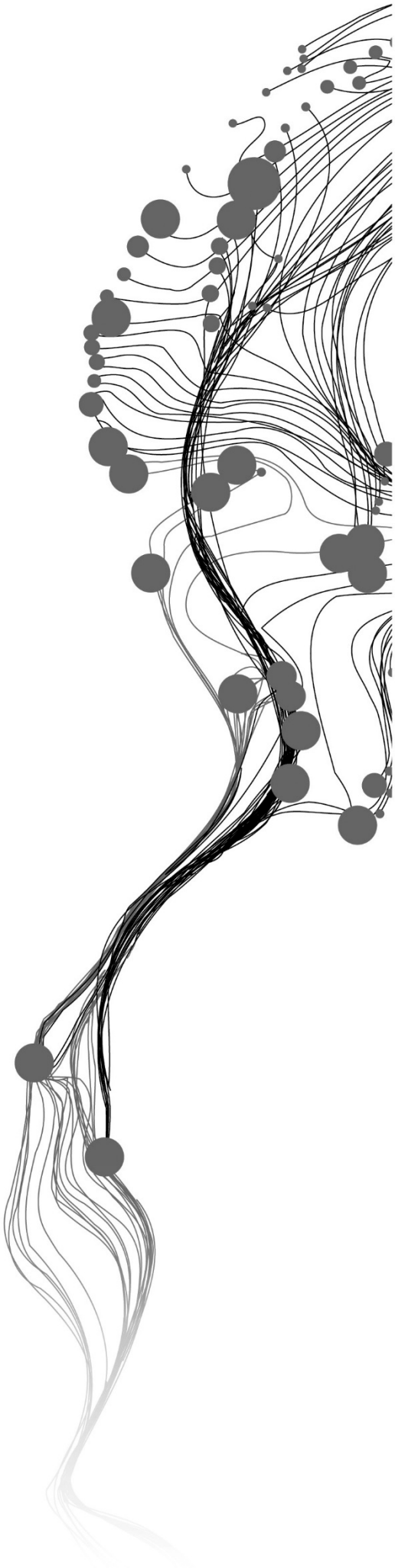
FULLY CONVOLUTIONAL NETWORKS FOR STREET FURNITURE IDENTIFICATION IN PANORAMA IMAGES

YING AO
February 2019

SUPERVISORS:
Dr. Y. Yang
Dr. R. C. Lindenbergh



FULLY CONVOLUTIONAL NETWORKS FOR STREET FURNITURE IDENTIFICATION IN PANORAMA IMAGES

YING AO
Enschede, The Netherlands, February 2019

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.
Specialization: Geoinformatics

SUPERVISORS:
Dr. Y. Yang
Dr. R. C. Lindenbergh

THESIS ASSESSMENT BOARD:
Prof. Dr. ir. M. G. Vosselman
Prof. Dr. techn. F. Rottensteiner, Leibniz Universität Hannover, Institut für Photogrammetrie und Geoinformation, Germany
etc

ABSTRACT

Panoramic images are used increasingly wide in the past years. They provide users with broader viewing angle than normal perspective images and the cost is relatively low, which makes them suitable 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 extract street furniture information from omnidirectional images and the transformed images of them. The transformation is implemented from four directions using Gnomonic projection.

To separate light poles and traffic signs from background, scene understanding methods are needed. In this study, semantic segmentation is performed on the images and the pixel-level task is 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. Then the focal loss function, which is known for solving the class imbalance problem, will be introduced in the FCN model to improve the results.

In the experiment, we use FCN-8s model that pre-trained on cityscape dataset and finetune it by our own data. Then replace cross entropy loss function with focal loss function in the FCN model and train it again to produce the predictions.

The evaluation of predictions shows that in all results from pre-trained model, fine-tuning, and FCN model with focal loss, the transformed images have better performance than panoramic images. And the average accuracy and mean IoU of the results are gradually improved in the three phases.

Keywords

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

ACKNOWLEDGEMENTS

I'd like to appreciate my first supervisor Dr. M. Y. Yang for his beneficial guidance, valuable suggestions, and kind support during my MSc thesis research. And I also want to thank my second supervisor Dr. R. Lindenbergh for the critical comments during the proposal and midterm defenses which helped me progress.

I want to send my sincere gratitude to all teachers in the GFM program, for their wonderful lectures and patient guidance in exercises provide me with sufficient knowledge and operational ability to achieve this study.

I should say thank you to all my friends Yanwen, Yiwen, Shan, Li and my senior Ye, Yaping, Zhenchao as they back up me in my hard time and care for me when I was sick. I'm very glad to have you by my side during the study.

In the end, I want to express my love to my parents. They not only make me worry-free in the materials but also give me mentally supports that encourage me to complete the study in the abroad.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. Motivation and problem statement

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 the object detection field since various kinds of furniture are contributing directly to 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). And urban planners require 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 images, which is infrequently used data for research, to experiment automatically perform the above-mentioned tasks.

Panoramic images are more and more noticeable since their omnidirectional vision with a 360° perspective, 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 et al., 2017) and 3D structure reconstruction (Pintore et al., 2016) for their inner three-dimensional information. But they have been rarely used in the conventional object identification in street views.

In order to achieve street furniture recognition, many approaches have been developed in the past years. Surveying, satellite-based remote sensing, and photogrammetry are worldwide spreading technologies. But for street details, they are not appropriate because they aim at mapping, not object detection. Airborne laser scanning and aerial nadir or oblique imagery are also not suitable for this detection since they must have a centimetric resolution which needs to be acquired from a low altitude with thus high costs due to the multi-flight for overcoming vertical occlusions (Paparoditis et al., 2012). There are mainly two kinds of ways to deal with street furniture detection. One is using 3D point clouds from mobile laser scanning, and the other one is working with 2D images. MLS systems can avoid some occlusions because of mobile characteristic and diverse sensors for different scanning plane. Nevertheless, MLS just consider the objects' spatial characteristics and spatial relations which make it difficult to identify close by and similar objects, e.g. a tree connected to a lamp pole. Also, considering the cost of acquiring MLS data is much more than capturing images from the camera, MLS is not a good choice for this study. For 2D images such as street views and colour images captured from cameras that placed on the vehicles, image segmentation is what researchers 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 algorithms such as support vector machines, conditional random field, and fully convolutional 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). Furthermore, the semantic segmentation results may need refinement, such as applying a post-processing stage using a fully connected pairwise Conditional Random Field (CRF) (L.-C. Chen et al., 2018).
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 (FCN) 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 omnidirectional 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. In addition, we will use the focal loss to deal with the class imbalance problem that the majority of the images is unwanted background. The focal loss function will be introduced in the FCN model and the FCN model will be trained with both datasets again to see if their predictions have been improved.

