Efficient 3D City Modelling from Large-scale Aerial Images by Deep Learning

Michael Shell, *Member, IEEE,* John Doe, *Fellow, OSA,* and Jane Doe, *Life Fellow, IEEE*

***Abstract*—Extracting buildings from remote sensing images plays an important role in urban applications (e.g., urban planning and digital city). However, this task is quite difficult due to great diversity of buildings and similarities between buildings and other categories. Recent approaches have attempted to harness the capabilities of deep learning techniques for building extraction. In this paper, we propose a robust system which can extract buildings from large-scale remote sensing images and build 3D models for extracted building areas. Learning low-level information of images becomes as important as learning high-level semantic information since buildings in remote images possess various scales and aspect ratios. So we propose a novel hierarchically fused fully convolutional network (HF-FCN). The proposed network generates the final prediction results in a fusion manner through making full use of the information extracted from each layer. Using modified VGG16 network, our method achieves state-of-the-art performance on several available remote sensing image datasets. In addition, we add the corresponding Digital Surface Model (DSM) map and extract the segmented building area to generate the point cloud of its roof. Then, based on the generated point cloud, the 3D modeling of buildings is implemented.**

***Index Terms*—building extraction, Hierarchically Fused Fully Convolutional Network (HF-FCN), 3D city modelling**

I. Introduction

BUILDING extraction, which aims to extract rooftop in a large-scale remote sensing image, remains one of the main challenges in the field of remote sensing for several decades. In addition, automatic extraction of building rooftop from aerial and satellite imagery is an important step in many applications, such as: urban planing, automated map making, 3D city modeling, updating geographical dataset and military reconnaissance. It is particularly difficult to extract rooftop from remote sensing images at the pixel level because of the following three reasons: i) Density of the structures in the scene. A rural scene has low density but an urban scene has high density, with a suburban scene in between (medium density). ii) Shape of the structure. Buildings come in many shapes from simple rectangular blocks with flat roof to complex shapes with intricate, multi-based roof structure. iii) Image quality. Images vary in terms of contrast, resolution, and visibility [1]. Some remote sensing patches are shown in Figure 1, which illustrate the difficulties of building extraction task.

M. Shell was with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332 USA e-mail: (see http://www.michaelshell.org/contact.html).

J. Doe and J. Doe are with Anonymous University.

Manuscript received April 19, 2005; revised August 26, 2015.

(a) (b) (c) (d)

Fig. 1. Examples of remote sensing patches with different kinds of challenges. (a) Shadow occlusion in green frame. (b) Low inter-class differences. (c) High intra class variance. (d) A lot of tiny buildings close to each other.

In the past decades, many investigators made some experimental investigations to extract buildings automatically. In the early days, many knowledge-based methods were put forward by [1], [2], [3], [4], [5]. Their basic ideas are derived from prior knowledge of buildings, for example buildings are closed polygons made up of some straight lines. Some other methods are based on energy functions, mainly including the variational level set evolution, improved snake model and graph cut [6], [7], [8].

In recent years, with the development of machine learning, some methods via machine learning are gradually penetrating into remote sensing domain. At first, some shallow networks were proposed for multiple object extraction [9], [10], [11], [12]. Afterwards, with the improvement of hardware and computation ability of computer, methods based on deep learning in computer vision are introduced into the field of remote sensing images. Some researchers tried convolutional neural networks for aerial images classification and semantic pixel labelling [13], [14], [15], [16], [17].

In this work, we propose a modified Convolutional Neural Network (CNN) architecture to extract buildings from satellite imagery by adding fusion sides. Not only the final prediction result obtained by CNN is used, but also the features of other layers are combined. We make full use of low-level appearance information on the basis of high-level semantic information. Numerous experiments on three remote sensing image datasets all obtained fairly good results. After rooftop extracting, the depth map is used to create the point cloud of rooftop. Based on the point cloud, the 3D models are carried out using the method proposed by Zhou [18]. Figure 2 illustrates pipeline of our work. Our technical contributions are:

1) A robust CNN which is specially designed for multi-scale building extraction is proposed. It can deal with the problems of different sizes, diverse appearance and mutual occlusion of buildings and etc. The overall accuracy based on HF-FCN exceeds the state-of-art algorithms.

2) Our approach has less computational cost than other

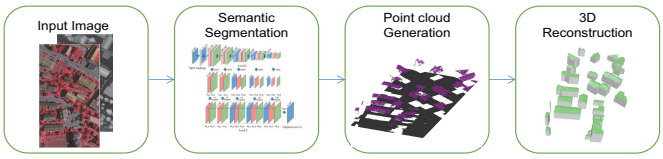


Fig. 2. An overview of proposed urban 3D modelling framework. The inputs of our system are multi-channel images. Semantic segmentation that extracts building areas from aerial images is the first step. After it, we generate the point cloud according to the DSM. Based on the point cloud, the 3D reconstruction is implemented.

methods.

3) A complete system for building extraction and 3D modelling of large-scale urban areas is proposed which makes the process of building extraction efficiently. And the building extraction part is a end-to-end network that does not need any post processing.

A. Paper Organization

The remainder of this paper is organized as follows. Section II summed up the related past works. In Section III we introduce the architecture of our network and show the training steps of our HF-FCN. And in Section IV, a brief description of the dataset used for our task is provided. HF-FCN training strategies, details and its evaluation metrics are described. In Section V we present the experimental results and the results of 3D modeling. Finally, the conclusion and discussion are discussed in SectionVI.

II. Related Work

Building extraction is one of most fundamental problems in remote sensing domain, which has been studied for nearly 30 years. Meanwhile, a lot of research achievements have sprung up. As time goes by, we can roughly divide these methods into three groups: one is based on the shape prior, another is based on the energy function and machine learning third. Here we briefly review some representative methods that have evolved in the past decades.

During early days, methods are mainly based on the hypothesis of prior knowledge. Huertas et al [1] assumed that buildings are rectangular or composed of rectangular components. Based on it, the approach detected lines and corners, traced object boundaries and used shadows to verify the hypotheses finally. Noronha [2] later proposed a system for building detection and modelling with the assumption that the roofs were flat or symmetrical and walls were vertical. By known ground height and detected rooftop, the reconstructed models could be soon obtained. Further, Nosrati [3] transformed the line and intersection points of the image into a graph presentation, and turned the problem of polygon finding into the one finding loop in the graph. However, it was still built on assumption that the buildings are polygonal. Izadi [4] presented a complete system for building detection and modelling. During building detection, a tree consisting of intersection points of lines was created and refined based on the found hypotheses. The sun azimuth and elevation angles were used to estimate the height with existing shadows. With the height of buildings estimated, the three-dimensional polygonal building model estimation was done. In recent years, very high resolution (VHR) optical satellite imagery could be obtained easily. Wang [5] proposed an efficient method for automatic rectangular building extraction from VHR remote sensing images by detecting line segments and grouping lines based on path completion and closed contour search.

