Fully Convolutional Networks for Semantic Segmentation of Very High Resolution Remotely Sensed Images Combined With DSM

Weiwei Sun and Ruisheng Wang

***Abstract*— Recently, approaches based on fully convolutional networks (FCN) have achieved state-of-the-art performance in the semantic segmentation of very high resolution (VHR) remotely sensed images. One central issue in this method is the loss of detailed information due to downsampling operations in FCN. To solve this problem, we introduce the maximum fusion strategy that effectively combines semantic information from deep layers and detailed information from shallow layers. Furthermore, this letter develops a powerful backend to enhance the result of FCN by leveraging the digital surface model, which provides height information for VHR images. The proposed semantic segmentation scheme has achieved an overall accuracy of 90.6% on the ISPRS Vaihingen benchmark.**

***Index Terms*— Fully convolutional networks (FCN), deep learning, semantic segmentation, remote sensing, very high resolution (VHR).**

I. Introduction

VERY high resolution (VHR) remotely sensed images, whose ground sampling distance is smaller than 30 cm, is an important sort of data for a wide range of applications such as mapping [1] and earth observation [2]. Semantic segmentation aims to assign pixels in VHR image with category labels [3], [4], which is an important but unsolved problem in remote sensing [5]. To date, numerous different methods for semantic segmentation of VHR images have been proposed, including random forest [6], support vector machine [2], and probabilistic graphical model [7]. Among these approaches, handcrafted features [1], [7] are used to feed classifiers.

More recently, deep convolutional neural networks (CNN) [8], [9] have achieved great success in the field of computer vision [10]. For semantic segmentation, FCN proposed in [3], which harnesses the strength of end-to-end styles, has become a powerful and promising scheme. By now, FCN-based schemes have been the top performers on social media images in PASCAL-VOC Challenge [11] and VHR images in ISPRS Vaihingen benchmark [12]. Nevertheless, due to downsampling operations such as pooling layers, blurry object boundaries, which are undesirable for semantic segmentation [5], [13], are very common in the result of FCN. Long *et al.* [3] introduced that FCN confronted an inherent tension between "what" from deep layers and "where" from shallow layer. For clarification, we define "what" and "where" as semantics and details, respectively.

In the framework of FCN, deeper layers contain more semantics but fewer details. In order to recover the detailed information washed out by downsampling operations, Long *et al.* [3] proposed the average fusion strategy (AFS) where deep layers resolving semantics and shallow layers resolving details are averagely fused. However, in the classification process of pixels, the performance of semantics and details is dependent on area types. For instance, semantics are more appropriate than details for homogeneous areas like interiors of objects. Meantime, details outperform semantics for heterogeneous areas like object contours. Therefore, it is unreasonable to consider an equal contribution by semantics and details in semantic segmentation. In view of this, we propose a new fusion strategy based on the max operation that extracts maximums between deep layers and shallow layers. Within the gating mechanism, the proposed maximum fusion strategy (MFS), which aims to obtain more useful information from semantics and details for semantic segmentation, shows improved performance.

The idea of leveraging geometry information to enhance the interpretation of color images is increasingly popular in the computer vision community [14]. A few researchers reported that the classification result could be significantly improved by using the digital surface model (DSM) (height information) [7]. Our observation is that FCN purely relying on color images tends to make mistakes in black areas like shadows, which are ubiquitous in remotely sensed images. Height information provided by DSM can correct some of these mistakes. Sherrah [5] utilized DSM as an input channel of FCN. Since DSM and images have different statistics, Marmanis *et al.* [4], [13] processed color images and DSM in two separate networks. In this letter, we propose a DSM backend based on morphological operations to extract complementary information from both color images and DSM. The experiments show that the proposed method achieves improved results.

Manuscript received August 23, 2017; revised December 4, 2017 and January 5, 2018; accepted January 15, 2018. Date of publication February 5, 2018; date of current version February 23, 2018. This work was supported by the Natural Sciences and Engineering Research Council of Canada. *(Corresponding author: Ruisheng Wang.)*

W. Sun is with the Department of Geomatics Engineering, University of Calgary, Calgary, AB T2N 1N4, Canada (e-mail: weiwei.sun@ucalgary.ca).

