Semantic Segmentation based Building Extraction Method using Multi-source GIS Map Datasets and Satellite Imagery

Weijia Li[1*], Conghui He[2*], Jiarui Fang[2], and Haohuan Fu[1]

[1]Department of Earth System Science, Tsinghua University, China
[2]Department of Computer Science, Tsinghua University, China
{liwj14,hch13,fjr14}@mails.tsinghua.edu.cn
haohuan@tsinghua.edu.cn

Abstract

This paper describes our proposed building extraction method in DeepGlobe - CVPR 2018 Satellite Challenge. We proposed a semantic segmentation and ensemble learning based building extraction method for high resolution satellite images. Several public GIS map datasets were utilized through combining with the multispectral WorldView-3 satellite image datasets for improving the building extraction results. Our proposed method achieves the overall prediction score of 0.701 on the test dataset in DeepGlobe Building Extraction Challenge.

1. Introduction

Building extraction from remote sensing images is a popular research issue with wide attention. Deep convolutional neural networks (DCNNs) based semantic segmentation methods have been used in several building extraction or semantic labeling studies and achieved state-of-the-art performance. For instance, Audebert et al [3] proposed an Fully Convolution Network (FCN) based method for semantic labeling of Earth Observation images. Iglovikov et al [11] proposed a U-Net based method for semantic segmentation of different classes in multispectral satellite imagery.

Many studies investigated data fusion methods to utilize various data sources for building extractions. The LiDAR dataset is one of the most widely used earth observation data in many building extraction studies which provides useful point cloud based 3D features [9]. However, the LiDAR datasets that can be obtained for free are very limited. The OpenStreetMap is a very popular open source Map dataset that have been used in several building extraction or semantic segmentation studies [5, 13]. Although the effectiveness of OpenStreetMap has been proved in many studies, it still cannot provide enough useful information for building extraction in many regions, such as the selected areas of Shanghai, Las Vegas and Khartoum in this challenge [8].

In this paper, we proposed a semantic segmentation and ensemble learning based building extraction method using multiple data sources. We utilized a data fusion method through combining the multispectral satellite image datasets with several public GIS map datasets. We explored the capabilities of several deep learning methods for the building extraction task. The building extraction results were analyzed comprehensively based on the actual situation of each city.

2. Datasets

In this challenge, we use both WorldView-3 satellite datasets and public GIS map datasets for each city, as we found the GIS map datasets provide useful information for building extraction. We select the 8-band multispectral WorldView datasets with pan sharpening (MUL-PanSharpen) as the satellite datasets used for four cities. For each city, we select the most informative map from several public map datasets. We use the OpenStreetMap datasets for Paris and Khartoum. We use the Google Map datasets for Vegas. For Shanghai, there often exists a coordinate shifting between the satellite images and most map datasets (e.g. Google Map and Baidu Map). Moreover, the OpenStreetMap in Shanghai provide few information of buildings. For these reasons, we choose the MapWorld datasets for Shanghai city.

The OpenStreetMap [10] used in this work can be downloaded from its official website. The Google Map datasets

*Equal contribution

and the MapWorld datasets can be dowloaded through their Static Map API respectively [1, 2]. For each satellite image in both training and test datasets, we collect its corresponding map image in the same location as the satellite image according to its geospatial information (e.g. longitude, latitude and spatial resolution). All the map image datasets are resized to 650×650 pixels for further combination with the satellite images.

3. Methods

The overall workflow of our proposed method is illustrated in Figure 1. During training phase, the provided annotation files will be transformed into pixel labels for conducting the supervised learning process. Models based on DCNNs are trained for classifing each pixel of the input image. Our adopted DCNNs based semantic segmentation models combining with the pre-processing and the post-processing techniques are illustrated as follows.

3.1. Data pre-processing

In this work, we proposed a two-level data pre-processing method to mitigate the lack of available training data for the semantic segmentation models training. In the first level, the training and test datasets of each city are pre-processed into two collections of datasets. The first dataset collection consists of the original 8-band multispectral satellite images. As mentioned in section 2, map datasets can provide extra useful information for building extraction. However, it is not reasonable if we merely use the 3-band map datasets to train a separate network, because in some regions the building information in map datasets is missing or does not match the one in the corresponding satellite images. Consequently, the second dataset collection consists of 8-band images, which combines the first five bands of the original satellite images with the three bands of map images.

In the second level, each of the two above dataset collections is further pre-processed into two formats of input image for each semantic segmentation model respectively. First, the 650×650 images are scaled to the size of 256×256 pixels. Second, the 650×650 images are sliced into 3×3 sub-images and all sub-images are used as the input of the network. Consequently, for each city, we finally achieved four groups of pre-processed datasets. The whole training dataset is randomly splitted into two parts: 70% as training data and the remaining 30% as validation data. Before feeding our dataset collections into deep neural network, we also increase the training data by four times through four 90-degree rotations.

3.2. Semantic image segmentation

Two Deep Learning models commonly used in computer vision image segmentation tasks have been investigated for the building extraction task.