The aforementioned shape-based methods seem to have a good performance in rural scenes with low density of buildings. However, there are some limitations of these methods. Firstly, the shape-based methods inherently limited to handle buildings of arbitrary shape. Secondly, these methods may failed to deal with complicated cases e.g. buildings are close to each other, which thereby is difficult to adapt to today's applications. Thirdly, the method based on shadow to verify corners and calculate height is greatly limited by obvious shadow and sparse building environment.

Later, some energy-based methods have been applied to automatic rooftop extraction. Cote [6] employed corner detection as initial rooftop estimates, and refined with level set evolution. Peng and Liu [7] proposed an approach that segments remote sensing images into high objects, ground and shadow regions, with further refined by an improved snake model. The urban-region-detection problems were casted as one of multiple subgraph matching by Sirmacek et al [8]. They considered each SIFT keypoint as a vertex, neighborhood between different vertex as edge of the graph and formulated the problem of building detection in terms of graph cut.

Over the past decade, CNNs have achieved great success in the field of computer vision. There are significant amount of efforts on semantic pixel-level classification for extraction buildings in remote sensing imagery. Mnih [9] proposed a shallow patch-based network which has only five layers. The input to the model was a 64 by 64 aerial image patch and the output of the network was processed by conditional random fields (CRFs). Satio et al. [10] applied two major strategies to improve the performance of the network proposed by Mnih. One was a channel-wise inhibited softmax (CIS) for getting a multi-label prediction result, the other was model averaging with spatial displacement (MA) for enhancing the prediction result. Alshehhi et al. [11] adjusted the architecture of network proposed by Mnih through changing the kernel size of convolutional layers and replacing the fully connection layer of the last layer with the average pooling layer. Alternative post-processing stages such as CRFs and multi-scales were used to improve the final result. Some methods took advantage of the feature extraction capability of CNNs to generate feature descriptions of patches. Paisitkriangkrai et al. [13] made use of both the CNN and hand-craft features extraction,which were combined together to generate final predicted labels of each patch. They also used CRFs as post-processing to get a sound result. Zhao et al. [12] proposed a method using edge information of VHR to guide semantic segmentation which combined by CNN-based deep features and semantic-free segments. Unlike Paisitkriangkrai, [17] proposed a multi-label pixelwise classification method using the feature vector

extracted by a CNN to train a Support Vector Machine (SVM) for classification.

More recently, Long [19] illustrated that Fully Convolutional Networks (FCN) could better handle the problem of multi-label pixel-wise classification. By up-sampling, final predicted result could be the same resolution of the input. Liu et al. [14] did a further research on the formulation proposed by Paisitkriangkrai [13] but used FCN as the branch of CNN and applied a higher-order CRFs as post-processing. Unlike traditional CRFs, the label consistency for the pixels within the same segment were enforced by higher-order CRFs. SegNet [20] delivered pooling indices computed in the max-pooling to the decoder. It eliminated the need of learning during the up-sample stage while achieving good segmentation performance. The SegNet architecture was used by Audebert et al. [15] for semantic labeling of remote sensing and got better prediction result compared to the traditional methods. Kampffmeyer at al. [16] proposed a novel idea that using CNN with missing data for urban land cover classification. Its idea came from a modality hallucination architecture proposed by Hoffman et al. [21] which learned with side information during training stage.

Although above-mentioned CNN-based models have exceeded the traditional methods significantly, all of them lost some important hierarchical features extracted from shallow layers of CNNs. These methods usually apply the CNN features from the last layer to get a segmentation result. It is possible to omit tiny objects during the process of pooling, and could not handle the situation when the size of buildings have great difference in distribution. Aiming at this case, a hierarchically fused fully convolutional network is proposed to combine CNN features extracted form each convolutional layer to capture detailed information of input. We will show the details of our paper below.

III. Hierarchically Fused Fully Convolutional Network

A. Network Architecture

VGG16 network which is widely used in computer vision field is regarded as the backbone network of our HF-FCN. It consists of 13 convolutional (conv) layers and 3 fully connected (fc) layers while its convolutional layers are divided into five groups with a pooling layer after each group. With the deepening of network, the receptive field (RF) of each activation units is increasing. Detailed receptive field and stride size of different layers are shown in Table I . The F1_1 in Table I denotes the feature map generated by conv1_1. Some modifications are made to apply to our roof extraction task including removing its fc layers and last pooling layer. The reasons of these operations are 1) The fc layer generates many parameters and takes up too much memory. 2) The existence of fc layer limits the size of input image. 3) After the last pooling layer, the resolution of the feature map is reduced to 1/32 of the input, which is too small to building extraction task.

In order to use the information extracted from different layers, we add the fusion branch on the backbone network. It fuses the prediction results obtained by different feature maps. This idea is similar to getting the response of scale function of images when looking for the SIFT feature points. After getting the responses of different scale functions, the biggest response is selected between adjacent scales of each feature point. In our network, choosing the most suitable scale is determined by weights of fusion layers. Extracting the information from different scales of receptive field as well as different levels of semantic layer is fused into a final prediction result. Meanwhile, all of our fusion operations are learned from the network indicating that network automatically learns the connection between feature maps. The noval network proposed by us is shown in Fig. 1. The trimmed VGG16 Net is denoted as Level 1 in our HF-FCN. Each convolutional layer in Level 1 is connected to a convolutional layer with kernel size 1. The outputs of these layers are upsampled and cropped into the same resolution of input image. All the up-sampling feature maps are concatenated and connected by an 1×1 convolutional layer. These parts make up our level 2.

In Level 1, with the growing of the receptive field, the detailed information is captured by each convolutional layer from fine-grained to coarser while the semantic information is captured from low level to high level. The first fusion operation is aimed to eliminate the redundancy within the feature maps of the same size. For the task of rooftop extraction, we need not only the information of the appearance of the building captured by shallow layers, but also the information of the line and corner extracted by middle layers and the high-level semantic information is needed which mainly comes from deep layers. Hence, in Level 2, we combine hierarchical features from whole up-sampling layers into a final prediction. Figure 4 shows the upsampled feature maps from different convolutional layers. Given an aerial image, the U1_1 in Fig. 4(b) with small receptive field extracts low-level features like edges. In Fig. 4(c), the U1_2 looks like an over-segmentation which groups pixels with similar color or texture into a subregion. In the U2_1, as Fig. 4(d) shows, shape information is augmented. From the U3_3, we can see that regions with significantly varying appearance are merged into an integrated building by considering high-level features. In U4_3 and U5_3, more semantic information of rooftop is got, which can distinguish the rooftop and the roads with similar color and deal with problems caused by shadow. Since all the upsampled feature maps are fused, it is expected to achieve a boost in rooftop segmentation. The final prediction result are shown in Fig. 4(h).

