Efficient Building Detection from Satellite Images Using Multi-scale Networks

Feiyu Qin, Tongcun Zuo, Xuejin Chen

***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 the 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 method which extracts buildings from large-scale remote sensing images efficiently via deep learning. And further we extend our method to the 3D building reconstruction to accelerate the overall process. Our study demonstrates that learning low-level appearance information and high-level semantic information are equally important in building extraction task since buildings possess various scales and aspect ratios in the scene. Hence, to make full use of the information extracted from each layer, we propose a simple but effective hierarchical fusion operation which fuses the feature maps between channels stage by stage. In this paper, a novel network named hierarchical fused fully convolution network(HF-FCN) is also described which fuses the information through combining the fusion operations to the general networks. The experiments on several available remote sensing image datasets show that our method achieves state-of-the-art performance.**

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

I. Introduction

BUILDING extraction, which aims to extract building regions in a high resolution satellite image, remains as one of the fundamental challenges in the field of remote sensing. Extraction of building rooftops 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 so on. However, it is particularly challenging to automatically extract rooftops at the pixel level for the following reasons:

- Different density of buildings in different type of scenes. A rural scene has low density, while an urban scene has high density, and a suburban scene has medium density in between.
- Diverse shapes of the buildings. Buildings appear in a wide range of shapes, varying from simple rectangular blocks with flat roof to complex shapes with intricate roof shapes.
- The quality of remote sensing images. Images vary in terms of contrast, resolution, and image principle [1]. (xuejin:Do we have problem with image qualities? Or do you have experiments on that?)

We show several typical patches of satellite images in Fig. 1 to illustrate the above challenges in the building extraction task. Fig. 1(a) shows that the back face of the roof covered in shadow. And Fig. 1(b)(c) demonstrates the variance between and within different classes. Additionally, the dense and tiny buildings are presented in Fig. 1(d).

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

Fig. 1. Examples of satellite image patches with different kinds of challenges. (a) Shadow occlusion in green frame. (xuejin:But you only extract building roofs..So the shadow is good for large contrast.)(b) Low inter-class differences. (c) High intra class variance. (d) A large amount of tiny and dense buildings.

In the past decades, many researchers have made enormous effort to extract buildings automatically from satellite images. At first, many simple knowledge-based methods were put forward [1]–[5]. Their basic ideas are derived from prior knowledge that buildings are closed polygons made up of some straight lines. Some others are energy-based methods including the variational level set evolution, improved snake model and graph cut [6]–[8]. Due to that early methods depend heavily on prior knowledge and initialization, they do not work well for building extraction of complex scenes.

In recent years, with the development of machine learning, many deep-learning-based techniques are gradually introduced into the remote sensing area. At first, some shallow networks were proposed for multiple geographic object extraction [9]–[12]. These methods typically use image patches with a fixed size feeded into a shallow CNNs for segmentation, they are not efficient or accurate enough for the pixel-wise segmentation task for large-scale satellite images. (xuejin:Do these method use deep learning or just machine learning?) Later on, with the rapid growth of computational powers, deep learning technology developed rapidly and some researchers tried deep learning for aerial images classification and semantic pixel labeling [13]–[17]. Unfortunately, while ignoring the hierarchical information extracted by the network, they could not deal with the scenes which contains close-packed buildings well.

This paper propose a relatively simple, but very effective strategy for fusing multi-scale features in neural networks. It could be combined with a general CNN architecture easily for robust building extraction. Differ from above mentioned methods, we take full advantages of *the low-level appearance information as well as high-level semantic information* by the

novel fusion operation in a way of stage by stage. Inspired by the FCN [18], whose output is in the same resolution of its input, we propose a novel hierarchically fused FCN, named HF-FCN for buildings pixel-wise classification. Differ from the traditional FCN, a set of hierarchical fusion operations are used to fuse the intra layer information and inter layer information respectively which improve the performance of FCN greatly. And numerous experiments conducted on three remote sensing image datasets all obtain fairly good results. Further, we integrate our method into the pipeline of building reconstruction as the part of building detection. (xuejin:3d modeling is not part of building detection.) (xuejin:Do you have any quantitative measure for the quality of 3D reconstruction?)

Overall, our technical contributions are:

1) We propose an effective hierarchical fusion strategy which is specially designed for multi-scale building extraction in high-resolution satellite images. Combining with a generic FCN, a novel network is presented, named HF-FCN that can deal with the problems of different sizes, diverse appearance and mutual occlusion of buildings and etc.
2) HF-FCN is an end-to-end network that does not need any post processing. The approach is much more computationally efficient than existing techniques. Besides, the overall accuracy of the proposed HF-FCN exceeds the state-of-art algorithms. (xuejin:Do you compare with recent papers?)

The remainder of this paper is organized as follows. Sec. II sums up the related works in the past. In Sec. III, we introduce the fusion operation and architecture of HF-FCN, as well as the training details. In Sec. IV, a brief description of the dataset used for our task is provided. HF-FCN training strategies, details and its evaluation metrics are also described. In Sec. V, we display and analysis the experimental results. Extension in 3D building modeling are presented in Sec. VI. Finally, the conclusion is discussed in Sec. VII.

II. Related Work

Building extraction is one of the most fundamental problems in the remote sensing area, which has been studied for nearly 30 years. As time goes by, many research achievements have sprung up. We roughly divide these methods into three groups: shape-prior-based, energy-based, and machine learning methods. Moreover, several related work which are popular in compute vision domain similar to our task are also introduced.

Methods Based on Shape Priors. During early days, without plenty of data and efficient learning strategy, early methods are mainly based on prior knowledge that describes the roof shapes and appearances. Assuming that buildings are typically composed of rectangular components, Huertas and Nevatia [1] detected lines and corners, traced object boundaries and used shadows to verify building hypotheses. Later, a system [2] for building detection and modeling was proposed using a hypothesize and verify approach. The approach presumed that the rectangular roof components were formed by lines and verified by finding evidence of visible walls and shadows. With known ground height and detected rooftop, the reconstructed models could be soon obtained. Further, Noronha and Nosrati [3] transformed the line and intersection points in remote sensing images into a graph presentation, and turned the problem of polygon detection into cycle detection in the graph. Later, Izadi and Saeedi [4] presented a complete system for building detection and modeling. (xuejin:what is its relation with [2]?) Differ from the [2], in the stage of extracting buildings, a tree consisting of intersection points of lines was created and refined according to the found hypotheses. The sun azimuth and elevation angles were used to estimate building heights with the shadow afterwards. In recent years, very high resolution (VHR) optical satellite imagery could be obtained easily. Wang et al [5] proposed an efficient method for automatic rectangular building extraction by detecting line segments and grouping them based on path integrity and closed contour. Nevertheless, the method depends on the clear remote sensing images and accurate line segmentation heavily.

