BUILDING CLASSIFICATION OF VHR AIRBORNE STEREO IMAGES USING FULLY CONVOLUTIONAL NETWORKS AND FREE TRAINING SAMPLES

Y. Chen [a*], W. Gao [b*], E. Widyaningrum [c*] , M, Zheng [d*], K. Zhou [c*,**]

[a] Dept. Computer Science, Delft University of Technology, The Netherlands -. Y.Chen-35@student.tudelft.nl
[b] Dept. Urbanism, Delft University of Technology, The Netherlands - W.Gao-1@tudelft.nl
[c] Dept. of Geoscience and Remote Sensing, Delft University of Technology, The Netherlands - (E.Widyaningrum, K.Zhou-1)@tudelft.nl
[d] Dept. OTB, Delft University of Technology, The Netherlands - M.Zheng-1@tudelft.nl

Commission IV, WG IV/3

KEY WORDS: Building Classification, VHR Airborne Stereo Images, FCN, Base Map, Mislabels, Free Training Samples, Fine Tuning, Atrous Convolution.

ABSTRACT:

Semantic segmentation, especially for buildings, from the very high resolution (VHR) airborne images is an important task in urban mapping applications. Nowadays, the deep learning has significantly improved and applied in computer vision applications. Fully Convolutional Networks (FCN) is one of the tops voted method due to their good performance and high computational efficiency. However, the state-of-art results of deep nets depend on the training on large-scale benchmark datasets. Unfortunately, the benchmarks of VHR images are limited and have less generalization capability to another area of interest. As existing high precision base maps are easily available and objects are not changed dramatically in an urban area, the map information can be used to label images for training samples. Apart from object changes between maps and images due to time differences, the maps often cannot perfectly match with images. In this study, the main mislabeling sources are considered and addressed by utilizing stereo images, such as relief displacement, different representation between the base map and the image, and occlusion areas in the image. These free training samples are then fed to a pre-trained FCN. To find the better result, we applied fine-tuning with different learning rates and freezing different layers. We further improved the results by introducing atrous convolution. By using free training samples, we achieve a promising building classification with 85.6% overall accuracy and 83.77% F1 score, while the result from ISPRS benchmark by using manual labels has 92.02% overall accuracy and 84.06% F1 score, due to the building complexities in our study area.

1. INTRODUCTION

1.1 Background

Remote sensing images have been widely used and play an important role in various applications, especially for mapping purposes. The VHR (Very High Resolution) airborne images with very detailed geographic information provide an opportunity to create large scale maps by detecting and classifying buildings, roads, water bodies, vegetation, etc. Rapid progress on deep learning techniques in the last few years has drawn geo-scientists attention to implementing artificial geointelligence for an urban mapping application. Cheng et al. (2017) reviewed several benchmarks that contain remote sensing scene classification datasets used for neural networks, namely: UC Merced Land-Use, WHU-RS19, SIR-WHU, RSSCN7, RSC11, Brazilian Coffee, and NWPU-RESISC45. However, the benchmark dataset in remote sensing often has limited training samples in a specific area. The generalization ability for another area of interest is adversely affected. In this study, an existing high precision map is used to label the airborne VHR stereo images to provide large scale training samples. However, maps and images often have time difference, so a small amount of mislabels will be introduced. In this study, whether the classification is robust to these mislabels, in the sense of changes (such as building changes, trees growth, road changes, etc.) should be evaluated. Apart from these mislabels, unlike Toronto dataset (Wang et al., 2016) with perfect matches between maps and aerial images, there are three main resources, which may introduce mislabels are addressed . The three main resources are:

a. The relief displacement of high objects in the image, especially for buildings, resulting in a serious positional mismatch between the image and the map.
b. The map and images have different object representation. In the map, a building is defined based on the footprint of the walls on the map, while in the image, building only can be shown based on the roof. In this case, an overhanging roof creates the mismatched areas due to different shape and size of a building in the map and the image.
c. The airborne stereo image is often has less accurate colors in the occluded area which can be seen from one image but not others.

