FULLY CONVOLUTIONAL NETWORKS FOR STREET FURNITURE IDENTIFICATION IN PANORAMA IMAGES

Y. Ao[1,*], J. Wang [2], M. Zhou [3], R. C. Lindenbergh [2], M.Y. Yang [1]

[1]Dept. of Earth Observation Science, Faculty ITC, University of Twente, The Netherlands -y.ao@student.utwente.nl, michael.yang@utwente.nl

[2]Dept. of Geoscience and Remote Sensing, Delft University of Technology, The Netherlands – (jinhu.wang, R.C.Lindenbergh)@tudelft.nl

[3]Academy of Opto-Electronics, Chinese Academy of Sciences, China – zhoumei@aoe.ac.cn

Commission II, WG II/4

KEY WORDS: Panoramic Images, Semantic Segmentation, Street Furniture, Object Identification, Fully Convolutional Networks

ABSTRACT:

Panoramic images are widely used in many scenes, especially in virtual reality and street view capture. However, they are new for street furniture identification which is usually based on mobile laser scanning point cloud data or conventional 2D images. This study proposes to perform semantic segmentation on panoramic images and transformed images to separate light poles and traffic signs from background implemented by pre-trained Fully Convolutional Networks (FCN). FCN is the most important model for deep learning applied on semantic segmentation for its end to end training process and pixel-wise prediction. In this study, we use FCN-8s model that pre-trained on cityscape dataset and finetune it by our own data. The results show that in both pre-trained model and fine-tuning, transformed images have better prediction results than panoramic images.

1. INTRODUCTION

Object detection for street details has been a popular research topic for its wide applications. The rapidly developing autonomous driving requires highly accurate objects recognition on the street scenes to achieve a satisfying performance and safety for self-driving cars (X. Chen et al., 2016). Also, object detection serves the tracking task as an observation and a successful object trajectory determination needs a large amount of the observations (Ess et al., 2010). Furthermore, in terms of the hot robotics field, the identification of the street details is necessary for outdoor mobile robots navigation (Benavidez & Jamshidi, 2011). In addition, as for the social aspect, street objects detection helps to inventory the real-world targets in an efficient way, which were annotated manually in the previous days (Creusen et al., 2012).

Street furniture identification plays an essential role in object detection field since various kinds of furniture are contributing directly to the people's daily safety. For examples, street lamps give cars and pedestrians sight support in the night and traffic signs warn the road users for upcoming dangerous situations which guarantee the smooth traffic flows. Street furniture identification has many applications including road maintenance and urban plannings. Moreover, the government needs position and type information of street furniture in order to maintain it when it's broken or its visibility degrades over time (Hazelhoff et al., 2014). Urban planners require the existing street furniture information to locate the new ones. Those works are usually completed either by acquiring street furniture information manually or from laser scanning data, or by 2D image data. In this study, we propose to use panoramic image, which is infrequently used data for research, to experiment automatically perform the above-mentioned tasks.

Panoramic images are more and more noticeable since their 360° perspective vision providing users with a broader view than normal images, while the cost of acquiring them is relatively low. They are increasingly used in many scenarios in the past few years, such as street view capture and indoor monitor. Panoramas also played an efficient role in the scientific researches like robot localization field (Marinho, Almeida, Souza, Albuquerque, & Rebouças Filho, 2017) and 3D structure reconstruction (Pintore, Ganovelli, Gobbetti, & Scopigno, 2016) for their inner three-dimensional information. But they have been rarely used in the conventional object identification in street environment.

To recognize street furniture from the panoramic images, this study proposes to apply semantic segmentation on the images. The pixel-leveled method is achieved by an end-to-end deep learning model that is Fully Convolutional Networks and produces dense per-pixel labeled predictions. It will take a long time and demand a lot of resources to train an FCN model from the beginning. In view of our own small panorama dataset, we decided to fine-tune a pre-trained FCN model which was trained on a big dataset with a similar scenario to ours. In addition, we transform the panoramic images to conventional perspective images, since there is no available street-view panorama dataset that can be used by a pre-trained model. In this study, we will fine-tune the pre-trained FCN model with both panoramic images and transformed images and compare their predictions to see if the panoramic properties affect the training and predicting results.

Below the related works of street furniture identification is presented in Section 2. Methods for image mapping, image pre-processing and semantic segmentation by FCN model are explained in Section 3. Experiment procedure and results are given in Section 4. Conclusions are described in Section5.

2. RELATED WORK

Many approaches have been developed for street furniture identification in the past years. Surveying, satellite-based remote sensing and photogrammetry are worldwide spreading technologies. But for street details, they are inappropriate because they aim at mapping, not object detection. Airborne laser scanning and aerial nadir or oblique imagery are also not suitable for this application since they must have a centimetric resolution which needs to be acquired from a low altitude with a thus high cost due to the multi-flight for overcoming vertical occlusions (Paparoditis et al., 2012).