1.2. Research identification

1.2.1. Research objectives

The study aims at identifying street furniture such as the lamp poles and traffic signs poles along the road from panorama images captured by mobile mapping system. The main aim can be divided into three sub-objectives:

1. Propose a pre-trained FCN model to perform semantic segmentation of the panorama images.
2. To improve the results by fine-tuning the FCN model and applying focal loss function.
3. Compare the predicting results from each step.

1.2.2. Research questions

Sub-objective 1:

- How many classes needed for the semantic segmentation?
- How can the panorama images be used in an FCN model, which means how to deal with the panoramic characteristics in the FCN model?
- Which pre-trained FCN model is the most appropriate for this study?

Sub-objective 2:

- How to fine-tune the classifier with part of the panorama images?
- How to set the appropriate parameters for fine-tuning?
- How to decide the optimal parameters of the focal loss function?

Sub-objective 3:

- Which method is chosen to evaluate the results?
- Are the results improved by finetuning? If so, how much does it improve?
- Are the results improved by adding focal loss function? If so, how much does it improve?

1.2.3. Innovation aimed at

Most of the existing research use only mobile laser scanning data or normal street view data to identify the street furniture, while this study aims at extracting street furniture from panorama images. And the proposed study not only perform the semantic segmentation with a pre-trained FCN model followed by the fine-tuning process but also bring in the focal loss function to solve the class imbalance in the images.

1.3. Research identification

The thesis is organized with seven chapters. Introductions consisting of the motivation, problem statement, and research identification are presented in Chapter 1. Chapter 2 describes the related works of this research. Chapter 3 explained the methodology used in the study. Chapter 4 address the experiment implementation process. Chapter 5 shows the semantic segmentation results. Chapter 6 gives a short conclusion of this research and expected future work.

2. LITERATURE REVIEW

This chapter gives short reviews of related works. At first, the applications of panorama images are briefly described in section 2.1. Then, the various methods for street furniture identification including the use of 3D point cloud data and 2D images are reviewed in section 2.2. Section 2.3 introduces image semantic segmentation and how deep learning techniques are applied to it. In addition, the associated works of Fully Convolutional Networks which is the main approach in this study, and the reviews of loss functions are also explained in this section.

2.1. Applications of panorama image

Panorama images are captured by special cameras or by stitching multiple images in the post-processing period. They present a wide-angled view that is required by many daily life scenarios as well as researches. Back at the end of the century, Pajdla and Hlaváč (1999) presented their research which is taking advantage of panoramic images' rotation invariant representation characteristic to achieve the image based localization. In the following years, Argyros et al. (2001) utilized panoramic images to assist robot returning to its initial location based on corner tracking, and Labrosse (2007) performed robot's long-range navigation by composing a series of short-range homing steps using panoramic images. Until recent years, panoramic images are still popular experimental data in the images based localization and navigation field. Bhongale and Gore (2017) designed an autonomous robot navigation monitoring system benefiting obstacles detection from the 360-degree vision of omnidirectional images.

Another research field has preference for panoramic images for a long time is three-dimensional reconstruction. Sturm (2000) presented a method to reconstruct 3D piecewise planes from single panorama images. Then a whole 3D metric reconstruction of the around scenarios was achieved from multiple omnidirectional images (Mičušík, Martinec, & Pajdla, 2004). Panoramic images with unique broad viewing direction also played an important role in the hot virtual reality field to construct virtual cities (Ikeuchi et al., 2004). And recently, the omnidirectional images were used to improve the performance of Structure from Motion (SfM) (Song, Watanabe, & Hara, 2018).

Although panorama image is frequently appearing in the various studies, it has been rarely used in the object identification field, which makes it worth researching.

2.2. Street furniture identification

For street furniture identification, there are mainly three approaches, 3D data, 2D images and the combination of them. The following subsections briefly review the associated researches of this street-level work by the above three ways.

2.2.1. 3D data

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 scenes structure from monocular images, while Hu He and Upcroft (2013) reconstructed 3D street view from stereo image pairs. However, their works focus on using geometric characteristics within those images. In this respect, the spatial relationships within the images can be easily affected by undesired occlusions or by the change of shooting angle. Therefore, a method that can obtain more integrated spatial information and build the 3D model from variable locations is in need.

Point cloud data is an appropriate choice for acquiring 3d data. It can be generated from Airborne Laser Scanning (ALS), Terrestrial Laser Scanning (TLS) or Mobile Laser Scanning (MLS). ALS is usually used to record the earth surface, like the topography of a large area and it's not accurate enough for small objects. TLS has a static scanning process with ground-based equipment. It can create high-resolution point cloud data of the surrounding environment and objects but lack of mobility, which means it just suits for fixed spots. MLS data which can avoid some occlusions on account of mobile characteristic and diverse sensors for different scanning plane can perform better in street-level work by directly providing accurate spatial information (Alho et al., 2011). Pu et al. (2011) classified the street objects by introducing an initial method which is to roughly classify the on-ground segments and then recognize the road inventories by shape. Cabo et al. (2014) and Wang et al. (2017) pushed forward the extraction of pole-like street furniture from MLS with voxel-based approaches. The difference lies in that the former uses a square cube while the latter uses icosahedron to build the 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. Furthermore, PoingNet is a pioneering work for 3D point cloud classification and segmentation which benefit from deep learning (Schwarz & Behnke, 2017). Nevertheless, MLS mainly consider the objects' spatial characteristics and spatial relationships which make it significantly difficult to identify close by and similar objects (Wang et al., 2017). Also, considering the cost of acquiring MLS point cloud data is much higher than capturing images from the camera, MLS is not an optimal choice for this study.

2.2.2. 2D data

For 2D images such as street views and colour images which are captured by cameras mounted on vehicles, they can be used to recognize street furniture by object detection or image segmentation technique. Although pole-like street furniture like lamp poles with features of narrow and long is hard to be identified only in 2D data, the images are often used to recognize other regular shaped street furniture like traffic signs. Greenhalgh and Mirmehdi (2012) detected road traffic signs by finding candidate regions as maximally stable extremal regions (MSERs) followed by support vector machine (SVM) to classify the signs. But this detection method results in a series of rectangular regions bounding the signs, while our study wants to identify the signs segments. Khan, et al., (2011) proposed an automatic approach to recognize the traffic signs by image segmentation and joint transform correlation (JTC) integrated with shape analysis. The segmentation is implemented using colour feature extraction with a Gabor filter and K-means algorithm to cluster the pixels. This method can segment the correct shape of the signs but sometimes the contents of signs are also separated, which is not what we expect. In our study, we want to separate the whole lamp poles and traffic signs from the background in 2D images.