In classification, the pretrained FCNs (Long et al.) are proved with very good performance and high computational efficiency for semantic segmentation. However, many hyper-parameters in FCNs have significant impacts on the performance. Moreover, the FCNs structures are constructed to extract features for images in ImageNet. However, the capability of the structure for extracting valuable features for VHR images are questionable.
In this paper, we propose a novel and fully automatic approach to classify buildings from VHR stereo images by using existing maps to provide free training samples. The scientific contributions are as follows:

a. We provide an approach to reduce the mislabels from relief displacement, different representation between the base map and the image, and occlusion areas in the image.

*Authors contributed equally and are ordered alphabetically
** Corresponding author

b. We test different learning rates and freeze different layers to select the best hyper-parameters for classification. We replace the convolutions in FCN with Atrous convolutions to extract more global features for our VHR data.
c. We compare our building classification results with those from ISPRS benchmark. The comparable results show that FCNs are robust to a small amount of mislabels.

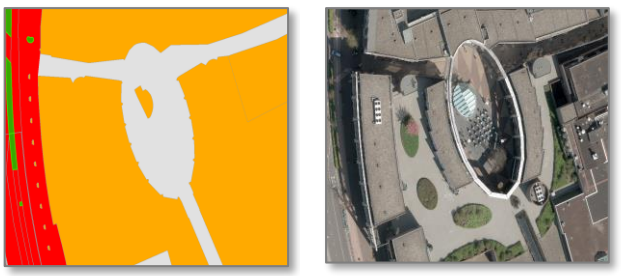
1.2 Data specification and study area

In this study, we aim to do VHR airborne images classification by using an updated base map version as the training samples. We use datasets that has five years time difference. However, the choice of the two datasets may make our study more challenging. However, it does not reduce the effectiveness of the provided approach.

The datasets used in this study are specified as below:

a. The Dutch large-scale base map which is called as BGT map. It has a vector format. We use the updated version, the Year 2016. This map is composed of several objects/map layers such as traffic area, bridge, building, terrain, plant cover, solitary vegetation, fence, road, tunnel, water-body, etc.
b. The VHR airborne images have a spatial resolution of 3.5 cm and are georeferenced with one pixel positional accuracy. These aerial images were acquired in 2011.
c. The validation dataset for classification is obtained by manual delineation especially to obtain the changed building parts in the base map according to the airborne images.
d. The ISPRS Benchmark dataset, the Vaihingen – Germany. It comprises of 16 raw airborne images and 16 semantic label images. The comparable results show that FCNs are robust to a small amount of mislabels.

Our study area is located in Amersfoort city of The Netherlands. The total size of the study area is 611 x 1050 m. A subset of our dataset is shown in Figure 1.



a. The BGT map b. The VHR airborne images

Figure 1. A subset of Amersfoort dataset

2. RELATED WORK

One of the reasons that CNNs are very powerful in computer vision task is that they can automatically extract the deep feature of the image instead of the man-craft features. Although the CNN have superior classification performance, it could only provide the "image-label" which means one image can only be classified into one class. In order to delineate the boundary for mapping purpose from remote sensing images, semantic segmentation should be performed. Many types of research (Farabet et al. 2013, Pinheiro et al. 2014) employed patch based CNNs to derive semantic segmentation by classifying the image patch centered in that pixel. However, this approach is computationally intensive. In order to keep the advantages of CNN but saving more computation budget, FCNs (Long, 2015) is proposed with "pixel-label" classification by replacing the last fully connected layer with convolutional layers. By keeping convolutional layers from CNN, deep feature extraction still exists. Another advantage of the FCN architecture (Figure 2.) is that it has the ability to accept any size of the image and output a classification map with the same size. Fully convolutional networks (FCN) has been successfully applied in remote sensing images. It is proved with good accuracies and efficiency computations (Kampffmeyer et al., 2016, Bittner et al., 2017, Fu et al., 2017).

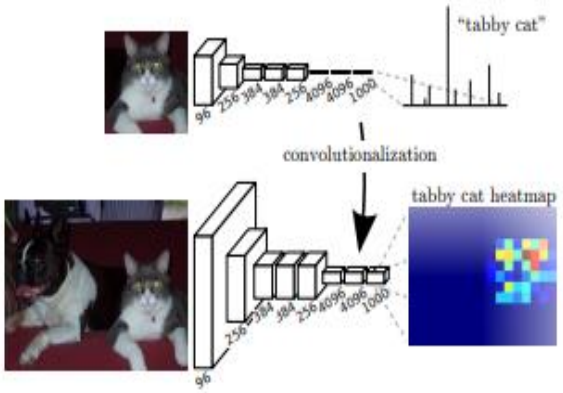


Figure 2. The architecture of the FCN (source: Long et al., 2015)