For this street-level work, there are mainly two approaches, 3D data or 2D images. The 3D data usually acquired via two ways, 3D models estimated from 2D images or 3D point clouds from laser scanning system. Saxena et al. (2009) made 3D urban scene structure from monocular images, while Hu and Upcroft (2013) reconstructed 3D street view from stereo image pairs. However, their work focuses on using geometric characteristics extracted from those images. In this respect, mobile laser scanning (MLS) data which can avoid some occlusions on account of mobile characteristic and diverse sensors for different scanning platforms can directly acquiring more accurate spatial information (Alho et al., 2011). Pu et al. (2011) classified the street objects by introducing an initial vertically slicing method to recognize the shape of street furniture. Cabo et al. (2014) and Wang et al. (2017) extracted pole-like street furniture from MLS point cloud data with voxel-based approaches. The difference lies in that the former uses square cube while the latter uses icosahedron to build a shape descriptor. Also, Rodríguez-Cuenca et al. (2015) applied the pillar structure to organize the point cloud and detect vertical urban elements by means of an anomaly detection algorithm. Nevertheless, MLS point cloud based approaches mainly consider the objects' spatial characteristics and spatial relations which make it really difficult to identify close by and similar objects (Wang et al., 2017). Also, considering the cost of acquiring MLS point cloud data is much more than capturing images from the camera, MLS is not an optimal choice for this study.

For 2D images such as street views and color images which are captured by cameras mounted on vehicles, image segmentation is regard as one of the most essential tasks for object extraction. A number of methods have been developed to solve this problem, from elementary pattern analysis like Hough transformation, via feature extraction-based tools like boosting, to more advanced machine learning and deep learning algorithms such as Support Vector Machines, Conditional Random Field, and Convolutional Neural Networks (Krylov et al. 2018). To detect specific objects effectively, researchers proposed their new machine learning descriptors or improve the results by integrating with someone else's machine learning model, such as to extract utility poles (Zhang et al., 2018) by RetinaNet object detector (Lin et al., 2017). Thanks to the relentless success of machine learning algorithms, a lot of image processing methods have achieved satisfying semantic labeling results (Cordts et al., 2016).

Taking advantages of machine learning method to perform semantic segmentation and classification have become a general trend. Convolutional Neural Networks (CNN) is the cornerstone of various state-of-the-art approaches. Krizhevsky et al. (2012) used CNN to classify the large ImageNet dataset, which motivated many researchers to explore the capabilities of the networks for semantic segmentation. In the followed research, Fully Convolutional Network (FCN) presented by Long et al. (2015) is one of the most significant and popular methods among the subsequent techniques (Garcia-Garcia et al., 2017). FCN replace the fully connected layers of those existing classification model like AlexNet, VGGnet, and GoogLeNet with convolutional ones and transfer their classification scores into fine-tuning segments. Standing on the shoulder of the giants, many researchers have done further works in this field. Zeng et al. (2017) feed FCN with multi-view RGB-D data to do the segmentation and label job and Shelhamer et al. (2016) proposed an adapted FCN deal with video sequences data.

In this paper, the baseline is also Fully Convolutional Networks but experimented with different data which is panoramic images from mobile mapping system.

3. METHODOLOGY

The proposed street furniture identification workflow is shown in Figure 1. Panoramic images are inputs of the study and they are firstly transformed into perspective images consecutively. Secondly, pre-processing is conducted on both panoramic images and transformed images. Then, a pre-trained FCN model is introduced to produce predictions directly, The next step is semantic segmentation by fine-tuning the FCN model with training images and produces predictions for testing images.

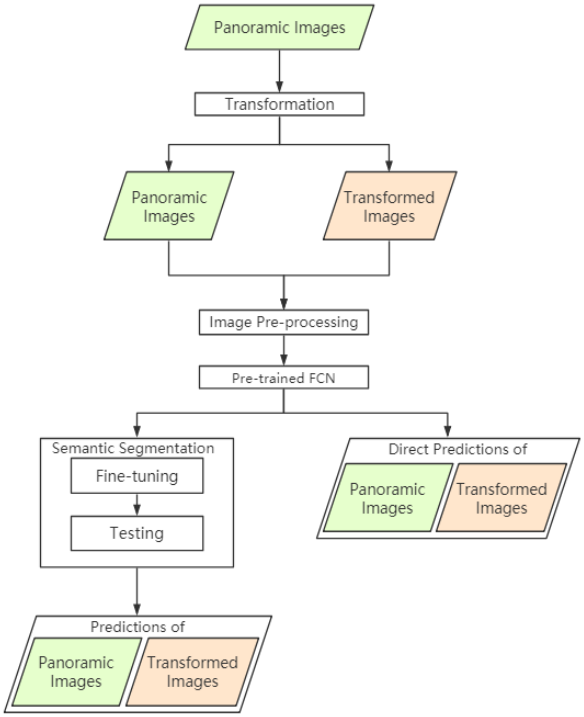


Figure 1. Workflow of the proposed street furniture identification.

3.1 Transformation

The images used in this study are panoramic which are different from the normal images. Its 360° vision presents users a wide viewing angle as well as suffering from distortions especially at the top and bottom of the images. In order to make them similar to the training data of the pre-trained FCN model, the panoramic images need to be transformed to into normal perspective images.

Panoramic images are captured by 360° cameras. They are transformed from the spherical image (Figure 2) into planar image by equirectangular projection which is a cylindrical

equidistant projection and the output is equidistant along the horizontal and vertical direction (Su & Grauman, 2017).

Figure 2. A panoramic image on a sphere. The lines represent latitude and longitude.