The aforementioned prior-based methods achieves good performance in rural scenes with sparse buildings. Nevertheless, there are several limitations of these methods. First, shape prior-based methods have difficulties on handling buildings of arbitrary shapes. Second, they may fail to deal with complicated cases, for instance, buildings are close to each other. Third, the algorithms using shadows to verify corners and estimate height greatly rely on clear shadows and sparse building environment.

Energy-based Methods. Meanwhile, several energy-based methods in image segmentation domain have been applied to automatic rooftop extraction. At first, Peng et al [7] used an improved snake model to refine the coarse segmentation results. The urban-region-detection problems were casted as one of multiple subgraph matching by Sirmacek and Unsalan [8]. They considered each SIFT keypoint as a vertex, neighborhood between vertexes as edges of the graph and converted the original image segmentation problem into a graph-cut optimization process. Later, a level set evolution method was proposed by Cote and Saeedi [6] which employed detected corners as endpoints of the initial curves and refined by level set evolution. (xuejin:what is the order of the related papers? year? method?) The most crucial part of energy-based methods is a good initialization, which is very sensitive to image contents. Therefore, energy-based methods are not applicable to the high-altitude remote sensing images of intensive buildings with severe shadow occlusions and diverse building appearances.

Shallow Networks. While machine learning have achieved great success in the field of computer vision, a large amount of effort has also been put on the building extraction task using machine learning technique. At first, Mnih [9] proposed a shallow patch-based network which has five layers with a 64×64 aerial patch as input. The output of the network was processed by conditional random fields (CRFs) to constrain the segmentation continuity. Afterwards, Satio et al. [10] putted forward two major strategies to improve the performance of [9]. One was a channel-wise inhibited softmax (CIS) for getting a multi-label prediction result, the other was model averaging (MA) with spatial displacement for enhancing the

prediction result. Alshehhi et al. [11] adjusted the architecture of [9] through changing the kernel size of convolutional layers and replacing the last fully connection layer with the average pooling layer. Alternative post-processing strategies such as CRFs and multi-scales were also used to improve the final prediction results. At the same time, some other methods took advantage of the feature extraction capability of CNNs to generate feature descriptions of patches for further segmentation. For instance, Paisitkriangkrai et al. [13] combined CNN-extracted features and hand-crafted features together to generate predicted labels for each patch. CRFs were used as post-processing to get a sound result. Unlike [13], [17] put forward a multi-label pixel-wise classification method using the feature vector extracted by a CNN to train a Support Vector Machine (SVM) for classification. Other appearance information, such as edges, are also harnessed to guide the shallow network to extract buildings [12].

Although these methods outperformed traditional methods, there are still several disadvantages. (a) The methods using shallow networks always cast the problem of building segmentation as a patch classification problem. It greatly reduces the segmentation accuracy. (b) Most of them are processed by at least one kind of post-processing, which is time-consuming.

Deep Learning Methods. More recently, Long et al. [18] illustrated that Fully Convolutional Networks (FCN) handle the problem of multi-label pixel-wise classification better. Hence, Liu et al. [14] conducted a further research on the formulation proposed by Paisitkriangkrai [13], but with FCN instead of shallow networks. A higher-order CRF was applied as post-processing. In order to reduce the information loss in pooling layers and accelerate the decodeing of FCN, SegNet [19] delivers pooling indices(xuejin:indexes?) computed in the max-pooling stage to the decoder. (xuejin:Is it a common strategy?) By using preserved pooling indices mask, it eliminated the need of learning and recovered lost information during unpooling stage.(xuejin:unclear here.) Accordingly, Audebert et al. [15] explored how the deep learning methods could be used in remote sensing. They applied SegNet (xuejin:without any modifications?) to semantic labeling of remote sensing images and got better prediction results compared to traditional methods. Later, Kampffmeyer at al. [16] proposed a novel idea that using CNN with missing data for urban land cover classification. The idea came from a modality hallucination architecture proposed by Hoffman et al. [20] which solved the problem that missing some kinds of data during test process.

Above-mentioned deep learning models have exceeded the traditional methods significantly. However, most of them completely ignored important hierarchical features encoded in the CNNs. Because the building size varies largely in different area, both low-level and high-level features are important to accurately extract building boundaries.

Common Semantic Segmentation. In the pioneering FCN network [18], the feature maps in deeper layers, after a series of downsampling, lost many fine structures. In order to weaken the detail loss, many new networks have been proposed for semantic segmentations in computer vision. Chen et al. [21] proposed an atrous (xuejin:Is this word right?) convolution which enlarged the receptive fields and reduced the number of pooling layers at the same time. Vemulapalli et al. [22] further extended the Deeplab [21] with a pairwise network and proposed a Gaussian Conditional Random Field Network for more continuous segmentation results. Afterwards, with the advent of the powerful networks such as ResNet [23], GoogLeNet [24] and their variants [25] [26] [27], a large amount of literature made use of these networks as their main structure for semantic segmentation. Zhao et al. [12] recently developed a pyramid pooling module following the ResNet [23] to get multi-scale feature maps and connected them with the feature maps which before pyramid pooling to create the final prediction. Zuo et al. [28] described a hierarchically fused fully convolutional network, which combined the feature maps from each group of VGG16 Net to generate the final prediction. In this paper, we extend the work of [28] and proposed a simple but effective fusion operation that could be easily combined to the general network. We also explore the effect of different layers of features on the final result. The details of our idea will be described in the next section.

