
U-Net	747	901
Tiramisu	1050	1185
FRRN	1065	1090
USPP	775	910

TABLE 6. Comparison of model-testing time in seconds of SegNet, FCN, U-Net, Tiramisu, FRRN, and USPP for each image on the two datasets. For each dataset, the minimum value is highlighted in bold.

	Massachusetts dataset	INRIA dataset
SegNet	**7**	9
FCN	20	22
U-Net	9	**8**
Tiramisu	14	12
FRRN	15	12
USPP	10	9

For the model-testing process, the SegNet, FCN, U-Net, Tiramisu, FRRN, and USPP require 7 sec, 20 sec, 9 sec, 14 sec, 15 sec, and 10 sec for the Massachusetts dataset, and 9 sec, 22 sec, 8 sec, 12 sec, 12 sec, and 9 sec for the INRIA dataset, respectively. The FCN model is the most time-consuming and two to three times slower than the fastest models SegNet. By contrast, SegNet, U-Net, and USPP are quite similar in time consumption during the testing procedure.

These findings demonstrate that USPP delivers building segmentation with consistently better performance scores except for Precision at computational costs that are comparable to established CNN methods.

V. DISCUSSION

A. ABOUT THE PROPOSED USPP MODEL

In recent years, deep learning, especially convolutional neural networks, have been widely applied in computer vision