Large-scale benchmarks are significant to train the large-scale convolutional neural networks (CNN). ImageNet with tremendous amount of training samples facilitates the success of CNN in most of the computer vision tasks. However, ImageNet contains very limited training examples for remote sensing and urban mapping applications. Even for transfer learning, fine-tuning still needs many training examples from the area of interest. To solve the issue above, Wang et al. (2016) create a large-scale benchmark dataset, TorontoCity, by using a high precision map to create ground truth for labelling airborne and mobile images in order to test different neural networks for various vision tasks. However, the problems of their research is that maps and aerial photos assume perfectly matched and it is not true in most of cases. For example in Section 1.1, it illustrates many mislabeling. Instead of creating benchmarks for different applications by using maps, we directly apply semantic segmentation on VHR images by using maps to provide free training samples. As a result, the classification results will be used for change detection on the map to update the map.

3. METHODOLOGY

In this section, we describe the components of our approaches to conduct the VHR airborne image segmentation into three parts. The first step is to generate the labels from the base map (BGT). The second step consists of some procedures to provide clean training samples by removing the detected mislabeled areas. In the last step, the clean training samples are fed into the fully convolutional network. The workflow of our study is illustrated in Figure 3.

3.1 Generate the label

One of the factors for a successful FCN is depended on the quality and quantity of the training data. Thus, we aim to provide training samples with noise or mislabeled pixels as less as possible. A class aggregation is a necessary step since the BGT

map has several map layers or object classes. For creating the labels, we aggregate the BGT map into two classes: building and non-building.

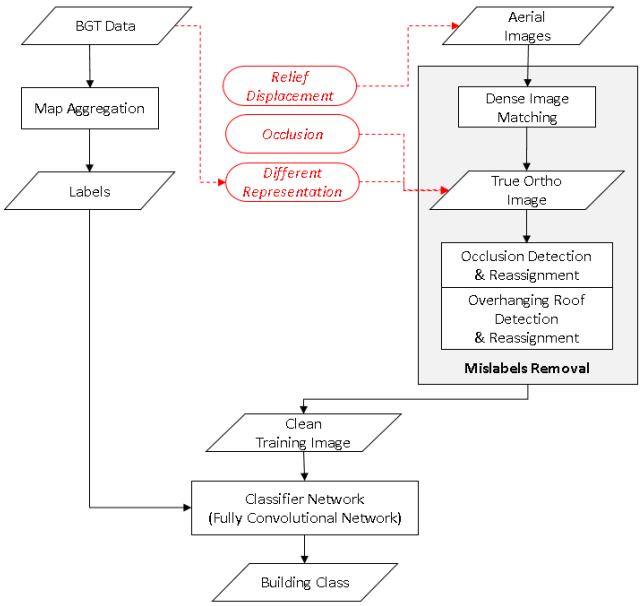


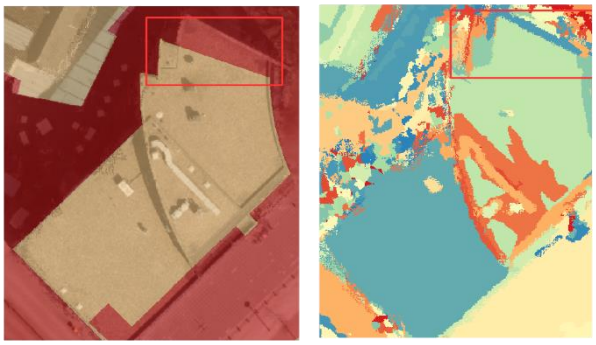
Figure 3. The Research Workflow

3.2 Reduce the mislabel

We conduct a series of procedures to provide clean training samples for anticipating some problems that may arise due to problems above in chapter 1.1.

3.2.1 Removing the relief displacement: In the image, a high building may have a serious relief displacement due to aerial acquisition angle and height differences. The relief displacement causes a positional error of the building roof in the VHR images, which means that some pixels of the building roof are shifted from its true location. This relief displacement causes a building misalignment between the images and the base map. In most of the cases, the relief displacement of the airborne images can be removed by generating the true orthophotos. We apply a straightforward way to provide the true ortho-images by triangulation interpolation from colored point clouds generated by dense image matching algorithm from stereo images. We use Pix4D software to construct the RGB point clouds. These images are then used to provide training samples labelled by maps and also used for further reducing mislabels.