We can map the equirectangular images with a simple gnomonic projection for knowing every pixel 's latitude and longitude on the sphere (Coors et al., 2018). The geometric principles of Gnomonic Projection are illustrated in Figure 3.

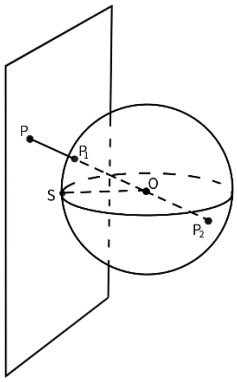


Figure 3. Gnomonic Projection

The principal point O is the center of the sphere, and every point on the sphere can be projected onto the plane that represents the 2D perspective image through the radial from the center point. For example, point P on the plane is the projection of point P_1 on the sphere. The projection equations (Weisstein, n.d.) are given as below:

$$x = \frac{\cos\phi \sin(\lambda-\lambda_0)}{\cos c} \quad (1)$$

$$y = \frac{\cos\phi_0 \sin\phi - \sin\phi_0 \cos\phi \cos(\lambda-\lambda_0)}{\cos c} \quad (2)$$

$$\cos c = \sin\phi_0 \sin\phi + \cos\phi_0 \cos\phi \cos(\lambda-\lambda_0) \quad (3)$$

The transformation equation is for the plane tangent at the point having latitude and longitude (ϕ_0, λ_0) which is the Point S in Figure 3. The point with latitude and longitude (ϕ, λ) will be located on the plane with position (x, y). In the transformation procedure, we usually fix the output image size and then find the corresponding point P (x, y) of the point P_1 (ϕ, λ). Therefore, we need to use the inverse equation of the above equation, which are given as below:

$$\phi = \sin^{-1}(\cos c \sin\phi_0 + \frac{y \sin c \cos\phi_0}{\rho}) \quad (4)$$

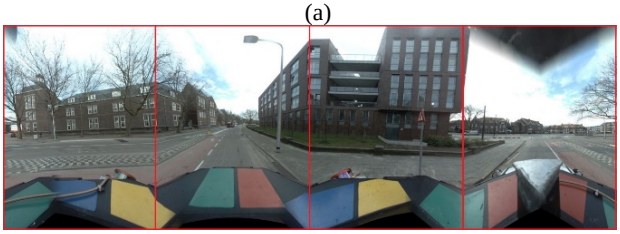
$$\lambda = \lambda_0 + \tan^{-1}(\frac{x \sin c}{\rho \cos\phi_0 \cos c - y \sin\phi_0 \sin c}) \quad (5)$$

$$\rho = \sqrt{x^2 + y^2} \quad (6)$$

$$c = \tan^{-1}\rho \quad (7)$$

It is not possible to map the whole panoramic image in one direction. Therefore, we choose four directions each range 90 degrees along the great circle of the sphere which is also the horizontal middle line of the panoramic image. In vertical direction, the mapping range is 120 degrees, which is ±60 degrees of the great circle. We do not project the whole content in the vertical direction because the top and bottom parts are not region of interest and the projection of 120 degrees is the best-performing one. The transformed image is shown in Figure 4 (b) and the projection of four directions is margined with red lines.

(a)



(b)

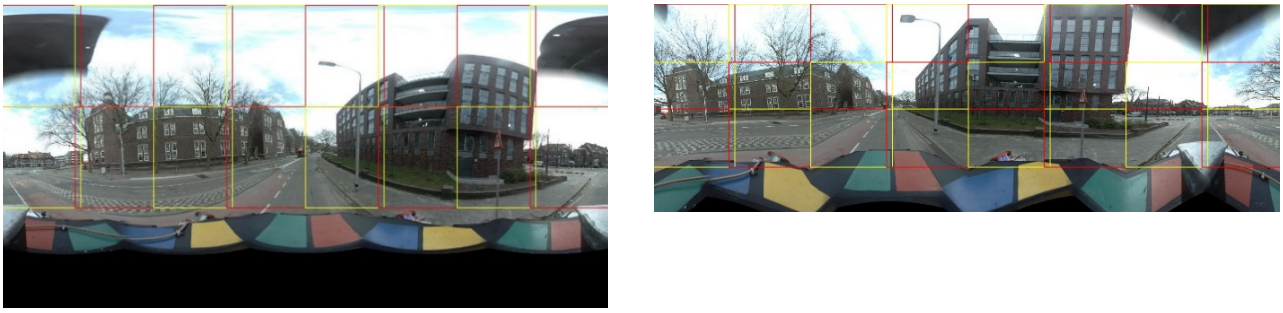
Figure 4. Transformation of the panoramic image. (a) The original panorama. (b) The transformed image. The red lines represent the margin of each direction's projection.

3.2 Image Pre-Processing

Image pre-processing in this study consists of two aspects, i.e. cropping and data augmentation. The image size of the panoramic image is 5400 x 2700 pixels and the size of the transformed image is 5400 x 1800 pixels. Both images are too large to train in the model for our hardware condition constraint, hence cropping is needed. Data augmentation aims to enhance the contrast of the transformed images since low-contrast details in the original image may affect the training results and increase the images for training.