U-Net [14] is one of the most successful and popular DCNN architecture for semantic segmentation. It is a good choice for our task because it is designed for biomedical image segmentation, in which the amount of available training data is relatively low. We adopted a variant of U-Net proposed in [11] for DSTL Satellite Imagery Feature Detection challenge run by Kaggle. We modified its input layer to fit the size of our input image (256×256 pixels, with 8 channels), and we modified its output layer to generate output labels in 256×256 pixels. We also add a batch normalization layer [12] after each convolutional layer. Similar with the original U-Net model, we normalized every channel of the inputs and then use per-channel randomized linear transformations for each patch. We monitored the Jaccard coefficient as an indicator for early stopping during training to avoid over-fitting.

We also applied DeepLab[7] model in this competition. We adopted the DeepLabv3+ implementation and hyperparameters from official tensorflow repository repo[1] and finetuned it with our own data based on a pretrained model on VOC2012 datasets. However, we noticed that the evaluation accuracies of DeepLabv3+ are much lower than the accuracies obtained from U-Net. Thus we did not integrate its results in our final submission and left it for our future work.

3.3. Prediction ensembling

After the training and predicting phases of semantic segmentation model, we obtained the probability maps of each image in the test dataset, in which the grayscale value of each pixel indicates the probability that it belongs to the building class. We proposed a hierarchy of two-level ensemble method to combine the prediction results obtained from different models into the final integrated prediction result.

At the first level, we integrated the prediction results obtained from two image pre-processing methods. The probability map obtained from the first method is scaled back to 650×650 pixels, and the probability map of 9 sub-images obtained from the second method are merged into the whole map. The probabilities of each pixel obtained from the two methods are then averaged, resulting in the integrated probability map. At the second level, the integrated probability maps obtained from two dataset collections are further averaged into the final integrated results for post-processing.

3.4. Post-processing

After obtaining the integrated prediction results, we applied two post-processing strategies to optimize the final prediction results. On one hand, we adjusted the probability threshold (for separating buildings from backgrounds) from

[1] https://github.com/tensorflow/models/tree/master/research/deeplab

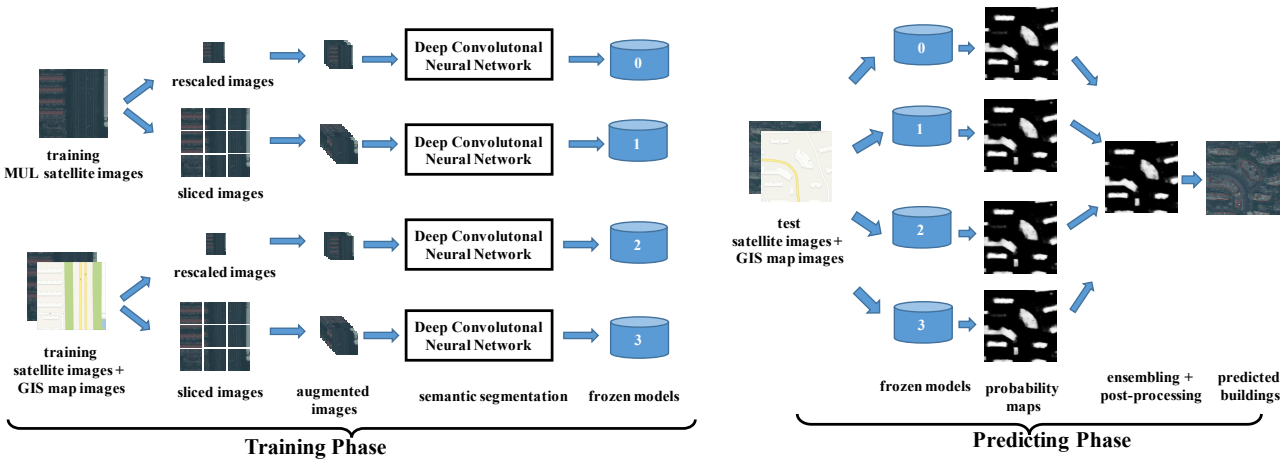

Figure 1. Workflow of our proposed method

0.45 to 0.55 for each city. On the other hand, we adjusted the minimal polygon size threshold (for filtering out noises in the prediction results) from 90 to 240 pixels for each city. The best probability threshold and the best minimal polygon size threshold for the validation dataset will be adopted to the test dataset for each city.

4. Results

4.1. Building extraction results

Table 1 shows the evaluation matrix of the building extraction results of each city in the development phase. Figure 2 shows the F1 scores of the baseline model of our proposed method and those obtained after three optimization strategies. For Vegas and Paris, three strategies bring the improvement of F1 score in similar extend. For Shanghai, the combination of Satellite and Map datasets brings remarkable improvement of F1 score. For Khartoum, the improvement of F1 score benefits greatly from the ensemble and post-processing strategy. In the final phase, the F1 scores of Vegas, Paris, Shanghai and Khartoum are 0.894, 0.740, 0.625 and 0.552, respectively.