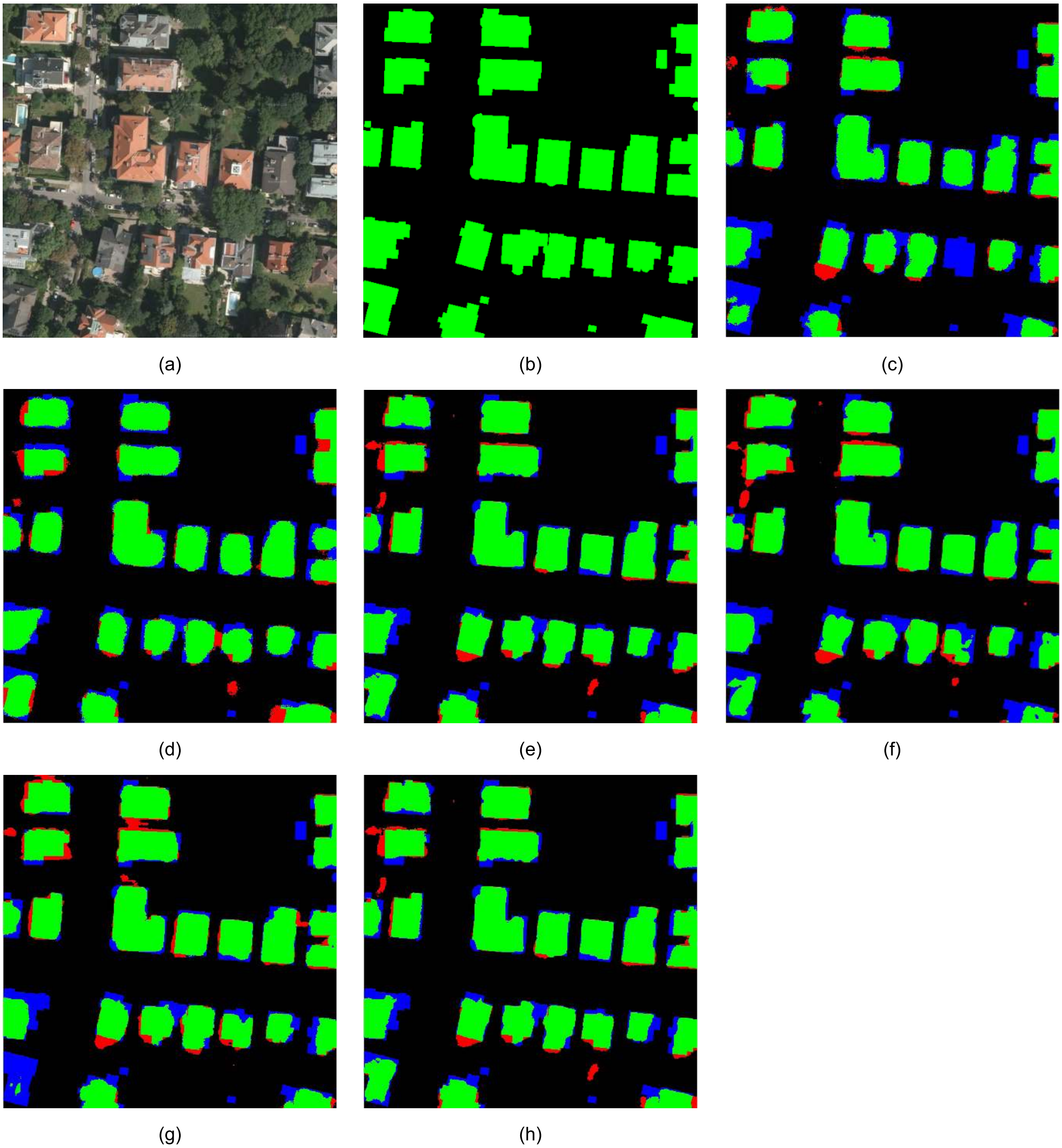


FIGURE 9. Segmentation results of different methods on the INRIA dataset. (a) Original input image. (b) Target map (ground truth). (c) The output of SegNet. (d) The output of FCN. (e) The output of U-Net. (f) The output of Tiramisu. (g) The output of FRRN. (h) The output of USPP. The green, red, blue, and black pixels of the maps represent the predictions of true positive, false positive, false negative and true negative, respectively.

and semantic segmentation. However, automatic building extraction from high-resolution remote sensing imagery is still a challenging task due to a large variety of appearing patterns and its spatial scale. As demonstrated in Liu *et al.* [38] and in Zhang and Wang [56], accurate building extraction depends on the acquisition of the unique morphological characteristics of the building. They also pointed out that a well-performing network for building segmentation requires large receptive fields and needs to consider the multi-scale context. To address these issues and to achieve an effective performance to extract buildings from remote sensing images, we proposed the novel USPP model. This model follows the basic structure of U-Net, with the significant improvement of an additional spatial pyramid pooling module included at the bottom of the encoder-decoder structure. The latter aggregates the spatial context information from the low convolutional layers to alleviate the problem of spatial information loss. USPP achieves satisfactory