3.2.1 Image Cropping

The size of cropping image set as 700 x 900 pixels, which is an appropriate size in consideration of our GPU capability and the remaining semantic information within single cropped image. Each image in the training data will be cropped in to 16 small images. Figure 5 shows how to crop the images with red and yellow lines enclosing the small images. The layout of the small images is organized in four directions with every four images in one direction and the four images have a little overlap at the edge. It can be observed that the cropping preserves significant fields and clips the unwanted parts.



(a) (b)

Figure 5. Image cropping. (a) Cropping arrangement of the original panoramic images. (b) The transformed images.

3.2.2 Data Augmentation

Image augmentation is applied on the cropped image with contrast enhancing degree from 1.3 to 1.8 as shown in Figure 6. The enhancement parameter is set based on experiments. When the parameter sets below 1.3 the images do not change obviously, and if it sets exceeding 1.8 it's overdone for the invisible dark details.

(a) (b) (c)

Figure 6. Image contrast enhancement. (a) The original image. (b) The adjusted image with enhancing parameter 1.3. (c) The adjusted image with parameter 1.8.

3.3 Fully Convolutional Networks

Fully Convolutional Networks is one of the cutting-edge architectures for semantic segmentation and have been carried out with many networks. As the cornerstone of the semantic segmentation field, FCN could be a stable architecture for a new used type of images. The chosen FCN-8s model in this study is based on VGG16 net, for it contains more details in prediction than FCN-16s and FCN-32s and it performs better on VGG16 net than AlexNet or GoogLeNet (Long et al., 2015). The FCN architecture is shown in Figure 7. It contains convolutional layers, max pooling layers, drop out layers, and deconvolutional layers. The prediction layers in the figure are convolutional layers with output channel number same as the class number, which means they can be interpreted as intermediate predictions. Furthermore, the net does not get results directly from the deconvolutional layers. It uses skip architecture which adds fuse operations aiming to take advantage of both predictions from pool3 and pool4 to optimize the results. The activation function used in the convolutional layers is ReLU (Rectified Linear Unit), which result in much faster training (Krizhevsky et al., 2017). The equation of ReLU is given below:

$$f(x) = \max(x, 0) \tag{8}$$

To calculate the loss of the net, we use softmax with cross-entropy. Softmax normalizes the classification to probability distribution which means transforming the output of the net as the probabilities of one pixel belonging to a class. Cross-entropy loss function acts as a measurement of loss between the probability distribution from softmax and the corresponding ground truth. The smaller the cross entropy, the more alike the

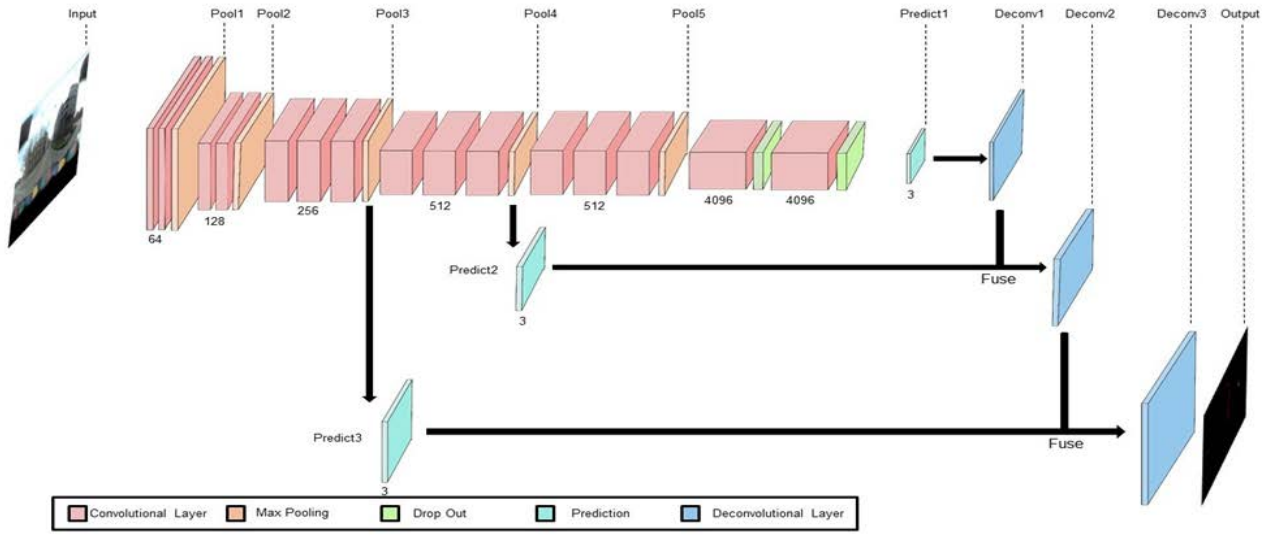


Figure 7. FCN Architecture

probability distribution. The formula of cross-entropy is given as below:

$$\mathrm{CE}(\mathrm{p},\mathrm{q}) = -\sum p(x) log q(x) \quad (9)$$

p(x) is the expected probability that is represented by binary indicator, and q(x) is the predicted probability. The sum is over the all classes.

The model accepts images of any size, therefore we can directly make predictions of our own images from the pre-trained FCN model. In the fine-tuning procedure, considering that our dataset is highly like the training dataset of the pre-trained model and our dataset is small. We decide to only adjust the top layer, such that the generic features and specific features can be kept as much as possible.