The most related work are U-Net [29] and FPN [30]. They both exploited the information from different layers. Differ from the U-Net which simple concatenated the feature maps from encoder to decoder, we apply a fusion operation firstly to fuse the feature maps created in the same convolution layers in the path of encoder to get more richer features. The main idea of FPN was leveraging the encoder part as a feature pyramid, with predictions made independently at all levels. During the top-down path of FPN, it only exploited the feature maps which were came from the last residual block of each stage. In comparison, we take advantage of all the feature maps in our network. Moreover, we upsample the feature maps from each stage to the same resolution of the input and apply a hierarchical fusion operation to fuse the upsampled feature maps for the final prediction.

(xuejin:Need a paragraph to discuss recent semantic segmentation networks.)

(xuejin:Also cite our accv paper and describe the relationship/difference of this journal paper with it.)

III. Hierarchically Fused Fully Convolutional Network

In this section, we introduce our hierarchical feature fusion architecture, and apply it to the common networks, VGG16 Net and ResNet. The overview diagram in Fig. 2 shows where the lateral connections and fusion operations take effect.

In order to leverage the feature pyramid while preserving fine-scale structure and high-level semantics, our network involves one bottom-up pathway for feature extraction in multiple scales, lateral connections at each individual scale, and one tunnel fusion (xuejin:another name? xx fusion?) for the final prediction.

Bottom-up Feature Extraction. The first part is a bottom-up pathway, which produces the hierarchical feature maps with convolutional layers. With the increase of the field of perception, the extracted semantic information is gradually from the lower level to the higher level. Each group of feature maps come from the same feature extractor contribute to the

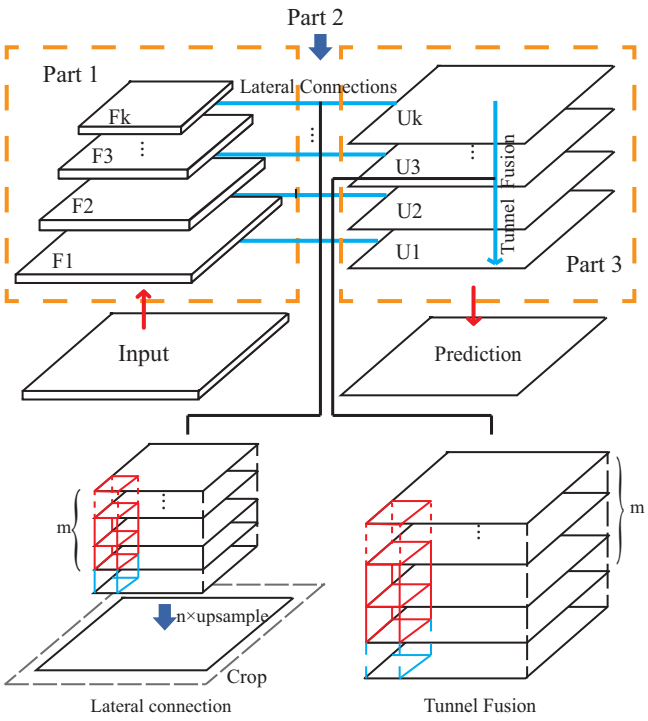


Fig. 2. The first line shows the overview of our network. The second row shows the details of lateral connection and tunnel fusion operation. Fk means the feature maps come from the kth layer. m for number of feature maps. n said the n times of up sampling.

$\{F_k\}, k = \{1, \ldots, K\}$ in Fig. 2, and K is group number of feature maps. $K = 13$ for VGG16 Net that we consider each convolution(conv) layer as a feature extractor. Specifically, for ResNets, we consider a ResBlock as a feature extractor, and $K = 16$.

Lateral Connections. A key component in the second part is a lateral connection which fuse the feature maps in the same level and map them to the finest resolution for the final prediction on the input image. A lateral connection consists of three steps: a 1×1 conv layer, a deconvolutional layer, and a cropping operation.

The 1×1 conv operation fuses the feature maps in the same group as:

$$Y(i,j) = \sum_{m=1}^{M} w_m X_m(i,j), \tag{1}$$

where M is the number of the feature maps in $\{X\}$. The w_m is the weight of conv kernels, which should be learned during the training process.

The output of $Conv\{F_k\}$ are then upsmapled by a transposed convolution to map the fused feature map to the resolution same as the input image. Contrary to the conv operation, the transposed conv is a process of mapping the semantics in the low-resolution feature maps to the original image resolution, which will be later combined to make the prediction directly on the finest scale. The deconvolution kernels of different groups are learned separately.

The $Crop(\{X\})$ operation is a center-aligned cropping which cuts the superfluous boundary of the upsampled feature maps. Therefore, the final output, $\{U_k\}_{k=\{1,\ldots,K\}}$, of each lateral connection in different level is a feature map in the same resolution with the input image.

Fusion for Prediction. The third part is a fusion stage which aims to fuse all the upsampled feature maps from all the lateral connections for the final prediction. This fusion operation plays a role of feature weighting. Using a 1×1 conv layer, a set of parameters w_k, $k : \Omega \to \{1, \ldots, K\}$ are learned to combine the hierarchical feature maps.

Finally, we use a sigmoid function to compute the pixel-wise probability of being the building class.

Network Training. The ground truth of a pixel in our dataset is labeled by 0 or 1 to indicate whether it belongs to a roof or not. When a remote sensing image X is fed into the network, the output is a prediction probability map $P(X;W)$ of roof, where W denotes all the parameters in our HF-FCN. We use the sigmoid cross-entropy loss function to penalize each pixel on the prediction map as:

$$L(W) = -\frac{1}{|I|}\sum_{i=1}^{|I|}[\tilde{g}_i \log P(X_i;W) + (1-\tilde{g}_i)\log(1-P(X_i;W))] \tag{2}$$

where $\tilde{g}_i$ is ground-truth label of X_i, and $|I|$ is the number of pixels in the input image X.

IV. Experiments

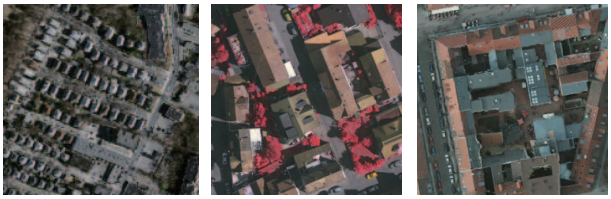
As a generic feature extraction and fusion solution, our method can be combined with any ConvNets. To verify the effectiveness of the proposed HF-FCN in the building extraction task, extensive experiments have been conducted on three remote sensing datasets. In this section, the experimental setups are described including details of datasets, training settings, and the evaluation criteria.