R. Wang is with the Department of Geomatics Engineering, University of Calgary, Calgary, AB T2N 1N4, Canada, and also with the Beijing Advanced Innovation Center for Imaging Technology, Capital Normal University, Beijing 100048, China (e-mail: ruiswang@ucalgary.ca).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LGRS.2018.2795531

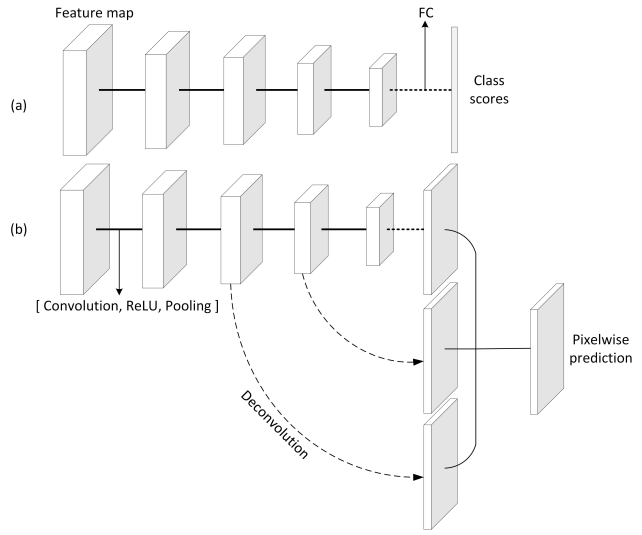

Fig. 1. Architectures of CNN and FCN. (a) Framework of CNN. With FC layers, CNN outputs a vector representing class scores. (b) Framework of FCN with fusion strategy. With deconvolution layers, FCN outputs per-class probability maps from units.

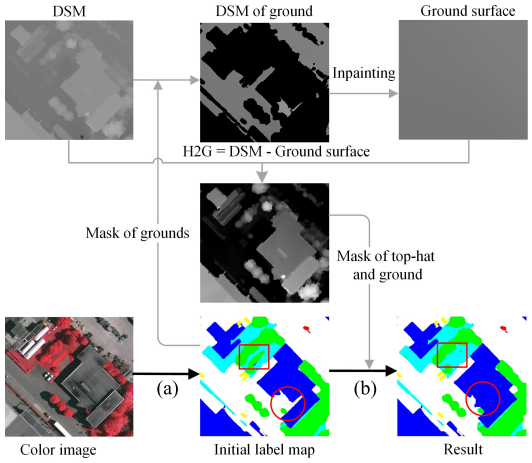

Fig. 2. Proposed scheme for semantic segmentation of VHR remotely sensed images combined with DSM. (a) FCN utilizes color images to compute initial label maps (white: impervious surfaces, blue: buildings, cyan: low vegetation, green: trees, and yellow: cars). (b) DSM backend refines the result of FCN by removing false top-hats (shown with red rectangle) and false grounds (shown with red circle) of initial label maps.

The main contribution of this letter is that we establish a promising semantic segmentation scheme for VHR remotely sensed images combined with DSM. The proposed scheme consists of FCN and DSM backend, which has achieved superior results on the ISPRS Vaihingen benchmark [12].

II. Background

A. Convolutional Neural Network

At present, CNN is one of the most successful methods in computer vision. The main layers that make up CNN are convolution, nonlinear activation function, and pooling. Each neuron in convolution layer inputs values from previous layers in a fixed-size and spatially localized window W, and outputs a weighted sum of the input values. Convolution layers are followed by nonlinear activation function like rectified linear unit (ReLU) [8]. Pooling layers perform statistic operations such as Max and Mean along spatial dimensions to downsample feature maps. By doing so, pooling layers can extract contextual information from a larger spatial context.

Popular CNN architectures such as AlexNet [8], VGGNet [9], and ResNet [15] are common in the number of downsampling operations like pooling. Fig. 1(a) shows that CNN includes five units that are sequentially downsampled by pooling layer. The layers between two units are stacked by convolution, ReLU, and pooling. To obtain the label of input image, FC layers shown in Fig. 1(a) convert feature maps into a vector representing class scores.

B. Fully Convolutional Network

To effectively utilize the end-to-end style of deep learning, Long *et al.* [3] converted CNN for imagewise classification to FCN that aims at pixelwise semantic segmentation. In the framework of FCN, FC layers are replaced by a convolution layer with 1×1 size kernel. To recover the size of feature maps, deconvolution layers that most commonly adopt bilinear interpolation are used to reverse the downsampling process. In fact, FCN for semantic image segmentation is a particular case of CNN, which is trained from end to end and pixel to pixel. Fig. 1(b) shows that FCN with a fusion strategy computes score maps among per-class probability maps interpolated from different-level units.