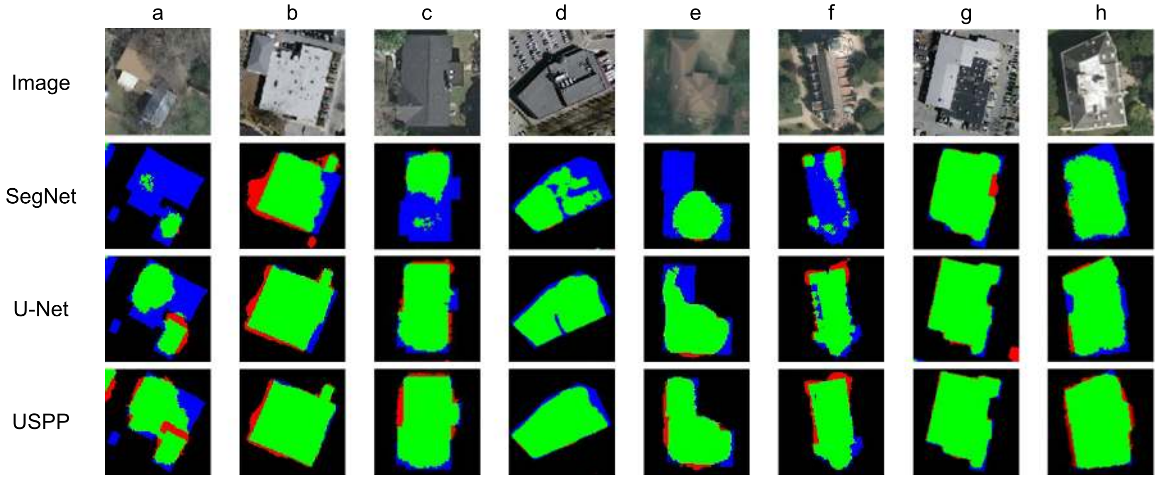


FIGURE 10. Segmentation results of randomly selected buildings from the two datasets using the SegNet, U-Net, and USPP model. The green, red, blue, and black pixels of the maps represent the predictions of true positive, false positive, false negative and true negative, respectively.

results with an Overall Accuracy of 91.3% and 90.9%, and an F1-score of 0.900 and 0.893 for the Massachusetts and INRIA dataset respectively. These scores are improvements over established methods in particular U-Net. It demonstrates that the USPP approach of expanding respective fields and considering multi-scale contexts are successful at a very small additional computational cost.

The accuracy and loss during the training phase reported in Section 3.3 and the experimental results reported in Section 4 show that the proposed approach achieved slightly higher accuracy and stability on the Massachusetts dataset than on the INRIA dataset. In fact, as discussed in Ji *et al.* [57], there are more wrong labels, high buildings, and shadows in the INRIA dataset that may substantially influence the discriminative ability of the deep learning model. Therefore, the differences in experimental results on the two datasets are acceptable and reasonable.

The USPP model proposed in this paper is light-weight and takes full advantages of the encoder-decoder structure and spatial pyramid pooling module. The qualitative and quantitative experimental results have proved that USPP achieved better performance than established models without a significant increase in training time. These findings demonstrated the applicability and efficiency of the USPP as a model in building extraction from high-resolution remote sensing images.

B. LIMITATIONS

Despite the performance improvements achieved by the proposed USPP, some issues remain to be considered. With the advancement of remote sensing technology, it has become much easier to collect high-resolution aerial images with abundant features and spectral information. This poses a major challenge for computer vision and image processing. The USPP model proposed in this research is able to improve the building segmentation, which demonstrates that enlargement of fields helps to improve the accuracy of semantic segmentation. However, the datasets used in this research do light-weight not cover images from different sensors, such as hyperspectral images and SAR images. 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. In future studies, we will try to expand the training dataset and further optimize the network architecture to be applicable to multi-spectral and point cloud data.

VI. CONCLUSION

Accurate and automatic building segmentation from remote sensing imagery is essential for application areas such as urban planning and disaster management. In this paper, we proposed a CNN framework, named USPP, to perform building segmentation on high-resolution remote sensing 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 demonstrating that the encoder-decoder and spatial pyramid pooling module are two powerful tools that need to be merged to take effect for building segmentation.

Experiments were conducted on two public building datasets: the Massachusetts and INRIA Aerial Image Labeling datasets. The results show that the proposed USPP model achieves high accuracy on these two datasets. Buildings were extracted successfully by USPP with fewer classification errors and with sharper boundaries. The qualitative and quantitative comparison with the established models SegNet, FCN, U-Net, Tiramisu, and FRRN have demonstrated that USPP outperforms these models. Compared with the standard U-Net, USPP gains 1.0% (0.913 vs. 0.904) and 3.6% (0.909 vs. 0.877) improvements in Overall Accuracy with the small increase of 3.6% and 1.0% in model-training time on the Massachusetts and the INRIA dataset respectively.

These findings showed the robustness of USPP and its ability to perform effective building segmentation from high-resolution remotely sensed images.

To further improve the extraction of building outlines, future work will combine the segmentation model with other image processing methods taking into consideration building edges and shapes. In addition, various data fusion strategies for the multi-scale remote sensing data (e.g., low-, medium-, high-resolution, and LiDAR) will be included, and some more heavy modeling structures will also be investigated.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers and associate editor for their constructive comments and suggestions. Also, we want to thank Volodymyr Mnih and l'institut national de recherche dédié aux sciences du numérique (INRIA) for their kind open-sourcing of the data. In addition, the first author wants to express his great acknowledgment to the China Scholarship Council (CSC) for providing financial support to study at The University of Queensland, Australia.