3.2.2 **Adjusting different buildings representation**: The building representation that shows in the BGT base map is different from that in the images. The base map presents a building as a footprint of the wall, while image presents a building as a roof. Due to its dissimilarity, the pixels in overhanging roof part (an area where a building in the base map is smaller than a building in the image) have wrong labels from maps. As shown in Figure 4.a., the mismatched of building footprint and the real building roof. To avoid mislabeling problem in the overhanging roof areas, we conduct a plane segmentation (Vosselman, 2010) in the point clouds. After point clouds triangulation, the plane-segmented image is obtained. A continuous plane is detected as the overhanging roofs often continue the planar trend from inner roof parts as shown in Figure 4.b. The overhanging roof is detected when the plane segment is partially located in the polygon. These plane pixels outside the building polygon is converted to black colors in the training image as the segments are not accurate in the building boundaries. Accordingly, the corresponding labels are also converted to non-buildings.



a. Overhanging roof not presented in the map

b. Overhanging roof detection

Figure 4. Building mismatched in the overhanging roof

3.2.3 Removing occlusion: In the urban scene, many objects can be seen from one image not from another due to relief displacement. Point clouds are hardly reconstructed from these occluded areas from dense image matching. Therefore, we may have some gaps or areas without any point clouds presence. As we produce the true ortho-images from these point clouds by triangulating interpolation, the occluded areas will make large triangles, and the interpolated color is deteriorated. As shown in the left image of Figure 5., the occlusion areas near to the building have blurred color (inside the red box).

Figure 5. The blurred pixels shows the occluded area (inside the red rectangle)

These stretched or blurred pixel contains wrong color information that may cause mislabels. Therefore, in this step, we detect the blurred pixels by checking whether the pixels are interpolated from a triangle with its edge larger than 70 cm (20 pixels). These detected pixels are converted to black colors, and the labels are marked as non-buildings.

3.3 The FCN classification

The convolutional layers from FCNs often borrow the pretrained networks (Long et al., 2015). In this study, the VGG-16 (Simonyan et al., 2014) is selected due to its superior performance. The VGG-16 is a convolutional neural network pre-trained by ImageNet. According to its state-of-art performance, many researchers use it for a basic building block to build their customize architecture. By using the applications, we do not have to train the whole network from the scratch but applying fine-tuning by feeding our training samples.

FCNs replace fully connected layers in CNNs with convolutional layers which preserves the spatial information of the input. However, when an image is passing through the convolutional layers, the pixel information of the image is losing significantly due to the pooling layers with downsampling process. FCNs use skip layers architecture to combine information from shallow layers and deep layers. By this way, we could combine global features from low layers with local features from high layers. The upsampling process is designed for regenerating image size as the input.

There are two critical issues should be considered as follows:
(1) How to define an architecture to derive features from remote sensing images?
Our dataset is very high-resolution images even to 3.5cm. As the accuracy of boundary delineation is related to resolution, we keep this high resolution for processing. The image size fed for FCN is often no more than 1000*1000 due to computational limits of a GPU. Therefore, the patch size is not more than 35m* 35m, that is often not enough for a large building. With a fixed size of convolutional layers in FCNs, max pooling and stride is then used to expand the receptive field of deep layers. The structure of VGG-16 in the FCN is very effective to extract local and global features for ImageNet images where the object of interest is in the center of the image. However, it may fail to extract effective global features to our dataset. Therefore, the atrous convolutions in Figure 6 are used to replace the convolutions in the VGG-16 block. The atrous convolution kernel is a new fashion of convolution, which are similar with 3*3 kernel but it has space between the kernels. The advantage of atrous kernel which enable to extract the features by changing the size of the kernel while keeping weights in the kernel the same.