III. Methods

Our major contribution is the scheme for semantic segmentation of VHR remotely sensed images combined with DSM. As shown in Fig. 2, the proposed scheme contains two parts: 1) FCN with the proposed MFS is utilized to compute initial label maps from color images and 2) with initial label maps and DSM, DSM backend is proposed to correct mistakes in the result of FCN. The DSM backend extracts ground surface from DSM according to initial label maps from FCN. Ground surface is used to refine the result of FCN. In this way, spectral information from color images and geometry information from DSM are effectively fused to improve semantic segmentation of VHR remotely sensed images.

A. Maximum Fusion Strategy

For pixel-to-pixel fusion, bilinear interpolation [3] recovers score maps from units to the same size of the input image. Interpolation of shallow units with a high resolution requires small strides and thus can keep relatively fine boundary information. Thus, scores from deep units contain more semantics but fewer details than scores from shallow units. The goal of fusion strategies is to combine semantics and details by fusing class scores evaluated from different units. It is noted that the performance of semantics and details is dependent on the area types. Thus, we propose MFS to selectively choose the semantics and details for different areas. In the forward process, the final score of class (c) is computed as

$$\text{score}^c = \text{Max}\left(\left[\text{score}_3^c, \text{score}_4^c, \text{score}_5^c\right]\right) \quad (1)$$

where score_i^c is computed from ith unit (i.e., 3, 4, 5) for c class. Max is a gating operation where merely the best score (i.e., the highest score) is allowed to flow into the final score. Within the gating mechanism, the best score representing the most appropriate score for a pixel is chosen among scores from different units. Compared with AFS where scores from different units are averagely fused, the advantage of MFS is that inappropriate scores and even noise of scores can be filtered out from the final score in the gating mechanism.

With the inclusion of increasingly shallow units, experimental results of both MFS and AFS achieve notable improvement for the third and fourth units. The fourth unit performs better than the third unit. However, for the first and second units whose scores contain many details but few semantics which are important for the classification of pixels, the performance of MFS and AFS is reduced. Therefore, fusion strategies do not consider the first and second units. Experiments show that MFS achieves better performance than AFS in various fusion styles.

In the backward process, we utilize the gradient descent method within the backpropagation [8] to optimize the loss function. To output the best scores from appropriate units, parameters in different units are updated according to the gradient of the loss function (L). In the framework of backpropagation, local gradients of MFS with respect to an input score is computed as

$$\frac{d\text{MFS}}{d\text{score}_i} = 1\left\{i = \arg\max_{j=3,4,5} \text{score}_j\right\} \tag{2}$$

where $1\{\cdot\}$ is the indicator function in which $1\{\text{a true statement}\} = 1$, and $1\{\text{a false statement}\} = 0$. Within the chain rule, local gradients of MFS (0 or 1) are multiplied by the local gradients of parameters in units. We utilize caffe [16] to implement the calculation of gated gradients. For each iteration, gated gradients do not update parameters in units with "0" gradients. By doing so, gated gradients merely update parameters of units that contribute the best scores.

B. DSM Backend

We observe and analyze representative mistakes in the result of FCN. They motivate the proposal of the DSM backend to refine the results. We denote top-hat as the objects (e.g., buildings and trees) with height (i.e., 3 m) above the surrounding ground (e.g., impervious surface and low vegetation). DSM backend ignores cars because DSM is not accurate. Inspecting the initial label maps of FCN-based methods, we summarize two common failures as illustrated in Fig. 2: 1) False ground which should be top-hat shown with the red circle and 2) false top-hat which should be ground shown with the red rectangle. Therefore, in this letter, we utilize DSM which can distinguish between top-hat and ground to correct these mistakes.

The DSM backend completes three tasks: 1) Calculates each pixel's height to the surrounding ground (H2G) image according to DSM and the corresponding initial label map computed by FCN; 2) removes false top-hats according to H2G images; and 3) removes false ground according to H2G images. Therefore, we propose three algorithmic strategies based on morphological operations as follows.