A. Datasets

We test our method on three datasets: Massachusetts dataset, Vaihingen dataset, and Potsdam dataset. These three datasets differ from each other on the building style, density, data channels, and spacial resolution.

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. The images are composed of red, green and blue channels. This dataset is built by Mnih [9] while the ground-truth is produced by Saito et al. [10]. The 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 by sliding a 256×256 window with 64 stride from right to left, top to bottom. The detailed description is shown in Table I.

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 consists of near infra-red, red, green imagery with corresponding digital surface models (DSMs). The dataset is divided into the training set, validation



(a) (b) (c)

Fig. 3. Sample patches on the three datasets. (a) Massachusetts dataset, (b) Vaihingen dataset, (c) Potsdam dataset.

TABLE I
COMPOSITIONS OF DATASETS

	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
Training patch size	256×256	256×256	256×256
Validation images	4	3	4
validation patches	2500	28376	25000
Validation patch size	256×256	256×256	256×256
Test images	10	2	3

set, and test set, which have 11 images, 2 images, and 3 images respectively. The same crop operations are done as the Massachusetts dataset.

c) Potsdam dataset: In the 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, the spacial resolution of the original image is reduced from 6000×6000 to 1500×1500. Each image in this dataset contains 5-channel information: infra-red, red, green, blue and DSM. We split the dataset into training, validation and test sets in a proportion of $7:2:1$.

Data augmentation including data rotation and mirror flipping is applied to each dataset, respectively. The detailed compositions of the three datasets are listed in Table I. Fig. 3 shows some patch samples of the three datasets. The characteristics and challenges of the three datasets are:

- The Massachusetts dataset has crowed buildings, which causes great difficulties for the separation of buildings.
- There is an obvious shadow occlusion in the Vaihingen dataset which may lead to a wrong segmentation in the part of the shadowed rooftop.
- Large intra class differences and small inter class differences are presented in the Potsdam dataset. One building consists of several materials while the roads and buildings appear in very similar colors.

B. Training Settings

We first train the proposed HF-FCN on the Massachusetts dataset due to its large amounts of training data. The ConvNets in bottom-up path are pre-trained on the ImageNet classification set and then fine-tuned on remote sensing dataset.(xuejin:on which dataset? ImageNet?) 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. When the HF-FCN converges on the dataset a), we use the trained model to initialize the other datasets. (xuejin:What do you mean by transfer? Fine-tuning? The input data has different number of channels.) All experiments in this paper are conducted using Caffe and trained on a single NVIDIA Titan 12GB GPU. The hyper-parameters for each dataset are listed in Table II.

TABLE II
PARAMETERS FOR NETWORK TRAINING

	Massachusetts	Vaihigen	Potsdam
mini-batch size	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

Several evaluation metrics are adopted in our work. For the Massachusetts dataset, the most common metrics are precision and recall. The standard (ρ=0) and relaxed (ρ=3) precision and recall scores are used to evaluate the prediction results. Here the relaxed precision means that the fraction of detected pixels are within ρ pixels of a true pixel while the relaxed recall is the fraction of true pixels are within ρ pixels of a detected pixel. (xuejin:Check how other paper explain this.) Moreover, the time cost is used to measure the efficiency of our HF-FCN. For the Vaihingen and Potsdam datasets, besides of the correctness and completeness, we use the F_1 score as an additional evaluation metric (xuejin:for what reason?)which is common used on these two datasets.

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}. \tag{3}$$

V. RESULTS AND DISCUSSION

In this section, we compare the proposed method on the three datasets with two categories of methods, including recent shallow-CNNs algorithms, such as Minh-CNN [9], Satio-multi [10] and Alshehhi-GAP [11], and deep-learning based approaches, including FCN [18], SegNet [19], Deeplab [21], U-Net [29] and FPN [30].

Moreover, for the HF-FCN itself, we expect to investigate which kind of information extracted from feature extractors and how the effects of extracted information on the final prediction. Thus, some up-sampled feature maps $\{U_k\}$ are presented. And several variants of HF-FCN which combine different up-sampling feature maps from lateral connections are proposed. For dataset b) and c), multichannel information is provided. Hence, to explore the impact of other information on the results, a simple experiment is presented which concatenate different kinds of information as input of HF-FCN. In addition, different feature extractor networks are tried as our bottom-up feature extraction network, including VGG16 Net

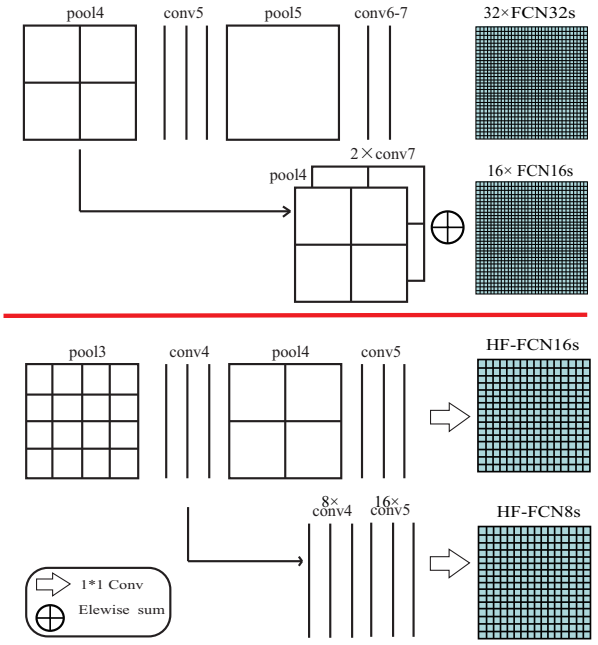


Fig. 4. FCN and HF-FCN variants. The feature maps generated from final group are fused into a coarse result, which is HF-FCN16s. The variant called HF-FCN8s concatenates the feature maps from the last 2 groups with the same fusion operation, and so on.

and ResNet. The results and discussion of above mentioned experiments are shown below.