TABLE I
The receptive field and the stride size of VGG16 net

layer	F1_1	F1_2	Pool1	F2_1	F2_2	Pool2
RF	3	5	6	10	14	16
stride	1	1	2	2	2	4
layer	F3_1	F3_2	F3_3	pool3	F4_1	F4_2
RF	24	32	40	44	60	76
stride	4	4	4	8	8	8
layer	F4_3	Pool4	F5_1	F5_2	F5_3	Pool5
RF	92	100	132	164	196	212
stride	8	16	16	16	16	32

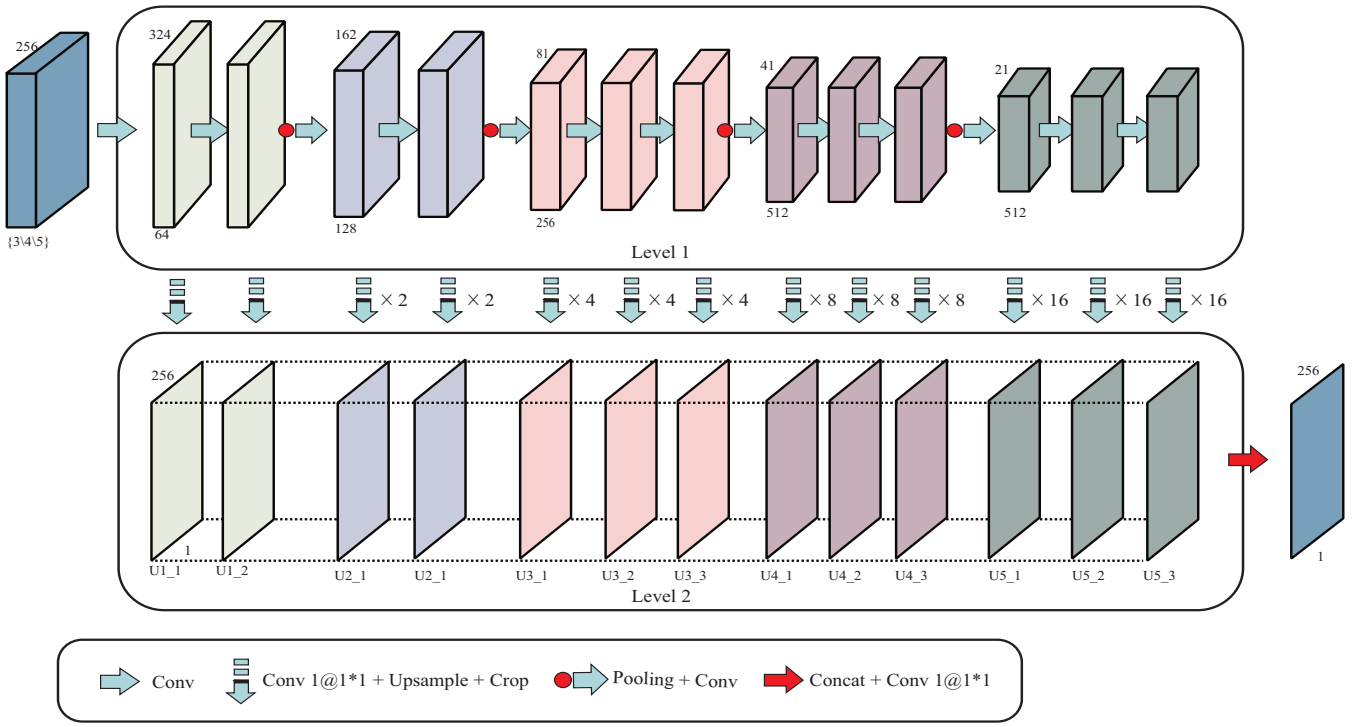


Fig. 3. Overall architecture of HF-FCN. The action of VGG16 network is extracting features from various layers. All feature maps from various layers merge into a feature map which eliminating redundancy in the same semantics and resolution. In Level 2, 13 feature maps are up-sampled and cropped to the same size of input. Finally the cropped feature maps fuse into a prediction result. All fusion operations used in HF-FCN are 1×1 convolution. The input channel could be 3, 4 or 5 for RGB, DSM, nDSM. ×2 means 2 times of upper sampling. U1_1 means the upsampling of F1_1, and so forth.

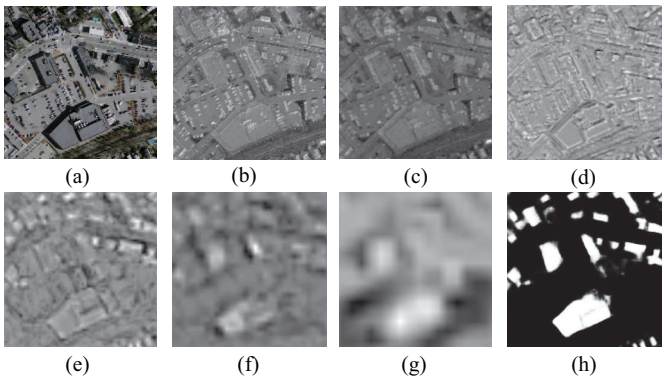


Fig. 4. (a) Input aerial image. (b-g) Feature maps generated from U1_1, U1_2, U2_2, U3_3, U4_3, U5_3, respectively. (h) Predicted labelling map

B. Network Training

The ground truth M in our dataset is labeled by 0 or 1. Here 1,0 means that this pixel belongs to a roof or not. When a remote sensing image X is inputted into the network, the output is a prediction probability map $P(X;W)$ of roof, and W denotes all the parameters that learned by HF-FCN. Each pixel value in $P(X_i;W)$ means the probability of this pixel belongs to rooftop. We use the sigmoid cross-entropy loss function formulated as

$$L(W) = -\frac{1}{|I|}\sum_{i=1}^{|I|}[\tilde{m}_i \log P(X_i;W) + (1-\tilde{m}_i)\log(1-P(X_i;W)], \tag{1}$$

where $\tilde{m}_i$ is label of X_i, $|I|$ is the number of pixels in the input image X.