1) *H2G Image:* First, based on the fact that FCN with MFS achieves overall accuracy (OA) of more than 90%, we extract the mask of large ground objects (i.e., grass and road). Second, we compute ground surface, which is the DSM after removing nonground areas. Inpainting method [17], [18] is applied to recover the height of nonground areas from DSM of grounds. Third, we obtain H2G image by subtracting the ground surface from DSM.
2) *False Ground:* With H2G images, we compute the mask of false grounds in initial label map. By encoding initial label map with greater number for top-hats but smaller number for ground, we perform dilation operation on initial label map to replace false grounds by surrounding top-hats in a fixed-size window. Before the calculation of false ground masks, we perform erosion operation with a small-size kernel (5×5) on H2G images to avoid mistakes of false ground masks on roof edges where DSM is not accurate. By doing so, the proposed algorithms are robust to noises such as jagged roof edges in DSM.
3) *False Top-Hat:* First, we perform dilation operation on H2G images. Second, masks of false top-hat are calculated with H2G images. Third, we perform erosion operation on encoded label map to remove false top-hat.

IV. Experiments

In this section, we investigate the effectiveness of the proposed scheme for semantic segmentation of VHR remotely sensed images. All networks are trained and tested with caffe [16] on NVIDIA GTX Titan-X GPUs.

A. Training Details

1) Data Set: We test proposed methods on the ISPRS Vaihingen benchmark [12]. This data set contains 33 images with associated DSM. 16 ground-truth images are provided for training. We use one of them as validation set and the remaining images as training models. The resolution of images is 9 cm which make small details visible. The ground truth contains five categories, including impervious surfaces, buildings, low vegetation, trees, and cars.

2) Data Set Preprocessing: For the limited memory of GPU, we split images of average size 2000×2000 into smaller patches of size 321×321, using a sliding window. To reduce overfitting on data sets, we employ two forms of data augmentation: 1) using overlapped window (sliding stride 160 pixel) to crop patches from images and corresponding ground-truth images and 2) rotating patches with $90°\,180°\,270°$. Thus, we have 8920 patches of the same size 321×321 for training networks.

3) Pretrained Networks: We converted traditional CNN for image classification into FCN by replacing FC with convolution layer of 1×1. Besides, 1×1 convolution layer is followed by the pyramid pooling module [11] that has shown excellent performance for semantics extraction. The feature maps generated by pyramid pooling are then fed to the deconvolution layer. In this letter, a residual neural network

TABLE I
RESULTS EVALUATED ON ISPRS VAIHINGEN BENCHMARK. PER-CLASS F1-SCORE (%) AND OA (%) ARE LISTED

Methods	Impervious surface	buiding	low vegetation	tree	car	Overall accuracy
FCN	91.6	93.4	80.0	87.9	85.1	88.6
FCN_AFS	92.0	94.7	82.6	89.4	85.6	89.9
FCN_MFS	92.2	94.7	83.1	89.5	86.4	90.1
FCN_MFS_DSMInput	89.6	92.1	79.5	87.1	70.8	87.2
FCN_MFS_DSMNet	92.0	94.8	83.0	89.2	85.9	89.9
FCN_AFS_DSMBackend	92.0	95.5	83.0	89.5	85.6	90.2
FCN_MFS_DSMBackend	**92.3**	**95.8**	**83.8**	**89.6**	**86.4**	**90.6**

TABLE II
ACCURACY OF SCORES ON THE VALIDATION SET

	5th unit	4th unit	3rd unit	Final
AFS	84.8	83.69	70.61	87.14
MFS	78.71	53.41	29.38	87.98

of 101 layers (ResNet101) [15] is the baseline for FCN. Fine-tuning is a practical strategy to train deep networks on a small data set like ISPRS Vahingen Benchmark [3]. Thus, we utilize ResNet-based models fully trained on the ImageNet data set [19], which contains millions of images.

4) Implementation Details: We use fixed learning rate policy where the rate equals the constant 10^{-5}. Momentum and weight decay are set at 0.9 and 5^{-4}. Doubled learning rates are set for biases. The number of iterations is 300 000.

B. Results

To predict labels of image, we apply FCN on patches cropped from images using a sliding window of size 321 × 321. Overlap is imposed on patches sampling when sliding window across image with stride of 80 pixels. We average predictions of overlapped area to avoid boundary effect [5] of patch sampling. The results are evaluated on the ISPRS Vaihingen test set using OA and F1 score [12].

We submit the results of FCN-based models to the Vahingen official test set. Table I shows per-class F1 score (%) and overall accuracy (%) of the submitted results. On the benchmark, the proposed scheme achieves an overall accuracy of 90.6%, which is among the top performers. In the following, we demonstrate the effectiveness of MFS and backend proposed in this letter.