REFERENCES

[1] 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.

[2] Y. Wang, B. Liang, M. Ding, and J. Li, "Dense semantic labeling with atrous spatial pyramid pooling and decoder for high-resolution remote sensing imagery," *Remote Sens.*, vol. 11, no. 1, p. 20, 2019.

[3] H. He, D. Yang, S. Wang, S. Wang, and Y. Li, "Road extraction by using atrous spatial pyramid pooling integrated encoder-decoder network and structural similarity loss," *Remote Sens.*, vol. 11, no. 9, p. 1015, 2019.

[4] X. Li, Z. Li, J. Yang, Y. Liu, B. Fu, W. Qi, and X. Fan, "Spatiotemporal characteristics of earthquake disaster losses in China from 1993 to 2016," *Natural Hazards*, vol. 94, no. 2, pp. 843–865, 2018.

[5] 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, 2019.

[6] 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.

[7] 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.

[8] Z. Huang, G. Cheng, H. Wang, H. Li, L. Shi, and C. Pan, "Building extraction from multi-source remote sensing images via deep deconvolution neural networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2016, pp. 1835–1838.

[9] Z. Guo, X. Shao, Y. Xu, H. Miyazaki, W. Ohira, and R. Shibasaki, "Identification of village building via Google Earth images and supervised machine learning methods," *Remote Sens.*, vol. 8, no. 4, p. 271, 2016.

[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, 2019.

[11] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. CVPR*, vol. 1, Dec. 2001, pp. 511–518.

[12] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[13] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[14] D. A. Clausi, "An analysis of co-occurrence texture statistics as a function of grey level quantization," *Can. J. Remote Sens.*, vol. 28, no. 1, pp. 45–62, 2002.

[15] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 886–893.

[16] 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, 2007.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1520–1528.

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

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

[25] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4700–4708.

[26] V. Mnih, *Machine Learning for Aerial Image Labeling*. Toronto, ON, Canada: Univ. Toronto, 2013.

[27] V. Mnih and G. E. Hinton, "Learning to detect roads in high-resolution aerial images," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2010, pp. 210–223.

[28] 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. 1–9, 2016.

[29] S. Saito and Y. Aoki, "Building and road detection from large aerial imagery," *Proc. SPIE*, vol. 9405, pp. 94050K-1–94050K-12, Feb. 2015.

[30] 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.

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

[32] 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, Dec. 2017.

[33] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2881–2890.

[34] 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.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[35] 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. 2017.

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

[37] R. Alshehhi, P. R. Marpu, W. L. Woon, and M. Dalla 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.

[38] 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, 2019.

[39] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? The Inria aerial image labeling benchmark," in *Proc. IEEE Int. Symp. Geosci. Remote Sens. (IGARSS)*, Jul. 2017, pp. 3226–3229.

[40] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional DenseNets for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jul. 2017, pp. 11–19.

[41] T. Pohlen, A. Hermans, M. Mathias, and B. Leibe, "Full-resolution residual networks for semantic segmentation in street scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4151–4160.

[42] Q. Liu, R. Hang, H. Song, and Z. Li, "Learning multiscale deep features for high-resolution satellite image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 1, pp. 117–126, Jan. 2017.

[43] 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 Observat. Remote Sens.*, vol. 11, no. 9, pp. 3252–3261, Sep. 2018.

[44] J. Nagi, F. Ducatelle, G. A. Di Caro, D. Ciresan, U. Meier, A. Giusti, F. Nagi, Jürgen Schmidhuber, and L. M. Gambardella, "Max-pooling convolutional neural networks for vision-based hand gesture recognition," in *Proc. IEEE Int. Conf. Signal Image Process. Appl. (ICSIPA)*, Nov. 2011, pp. 342–347.