2.2.3. Combination of 3d data and 2d data

Recently a line of researches combining MLS data and 2D image data to detect street objects have emerged, as sometimes the only one kind of dataset cannot achieve good enough results. Floros and Leibe (2012) accomplished joint 2D and 3D data by machine learning methods to semantic segment the street scenes. They proposed a Conditional Random Field (CRF) based framework that incorporates local 3D street scene geometry information into semantic segmentation algorithm to improve the segmentation quality. Xiao and Quan (2009) proposed an approach to segment the street views with pair-wise Markov Random Field (MRF) across multiple views. They extracted 2D and 3D features at a super-pixel level. Then they trained the classifiers for the unary data terms of MRF. It can be an efficient way to combine the two kinds of data, for 3D data's spatial features and 2D data's colour features can supplement each other in the object identification process.

2.3. Semantic segmentation

Image segmentation is regarded 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). 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).

2.3.1. Fully Convolutional Networks

Nowadays taking advantages of machine learning methods 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) did great work using CNN to classify the large ImageNet dataset, which motivated many successors 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 study, the baseline is also Fully Convolutional Networks but experimented with different data which is panoramic images from mobile mapping system.

2.3.2. Other networks for semantic segmentation

Currently, the networks for semantic segmentation can be considered as an encoder-decoder architecture. The encoder part that produces feature maps is generally a pre-trained network for classification such as VGG-16 and ResNet without their fully connected layers. The difference within the various networks lies on the decoder part that learns to map the features into high-resolution pixel-level segmentation. To achieve semantic segmentation effectively, researchers have proposed their new networks on the basis of CNN and FCN. Ronneberger et al., (2015) published their convolutional networks U-Net, which consists of a contracting path and a symmetric expanding path and forms a U-shape architecture, for biomedical image segmentation. It relies strongly on the use of data augmentation and takes advantage of the feature map from every stage of convolution, leading to effective learning from relatively small dataset. SegNet by Badrinarayanan et al., (2015) is another famous network for semantic segmentation. It uses a novel manner of upsampling the lower resolution feature maps which is to record the pooling indices computed in each max-pooling phase and perform the non-linear upsampling in the decoder part. This operation makes it free from learning to upsample. Following the upsampled maps is convolving with trainable filters and then results in dense feature maps. Other networks including E-Net for real-time segmentation (Paszke, et al., 2016), Pyramid Scene Parsing Network (PSPNet) with four-level pyramid pooling module (Zhao, et al., 2016) and RefineNet that refines high-level semantic features with fine-grained features from earlier convolutions and uses long-range residual connections along the downsampling process (Lin, et al., 2016) all make their own contribution to semantic segmentation development.

2.3.3. Focal loss function

Semantic segmentation can be refined with various methods, such as applying a post-processing stage using a fully connected pairwise Conditional Random Field (CRF) (L.-C. Chen et al., 2018). In this study, it will adjust the loss function to deal with the problem that the majority of the images is unwanted background while usually the networks use cross-entropy loss function. Lin et al. (2017) proposed to

address the foreground-background class imbalance problem by their novel focal loss and evaluate the loss by a designed RetinaNet. Yang et al. (2018) used focal loss function in CNN to detect vehicles. Apart from focal loss function, other loss function such as large margin softmax (L-Softmax) loss function that increases inter-class separability and intra-class compactness (Liu et al., 2016) and center loss function for face recognition (Wen et al., 2016) are also used for improve model performance. In consideration of class imbalance problem lying in this study, the focal loss function will be an appropriate choice.

3. METHODS

In this chapter, the main approaches used in the study will be explained. Section 3.1 presents the method of transformation which projects the panoramic images into normal perspective images. Section 3.2 describes the approaches for image pre-processing. In section 3.3, how the FCN model works for semantic segmentation is discussed.

The proposed street furniture identification workflow is shown in Figure 1. Panoramic images are inputs of the study and they are first 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 produces predictions directly, the next step is semantic segmentation by fine-tuning the FCN model with training images and produces predictions for testing images. Finally, the focal loss function is introduced in the FCN model. The FCN model has been trained again with the same training dataset and does the prediction.

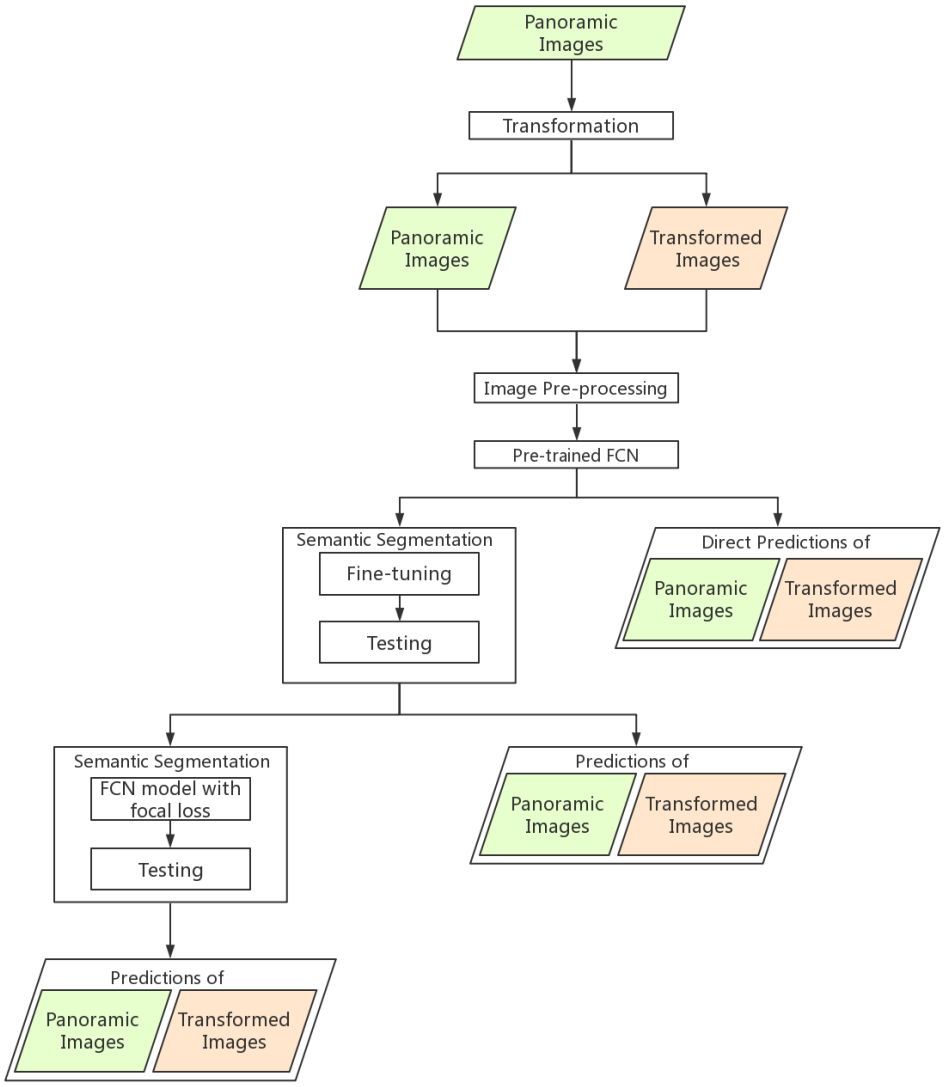


Figure 3.1 Workflow of the proposed street furniture identification method.