IV. Experiments

To verify the effectiveness of the proposed network, extensive experiments have been conducted on three datasets. Compared not only with other methods applied in the field of remote sensing image, but also semantic segmentation methods in computer vision field, our network all achieves good performance. The performance of various variants of HF-FCN are also shown below. In this section, the experimental setup is described and the variants of HF-FCN are illustrated.

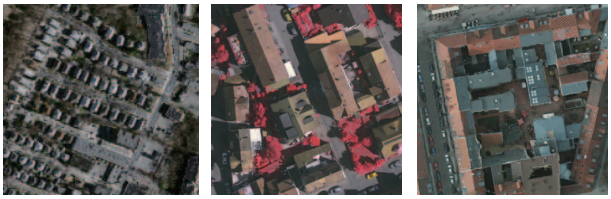
A. Dataset Description

a. Massachusetts dataset

Massachusetts dataset consists of 151 aerial images of the Boston area which covers roughly 340 square kilometers. The resolution of each image is 1500× 1500 with the spacial resolution of 1 meter per pixel, composed of red, green and blue channels. This dataset is built by Mnih while ground-truth of the images is produced by Saito et al. This dataset is split into three parts, a training set of 137 images, a test set of 10 images and validation set of 4 images. To train the network, we create a set of image tiles for training and validation. The detailed description is shown in Table II.

b. Vaihingen dataset

Vaihingen dataset is captured over Vaihingen which is a relatively small village with many detached buildings and small multi story buildings in Germany. This dataset contains 16 labeled images whose spacial resolution is 9cm per pixel. It contains near infra-red, red, green, blue imagery with corresponding normalized digital surface models (nDSMs)



(a) (b) (c)

Fig. 5. Sample patches on the three datasets (a) Massachusetts dataset (b) Vaihingen dataset (c) Potsdam dataset

TABLE II
COMPOSITION OF DATASET

	Masschusetts	Vaihingen	Potsdam
Labeled images	151	16	24
GSD	1m	9cm	5cm
Bands	R,G,B	IR,R,G,DSM	IR,R,G,B,DSM
Training images	137	11	17
Training patches	75938	115088	85000
Validation images	4	3	4
validation patches	2500	28376	25000
Test images	10	2	3

and row DSMs. The dataset is divided into training set, validation set and test set which has 11 images, 2 images and 3 images, respectively. The same crop operations are done as the Massachusetts dataset.

c. Potsdam dataset

In Potsdam dataset, there are 24 labeled images whose ground sampling distance is 5cm. This dataset shows a typical historic city with large building blocks. In order to grasp the global information of the building, we reduced the spacial resolution of the original image from 6000×6000 to 1500×1500. This dataset contains information about 5 channels: red, green, yellow, DSM and nDSM. We split the dataset into training, validation and test sets in proportion to 7:2:1.

Data augmentation is made on the dataset b and c. One reason is that methods use the dataset a do not extend the data. To make a fair comparison with other methods, we also do not extend it. Another reason is the data quantity of dataset b and c is not enough. It may lead to inadequate training of our network. Hence, some simple augmentation operations are made in dataset b and c including data rotation and mirror filp. Component of the datasets are shown in Table II. Figure 5 shows the sampled patches of dataset a, b, c.

B. Training Settings

HF-FCN first trained on dataset a because of large amounts of training data. The pre-trained VGG16 Net model is used to finetune our HF-FCN. We Use the stochastic gradient descent algorithm with the learning rate divided by 10 for each 8000 iterations to train our network. The drop-out ratio is set to 0.5 which avoids overfitting. When the HF-FCN converges on the dataset a, we transfer it to the other datasets. All experiments in this paper are performed using the deep learning framework Caffe and train on a single NVIDIA Titan 12GB GPU. The hyper-parameters are listed in Table III.

TABLE III
PARAMETERS FOR NETWORK TRAINING

	Massachusetts	Vaihigen	Potsdam
mini-bachsize	18	15	15
initial learning rate	10^-5	10^-6	10^-5
test_interval	1000	1000	1000
training iteration	10000	10000	10000
momentum	0.9	0.9	0.9
clip_gradients	16000	10000	10000
weight_decay	0.02	0.005	0.005

C. Evaluation Metrics

Some evaluation metrics are adopted in our work. For dataset a, the most common metrics are correctness (precision) and completeness (recall). We use the stantdard precision and recall scores (ρ=0) and the relaxed precision and recall scores with ρ=3 to evaluate the prediction results. Here the relaxed precision means the predicted pixels are within ρ pixels of a true pixel while the relaxed recall is the true pixels are within ρ pixels of a predicted pixel. The time cost is used to measure the efficiency of our HF-FCN. For dataset b and c, we use correctness, completeness and F1 score as evaluation metrics.

$$completeness = \frac{TP}{TP+FN}, \tag{2}$$

$$correctness = \frac{TP}{TP+FP}, \tag{3}$$

$$F1\_score = 2 \cdot \frac{completeness \cdot correctness}{completeness + correctness} \tag{4}$$

where TP indicates the true positives, FP indicates the false positive, TN indicates the true negatives and FN indicates the false negatives.

V. RESULTS AND DISCUSSION

In this section, the proposed method using dataset a, b, c are compared to the recent non-deep-learning algorithms, such as Minh-CNN [9], Satio-multi [10], Context [15]. In addition, our method is compared with some recent deeplearning based approaches, including FCN [19], SegNet [20] and etc. Moreover for HF-FCN itself, we expect to investigate the effects of extracted information from different layers on the final prediction. Hence, some variants which combine different upsampling feature maps from Level 2 are proposed with details shown in Figure 6.