A. Massachusetts Dataset

On the Massachusetts dataset, both the shallow-CNNs algorithms and deep-learning based approaches are compared with our method. Table III presents the quantitative analysis results with the standard and relaxed precisions, and recall as our evaluation criteria. Our method shows obvious superiority in terms of speed and precision. Compared with Satio-multi-MA&CIS [10], the best among the shallow-CNNs algorithms, the VGG version of our HF-FCN achieves 5.5% and 1.3% higher standard and relaxed recall. Meanwhile, our HF-FCN reduces the time cost from 67.84s to 1.07s, and gets a 63× speedup. These significant improvements demonstrate that HF-FCN achieves better performance in both effectiveness and efficiency.

While comparing our HF-FCN with recent semantic segmentation methods, as shown in Table III and Fig. 8, we get about 3× speedup without sacrificing the accuracy, compared with the most state-of-the art method U-Net [29]. From the visual results, our method preserves the details and integrity of the building better than others.

In addition, to explore which kinds of information extracted by hierarchical fusion operation in lateral connections. Some upsampled feature maps $\{U_1, U_2, U_3, U_7, U_{10}, U_{13}\}$ are shown in Fig. 5. The U1_1 (U_1) in Fig. 5(b) means the upsampled feature map from F1_1 (F_1) which are feature maps generated from conv1_1 in VGG16 Net. Due to small receptive field of conv1_1 and conv1_2, they extract low-level features like edges. And the U1_2 (U_2) looks like an over-segmentation which groups pixels with similar color or texture into a subregion. With the deepening of the network, in the U2_1 (U_3), as Fig. 5(d) shows, shape information is augmented. And from the U3_3 (U_7), we can see that regions with significantly varying appearance are merged into an integrated building by considering high-level features. In U4_3 (U_{10}) and U5_3 (U_{13}), more semantic information of rooftop is got, which can distinguish the rooftop and the roads with similar color and deal with the problem caused by shadow. The final prediction results are shown in Fig. 5(h).

TABLE III
CORRECTNESS AT BREAKEVEN OF HF-FCN V.S. [9] [10] [11] [18] [19] [29] [21] [30]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	–	–
FCN_4s [18]	0.839	0.6147	4.20
SegNet [19]	0.7710	0.5675	2.39
FPN [30]	0.9504	0.7662	3.78
U-Net [29]	**0.9638**	**0.8357**	3.17
DeepLab_V2 [21]	0.9620	0.7575	**1.89**
HF-FCN(VGG16 Net)	0.9643	**0.8424**	1.07
HF-FCN(Ele_Sum)	0.9639	0.840	1.15
HF-FCN(VGG+data aug)	**0.9650**	0.8357	1.38
HF-FCN(ResNet)	0.9588	0.8175	2.42
HF-FCN16s	0.9330	0.7233	0.85
HF-FCN8s	0.9643	0.8171	0.93
HF-FCN4s	0.9632	0.8394	**0.99**

Secondly, to explore the effects of the feature maps generated from each feature extract stage $\{F_k\}$ on the final result, variants of HF-FCN which are counterpart of FCN are designed. Fig. 4 shows the contrast diagram of variants of FCN and HF-FCN. Unlike FCN, a fusion operation rather than summation is leveraged to build our HF-FCN 16s, 8s and 4s. The precision-recall(PR) curves, prediction results and quantitative results of HF-FCN variants are shown in Fig. 6, Fig. 7 and Table III respectively. (xuejin:Figure 8, Figure 9 and Table V). From the diagrams, we can get the following conclusions easily:

- The prediction result obtained from the last layer gets a coarse result, which loses much of location information that are mainly encoded in the shallow feature maps.
- 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.
- The PR curves of HF-FCN4s and HF-FCN almost coincide. It illustrates the low-level information has little effect on the prediction results.
- 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 feature maps contained useful hierarchical information that is critical to the final prediction.

In the end, we want to prove that our lateral connections and tunnel fusion operations learn the connections between feature maps. The connection weights of F1_1, F4_1 and tunnel

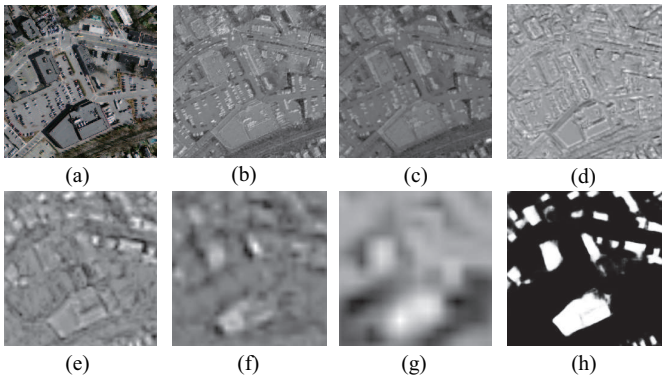


Fig. 5. (a) Input aerial image. (b-g) Feature maps of U1_1, U1_2, U2_2, U3_3, U4_3, U5_3, respectively. (h) Predicted label map. All the images are normalized to the range of $0 - 255$.

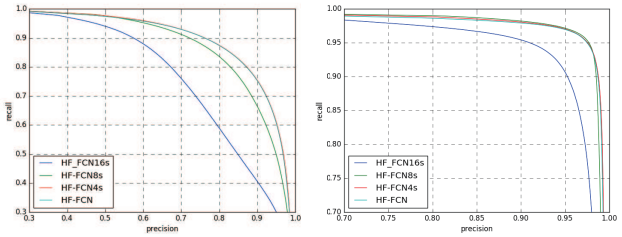


Fig. 6. The relaxed PR curves from HF-FCN variants with two slack parameters. The slack parameter ρ on the left is 0. And ρ = 3 on the right.

fusion are shown in Fig. 9. The weights are not the same, which means that fusion operations have effect on feature combination. From the Fig. 9(a) to Fig. 9(c), the range of weights increases gradually. And from the Fig. 9(c), we can arrive at the conclusion that the different layers have virous effects on the final result. For example, the U1_1 has little effect on the prediction while the U3_2 and U4_3 play more important roles on the final prediction. It also in accordance with our experimental results that middle layers provide more information. In addition, we do an extra experimentation combining the feature maps by element-wise addition. The prediction results are shown in Table III. It is 0.2% lower than our best result. (xuejin:Weight for what? to fuse feature map? The distribution does not make too much sense.)