3.1. Transformation

The images used in this study are panoramic images which are different from the normal perspective images. The omnidirectional 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 into normal perspective images.

Panoramic images are captured by 360-degree cameras. They are initially spherical images and transformed into planar images by equirectangular projection which is a cylindrical equidistant projection and the output is equidistant along the horizontal and vertical direction (Su & Grauman, 2017). Below in Figure 3.2 shows an example of it.

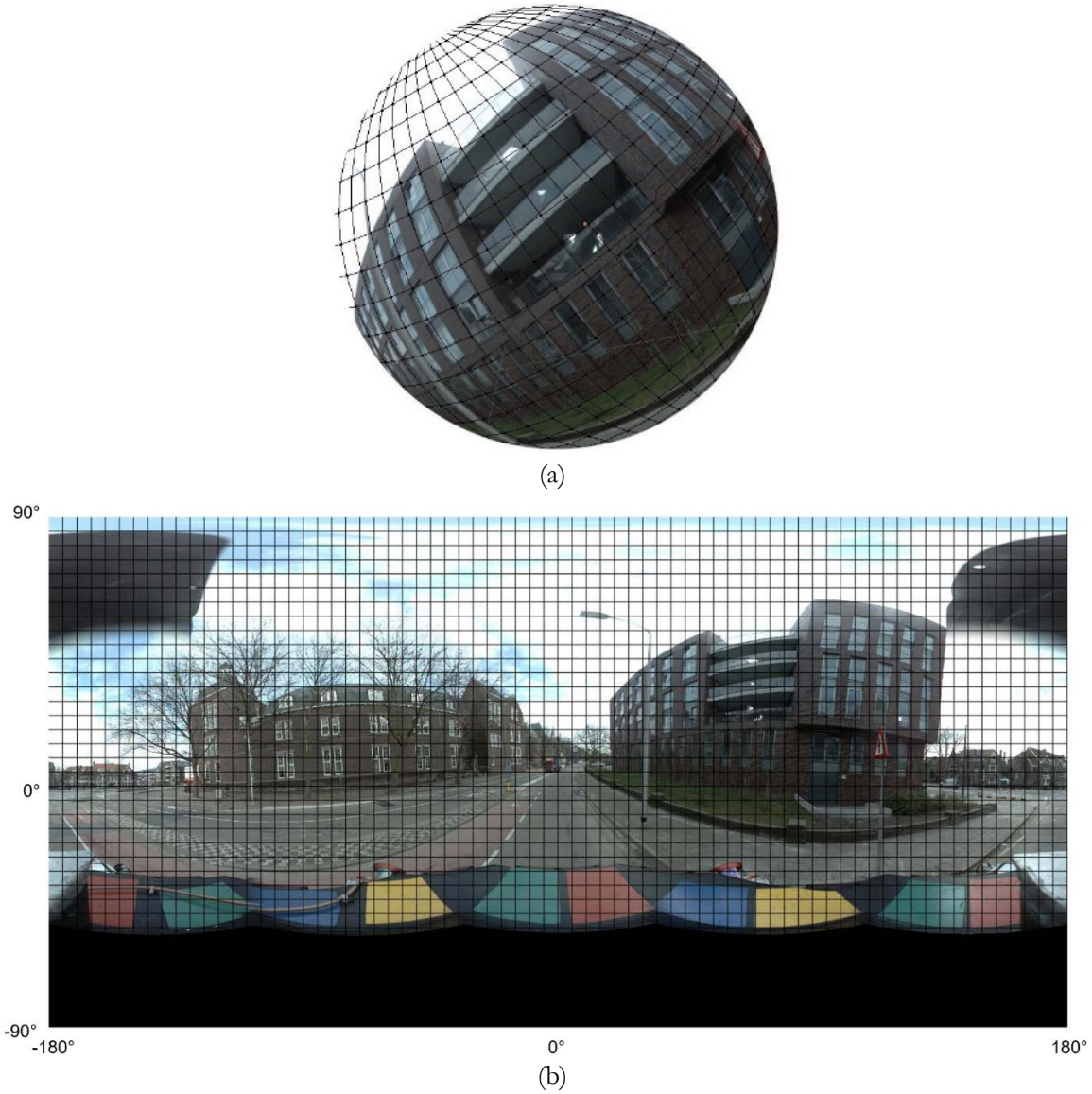


Figure 3.2 An example of projection from spherical image to planar image. (a) A panoramic image on a sphere. (b) A panoramic image. The lines represent latitude and longitude.

We can map the equirectangular images with a simple gnomonic projection for knowing every pixel 's latitude and longitude on the sphere (Coors, Condurache, & Geiger, 2018). The geometry relationship within mapping is based on the graph shown in Figure 3.3.

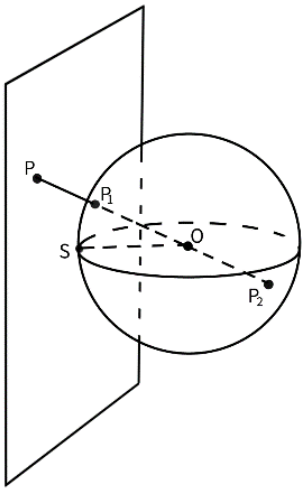


Figure 3.3 Gnomonic projection.

The principal point O is the center of the sphere, and every point on the sphere can be projected to 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 P1 on the sphere. The projection equations (Weisstein, n.d.) are presented as below:

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

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

$$\cos c = \sin\phi_0 \sin\phi + \cos\phi_0 \cos\phi \cos(\lambda - \lambda_0) \tag{3}$$

The transformation equation is for the plane tangent at the point having latitude and longitude (ϕ0, λ0) which in Figure 3 is point S. 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 P1 (ϕ, λ). Therefore, we need to use the inverse equation of the above equation, which is given as below:

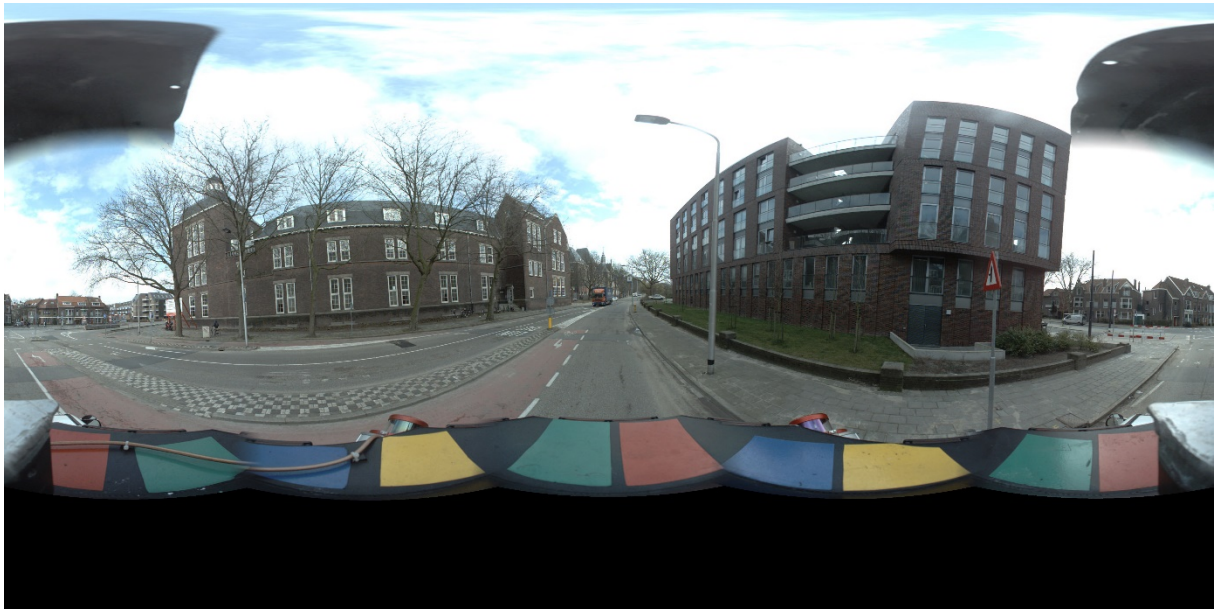
$$\phi = \sin^{-1}(\cos c \sin\phi_0 + \frac{y \sin c \cos\phi_0}{\rho}) \tag{4}$$

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

$$\rho = \sqrt{x^2 + y^2} \tag{6}$$

$$\mathrm{c} = \tan^{-1}\rho \tag{7}$$

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 the 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 regions of interest and the projection of 120 degrees is the best-performing one. The source code of transformation is publicly available (june-choi, 2017). The transformed image is shown in Figure 3.4 (b) and the projection of four directions is margined with red lines.



(a)

(b)

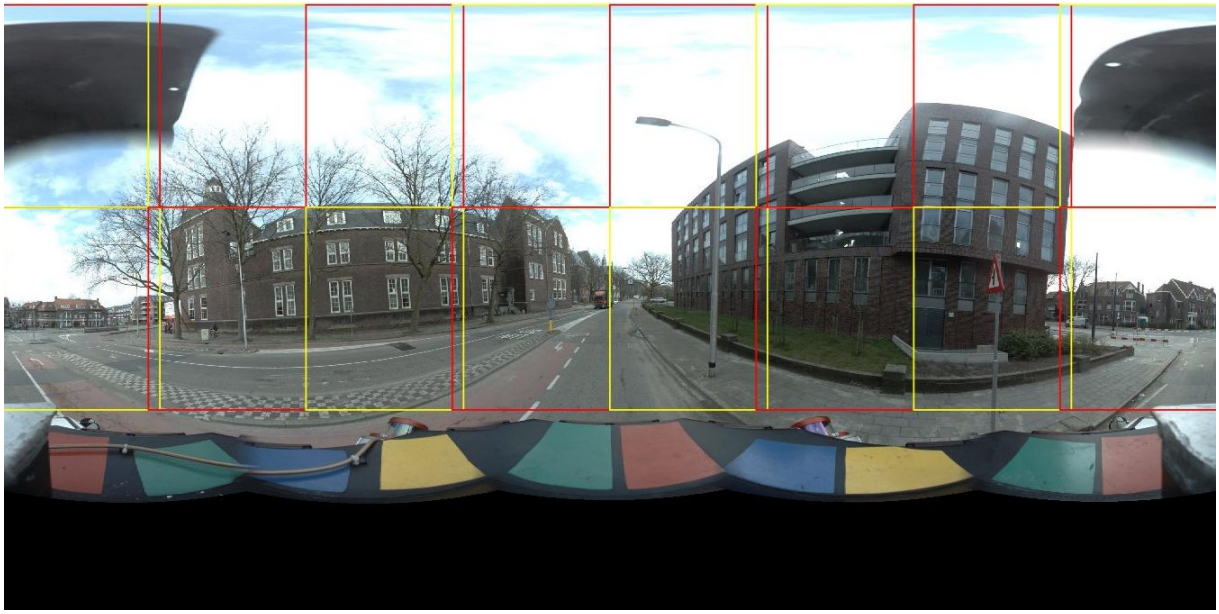
Figure 3.4 Transformation of the panoramic image. (a) The original panorama. (b) The transformed image. The red lines represent the margin of each direction's projection.

3.2. Image pre-processing

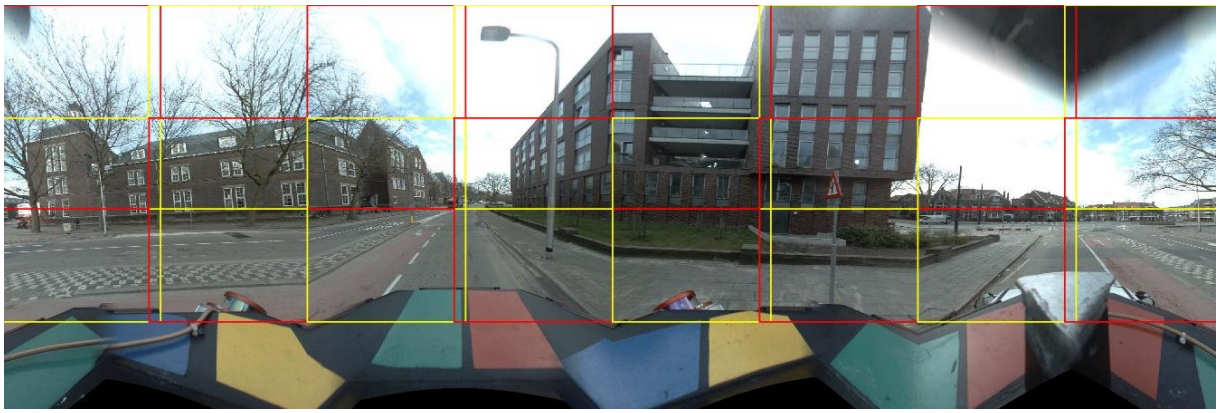
Image pre-processing in this study consists of two aspects, i.e. cropping and data augmentation. The size of the panoramic image is 5400 x 2700 pixels and the size of the transformed image is 5400 x 1800 pixels. Both images are too large to train the model because of our hardware condition constraint, 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.

