An aerial image segmentation approach based on enhanced multi-scale convolutional neural network

Xiang Li†, Yuchen Jiang†, Hu Peng†, Shen Yin†
†School of Astronautics, Harbin Institute of Technology, Harbin, China, 150001
Email:lixianghit@yeah.net, yc.jiang2016@foxmail.com, penghu0522@foxmail.com, shen.yin2011@googlemail.com

***Abstract*—Aerial images are the images captured from high attitudes above the ground. Processing and analyzing aerial images play central roles in terrain modeling, agricultural monitoring, city planning, environmental surveillance, etc. Aerial images are developing towards high resolution and large size, which poses a major challenge in pixel-level image segmentation. With the rapid development of deep learning technology, the application of deep learning to image semantic segmentation has obtained satisfactory effect. In this paper, we propose a novel aerial image segmentation method based on convolutional neural network (CNN). The main structure of the proposed network adopts U-Net. In order to capture objects of different scales in the deep features, a group of cascaded dilated convolution is inserted at the bottom of U-Net which has different dilation rates. Furthermore, to better optimize the network at different scales, an auxiliary loss function is proposed to be integrated in the cascaded dilated convolution. The effectiveness of the proposed method is evaluated on the Inria Aerial Image Labeling Dataset. Experiment results show that the proposed method has better segmentation performance than existing approaches.**

***Index Terms*—Convolutional neural network, semantic segementation, aerial images, deep learning**

I. INTRODUCTION

Using aerial images to identify objects on the earth's surface has attracted great attention. Previously, only coarse resolution multispectral images could be obtained to apply to this problem. For example, the images of 30m ground sampling distance (GSD) taken by Landsat and the images of 2.2m ground sampling distance taken by Quickbird. With advances in aviation and photography, a large number of aerial remote sensing images are generated every day through satellites. These high-resolution aerial images are with a ground sampling distance of 5-10cm [1], which can be used to clearly identify buildings, cars, trees and other objects. So aerial images contain a lot of surface information. Nowadays, the development of science and technology tend to be intelligent and automated [2] [3]. Automaticly semantic labeling of these surface information(e.g., labeling the building area and non-building area) has been a hot topic in remote sensing. Image labeling is essentially image segmentation at the pixel level, different objects in the segmented image are presented in different ways.

The labeled aerial images can provide a lot of valuable information in many fields. In agriculture and geology, aerial images can be used to calculate land use, forest cover and the distribution of crops. In the field of energy and mining, aerial images can be used for mineral exploration and energy detection. Aerial images can also be used for urban planning, defense, navigation, autonomous driving and other aspects [4]. However it takes expert knowledge and plenty of time to implement image segmentation by manual labeling. With the breakthrough of deep learning, especially convolutional neural network, deep learning has played an indispensable role in image classification, segmentation and detection. Deep learning method for image segmentation is a supervised learning method driven by data [5] [6] [7]. Now convolutional neural networks are able to segment images end to end without manual features.

Image segmentation using convolution neural network mainly involves two processes. Firstly, after a series of convolution layers and pooling layers, the input image is transformed into feature maps whose length and width are both less than the input image while depth is greater than the input image. Feature maps represent the deep semantic feature of input image, for example, feature maps can represent which categories of objects are included in the input image. The second stage is to map the feature maps back to the input image size through multiple up-sampling. At this point, each pixel of the output image represents the category of pixel in the input image. Generally speaking, the two processes can be viewed as encoding and decoding.

In this paper, buildings and non-buildings in aerial images are segmented. It is difficult to segment buildings and non-buildings because the terrain of aerial images are complex [8]. Buildings may be obscured by trees and other objects and buildings in the image are varied in color, shape and reflectivity. As can be seen from Fig. 1, buildings are different in size, it is obviously unreasonable to grasp them with convolution of the same receptive field. So this puts forward very high requirements for the network.

The main contributions can be summarized in three aspects:

1) A U-Net is constructed as the main network, and the bottom convolution layer of U-Net is replaced by a set of cascaded dilated convolution with different dilation rates. Each cascaded dilated convolution corresponds to a specific receptive field.

2) In order to overcome gradient disappearance and to better optimize the network, it is proposed to add an auxiliary loss function after the cascaded dilated convolution.