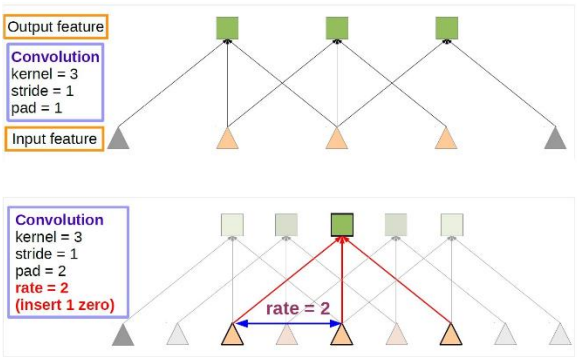


Figure 6. The 1-D Atrous convolution representation (Liang et al., 2018)

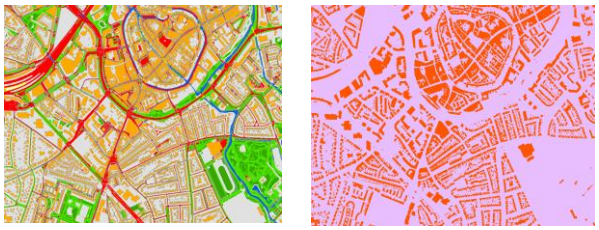
(2) How to fine-tune hyperparameters from pre-trained weight to fit for our dataset?
We focused on ISPRS benchmark dataset (Cramer, 2010) for fine-tuning parameters. These parameters which result in the best classification performance are applied for our own dataset. In this study, the most critical parameter to tune is the learning rate. We tried different learning rates to find the best one. In addition, layer freezing is also considered in the paper. The previous layers in the deep neural network extract the low level features with respect to the images such as edges and corners, while the deep layers extract the high level features specific to training images. Therefore, in the fine tuning process, we freeze the previous layers in order to prevent updating the low level features from back propagation, at the same time, update the weights in the deep layers which is not frozen.

4. RESULT AND DISCUSSION

We applied building classification on Amersfoort dataset with FCNs using free training sample by reducing mislabels. The result is compared with that of FCN for ISPRS benchmarks.

4.1 Generate the label

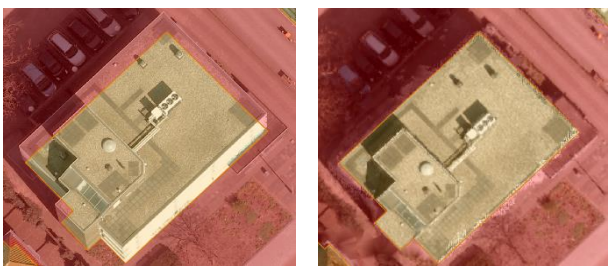
The extracted labels from the BGT for the FCN classification is shown in Figure 7. The labelled image is extracted from the aggregated map that has the same spatial resolution 3,5 cm.



a. The BGT Map b. The training image
Figure 7. The BGT map is aggregated into labelled image

4.2 Reduce the mislabel

4.2.1 Removing the relief displacement: By using the stereo-images, we solve the misalignment problem between the base map and the images. In Figure 8.a. we can still see the building wall on the right side of the building, which means that there is a building misalignment in the base map (yellow area) with the image due to relief displacement. As shown in Figure 8.b., the building in the true ortho image is perfectly matched with the base map (yellow area).



Building misalignment between the map (yellow) and image b. A matched building in the map and image
Figure 8. Removing the misalignment problem

4.2.2 Adjusting different building representation: In Figure 9., the pixels of the overhanging roof in the image (inside the blue ellipse of Figure 9.a.) are detected and then assigned as zero value with black color (inside the blue ellipse of Figure 9.b.). In the end, these pixels are assigned as non-building pixels in the training samples.

a. Image contains overhanging roof b. Image without overhanging roof
Figure 9. Solving the building representation differences.

4.2.3 Removing the occlusion: Figure 10. illustrates the differences between the unclean and clean training image. In Figure 10.b., the occluded part of the image is removed and assigned as a non-building class.

a. An occluded image b. A clean image

Figure 10. Removing the occluded areas to provide clean training samples

4.3 Fully convolutional classification

4.3.1 Fine-tuning with learning rate and freeze layers: To test the FCN architecture and parameter setting, we use the ISPRS benchmark dataset (Cramer, 2010), the Vaihingen. The dataset consists of VHR orthorectified airborne images including the classified image. This datasets include 16 images with ground truth label. We use 12 images as the training samples and four images as the test samples. The result is evaluated based on the ground truth data. The classification accuracy is measured by four metrics: recall, precision, overall accuracy (OA) and F1 score. Figure 11. shows the result of classification by tuning learning rates. We tested six learning rates on four quantitative measurements. When learning rate is 1e-5, the result is the best. Figure 12. shows the result by changing the freezing layers. We separately freeze the third layer, fourth layer and do without a freeze layer, and then compare the classification results still on four quantitative measurements. The test with a non-freeze layers presents the best performance.