A. Massachusetts dataset

On the Massachusetts dataset, our method is compared to three state-of-the-art approaches. Table 4 and Figure 7 present the quantitative analysis and precision-recall curves respectively. A standard and relaxed precision and recall are amply to make a comparison. From the reuslt, our method shows obvious superiority in terms of speed and precision. When comparing with SatiomultiMA&CIS, standard and relaxed recall are 5.5% and 1.3% higher than it, respectively. And, at the same time, the time coat is reduced from 67.84s to 1.07s, the speed is up about 63 times. These significant

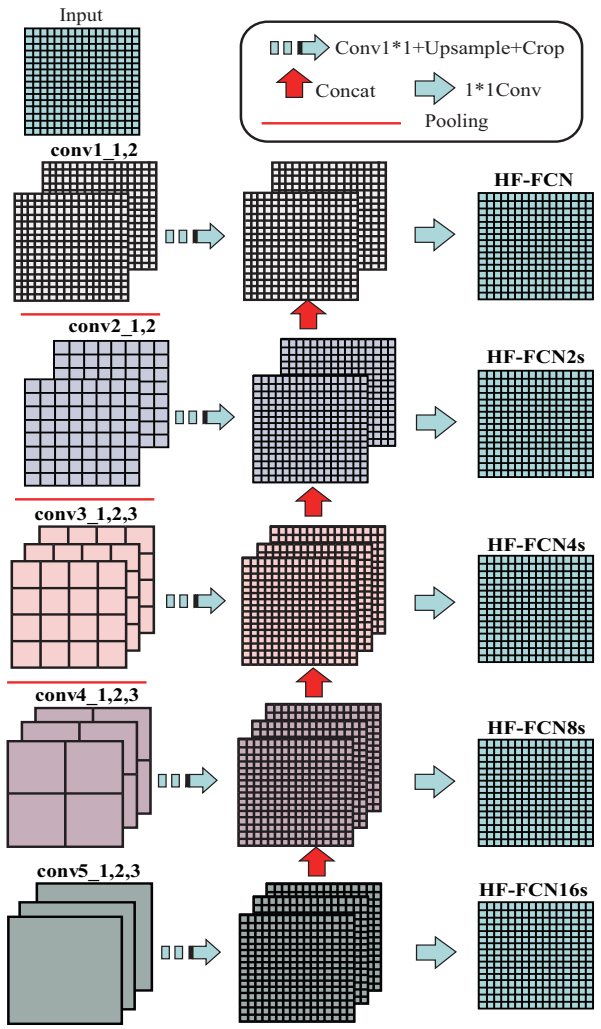


Fig. 6. HF-FCN variants. The feature maps generated from final group are fused into a coarse result, which is HF-FCN16s. The variant called HF-FCN8s concatenates the feature maps from the last 2 groups with the same fusion operation, and so on.

TABLE IV
CORRECTNESS AT BREAKEVEN OF HF-FCN V.S. [9] [10] [11] ON MASSACHUSETTS TEST SET. COST TIME IS COMPUTED IN THE SAME COMPUTER WITH A SINGLE NVIDIA TITAN 12GB GPU

	Recall (ρ = 3)	Recall (ρ = 0)	Time (s)
Mnih-CNN [9]	0.9271	0.7661	8.70
Mnih-CNN+CRF [9]	0.9282	0.7638	26.60
Satio-multi-MA [10]	0.9503	0.7873	67.72
Satio-multi-MA&CIS [10]	0.9509	0.7872	67.84
Alshehhi-GAP+seg [11]	0.955	–	–
Proposed model (HF-FCN)	0.9643	0.8424	1.07

improvements demonstrate that HF-FCN achieves best performance in effectiveness and efficiency.

Meanwhile, extensive comparisons are made between HF-FCN and other mainstream methods in semantic segmentation domain. The visual results are shown in Figure 10. From it, we can see that details and integrity of the building are well preserved by using our method.

To explore the effects of the feature maps at different scales on the final result, variants of HF-FCN which are counterpart

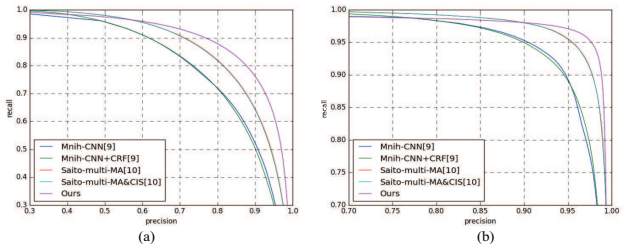


Fig. 7. The relaxed precision-recall curves from different methods with two slack paramters, (a) ρ = 0; (b) ρ = 3. All curves of our model are located above others.

TABLE V
PERFORMANCE COMPARISON BETWEEN HF-FCN VARIANTS ON MASSACHUSETTS TEST SET.

	Recall (ρ = 3)	Recall (ρ = 0)
HF-FCN16s	0.9330	0.7233
HF-FCN8s	0.9643	0.8171
HF-FCN4s	0.9632	0.8394
HF-FCN	0.9643	0.8394

of FCN are designed. Unlike FCN, a fusion operation rather than summing up components in respective locations are leveraged to build our HF-FCN 16s, 8s, 4s. The performance of these variants are shown in Figure 8, Figure 9 and Table V. From the disgrams, we get the following conclusions. First, the prediction result obtained from the last layer gets a coarse result, which losts much of location information that mainly encoded in the shallow feature maps. Second, the largest gap presented between HF-FCN16s and HF-FCN8s about 9% in recall rates, it may suggest that the most information supplement to the HF-FCN is got in middle layers. Third, the PR curves of HF-FCN4s and HF-FCN almost coincide. It illustrates the low-level information has little effect on the prediction results. Forth, with the addition of the shallow feature map, the network is more distinct for the segmentation of tiny buildings, which solves the problem of easy adhesion to adjacent buildings. Since, all the Conv layers contained useful hierarchical information that is critical to the final prediction.

B. Vaihingen dataset

On Vaihingen dataset, three experiments are undertaken to explore the effects of different inputs, diverse variants and various methods. We utilize three kinds of combinations of

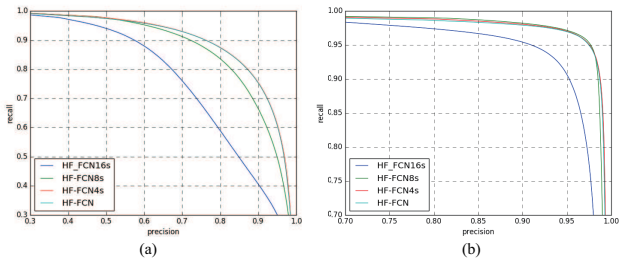


Fig. 8. The relaxed precision-recall curves from HF-FCN variants with two slack paramters. The biggest gap occurs between HF-FCN16s and HF-FCN8s, which indicates the most additional information coming from middle layers.