Table 1. Evaluation matrix of each city in the development phase, including true positive (TP), false positive (FP), false negative (FN), and F1 score.

	Vegas	Paris	Shanghai	Khartoum
TP	29705	3807	10792	3503
FP	1736	523	3502	960
FN	5928	2030	9859	4690
F1 Score	0.886	0.749	0.618	0.554

4.2. Results analysis and discussion

Shanghai has the second lowest F1 Score among the four cities. Compared with the other three cities, Shanghai has

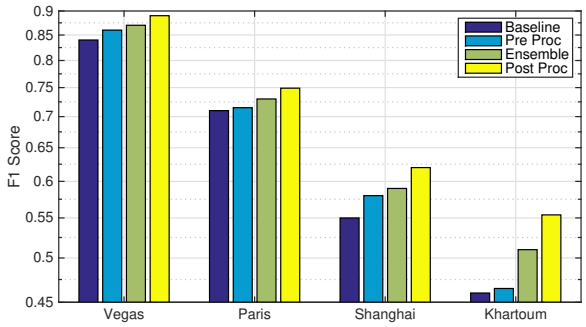

Figure 2. The F1 Scores of four cities with different optimization strategies in the development phase.

the richest diversity of buildings with respect to construction areas, heights, and architectural styles, etc. The distance between the building roof and the corresponding labeled polygon is large for high-rise buildings and small for low-rise buildings. The buildings located in residential areas are much easier to be extracted correctly compared with those located in gardens, agricultural areas and industrial areas. It's often difficult to correctly extract the buildings with green roofs, the low-rise buildings covered by trees, and the tiny buildings (e.g. located in the entrance of parking lots or gardens), etc. when only using the satellite datasets. The utilization of Map datasets can solve these problems to a great extent.

Khartoum has the lowest F1 Score among the four cities, resulting from many possible reasons. For instance, most of the public Map datasets provide little useful information for building extraction in Khartoum. Moreover, the building areas vary greatly in Khartoum and it's hard to decide whether a group of neighbouring buildings should be extracted together or separately.

Paris achieves the second highest F1 score among the four cities. The buildings in Paris dataset are in a relatively unified style. The buildings with green roofs and the buildings surrounded by forests are harder to be extracted correctly. Vegas achieves the best building extraction results among the four cities. Most of the buildings in Vegas dataset locate in residential areas and their architectural styles are more unified than those of other three cities. The buildings in countryside areas in different architectural styles and with fewer training samples are relatively harder to be extracted correctly.

5. Conclusions and future work

In this work, we proposed a semantic segmentation and ensemble learning based building extraction method using both satellite imagery and multi-source GIS map datasets. Our proposed method achieves the overall prediction score of 0.701 on the test dataset in the building extraction challenge.

In the future we will try more latest proposed DCNN models, such as LinkNet [6] and SegNet [4]. Some traditional computer vision techniques can be used as complementary of DCNN for Shanghai and Khartoum datasets.

6. Acknowledgement

This work was supported in part by the National Key Research and Development Plan of China (Grant No. 2016YFA0602200), the National Natural Science Foundation of China (Grant No. 61303003 and 41374113), and the China Postdoctoral Science Foundation (Grant No. 2016M601031). We would also like to thank NVIDIA for providing the GPU devices.

References

[1] Google map api. `https://developers.google.com/maps/documentation/static-maps/`.

[2] MapWorld. `http://lbs.tianditu.com/staticapi/static.html`.

[3] N. Audebert, B. Le Saux, and S. Lefèvre. Joint learning from earth observation and openstreetmap data to get faster better semantic maps. In *EARTHVISION 2017 IEEE/ISPRS CVPR Workshop. Large Scale Computer Vision for Remote Sensing Imagery*, 2017.

[4] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[5] R. Cabezas, J. Straub, and J. W. Fisher. Semantically-aware aerial reconstruction from multi-modal data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2156–2164, 2015.

[6] A. Chaurasia and E. Culurciello. Linknet: Exploiting encoder representations for efficient semantic segmentation. *arXiv preprint arXiv:1707.03718*, 2017.

[7] 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*, 40(4):834–848, 2018.

[8] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. *arXiv preprint arXiv:1805.06561*, 2018.

[9] S. Du, Y. Zhang, Z. Zou, S. Xu, X. He, and S. Chen. Automatic building extraction from lidar data fusion of point and grid-based features. *ISPRS Journal of Photogrammetry and Remote Sensing*, 130:294–307, 2017.

[10] M. Haklay and P. Weber. Openstreetmap: User-generated street maps. *IEEE Pervasive Computing*, 7(4):12–18, 2008.

[11] V. Iglovikov, S. Mushinskiy, and V. Osin. Satellite imagery feature detection using deep convolutional neural network: A kaggle competition. *arXiv preprint arXiv:1706.06169*, 2017.

[12] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[13] G. Máttyus, S. Wang, S. Fidler, and R. Urtasun. Hd maps: Fine-grained road segmentation by parsing ground and aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3611–3619, 2016.

[14] 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*, pages 234–241. Springer, 2015.