3) Performance of the proposed method is compared with several existing methods, which illustrates superiority of the proposed approach and can serve as a reference for further

978-1-5386-8500-6/19/$31.00 ©2019 IEEE

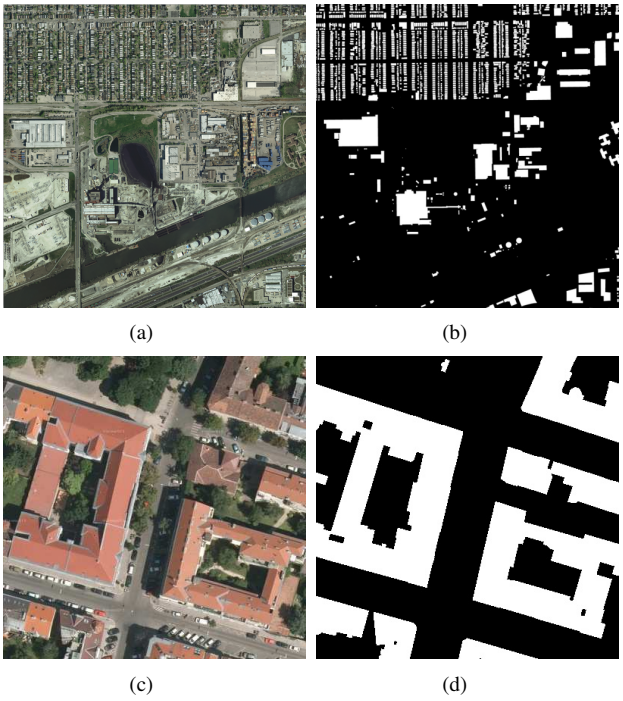


Fig. 1. Examples of the Inria Aerial Image Labeling Dataset, (a) raw image (global view), (b) groundtruth corresponds to the image from the global view, (c) raw image (local view), (d) groundtruth corresponds to the image from the local view

researches.

II. RELATED WORK

Semantic segmentation of images is a very important field in computer vision, which refers to mark the category of each pixel in image. Before deep learning was applied to semantic segmentation, people mainly use traditional machine learning methods. Jake Porway et al. [9] proposed a hierarchical contextual model based on the statistical framework. The model learns the hierarchy of scenes and objects to handle the representation of scenes at different scales, and adds context constraints to resolve ambiguities in scene interpretation. There are a lot of works make use of the contextual model [10] [11] [12]. Wang et al. [13] used scale invariant feature transformation to extract the corner points which is considered as a sign of the building. Rau et al. [14] built an expert decision system using different kinds of spectral, geometric and topological knowledge. Senaras Caglar et al. [15] extracted a large number of color, texture and shape features from segmented and preprocessed image fragments, then combined the results of multiple classifiers.

The first successful application of deep learning to image semantic segmentation is the full convolution neural network [16]. It replaces the fully connected layer with convolution layer, so that the network can accept images of any size and output segmentation images of the same size as the original. Furthermore, U-Net [17] is the first to adopt encoder-decoder architecture and have a symmetric down-sampling and up-sampling process, which is much more elegant. Different from FCN, U-Net merges the feature maps of up-sampling and down-sampling on the channel dimension. Later semantic segmentation methods have produced many variations based on these two frameworks. The series of deeplab (v1,v2,v3) [18] [19] [20] used dilated convolution, which is essentially adding holes to the convolution kernel. Dilated convolution can make perceptive field larger while keeping the same computation as the ordinary convolution. Deeplab v3+ [21] and pyramid scene parsing network [22] introduced global pyramid pooling, global pyramid pooling scales the feature maps to several different sizes so that the feature has better global and multi-scale information.

Specific to semantic segmentation of aerial images, aerial images have high resolution and large size. Yuan [23] constructed a fully convolutional neural network to predict the distance between pixels and boundaries, and used a large amount of architectural footprint data provided by geographic information system (GIS) to train the network. Huang et al. [24] used a convolutional neural network with two branches, the input of one branch being the original RGB image and the input of another branch is pan-sharpened images. Bischke et al. [4] proposed a cascaded multi-task loss along with using a deeper network architecture, which adds a loss function specifically for learn per pixel information about the location of the boundary and capture implicitly geometric properties. Paisitkriangkrai et al. [25] used the combination of CNN features and manual features to conduct aerial image segmentation, and used CRF as the back end to smooth the image. Sun et al. [1] focused on the diversity, FCN ensemble learning strategy is proposed. Sherrah [26] used a non-downsampling approach and a FCN was introduced that preserves the full input image resolution at every layer. Marmanis et al. [27] in order to use both image and DEM data as input, a delayed fusion method is adopted to use two parallel processing chains with the same structure in the network.