1) MFS: As shown in Table II, scores from units individually make pixelwise predictions through softmax function and obtain reasonable results. Table II illustrates that MFS achieves lower accuracy in scores directly from different units, but better results in fused score than AFS. Scores fused in MFS contain more noise than scores fused in AFS. We conclude that MFS outperforms AFS in filtering out noise from different scores. Since AFS averagely fuses scores from different units, all the scores are equally optimized. Thus, AFS performs better than MFS in the scores directly from units. We initialize MFS and AFS with parameters of FCN without fusion strategy. Therefore, scores from the fifth units are well optimized and thus have smaller training loss than scores from shallow units. Since the best scores are largely from fifth units, as illustrated in Fig. 3, MFS which merely needs to optimize the best scores from shallower units converges faster than AFS, which needs to optimize all the scores.

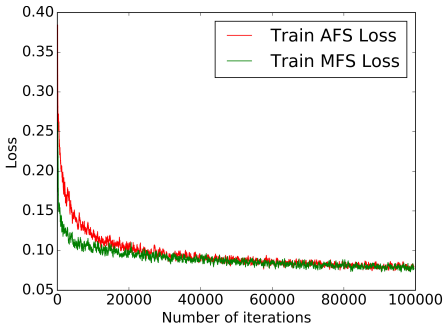

Fig. 3. Training loss of MFS and AFS. We initialize MFS and AFS with parameters of FCN without fusion strategy.

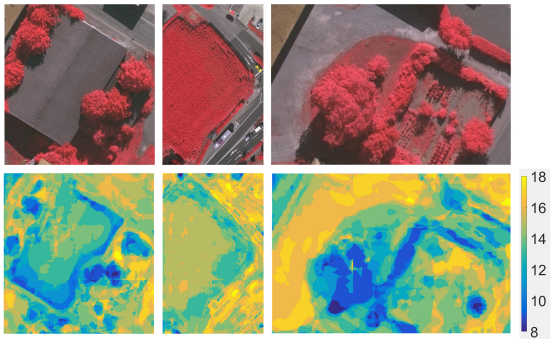

Fig. 4. Visual close-ups of distribution maps over different units contributing the best scores: Every pixel has six class scores (five for categories and one for clutter) from different units. We encode source units contributing the best scores with digits 1 (third unit), 2 (fourth unit), and 3 (fifth unit). The value of each pixel is the sum of six digits representing the units giving six best scores. With the encoding, greater values of pixels indicate deeper units contributing the best scores.

Fig. 4 shows the potential of MFS to selectively choose the semantics and details for different areas. It demonstrates that best scores of homogeneous areas (e.g., interiors of big objects) are from deeper units. When details are required by heterogeneous areas (e.g., small objects and contours of big objects), the best scores are mainly from shallow units.

Fig. 5 illustrates the visual prediction of FCN, FCN_AFS, and FCN_MFS. Both AFS and MFS significantly boost the performance of FCN by fusing the details and semantics. However, in homogeneous areas, AFS misclassifies some pixels in the interior of building or grass. Inspecting the

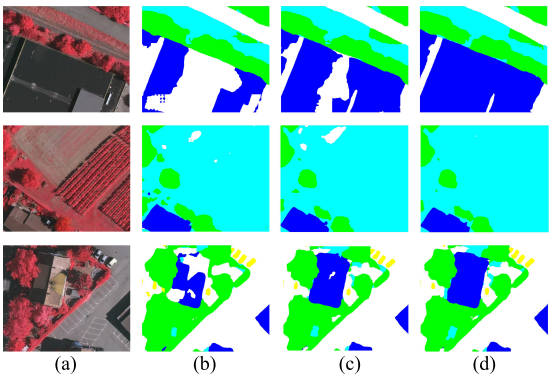

Fig. 5. Visual close-ups of FCN results on test set (white: impervious surfaces, blue: buildings, cyan: low vegetation, green: trees, yellow: cars). (a) Input image patch. (b) FCN without fusion strategy. (c) FCN with AFS. (d) FCN with MFS.

images, we observe that the interior of buildings and grass have a similar color as impervious surfaces. In this case, details from shallow layers degrade the performance of FCN. In MFS, more semantics is considered for homogeneous areas. Fig. 5 illustrates that MFS outperforms AFS in homogeneous areas.

As shown in Table I, the proposed MFS achieves better improvement for FCN than AFS on the benchmark.