B. Vaihingen dataset

On Vaihingen dataset, three experiments are undertaken to explore the effects of different inputs, diverse variants and various methods. Three kinds of combinations of image channels are utilized as inputs. The inputs of the 3 channels are IR, R, G and adding the nDSM(normalized Digital Surface Model) as the forth channel. Based on it, DSM is added and made up 5-channel input. Three standards are used to make a more comprehensive evaluation. The evaluation results are shown in Table IV, which illustrate that 3-channel input performed better than the others. The Rec and Pre in Table IV means the recall and precision of prediction results. And F1 indicates the F1 score of results. The number in bold shows the best results of our methods and other methods. Visual results of our methods with different kinds of input are shown in Fig. 12.

(xuejin:Do you compare with others?)

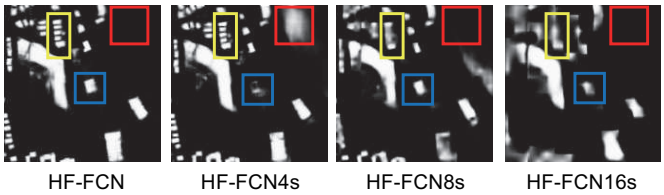


Fig. 7. Prediction results of HF-FCN, HF-FCN4s, HF-FCN8s and HF-FCN16s. The yellow box shows the continuous refinement of the tiny buildings. The red and blue boxes show the mutual promotion and contradiction between different layers.

TABLE IV
PERFORMANCE COMPARISON OF THE RESULTS OF DIFFERENT INPUTS AND METHODS ON VAIHINGEN DATA SET.

	Pre	Rec	F1
FCN_4s [18]	0.871	0.884	0.878
SegNet [19]	0.917	0.861	0.887
U-Net [29]	0.848	0.737	0.789
FPN [30]	0.907	**0.907**	**0.907**
DeepLab_V2 [21]	**0.926**	0.881	0.903
HF-FCN_3in	**0.919**	**0.930**	**0.925**
HF-FCN_4in	0.907	0.872	0.888
HF-FC_5in	0.858	0.900	0.878
HF-FCN16s	0.886	0.854	0.870
HF-FCN8s	0.911	0.864	0.887
HF-FCN4s	0.910	0.861	0.885

Many experiments are done to compare with other methods, including methods using the same dataset and the other deep-learning-based methods. The detail comparison results using the same dataset are shown in Fig. 13. From a visual perspective, our method gets a much more refined roof region, both on continuity of labels and integrity of structural. The quantitative results of deep learning methods are shown in Table IV. Compared to FPN [30], the F1 score of our method is 1.8% higher.

The results of diverse variants are shown in Fig. 10. The HF-FCN_1 in Fig. 10 indicates that the last conv layer in tunnel fusion does not use the previous trained model to initialize. And HF-FCN means that the whole layers use the pre-trained model to initialize. From the curves, the performance of HF-FCN exceeds the variants and gets a excellent result. Additionally, using the pre-trained weights of tunnel fusion has a significance in the final results.

C. Potsdam dataset

The same experiments are implemented on Potsdam dataset, including effects of different channels of input, comparing with other methods and results of diverse variants of HF-FCN. Firstly, we utilize nDSM and IR information as extra inputs based on the RGB input. The specific quantitative evaluation and intuitive visual prediction results are shown in Table V and Fig. 14. In the validation process, the 4-channel input including RGB, nDSM gets better overall performance. Meanwhile, the 5-channel input including RGB, nDSM and IR seems perform

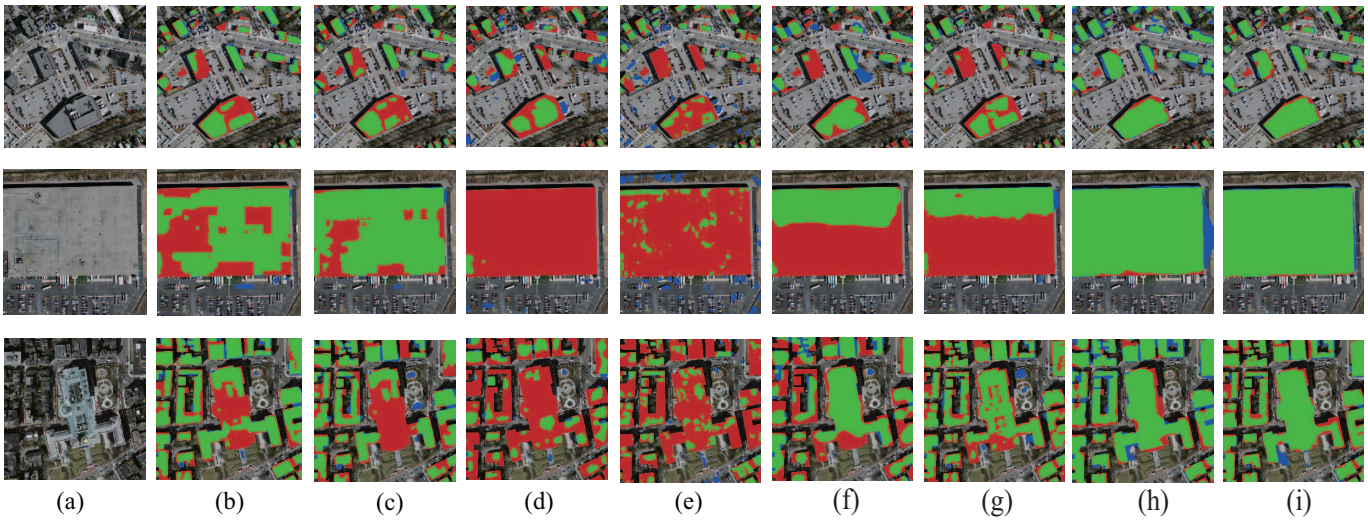


Fig. 8. From left to right: (a) input images, results of (b) Mnih-CNN+CRF, (c) SatiomultiMA&CIS, (d) FCN4s, (e) SegNet, (f) DeepLab_V2, (g) U-Net, (h) FPN, and (i) Our HF-FCN. We show the true positives in green, false positives in blue, and false negatives in red.