III. PROPOSED METHOD

A. Architecture of the multi-scale network

The proposed multi-scale network architecture is shown in Fig. 2. It is mainly composed of three parts: encoding, decoding and cascaded dilated convolution. The encoding process consists of 3×3 convolution layer and 2×2 maximum pooling layer, and the maximum pooling operation is carried out once for every three convolution operations. There are three down-sampling layers in the whole encoding process, after each down-sampling, the depth of the feature maps will be deeper, and the length and width will be half of the previous stage due to the down-sampling. The shallow network extracts some detailed features, such as lines, angles, points and so on. As the network deepens the extracted features are more abstract to express specific semantic information.

The decoding process is exactly opposite to the encoding process and symmetric, these features go through three up-sampling layers to change the length and width to be the same as input images, because the pixels in the input images are ultimately divided into building and non-building categories,

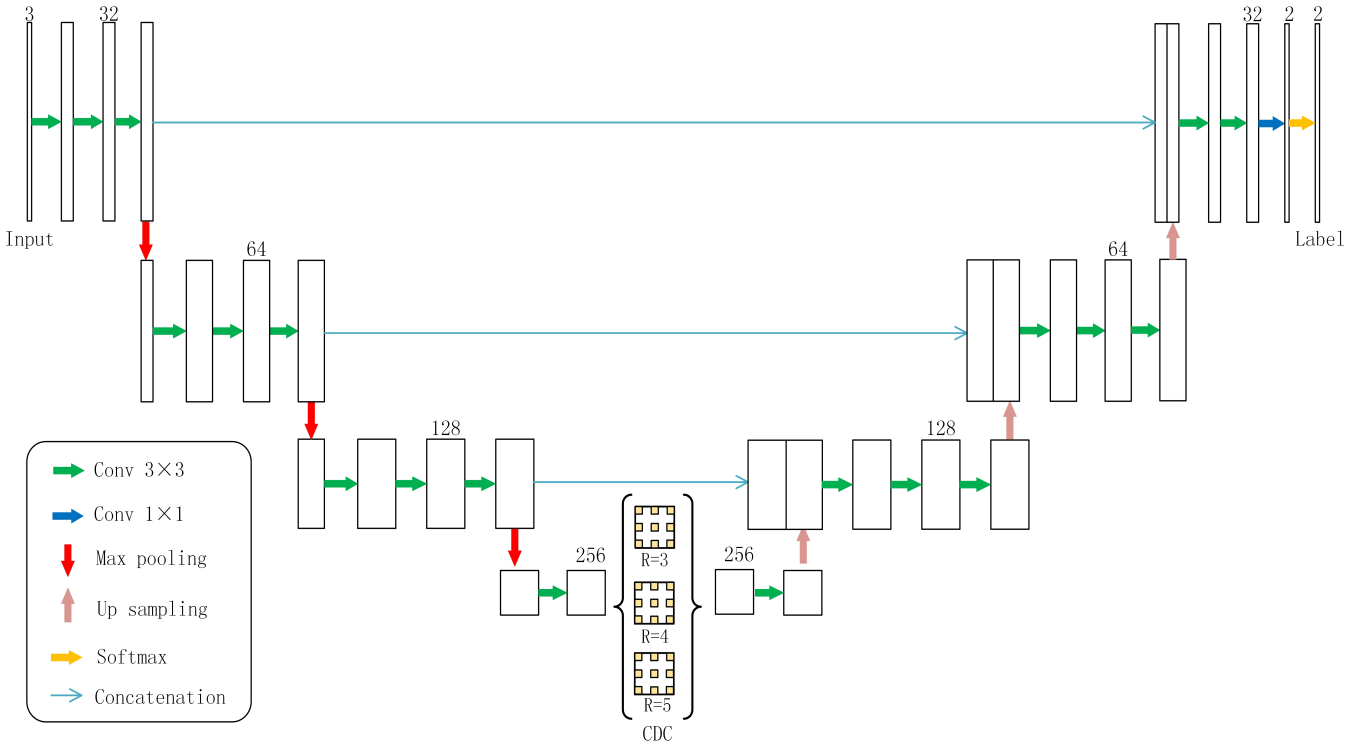


Fig. 2. Structure of the proposed multi-scale convolutional network

then a convolution layer of 1×1 is connected to change the depth to two. Finally, a sigmoid activation function is used to obtain the segmentation probability map, and the probability map is compared with groundtruth to complete the network optimization.