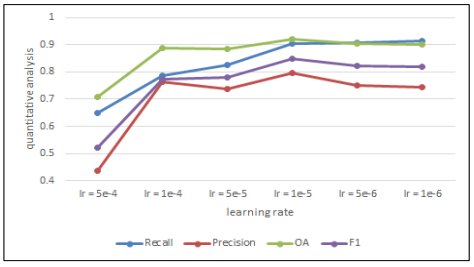


Figure 11. The learning rate of ISPRS dataset

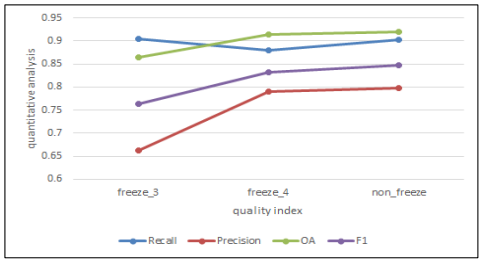


Figure 12. The learning rate of frozen convolutional layers

4.3.2 Apply the optimal paramaters: Since the tuning the parameter on our data set cost too much time, we adopt optimal parameters from ISPRS, that is learning rate equaling to 1e-5 and without freeze layers. Two experiments are applied on a clean data which contains 234 images and noisy data with 171 images. 63 images are used for validation. Clean data are applied with the three steps of reducing mislabels, while noise data don’t. As shown in Figure 13, the classification result for the same area on a noisy dataset is worse than the clean dataset. The reason is because the noisy data contains more mislabeled building pixels.

The result of noisy data has lower recall value (71.57 %) than the clean data, which means that it has 7% less building detection than that in the clean data. The overall accuracy of noisy data is 2% less and the F1 score is almost 4.4% less than the clean data.

We also applied the FCNs to ISPRS benchmark dataset with the manually labeled training samples. The comparison of the clean data with ISPRS benchmark result shows that the overall accuracy of our clean data (84.81%) is 7.2% less, but the F1 score (83.22%) is similar. It may happen due to our study area containing higher buildings complexities (such as building with grass or trees on the roof,high details roof, and roof – ground similarities as shown in Figure 1.b) than the ISPRS Vaihingen area.

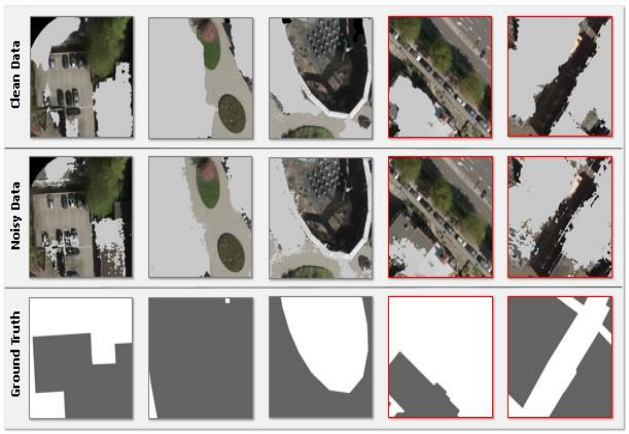


Figure 13. Comparison of building classification result from clean and unclean training samples (grey color represents the bulding area)

4.3.3 Apply an Atrous convolutions: We apply the atrous kernel on the clean dataset. As shown in Table 1., the result of atrous kernel slightly better than the previous kernel since the Atrous uses a larger receptive field. The overall accuracy increases from 84.81% to 85.60% and F1 score increases about 0.5%. This result gives a hint that the atrous convolution could help to get more global features in classifying our buildings.

	Recall	Precision	Overall Accuracy	F1 Score
ISPRS	90.27	79.75	92.02	84.06
Noisy data	71.57	90.22	82.73	78.82
Clean data	78.89	88.25	84.81	83.22
Clean data + Atrous	77.89	90.61	85.60	83.77

Table 1. Performance of the FCN classifier

Figure 14 and Figure 15 show an averall visual comparison of our FCN result with the ground truth data. Figure 14 shows the final output of the building classification results that is marked by light grey color. Figure 15 shows the ground truth data, where buildings areas are represented by dark grey color.