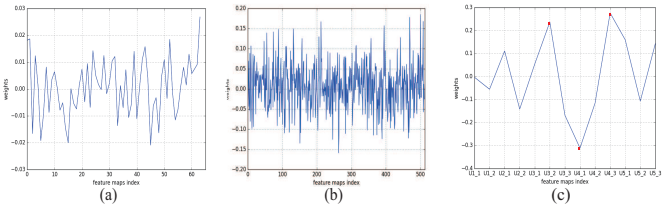


Fig. 9. (a) is weights learned by F1_1, (b) is weights learned by F4_1, (c) is weights learned by Part 3.

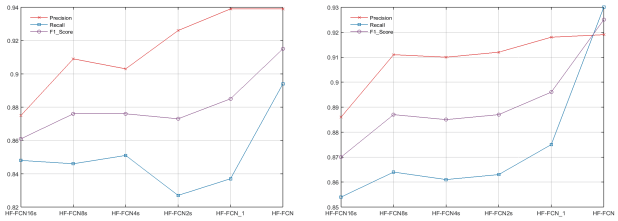


Fig. 10. Results of HF-FCN variants on Vaihingen dataset. (a) (b) shows the precision, recall and F1 score of validation set and test set of Vaihingen dataset respectively.(xuejin:Bigger font)

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 3-channel input to 5-channel input, the F1 score increases from 0.879 to 0.891 on the validation set and increases 0.031 on the test set. It indicates that the other information of geographical

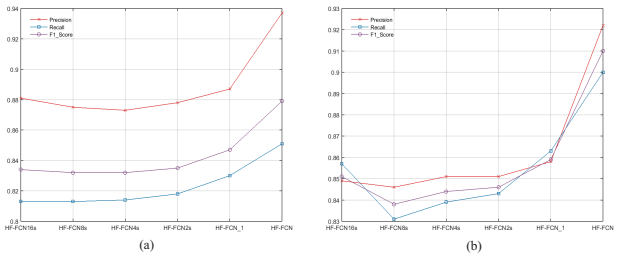


Fig. 11. Results of HF-FCN variants on Potsdam dataset. (a) (b) shows the precision, recall and F1 score of validation set and test set of Potsdam dataset respectively.

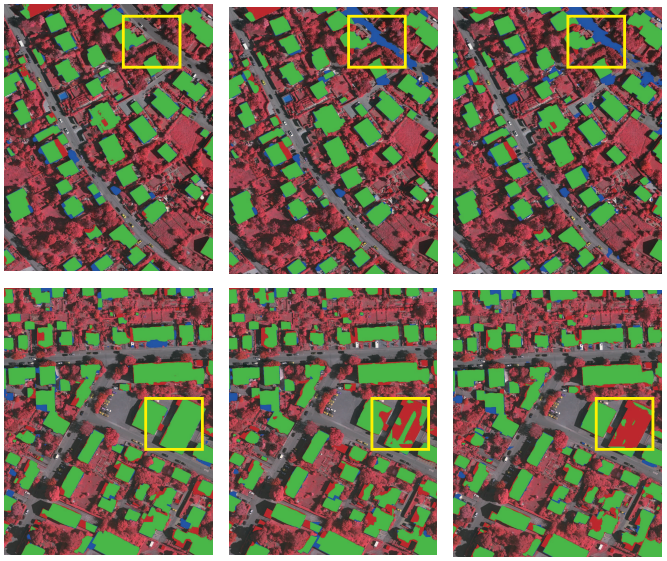


Fig. 12. Prediction results on Vaihingen dataset. (a) (b) (c) shows results of the 3-channel input, 4-channel input and 5-channel input of Vaihingen dataset respectively. Here, TP are shown in green, FP are shown in blue and FN are in red.

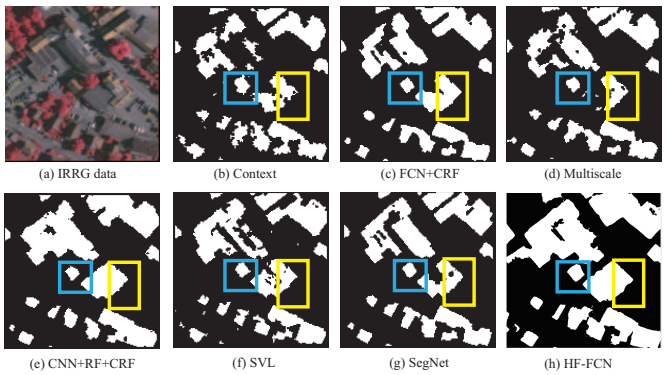


Fig. 13. Results of different methods. (a) is input image, (b)(d)(g) are results of [15], (c) is result of [31], (f) is result of [32], (g) is our result. The blue and yellow frames show some details between these methods.

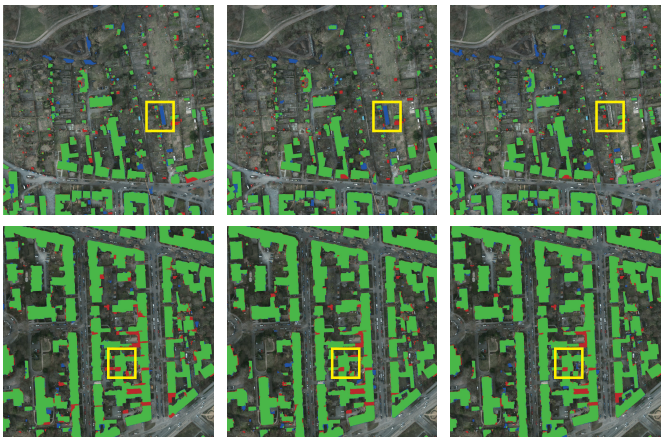


Fig. 14. Prediction results on potsdam dataset. (a) (b) (c) shows results of the 3-channel input, 4-channel input and 5-channel input of Vaihingen dataset respectively. Here, TP are shown in green, FP are shown in blue and FN are in red.

features has a certain effect on the final result.