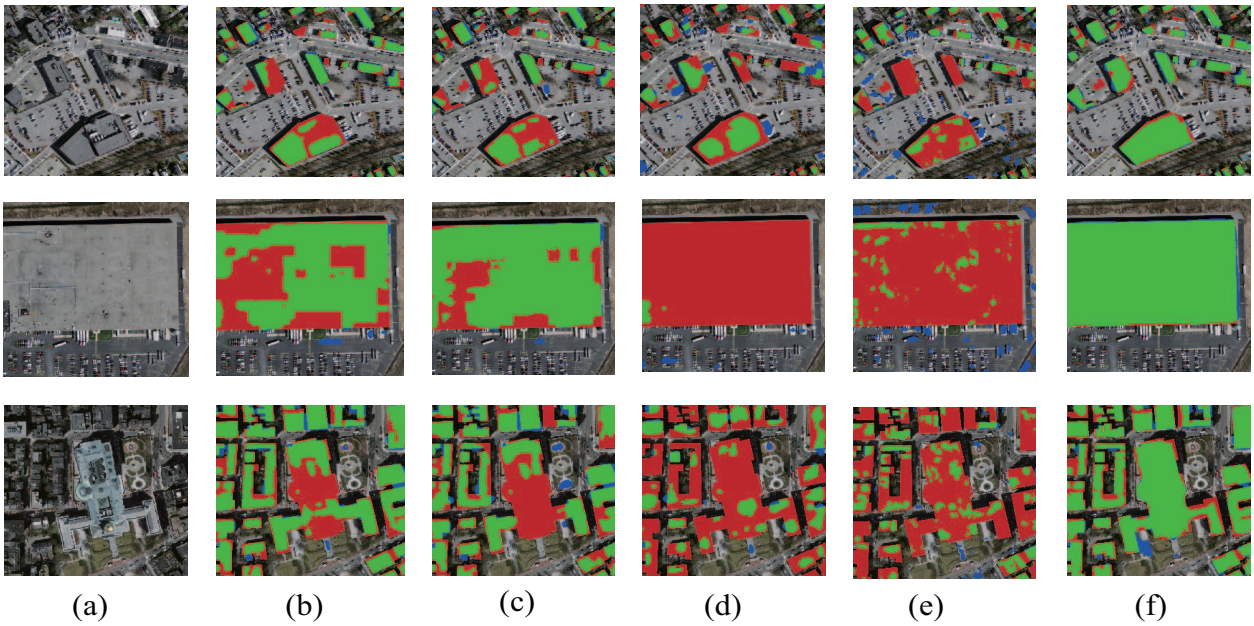


Fig. 10. (a) input images. (b) Results of Mnih-CNN+CRF. (c) Results of SatiomultiMA&CIS. (d) Results of FCN4s . (e) Results of SegNet. (f) Our results. TP are shown in green, FP are shown in blue and FN are in red.

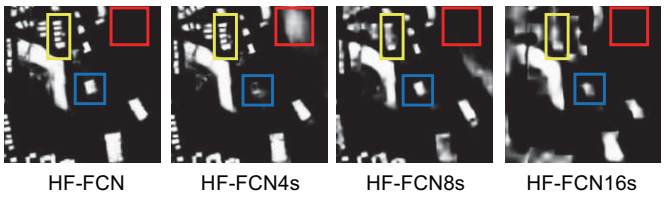


Fig. 9. Prediction results of HF-FCN, HF-FCN4s, HF-FCN8s and HF-FCN16s. The yellow box shows the continuous refinement of the tiny buildings. The red and blue boxes show the mutual promotion and contradiction between different layers.

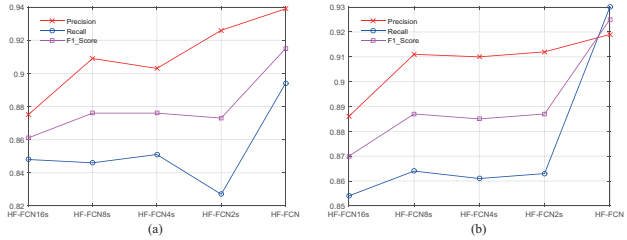


Fig. 11. Results of HF-FCN variants on Vaihingen dataset. (a) (b) shows the precision, recall and F1_score of validation set and test set of Vaihingen dataset respectively.

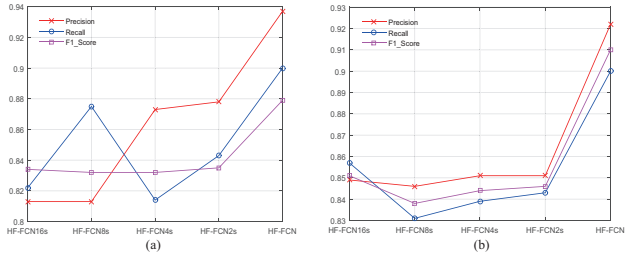


Fig. 12. Results of HF-FCN variants on Potsdam dataset. (a) (b) shows the precision, recall and F1_score of validation set and test set of Potsdam dataset respectively.

image channels as inputs. The inputs of the 3 channels are IR, R, G and adding the nDSM as the forth channel. Based on it, DSM is added and made up 5-channel input. We use three standards to make a more comprehensive evaluation. The evaluation results are shown in Table VI, which illustrates that 3-channel input performed better than the other. Corresponding visual results are shown in Figure 13.

The results of diverse variants are shown in Figure 11. From the curves, the performance of HF-FCN exceeds the variants and gets a excellent result. There are some other methods using this dataset. The comparison results are shown in Figure 14. From a visual perspective, our method gets a much more refined roof region. After getting the area of rooftop, the method proposed by [18] is used to generate the 3D models. Complete models and details are displayed in Figure 15.

C. Potsdam dataset

The same experiments are implemented on Potsdam dataset. First, We utilize DSM and IR information as extra inputs based on the RGB input. The specific quantitative evaluation and intuitive visual prediction results are shown in TableVII and Figure 16. In the validation process, the 4-channel input gets better overall performance. Meanwhile, the 5-channel input seems perform better in the course of testing. From the visual results, the 5-channel input network gets lower error detection rate which is shown on the image with small blue areas. And from the 5-channel input to 3-channel input, the F1 score increases from 0.907 to 0.915 on the validation set and increases 0.047 on the test set

As done on Vaihingen dataset, contrast experiments of HF-

TABLE VI
PERFORMANCE COMPARISON OF THE RESULTS OF DIFFERENT INPUTS ON VAIHIGEN DATA SET

	Img	3_in: IR, R, G			4_in: IR, R, G, nDSM			5_in: IR, R, G, DSM, nDSM		
		Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
Val	11	0.911	0.906	0.909	0.936	0.900	0.917	0.890	0.900	0.900
	28	0.94	0.875	0.906	0.96	0.792	0.868	0.952	0.823	0.883
	34	0.965	0.899	0.930	0.987	0.902	0.942	0.972	0.918	0.944
	Ave	0.939	0.894	**0.915**	0.961	0.865	0.909	0.939	0.880	0.907
Test	15	0.918	0.930	0.924	0.883	0.917	0.9	0.833	0.931	0.88
	30	0.921	0.929	0.926	0.931	0.827	0.876	0.875	0.877	0.876
	Ave	0.919	0.930	**0.925**	0.907	0.872	0.888	0.858	0.900	0.878