B. Cascaded dilated convolution

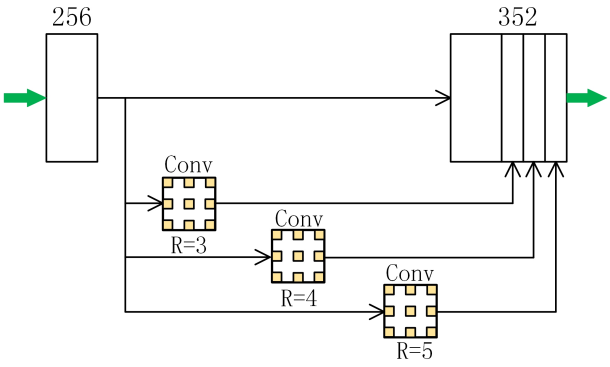


Fig. 3. Cascaded dilated convolution

The cascaded dilated convolution is inspired by Maoke Yang et al. [28]. They proposed to grasp objects of different scales with dilated convolution of different dilation rates and fuse features in a dense cascade. Because the current convolution network mainly adopts the convolution kernel of 3×3, if such convolution is used to extract large buildings, some relevant information will be lost. If increasing size of the convolution kernel, the computation will be greatly increased. Dilated convolution is a method of balancing the computation and the perceptive field. Dilated convolution regularly adds holes in the traditional convolution, and these holes do not participate in calculation, but the perceptive field of convolution increases with the addition of holes.

In the cascaded dilated convolution module, three convolution layers with the dilation rates of 3, 4 and 5 are used, and each convolution layer has 32 convolution kernels. The feature fusion uses the cascaded approach as shown in Fig. 3. So the number of channels after feature fusion is 352, and finally a 3×3 ordinary convolution is used to connect the decoding process.

C. Add auxiliary function to different scale features

The loss function of the proposed neural network is composed of two parts. The main part is the loss function of final output. The fusion features of cascaded dilated convolution module go through up-sampling and 1×1 convolution layer to obtain the auxiliary output, auxiliary loss function is constructed by auxiliary output. Total loss function can be described as:

$$L(W, w, w^{(i)}) = \alpha l_{fuse}(W, w^{(m)}) + \beta l(W, w^{(a)}) \quad (1)$$

Because the main part takes advantage of more networks, the main part should have more impact on the network. On the contrary, the auxiliary loss function only play a supplementary role, which can make the network converge faster and optimize

TABLE I
PERFORMANCE COMPARISON OF SEGMENTATION METHODS ON THE INRIA AERIAL IMAGE LABELING DATASET

Methods \ Sub-datasets		Austin	Chicago	Kitsap Co.	West Tyrol	Vienna	The whole dataset
Maggiori et al. [29]	IoU	61.20	61.30	51.50	57.95	72.13	64.67
FCN+MLP	Acc.	94.20	90.43	98.92	96.66	91.87	94.42
Bischke et al. [25]	IoU	76.49	66.77	72.69	66.35	76.25	72.57
SegNet(Single-Loss)	Acc.	93.12	99.24	97.79	91.58	96.55	95.66
Bischke et al. [25]	IoU	76.76	67.06	**73.30**	66.91	76.68	73.00
SegNet+MultiTask-Loss	Acc.	93.21	**99.25**	97.84	91.71	**96.61**	95.73
Khalel et al. [30]	IoU	**77.29**	68.52	72.84	75.38	78.72	**74.55**
2-levels U-Nets	Acc.	**96.69**	92.40	99.25	98.11	93.79	96.05
The proposed multi-scale network	IoU	73.09	**70.38**	72.45	**76.40**	**78.88**	74.24
	Acc.	96.43	92.92	**99.43**	**98.12**	93.98	**96.12**

IoU = Intersection over Union (see definition in Eq. (3)). Acc = Accuracy.

the first half of the network better. So in order to balance the relationship between the two, a coefficient is multiplied in front of each part. For the specific form of loss function, cross entropy loss function is adopted for each part:

$$\begin{aligned} L(W,w) = &\sum_{j=1}^{|X_+|} y_j log P(y_j = 1|X;W,w) \\ &- \sum_{j=1}^{|X_-|} (1-y_j) log P(y_j = 0|X;W,w) \end{aligned} \tag{2}$$

where$|X_+|$and$|X_-|$ denote the number of building and non-building pixels in the input images X, y_j represents the true label of pixel j. $P(y_j = 1|X;W,w)$ is the probability value that pixel j belongs to the building obtained through the network, so $P(y_j = 0|X;W,w)$ is the probability value that pixel j belongs to non-building.