3.2.1. Image cropping

The size of the cropping image is set as 700 x 900 pixels, which is an appropriate size in consideration of our GPU capability and the remaining semantic information within a single cropped image. Each image in the training data will be cropped into 16 small images. Figure 3.5 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 3.5 Image cropping. (a) Cropping arrangement of the original panoramic images. (b) The transformed images.

3.2.2. Data Augmentation

Image augmentation is applied to the cropped image, with contrast enhancing degree from 1.3 to 1.8, as shown in Figure 3.6. 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 3.6 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. 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, Shelhamer, & Darrell, 2015b). The FCN architecture is shown in Figure 3.7. It contains convolutional layers, max pooling layers, drop out layers, deconvolutional layers. The prediction layers in the figure are convolutional layers with output channel number the 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 probability distribution. The formula of cross-entropy is given as below:

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

p(x) is the expected probability that is represented by binary indicator, and q(x) is the predicted probability. The sum is over 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 similar to the training dataset of the pre-trained model and our dataset is small, we only adjust the top layer. Hence the final output channel fits the class number, and the generic features and specific features can be kept as much as possible.

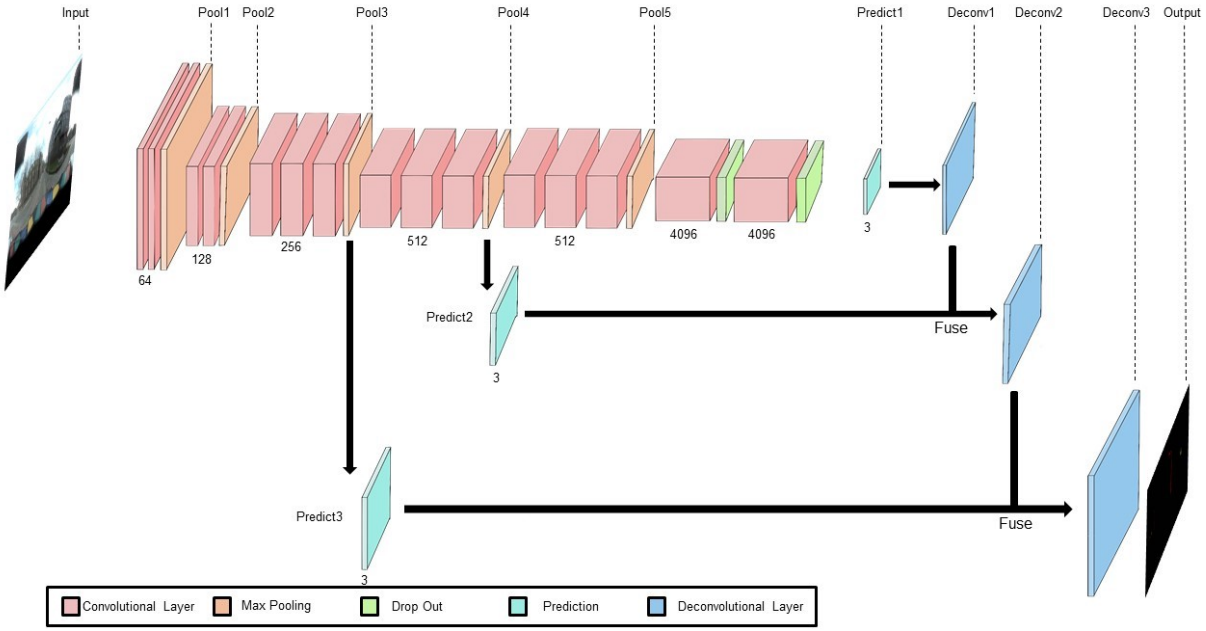


Figure 3.7 FCN Architecture.

3.4. Focal loss function

Focal loss function by Lin et al., (2017) is initially for object detection aiming at improving the accuracy of one-stage detectors and at the same time keep its speed. The authors think it is the class imbalance which leads to the lower accuracy of one-stage detector than the two-stage detector. When there are too many easy negative samples in the training dataset, it takes most of the loss and the model may degenerate. Previously, online hard example mining (OHEM) algorithm is proposed for class imbalance problem, however, it just adds the weight of the hard misclassified samples and ignores the easy well-classified samples (Shrivastava et al., 2016). Therefore, focal loss function taking the two kinds of samples into account is brought up. The formula is shown below:

$$\mathrm{FL}(p_t) = -(1-p_t)^{\gamma}\log(p_t) \tag{10}$$

p_t represents the probability of ground truth class, so that p_t range from 0 to 1. Focal loss function adds a factor $(1-p_t)^{\gamma}$ to the standard cross entropy loss function. Set customized focusing parameter $\gamma > 0$ to reduce the relative loss for well-classified examples, consequently, the model can focus on the hard-classified samples in the training process. In the Figure 3.8, the relationship between loss and probability with different γ is presented.

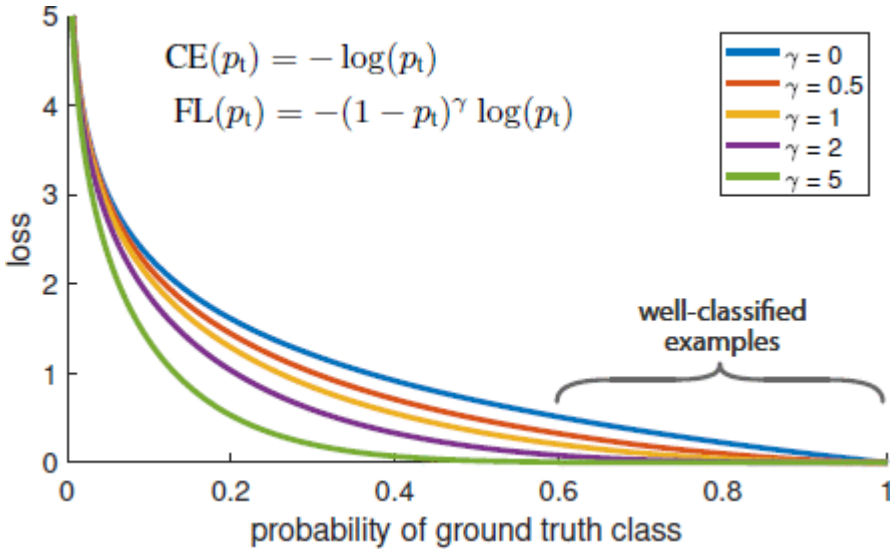


Figure 3.8 Relationship between loss and probability with parameter γ (Lin et al., 2017).

3.5. 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 below:

$$\text{Accuracy} = TP/(TP + FN) \quad (11)$$

$$\text{IoU} = TP/(TP + FN + FP) \quad (12)$$

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

In this chapter, how the experiment is implemented is described. Section 4.1 will show the dataset used in this study including the overview of it and its annotation. Section 4.2 shows how to get predictions from the pre-trained FCN model. Section 4.3 explains how the fine-tuning is working and section 4.4 describes how the focal loss function is playing a role in the FCN model.