3.4 Accuracy Assessment

Performance of the pre-trained FCN model and the fine-tuning are assessed by testing images with annotation. Two metrics will be used to evaluate the semantic segmentation results. Accuracy and IoU (Intersection over Union). The equations are given as below:

$$\text{Accuracy} = TP/(TP + FN)$$
$$\text{IoU} = TP/(TP + FN + FP)$$

Here, TP is the pixel number of true positive which means the correct predicted pixels. FN is the pixel number of false negative which is the unpredicted ground truth pixels. FP is the pixel number of false positive which represents the wrongly predicted pixels. The two metrics are calculated for every class and then averaged.

4. EXPERIMENTS AND RESULTS

4.1 Dataset

4.1.1 Overview

The data used in this study are provided by Delft University of Technology (TU Delft), which consists of 200 panoramic images. The images were captured on TU Delft campus by Ladybug3 panoramic camera of the Fugro Drive-Map mobile laser scanning system. The overall trajectory is shown in Figure 8.

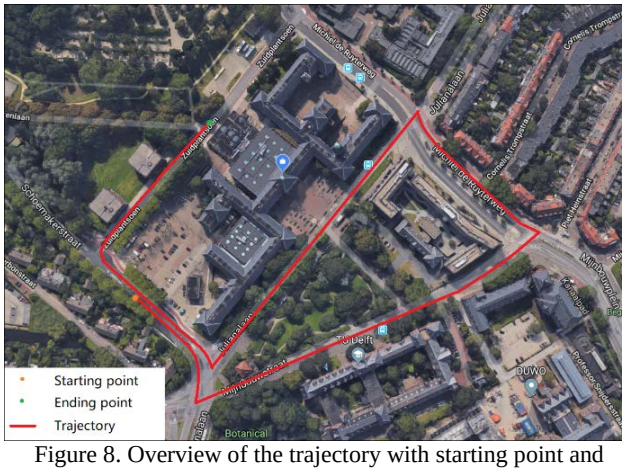


Figure 8. Overview of the trajectory with starting point and ending point.

4.1.2 Annotation

All training and testing images are annotated by MATLAB tool Image Labeler. We label the images to three classes, light poles, traffic signs, and background. Only the first two classes are target objects of interest in this study. The light-poles include the whole pole and the light part on the top, excluding the hanging advertisement boards and traffic signs. And the traffic signs class does not contain the poles that hold them. Figure 9 shows an example of the ground truth. For transformed images, the annotations are also projected to match them.

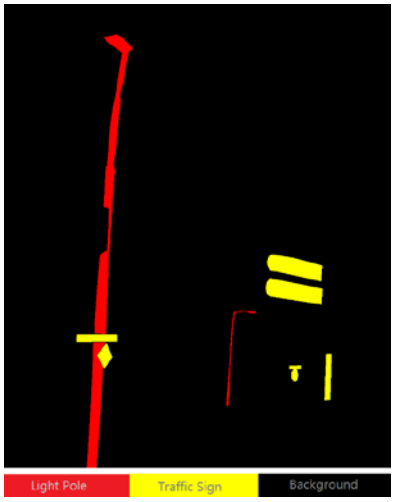


Figure 9. An example of annotation of object of interest. Light poles are in red and traffic signs are in yellow. Background is in black.

4.2 Pre-trained FCN Predictions

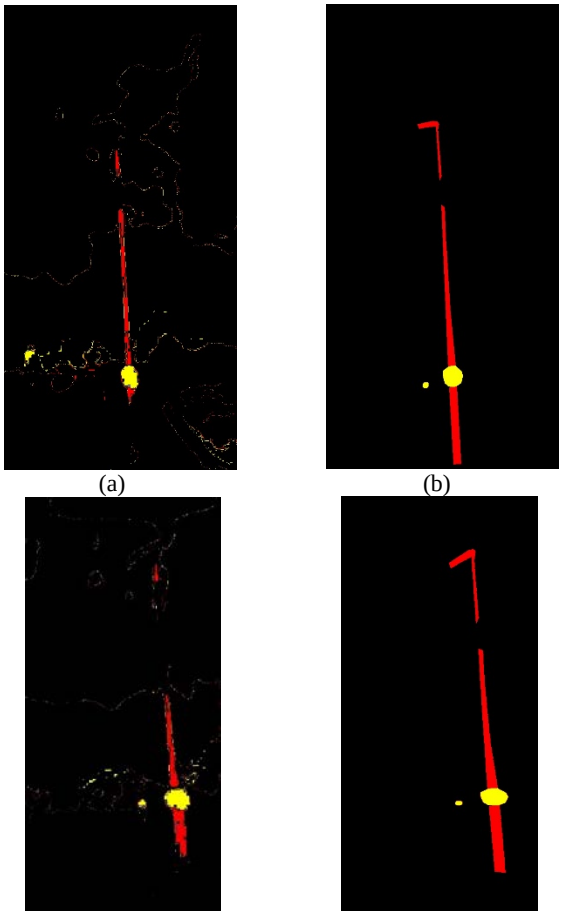
The pre-trained FCN model we use is FCN-8s model trained with cityscapes dataset. Cityscapes dataset is a large-scale street scenes dataset acquired from 50 cities in Germany with dense pixel annotations and it has 30 classes including our needed poles and traffic signs (Cordts et al., 2016). It contains urban scene captured from a car's angle of view which is very similar to our data. The source code is publicly available (Lyu, Vosselman, Xia, Yilmaz, & Yang, 2018).