IV. EXPERIMENT

A. Datasets

The proposed method is evaluated on the publicly available aerial image databases: Inria Aerial Image Labeling Dataset [29].

The training set of aerial image data set contains 180 orthogonal aerial RGB images, and the training set covers a total area of 405 square kilometers. The size of each image is 5000×5000 with a very small spatial resolution, only 0.3m. The pixels in groundtruth are labeled as two semantic classes: building (pixel value 255) and non-building (pixel value 0). The whole training set is divided into five subsets, corresponding to five cities such as Chicago, and each subset contains 36 images. Because the ground truth of the test set is not published, it can only be used to submit tests online. To facilitate comparison, we split the data set according to the description of Maggiori et al. [29], images 1 to 5 of each subset for validation, images 6 to 36 for training.

B. Experimental setup and technical details

Due to the large size of the original images, we randomly crop patches of size 200×200 from original images. At the same time, in order to enhance the robustness of the model and reduce overfitting, random flip in vertical and horizontal directions are applied to the patches. The batchsize is set to 16, and 50 epochs are performed. The optimizer uses Adam [31], learning rate initialized to 0.001, the strategy of linear attenuation is adopted, every epochs decrease once, and learning rate finally decrease to 0.0001. The coefficient used to balance the loss function: α is set to 1 , β is set to 0.2.

The multi-scale CNN experiment is implemented with Keras [32] and is performed on a computer equipped with two Intel E5 2678 CPUs and four NVIDIA GTX 1080Ti graphics cards. The whole training takes about thirty nine hours.

C. Evaluation metrics

By comparing the segmentation result with grundtruth, two evaluation metrics can be obtained to evaluate our method. The first evaluation metric is the Intersection over Union (IoU) for buildings. IoU is a concept used in target detection, which is the ratio of intersection and union between the building area in prediction and the building area in grundtruth. Perfect overlap is the ideal case, the ratio is 1, IoU can be defined as:

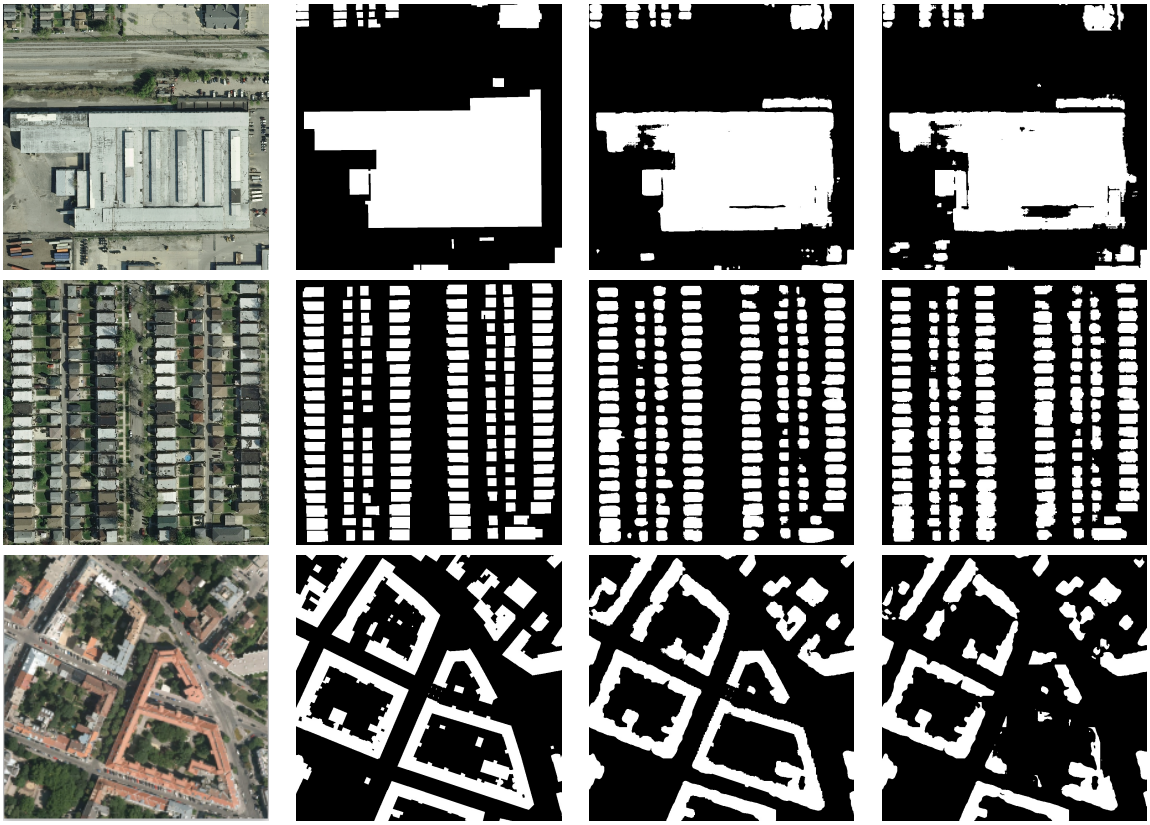
$$IoU = \frac{GT_{building} \cap P_{building}}{GT_{building} \cup P_{building}} \tag{3}$$