Figure 14. The building classification result

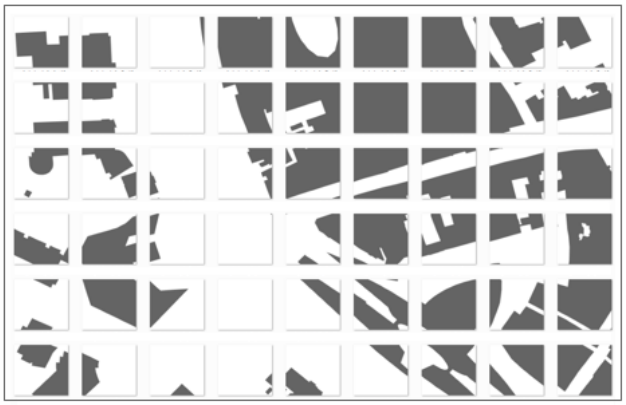


Figure 15. The ground truth

5. CONCLUSION AND RECOMMENDATION

We provide a novel and fully automatic approach to conduct a building classification by using free training samples. Based on our evaluation, it is critical that the three main mislabeling problems should be addressed. With the clean training samples, the overall accuracy and F1 score is increased by 2.1% and 4.4%. By applying the fine-tuning with the hyper-parameters, we obtained satisfying results of the building classification. The use of atrous convolution able to increase the overall accuracy and F1 score by 0.5%, which means that a larger atrous kernel may have better performance. The building classification of Amersfoort dataset using automatic labels has a less overall accuracy than the ISPRS dataset using manual labels. The possible reason is that our study area has higher building complexities than ISPRS Vaihingen area. Moreover, the similar F1 score also shows a promising result of our approach.
For a future work, adding the height information (such as Digital Surface Model) is worth to implement to increase the accuracy of FCNs, since some buildings have confusions with roads if only contexture features are considered.

ACKNOWLEDGEMENT

The authors acknowledge the NEO, Netherlands Geomatics & Earth Observation B.V. for providing the Amersfoort datasets.

REFERENCES

Bittner, K., Cui, S., Reinartz, P., 2017. Building extraction from remote sensing data using fully convolutional networks. The International Archives of Photogrammetry, Remote Sensing, and Spatial Information Sciences, 42.

Cramer, M., 2010. The DGPF test on digital aerial camera evaluation – overview and test design. Photogrammetrie – Fernerkundung – Geoinformation 2, pp. 73-82.

Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K. and Yuille, A.L., 2018. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence, 40(4), pp. 834-848.

Farabet, C., Couprie, C., Najman, L., LeCun, Y. , 2013. Learning hierarchical features for scene labeling. IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1915-1929.

Fu, G., Liu, C., Zhou, R., Sun, T., Zhang, Q., 2017. Classification for High Resolution Remote Sensing Imagery Using a Fully Convolutional Network. Remote Sensing, 9(5), pp. 498.

Cheng, G., Han, J., & Lu, X. 2017. Remote sensing image scene classification: benchmark and state of the art. Proceedings of the IEEE, 105(10), pp.1865-1883.

Liang, C., Papandreou, G., Kokkinos, I, Murphy, K., Yuille, A. 2018. DeepLab:Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 40, Issue 4, pp. 834-848.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431-3440.

Kampffmeyer, M., Salberg, A. B., Jenssen, R., 2016. Semantic Segmentation of Small Objects and Modeling of Uncertainty in Urban Remote Sensing Images using Deep Convolutional Neural Networks. In Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE Conference, pp 680-688.

Pinheiro, P. H., Collobert, R., 2014. Recurrent convolutional neural networks for scene labeling. In 31st International Conference on Machine Learning (ICML), No. EPFL-CONF-199822.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Vosselman, G., 2012. Automated planimetric quality control in high accuracy airborne laser scanning surveys. ISPRS Journal of photogrammetry and remote sensing, 74, pp. 90-100.

Wang, S., Bai, M., Mattyus, G., Chu, H., Luo, W., Yang, B., Urtasun, R., 2016. Toronto City: Seeing the world with a million eyes. arXiv preprint arXiv:1612.00423.