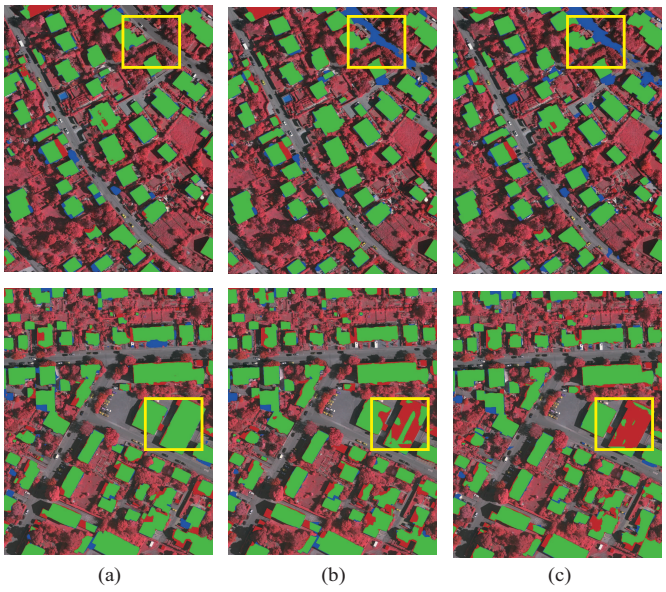


Fig. 13. Prediction results on Vaihingen dataset. (a) (b) (c) shows results of the 3-channel input, 4-channel input and 5-channel input of Vaihingen dataset respectively. Here, TP are shown in green, FP are shown in blue and FN are in red.

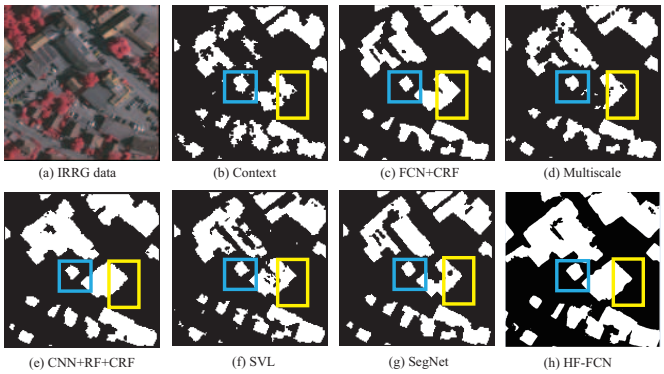


Fig. 14. Results of different methods. (a) is input image, (b)(d)(g) are results of [15], (c) is result of [22], (f) is result of [23], (g) is our result.

FCN variants are implemented. The performance curve of HF-FCN variants are shown in Figure 12. We compare HF-FCN with other methods applied to the Potsdam dataset. Some qualitative results are shown in Figure 17. HF-FCN got more remarkable segmentation results while edges and details segmentation of other methods do not perform so good. The same way of 3D city modelling is applied to the Potsdam

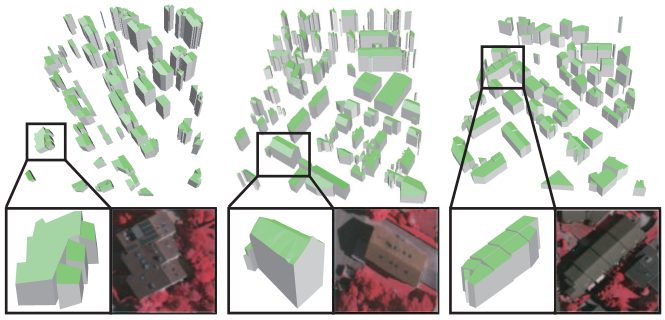


Fig. 15. The 3D modelling of Veihingen dataset. The single building model and its corresponding optical patch were shown together.

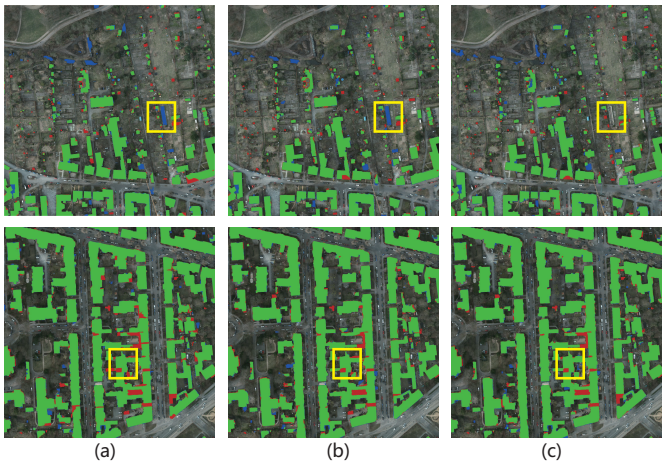


Fig. 16. Prediction results on potsdam dataset. (a) (b) (c) shows results of the 3-channel input, 4-channel input and 5-channel input of Vaihingen dataset respectively. Here, TP are shown in green, FP are shown in blue and FN are in red.

dataset. Some models of scenes are shown in Figure 18.

VI. CONCLUSION

In this paper, we propose a complete system for efficient 3D city modelling from large-scale aerial images via deep learning. A novel CNN architecture, HF-FCN, combines hierarchical semantic layers and multi-scale feature representation to implement final building detection. We design a fusion branch to fuse the feature map stage by stage. And the resulting HF-FCN method shows significant improvements over several previous methods. Distinct from the previous deeplearning based methods, we utilize the multi-scale inherent information within the CNN and get fine detail detection results. Unlike

TABLE VII
PERFORMANCE COMPARISON OF THE RESULTS OF DIFFERENT INPUTS ON POTSDAM DATA SET

	Img	3_in:RGB			4_in:RGB,IR			5_in:RGB,IR,nDSM		
		Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
Val	2_11	0.917	0.950	0.933	0.917	0.978	0.946	0.934	0.976	0.954
	4_10	0.937	0.945	0.941	0.926	0.943	0.936	0.947	0.946	0.946
	5_11	0.930	0.972	0.950	0.959	0.975	0.966	0.956	0.977	0.967
	7_10	0.964	0.536	0.689	0.950	0.590	0.728	0.939	0.554	0.697
	Average	0.937	0.851	0.879	0.937	**0.872**	**0.894**	**0.944**	0.864	0.891
Test	2_12	0.897	0.868	0.882	0.920	0.959	0.939	0.944	0.965	0.955
	6_7	0.894	0.902	0.898	0.915	0.909	0.912	0.901	0.918	0.909
	7_8	0.975	0.929	0.951	0.977	0.950	0.957	0.976	0.946	0.960
	Average	0.922	0.900	0.910	0.937	0.935	0.936	**0.940**	**0.943**	**0.941**