We directly predict 100 testing images of both panorama and transformed images by the pre-trained FCN model in the Google Cloud Platform with single GPU NVIDIA Tesla K80. The predictions are shown in Figure 10.

It can be seen from Figure 10 that performance of directly predicting from the pre-trained FCN model is not good in both panoramic images and transformed images. Not only the shape is poorly predicted with coarse edges, but also there are many noises in the prediction. In addition, the pole class in the pre-trained model represents various kinds of poles, which makes the predictions consist of many segments of pole class which are not labeled in the ground truth.

4.3 Fine-tuning

In order to make the pre-trained FCN model more appropriate for our dataset and eliminate the noises in the prediction, we modify the last convolutional layer with a new one that the output channel number equal to our class number. The new last layer's weights are initialized randomly and then train the net with the training images. Then the whole net is fine-tuned separately by both panoramic images and transformed images.



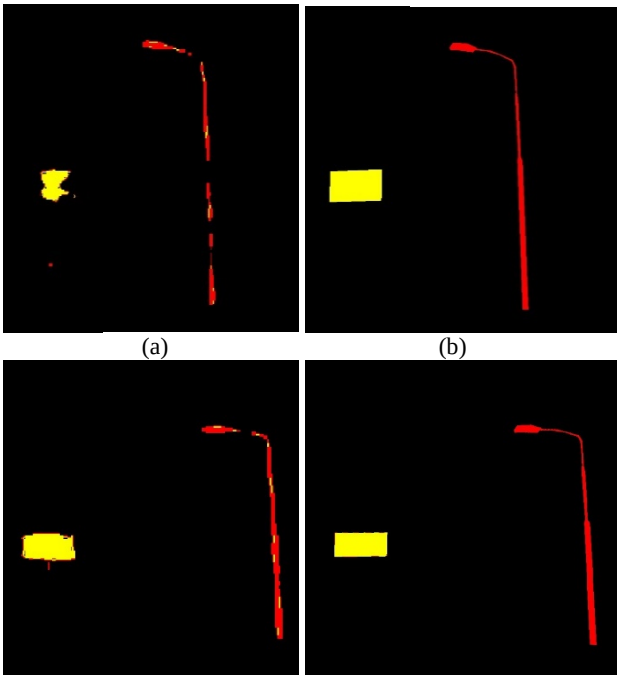
(a) (b) (c) (d)

Figure 10. Predictions of pre-trained FCN model. (a) (b) The predictions and labeled ground truth of panoramic images. (c) (d) The predictions and labeled ground truth of transformed images.

We have cropped the training images and augmented their contrast, hence our training image size decreased, and the quantity increased from 100 to 3200. The change of dataset makes the training images retain the details with no need to resize as well as enlarges the batch size from 1 to 8 under the hardware constraint. The base learning rate is set as $1e^{-4}$. We do not want the weights to update too fast to keep the meaningful information in the original weights from the pre-trained model.

We have trained the model for 20000 interactions, stopped when the loss has become extremely small and nearly do not change anymore. Then with the finetuning model, we predict the testing images again. The predictions in the same location are shown in Figure 11.

Figure 11 presents the details of prediction results of fine-tuning. Comparing the predictions with the corresponding labeled ground truth. Although they are still not smooth enough at the edges, the shapes of segments are very close to their label. The performance of the finetuning are better with transformed images than original panoramic images. The whole image predictions of fine-tuning are shown in the Figure 12. The light-pole is more completely predicted in the transformed image and more traffic signs are predicted in the transformed image than panoramic image.



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

Figure 11. Predictions of fine-tuning. (a) (b) The prediction and labeled ground truth of panoramic images. (c) (d) The prediction and labeled ground truth of transformed images.

4.4 Evaluation

The two results from pre-trained FCN model and fine-tuning are evaluated in aspects of accuracy and IoU, which are given in the Table1 and Table2 respectively.

It can be seen from the accuracy tables that the transformed images are predicted 0.62% more accurate than panoramic images. For light poles class, panoramic images have higher accuracy than transformed images, vice versa in the traffic sign class.