4.1. Dataset

4.1.1. Overview

The data used in this study are provided by the Delft University of Technology (TU Delft) consisting 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 including the starting point and ending point are shown in Figure 4.1.

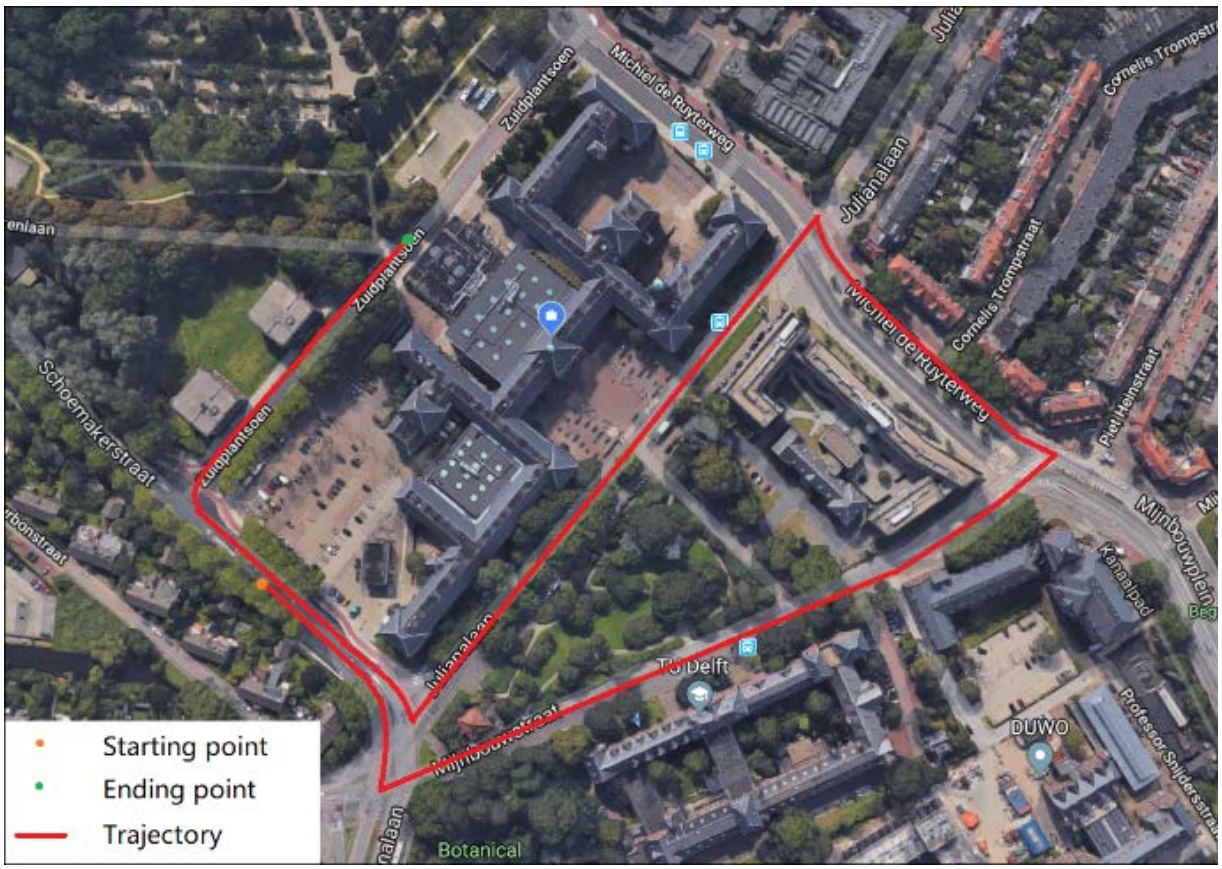
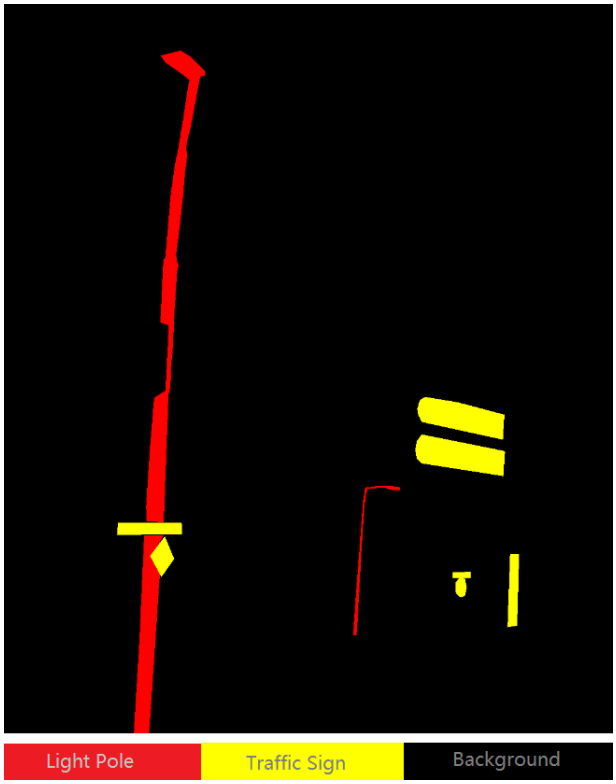


Figure 4.1 Overview of the trajectory with the starting point and ending point.

4.1.2. Annotation

All training and testing images are annotated by MATLAB tool Image Labeller. We label the images to three classes, light poles, traffic signs, and background. Here, only the first two classes are objects of interest in this study. Figure 4.2 shows an example of the panoramic image and its ground truth. For transformed images, the annotations are also projected to match them.

(a)



(b)