[45] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: https://arxiv.org/abs/1502.03167

[46] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 807–814.

[47] D. M. Hawkins, "The problem of overfitting," *J. Chem. Inf. Comput. Sci.*, vol. 44, no. 1, pp. 1–12, 2004.

[48] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proc. COMPSTAT*. Springer, 2010, pp. 177–186.

[49] Y. LeCun, "Handwritten digit recognition with a back-propagation network," in *Proc. Adv. Neural Inf. Process. Syst.*, 1990, pp. 396–404.

[50] J. Shore and R. Johnson, "Properties of cross-entropy minimization," *IEEE Trans. Inf. Theory*, vol. 27, no. 4, pp. 472–482, Jul. 1981.

[51] B. Bischke, P. Helber, J. Folz, D. Borth, and A. Dengel, "Multi-task learning for segmentation of building footprints with deep neural networks," 2017, *arXiv:1709.05932*. [Online]. Available: https://arxiv.org/abs/1709.05932

[52] A. Paszke, "Automatic differentiation in PyTorch," Tech. Rep., 2017.

[53] M. Isard, M. Budiu, Y. Yu, A. Birrell, and D. Fetterly, "Dryad: Distributed data-parallel programs from sequential building blocks," *ACM SIGOPS Oper. Syst. Rev.*, vol. 41, no. 3, pp. 59–72, 2007.

[54] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980

[55] M. Polak, H. Zhang, and M. Pi, "An evaluation metric for image segmentation of multiple objects," *Image Vis. Comput.*, vol. 27, no. 8, pp. 1223–1227, 2009.

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

[57] 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.

YAOHUI LIU was born in Yichun, Heilongjiang, China, in 1991. He received the B.S. degree in surveying and mapping from the Heilongjiang Institute of Technology, in 2013, and the M.S. degree in geographic information science from Yunnan Normal University, in 2016. He is currently pursuing the joint Ph.D. degree with the Institute of Geology, China Earthquake Administration and with The University of Queensland, Australia. His research interests include computer vision, image segmentation, deep learning, and risk management.

LUTZ GROSS received the M.Sc. degree in mathematics from the University of Hannover, Germany, and the Ph.D. degree in mathematics from The University of Karlsruhe, Germany, in 1996. He is currently an Associate Professor with The University of Queensland, Brisbane, Australia. His research interests include geophysical data processing and inversion, large-scale numerical modeling, and high-performance computing.

ZHIQIANG LI received the Ph.D. degree in geodynamics and tectonophysics from the Institute of Geology, China Earthquake Administration, Beijing, China, in 1997. He is currently a Professor with the China Earthquake Networks Center, Beijing. His research interests include earthquake emergency response and management, earthquake emergency basal database technology, earthquake disaster risk assessment techniques and application of GPS, GIS, and RS to earthquake emergency and earthquake resistance, and disaster relief.

XIAOLI LI received the master's degree in structural geology from the Institute of Geology, China Earthquake Administration, Beijing, China, in 2008. She is currently a Senior Engineer with the China Earthquake Networks Center, Beijing. Her research interests include earthquake emergency response and management, earthquake disaster risk assessment techniques, and application of GPS, GIS, and RS to earthquake emergency and earthquake resistance, and disaster relief.

XIWEI FAN received the Ph.D. degree in cartography and geographical information system from the Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, in 2015. He is currently an Associate Research Fellow with the Institute of Geology, China Earthquake Administration. His research interests include the retrieval and validation of land surface temperature/emissivity and earthquake damage estimation.

WENHUA QI received the bachelor's degree in hydrology and water resources engineering from the School of Water Resources and Environment, China University of Geosciences, Beijing, China, in 2008, and the master's degree in Tectonics from the Institute of Geology, China Earthquake Administration, Beijing, in 2011. His research interests include remote sensing, and citizen science for natural disaster risk assessment and governance.