where $GT_{building}$ is the region of the building in grundtruth and $P_{building}$ is the region of the building in prediction.

The second evaluation index is Accuracy (Acc), which is the percentage of correctly predicted pixel.

D. Results

Fig. 4 shows the segmentation results of three different types of buildings. From left to right are the original images, the groundtruth, the main output segmentation graph, and the auxiliary output segmentation graph. The first row displays the segmentation result for large buildings, which are generally easier to recognize, but the boundary and middle parts may be misaligned. The second row shows the segmentation results of the community buildings. The buildings in the community are smaller and denser. The algorithm can segment most of the



(a) Input (b) Groudtruth (c) Main output (d) Auxiliary output

Fig. 4. Aerial image segmentation results illustration using the proposed approach

buildings and only part of them are fused together. The third row shows the segmentation results of irregular buildings. It can be seen that the algorithm can still segment the contour of irregular buildings, but only the bulges on some boundaries are lost. From the visual effect, our algorithm has achieved good results and can be applied to various types of buildings.

Table I shows the specific performance measurements. In addition to the accuracy rate and the IoU of entire data set, the performance of each subset is counted. Compared with other people's results, we get the best accuracy, increasing the accuracy from 96.05 to 96.12. From the performance of each subset, better accuracy rate and IoU in most subsets are achieved. For example, in West Tyrol subsets, the accuracy rate is 98.12 and the IoU is 76.40, both of which are the best. Because the images of each subset are very different, most of the Chicago subset is urban buildings, and most of the Kitsap Co subset is countryside and forest, the algorithm cannot perform well on all subsets.

V. CONCLUSION

In this paper, an improved deep convolution network for aerial image segmentation of buildings from the background is proposed. It is proposed to replace the bottom layer of the U-Net with a cascaded dilated convolution module. The main purpose of the cascaded dilated convolution module is to obtain multi-scale features. At the same time, the auxiliary output and auxiliary loss function are obtained by directly up sampling the multi-scale features. From the input to the main output, the whole network architecture is U-Net, and from the input to the auxiliary output, the network architecture can be considered as FCN. From the aspect of design and training, a major advantage of the proposed approach is that it does not involve manual features and does not require specific pre-processing or post-processing, which can reduce the influence of subjective factors. Finally, the quantitive results (in TABLE I) and the visual effects (as shown in Fig. 4) illustrate that the overall performance of the proposed approach outweighs four existing approaches in image segmentation accuracy, and that it is applicable to all types of the building images.

REFERENCES

[1] X. Sun, S. Shen, X. Lin, and Z. Hu, "Semantic labeling of high resolution aerial images using an ensemble of fully convolutional networks," *Journal of Applied Remote Sensing*, vol. 11, no. 4, 2017.

[2] Y. Jiang and S. Yin, "Recursive total principle component regression based fault detection and its application to vehicular cyber-physical systems," *IEEE Transactions on Industrial Informatics*, vol. 14, pp. 1415–1423, April 2018.

[3] H. Yang, Y. Jiang, and S. Yin, "Fault-tolerant control of time-delay markov jump systems with$it\hat{o}$stochastic process and output disturbance based on sliding mode observer," *IEEE Transactions on Industrial Informatics*, vol. 14, pp. 5299–5307, Dec 2018.