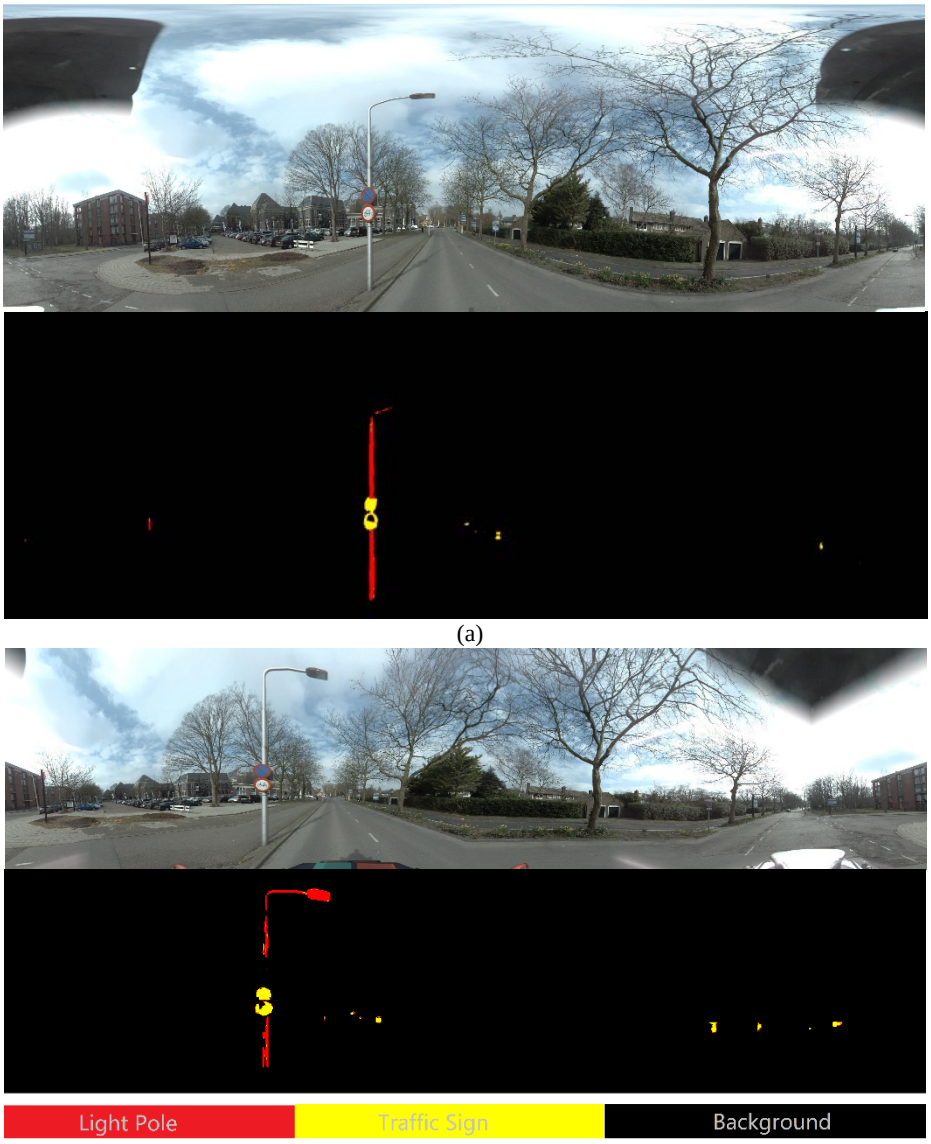
It varies more in IoU for it is a very sensitive metric. The mean IoU of pre-trained model is only several percentage which means there are many unexpected pixels are classified to the two classes including the noises and various poles. After fine-tuning, the mean IoU have increased by 26% and 33% in the prediction results of panoramic images and transformed images. Both light poles and traffic signs have higher IoU in the prediction of transformed images than panoramic images. Both predictions of light poles are better than predictions of traffic signs.

Class	Panoramic Images	Transformed Images
Light Poles	74.06%	68.30%
Traffic Sign	71.24%	78.16%
Average	72.65%	73.23%

Table 1. Accuracy of the results from pre-trained FCN model.

Class	Panoramic Images	Transformed Images
Light Poles	96.72%	94.13%
Traffic Sign	84.62%	88.47%
Average	90.68%	91.30%

Table 2. Accuracy of the results from fine-tuning.



(a)

(b)

Figure 12. (a) The panoramic image and its protections. (b) The transformed image and its predictions.

Class	Panoramic Images	Transformed Images
Light Poles	2.33%	2.26%
Traffic Sign	3.43%	3.88%
Average	2.88%	3.07%

Table 3. IoU of the results from pre-trained FCN model.

Class	Panoramic Images	Transformed Images
Light Poles	38.44%	39.66%
Traffic Sign	20.62%	33.16%
Mean	29.53%	36.41%

Table 4. IoU of the results from fine-tuning.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we compare the semantic segmentation results of pre-trained FCN model and fine-tuning. There are obviously big improvements in the fine-tuning results, which means although the datasets are very alike, the model can not be directly used on a new dataset to produce predictions. At the same time, we transform the panoramic images to perspective images and compare the performance of the state-of-the-art semantic segmentation model FCN implemented on the two very different kinds of images. Panoramic images have worse predicting results than transformed images in both pre-trained model and fine-tuning model. It indicates that the panoramic properties are not very fitted for normal deep learning model or it is because the pre-trained model we use was not trained on a panorama dataset. When apply deep learning model on panoramas, the images need to be pre-processed or the using network needs to be adjusted. We have done the method of pre-processing panoramic images in this paper. And also inspired by the appearance of SphereNet (Coors et al., 2018), modifying the network further by taking the special characteristic into consideration is the direction of our future work.

REFERENCE

Alho, P., Vaaja, M., Kukko, A., Kasvi, E., Kurkela, M., Hyyppä, J., Hyyppä, H., Kaartinen, H. (2011). Mobile laser scanning in fluvial geomorphology: mapping and change detection of point

bars. Zeitschrift Für Geomorphologie, Supplementary Issues, 55(2), 31–50.

Benavidez, P., & Jamshidi, M. (2011). Mobile robot navigation and target tracking system. In 2011 6th International Conference on System of Systems Engineering (pp. 299–304). IEEE.