We compare HF-FCN with other methods using the Potsdam dataset and several deep learning methods. Some qualitative results of methods using Potsdam datast are shown in Fig. 15. From the figure, we can easily see that HF-FCN got more remarkable segmentation results. And edges and structure of buildings are preserved better. The results of deep learning methods are shown in Table V. From the Table, the HF-FCN achieves the best result. And the F1 score far higher than the others.

As done on Vaihingen dataset, contrast experiments of HF-FCN variants are implemented. The performance curve of HF-FCN variants are shown in Fig. 11. The HF-FCN_1 in Fig. 11 indicates that the last conv layer in tunnel fusion does not use the previous trained model to initialize. And HF-FCN means that the whole layers use the pre-trained model to initialize. Initialization of parameters has a greater promotion on the final results.

TABLE V
PERFORMANCE COMPARISON OF THE RESULTS OF DIFFERENT INPUTS AND METHODS ON POTSDAM DATA SET

	Pre	Rec	F1
FCN_4s [18]	0.827	0.774	0.796
SegNet [19]	0.648	0.773	0.687
U-Net [29]	**0.924**	0.705	0.799
FPN [30]	0.873	0.868	0.870
DeepLab_V2 [21]	0.901	**0.876**	**0.887**
HF-FCN_3in	0.922	0.900	0.910
HF-FCN_4in	0.937	0.935	0.936
HF-FCN_5in	**0.940**	**0.943**	**0.941**
HF-FCN16s	0.849	0.857	0.851
HF-FCN8s	0.846	0.831	0.838
HF-FCN4s	0.851	0.839	0.844

VI. APPLICATION

The segmentation results are further used to 3D building reconstruction. We make use of the depth map and generate the point cloud of remote sensing images. After that, the 3D building reconstruction methods could applied to the generated point cloud. In this paper, the approach proposed by zhou [33] are used to generate the 3D models of buildings in the scene. Fig. 16 and Fig. 17 show the 3D models of Vaihingen and Potsdam dataset respectively. The details of a single building are also presented. From the figures, we can see that 3D models preserve the characteristics of buildings well whether the structure of the roof or the simplification of details.

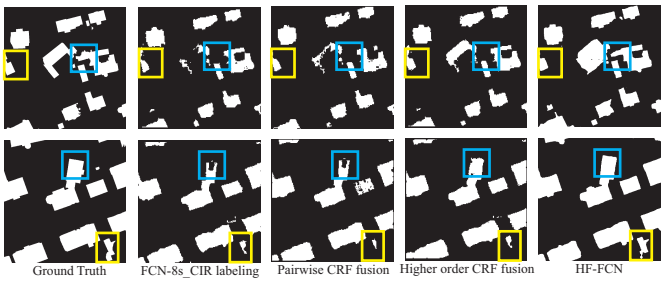


Fig. 15. Results of different methods. The second column is the results of using only the FCN with CIR(color-infrared image). Pairwise CRF fusion shows the result of fusing FCN-8s_CIR with LiDAR data in a pairwise CRF. The results of using higher-order CRF [14] as post processing are shown in third column. The last colunm shows our results.

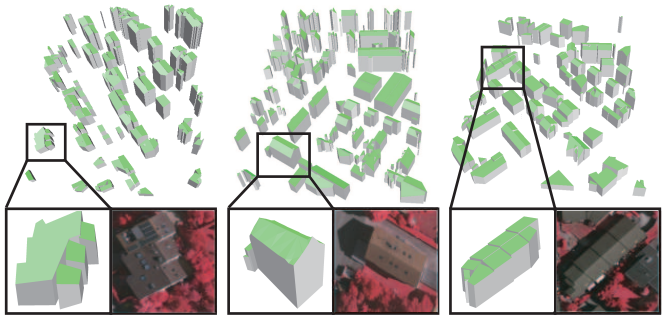


Fig. 16. The 3D modelling of Veihingen dataset. The single building model and its corresponding optical patch were shown together.

VII. CONCLUSION

In this paper, an efficient building detection approach is proposed and has a further application in building reconstruction. Using proposed feature fusion operations, a novel CNN architecture is presented for building extraction, named HF-FCN. Unlike previous shallow-CNNs algorithms, we provide an end-to-end network for building extraction. And it is robust to the different scales of buildings and efficient for a large-scale remote sensing images. On the other hand, distinct from the previous deep learning based methods, we utilize the

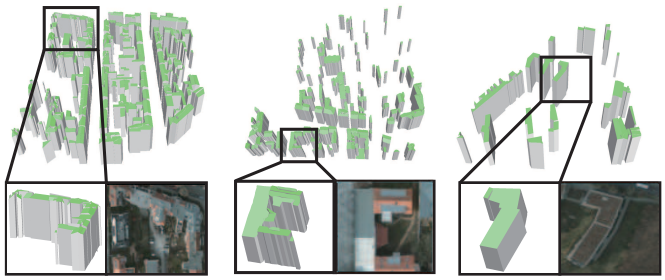


Fig. 17. The 3D modelling of Potsdam dataset. The single building model and its corresponding optical patch were shown together.

multi-scale inherent information within the CNN and refine the details by lateral connections and a tunnel fusion. In addition, an application of 3D building reconstruction depend on the segmentation results is implemented. 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 building extraction problem by 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] 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.

[19] 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.

[20] 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.

[21] 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," *arXiv preprint arXiv:1606.00915*, 2016.

[22] R. Vemulapalli, O. Tuzel, M.-Y. Liu, and R. Chellapa, "Gaussian conditional random field network for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3224–3233.

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

[24] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[25] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.

[26] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning." in *AAAI*, 2017, pp. 4278–4284.

[27] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 5987–5995.

[28] T. Zuo, J. Feng, and X. Chen, "Hf-fcn: Hierarchically fused fully convolutional network for robust building extraction," in *Asian Conference on Computer Vision*. Springer, 2016, pp. 291–302.

[29] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[30] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *CVPR*, vol. 1, no. 2, 2017, p. 4.

[31] 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.

[32] M. Gerke, "Use of the stair vision library within the isprs 2d semantic labeling benchmark (vaihingen)," 01 2015.

[33] 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.