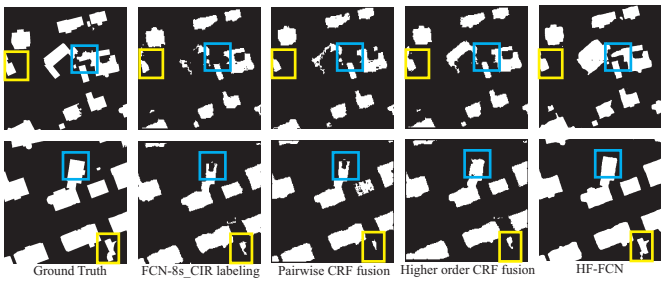


Fig. 17. Results of different methods. The second column is the results of using only the FCN with CIR. Pairwise CRF fusion shows the result of fusing FCN-8s_CIR with LiDAR data in a pairwise CRF. Higher-order CRF are used to generate the results shown in third column. Our results are shown in last colunm.

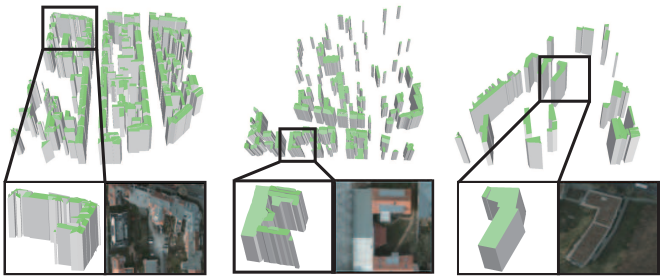


Fig. 18. The 3D modelling of Potsdam dataset. The single building model and its corresponding optical patch were shown together.

existing 3D reconstruction methods, our proposed approach relies on the HF-FCN to efficiently extract the area of buildings to build the 3D models of the geospatial objects. Finally, our study suggests that even with the powerful semantic expressive ability of CNNs and their good robustness to scale, it is still critical to address multi-scale problems utilizing hierarchical feature maps encoded in CNNs.

REFERENCES

[1] A. Huertas and R. Nevatia, "Detecting buildings in aerial images," *Computer Vision, Graphics, and Image Processing*, vol. 41, no. 2, pp. 131–152, 1988.

[2] S. Noronha and R. Nevatia, "Detection and modeling of buildings from multiple aerial images," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 5, pp. 501–518, 2001.

[3] M. S. Nosrati and P. Saeedi, "A novel approach for polygonal rooftop detection in satellite/aerial imageries," in *Image Processing (ICIP), 2009 16th IEEE International Conference on*. IEEE, 2009, pp. 1709–1712.

[4] M. Izadi and P. Saeedi, "Three-dimensional polygonal building model estimation from single satellite images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 6, pp. 2254–2272, 2012.

[5] J. Wang, X. Yang, X. Qin, X. Ye, and Q. Qin, "An efficient approach for automatic rectangular building extraction from very high resolution optical satellite imagery," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 3, pp. 487–491, 2015.

[6] M. Cote and P. Saeedi, "Automatic rooftop extraction in nadir aerial imagery of suburban regions using corners and variational level set evolution," *IEEE transactions on geoscience and remote sensing*, vol. 51, no. 1, pp. 313–328, 2013.

[7] J. Peng, D. Zhang, and Y. Liu, "An improved snake model for building detection from urban aerial images," *Pattern Recognition Letters*, vol. 26, no. 5, pp. 587–595, 2005.

[8] B. Sirmacek and C. Unsalan, "Urban-area and building detection using sift keypoints and graph theory," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 4, pp. 1156–1167, 2009.

[9] V. Mnih, "Machine learning for aerial image labeling," Ph.D. dissertation, University of Toronto (Canada), 2013.

[10] S. Saito, T. Yamashita, and Y. Aoki, "Multiple object extraction from aerial imagery with convolutional neural networks," *Electronic Imaging*, vol. 2016, no. 10, pp. 1–9, 2016.

[11] 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 Journal of Photogrammetry and Remote Sensing*, vol. 130, pp. 139–149, 2017.

[12] W. Zhao, S. Du, Q. Wang, and W. J. Emery, "Contextually guided very-high-resolution imagery classification with semantic segments," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 132, pp. 48–60, 2017.

[13] S. Paisitkriangkrai, J. Sherrah, P. Janney, V.-D. Hengel *et al.*, "Effective semantic pixel labelling with convolutional networks and conditional random fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 36–43.

[14] Y. Liu, S. Piramanayagam, S. T. Monteiro, and E. Saber, "Dense semantic labeling of very-high-resolution aerial imagery and lidar with fully-convolutional neural networks and higher-order crfs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 76–85.

[15] N. Audebert, A. Boulch, H. Randrianarivo, B. Le Saux, M. Ferecatu, S. Lefevre, and R. Marlet, "Deep learning for urban remote sensing," in *Urban Remote Sensing Event (JURSE), 2017 Joint*. IEEE, 2017, pp. 1–4.

[16] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen, "Urban land cover classification with missing data using deep convolutional neural networks," *arXiv preprint arXiv:1709.07383*, 2017.

[17] Y. He, S. Mudur, and C. Poullis, "Multi-label pixelwise classification for reconstruction of large-scale urban areas," *arXiv preprint arXiv:1709.07368*, 2017.

[18] Q.-Y. Zhou and U. Neumann, "2.5 d building modeling with topology control," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 2489–2496.

[19] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[20] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for scene segmentation," *IEEE transactions on pattern analysis and machine intelligence*, 2017.

[21] J. Hoffman, S. Gupta, and T. Darrell, "Learning with side information through modality hallucination," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 826–834.
[22] D. Marmanis, J. D. Wegner, S. Galliani, K. Schindler, M. Datcu, and U. Stilla, "Semantic segmentation of aerial images with an ensemble of cnss," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2016*, vol. 3, pp. 473–480, 2016.
[23] M. Gerke, "Use of the stair vision library within the isprs 2d semantic labeling benchmark (vaihingen)," 01 2015.