[4] S. Paisitkriangkrai, J. Sherrah, P. Janney, and A. van den Hengel, "Semantic labeling of aerial and satellite imagery," *IEEE Journal of*

Selected Topics in Applied Earth Observations and Remote Sensing, vol. 9, no. 7, pp. 2868–2881, 2016.

[5] H. Luo, S. Yin, T. Liu, and A. Q. Khan, "A data-driven realization of the control-performance-oriented process monitoring system," *IEEE Transactions on Industrial Electronics*, pp. 1–1, 2019.

[6] H. Luo, H. Zhao, and S. Yin, "Data-driven design of fog-computing-aided process monitoring system for large-scale industrial processes," *IEEE Transactions on Industrial Informatics*, vol. 14, pp. 4631–4641, Oct 2018.

[7] Y. Jiang, S. Yin, and O. Kaynak, "Data-driven monitoring and safety control of industrial cyber-physical systems: Basics and beyond," *IEEE Access*, vol. 6, pp. 47374–47384, 2018.

[8] Q. Zhang, Y. Wang, Q. Liu, X. Liu, and W. Wang, "Cnn based suburban building detection using monocular high resolution google earth images," in *Geoscience and Remote Sensing Symposium (IGARSS), 2016 IEEE International*, pp. 661–664, IEEE, 2016.

[9] J. Porway, K. Wang, B. Yao, and S. C. Zhu, "A hierarchical and contextual model for aerial image understanding," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8, IEEE, 2008.

[10] J. Verbeek and B. Triggs, "Region classification with markov field aspect models," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pp. 1–8, IEEE, 2007.

[11] F. Kor and W. Frstner, "Interpreting terrestrial images of urban scenes using discriminative random fields," 2008.

[12] Y. Tarabalka, M. Fauvel, J. Chanussot, and J. A. Benediktsson, "Svm and mrf-based method for accurate classification of hyperspectral images," *IEEE Geoscience and Remote Sensing Letters*, vol. 7, no. 4, pp. 736–740, 2010.

[13] M. Wang, S. Yuan, and J. Pan, "Building detection in high resolution satellite urban image using segmentation, corner detection combined with adaptive windowed hough transform," in *Geoscience and Remote Sensing Symposium (IGARSS), 2013 IEEE International*, pp. 508–511, IEEE, 2013.

[14] J.-Y. Rau, J.-P. Jhan, and Y.-C. Hsu, "Analysis of oblique aerial images for land cover and point cloud classification in an urban environment," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 3, pp. 1304–1319, 2015.

[15] C. Senaras, M. Ozay, and F. T. Y. Vural, "Building detection with decision fusion," *IEEE journal of selected topics in applied earth observations and remote sensing*, vol. 6, no. 3, pp. 1295–1304, 2013.

[16] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

[17] 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*, pp. 234–241, Springer, 2015.

[18] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *Computer Science*, no. 4, pp. 357–361, 2014.

[19] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

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

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

[22] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," 2016.

[23] J. Yuan, "Automatic building extraction in aerial scenes using convolutional networks," *arXiv preprint arXiv:1602.06564*, 2016.

[24] Z. Huang, G. Cheng, H. Wang, H. Li, L. Shi, and C. Pan, "Building extraction from multi-source remote sensing images via deep deconvolution neural networks," in *Geoscience and Remote Sensing Symposium (IGARSS), 2016 IEEE International*, pp. 1835–1838, IEEE, 2016.

[25] B. Bischke, P. Helber, J. Folz, D. Borth, and A. Dengel, "Multi-task learning for segmentation of building footprints with deep neural networks," *arXiv preprint arXiv:1709.05932*, 2017.

[26] J. Sherrah, "Fully convolutional networks for dense semantic labelling of high- resolution aerial imagery," *arXiv preprint arXiv:1606.02585*, 2016.

[27] 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," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 135, pp. 158–172, 2018.

[28] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "Denseaspp for semantic segmentation in street scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3684–3692, 2018.

[29] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark," in *IEEE International Symposium on Geoscience and Remote Sensing (IGARSS)*, 2017.

[30] A. Khalel and M. El-Saban, "Automatic pixelwise object labeling for aerial imagery using stacked u- nets," *arXiv preprint arXiv:1803.04953*, 2018.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[32] F. Chollet, "Keras https://github. com/fchollet/keras," 2017.