Bency, A. J., Kwon, H., Lee, H., Karthikeyan, S., & Manjunath, B. S. (2016). Weakly Supervised Localization using Deep Feature Maps. In ECCV 2016 Lecture Notes in Computer Science, vol 9905. Springer.

Cabo, C., Ordoñez, C., García-Cortés, S., & Martínez, J. (2014). An algorithm for automatic detection of pole-like street furniture objects from Mobile Laser Scanner point clouds. ISPRS Journal of Photogrammetry and Remote Sensing, 87, 47–56.

Chen, X., Ma, H., Wan, J., Li, B., & Xia, T. (2017). Multi-View 3D Object Detection Network for Autonomous Driving. CVPR, 1907–1915.

Coors, B., Condurache, A. P., & Geiger, A. (2018). SphereNet : Learning Spherical Representations for Detection and Classification in Omnidirectional Images, In CVPR.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B. (2016). The Cityscapes Dataset for Semantic Urban Scene Understanding. CVPR.

Creusen, I. M., Hazelhoff, L., & de With, P. H. N. (2012). A semi-automatic traffic sign detection, classification, and positioning system, Proceedings of the SPIE, Volume 8305, 83050Y–83050Y–6.

Ess, A., Schindler, K., Leibe, B., & Van Gool, L. (2010). Object Detection and Tracking for Autonomous Navigation in Dynamic Environments. The International Journal of Robotics Research, 29(14), 1707–1725.

Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., & Garcia-Rodriguez, J. (2017). A Review on Deep Learning Techniques Applied to Semantic Segmentation. ArXiv Preprint, 1–23.

Hazelhoff, L., Creusen, I. M., & de With, P. H. N. (2014). Exploiting street-level panoramic images for large-scale automated surveying of traffic signs. Machine Vision and Applications, 25(7), 1893–1911.

Hu He, & Upcroft, B. (2013). Nonparametric semantic segmentation for 3D street scenes. In 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (pp. 3697–3703). IEEE.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84–90.

Krylov, V. A., Kenny, E., & Dahyot, R. (2018). Automatic discovery and geotagging of objects from street view imagery. Remote Sensing, 10(5).

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollar, P. (2017). Focal Loss for Dense Object Detection. In ICCV (pp. 2999–3007). IEEE.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully Convolutional Networks for Semantic Segmentation. In CVPR, 3431–3440.

Lyu, Y., Vosselman, G., Xia, G., Yilmaz, A., & Yang, M. Y. (2018). The UAVid Dataset for Video Semantic Segmentation. ArXiv Preprint.

Marinho, L. B., Almeida, J. S., Souza, J. W. M., Albuquerque, V. H. C., & Rebouças Filho, P. P. (2017). A novel mobile robot localization approach based on topological maps using classification with reject option in omnidirectional images. Expert Systems with Applications, 72, 1–17.

Paparoditis, N., Papelard, J.-P., Cannelle, B., Devaux, A., Soheilian, B., David, N., & Houzay, E. (2012). Stereopolis II: A multi-purpose and multi-sensor 3D mobile mapping system for street visualisation and 3D metrology 3D city modeling. Revue Francaise de Photogrammetrie et de Teledetection 200(200), 69-79.

Pintore, G., Ganovelli, F., Gobbetti, E., & Scopigno, R. (2016). Mobile reconstruction and exploration of indoor structures exploiting omnidirectional images. SIGGRAPH ASIA 2016 Mobile Graphics and Interactive Applications on - SA '16, 1–4.

Pu, S., Rutzinger, M., Vosselman, G., & Oude Elberink, S. (2011). Recognizing basic structures from mobile laser scanning data for road inventory studies. ISPRS Journal of Photogrammetry and Remote Sensing, 66(6), S28–S39.

Rodríguez-Cuenca, B., García-Cortés, S., Ordóñez, C., & Alonso, M. (2015). Automatic Detection and Classification of Pole-Like Objects in Urban Point Cloud Data Using an Anomaly Detection Algorithm. Remote Sensing, 7(10), 12680–12703.

Saxena, A., Min Sun, & Ng, A. Y. (2009). Make3D: Learning 3D Scene Structure from a Single Still Image. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(5), 824–840.

Su, Y.-C., & Grauman, K. (2017). Learning Spherical Convolution for Fast Features from 360° Imagery. In Advances in Neural Information Processing Systems 30.

Wang, J., Lindenbergh, R., & Menenti, M. (2017). SigVox – A 3D feature matching algorithm for automatic street object recognition in mobile laser scanning point clouds. ISPRS Journal of Photogrammetry and Remote Sensing, 128, 111–129.

Weisstein, E. W. (n.d.). Gnomonic Projection. From MathWorld--A Wolfram Web Resource. http://mathworld.wolfram.com/GnomonicProjection.html

Zeng, A., Yu, K. T., Song, S., Suo, D., Walker, E., Rodriguez, A., & Xiao, J. (2017). Multi-view self-supervised deep learning for 6D pose estimation in the Amazon Picking Challenge. In IEEE International Conference on Robotics and Automation (pp. 1386–1393). IEEE.

Zhang, W., Witharana, C., Li, W., Zhang, C., Li, X., & Parent, J. (2018). Using Deep Learning to Identify Utility Poles with Crossarms and Estimate Their Locations from Google Street View Images. Sensors, 18(8), 2484.