2) DSM Backend: We demonstrate the effectiveness of the proposed DSM backend to enhance the result of FCN. Different improvements by utilizing DSM backend between FCN_AFS and FCN_MFS show that the accuracy of initial label map has a notable impact on the performance of the DSM backend. Thus, the baseline to test the DSM backend is FCN_MFS. To make use of DSM, Sherrah [5] utilize DSM as an input channel of FCN (DSMInput). The second method that trains an individual network (DSMNet) to process DSM is proposed in [4] and [13]. Due to the limited memory, we train a small FCN with a powerful pyramid pooling module to process DSM. The outputs of two FCNs processing DSM and color images are fused using MFS at deep layers.

In this letter, we test DSMInput, DSMNet, and the proposed DSM backend on the benchmark. Table I shows that the DSM backend achieves significant improvement in accuracy, while DSMInput and DSMNet both decrease the performance of merely color image, where DSMNet performs better that DSMInput. DSM, where objects with similar heights show a smooth pattern provides little texture information while CNNs are designed to learn deep features from texture information in the spatial domain. Thus, DSM reduces the performance of FCN. For the DSMNet, we observe that almost all outputs of DSMNet are smaller than the outputs of FCN for processing color images and thus have little effect on the maximum fusion process. Therefore, DSMNet outperforms DSMInput.

V. Discussion and Conclusion

In this letter, we propose a promising scheme based on FCN for semantic segmentation of VHR remotely sensed images. To more reasonably combine the semantics and details in FCN, we propose MFS, which performs better than the existing AFS. However, the edges (e.g., building edges) in the results of FCN with fusion strategy are not sharp enough. In the future, we need to preserve more edge information by combining object-based classification method and deep features from FCN [1]. In this scheme, the DSM backend imposed on the result of FCN is developed to obtain complementary information from both the images and DSM. However, the performance of the DSM backend is dependent on the accuracy of initial label map from FCN. Thus, in the future, we would like to improve the performance of using only color image by leveraging DSM within end-to-end networks.

References

[1] W. Zhao, S. Du, and W. J. Emery, "Object-based convolutional neural network for high-resolution imagery classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 7, pp. 3386–3396, Jul. 2017.

[2] M. Volpi, D. Tuia, F. Bovolo, M. Kanevski, and L. Bruzzone, "Supervised change detection in VHR images using contextual information and support vector machines," *Int. J. Appl. Earth Observat. Geoinf.*, vol. 20, pp. 77–85, Feb. 2013.

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

[4] 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 Ann. Photogram., Remote Sens. Spatial Inf. Sci.*, vol. 3, pp. 473–480, Jul. 2016.

[5] J. Sherrah. "Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery," unpublished paper, 2016. [Online]. Available: https://arxiv.org/abs/1606.02585

[6] A. Puissant, S. Rougier, and A. Stumpf, "object-oriented mapping of urban trees using random forest classifiers," *Int. J. Appl. Earth Observat. Geoinf.*, vol. 26, pp. 235–245, Feb. 2014.

[7] R. Qin and W. Fang, "A hierarchical building detection method for very high resolution remotely sensed images combined with DSM using graph cut optimization," *Photogramm. Eng. Remote Sens.*, vol. 80, no. 9, pp. 873–883, Sep. 2014.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

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

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[11] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 6230–6239, doi: 10.1109/CVPR.2017.660.

[12] M. Gerke, "Use of the stair vision library within the ISPRS 2D semantic labeling benchmark (Vaihingen)," Dept. Earth Observ. Sci., Univ. Twente, Enschede, The Netherlands, Tech. Rep., 2015.

[13] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla. "Classification with an edge: Improving semantic image segmentation with boundary detection," unpublished paper, 2016. [Online]. Available: https://arxiv.org/abs/1612.01337

[14] C. Couprie, C. Farabet, L. Najman, and Y. LeCun. "Indoor semantic segmentation using depth information," unpublished paper, 2013. [Online]. Available: https://arxiv.org/abs/1301.3572

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

[16] Y. Jia *et al.* "Caffe: Convolutional architecture for fast feature embedding," unpublished paper, 2014. [Online]. Available: https://arxiv.org/abs/1408.5093

[17] D. Garcia, "Robust smoothing of gridded data in one and higher dimensions with missing values," *Comput. Statist. Data Anal.*, vol. 54, no. 4, pp. 1167–1178, Apr. 2010.

[18] G. Wang, D. Garcia, Y. Liu, R. de Jeu, and A. J. Dolman, "A three-dimensional gap filling method for large geophysical datasets: Application to global satellite soil moisture observations," *Environ. Model. Softw.*, vol. 30, pp. 139–142, Apr. 2012.

[19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 248–255.