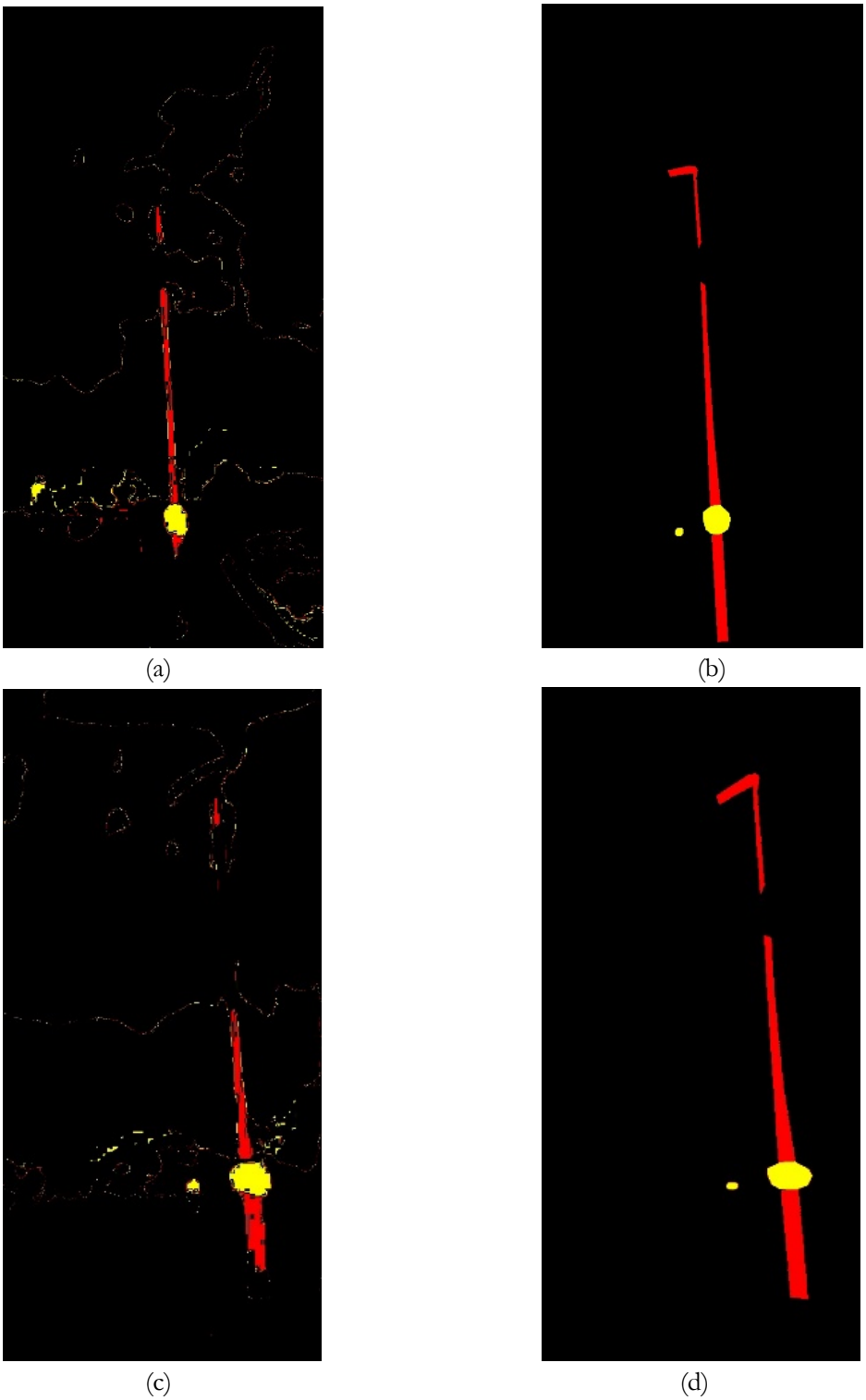
Figure 4.2 An example of annotation of objects of interest. (a) The original image. (b) The labeled ground truth. Light poles are in red and traffic signs are in yellow. The background is in black.

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 4.3.



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

Figure 4.3 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.

It can be seen from Figure 4.3 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 is 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 fine-tune the network by modifying 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.

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.

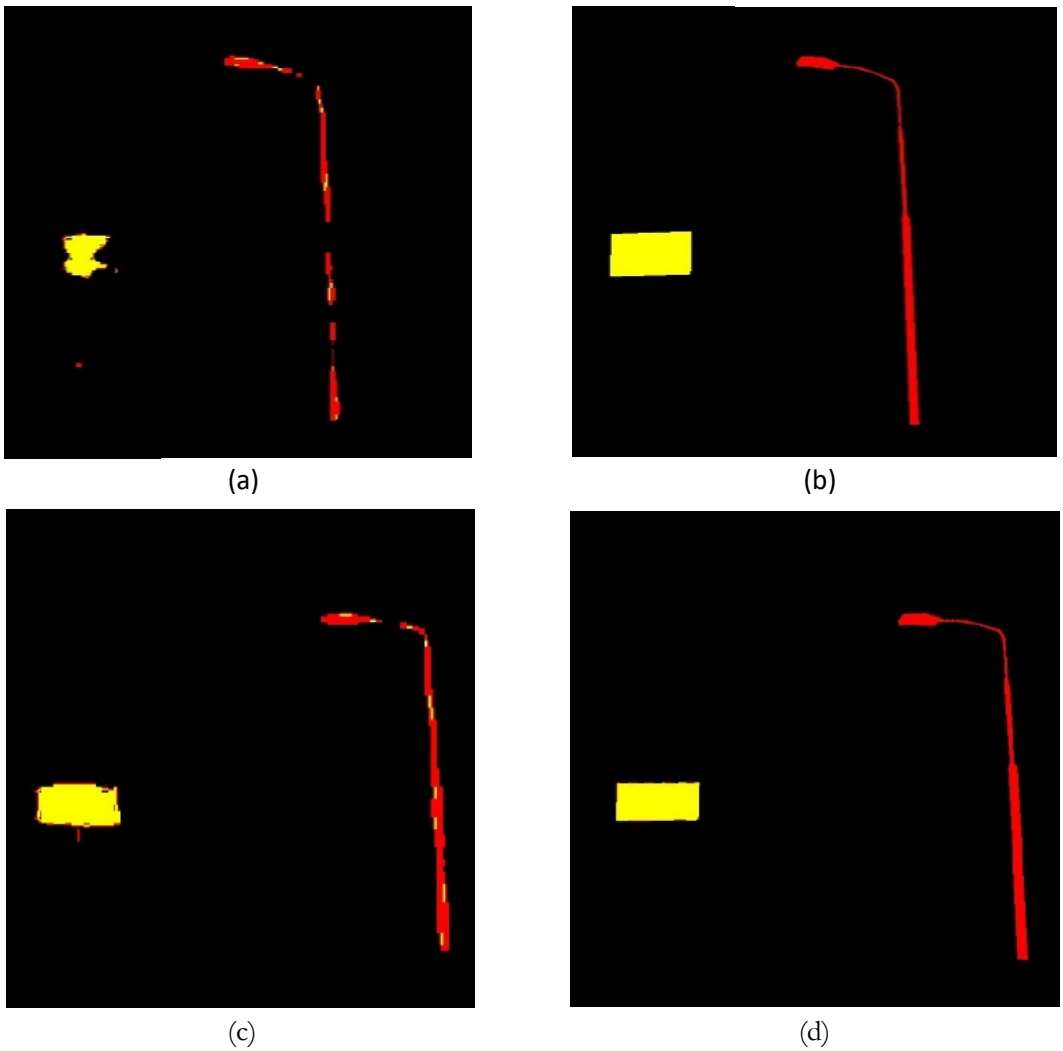


Figure 4.4 Predictions of fine-tuning. (a) (b) The prediction and labeled ground truth of panoramic images. (c) (d) The prediction and labeled ground truth of transformed images.

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 4.4. It presents the details of the prediction results of fine-tuning. Comparing the predictions with the corresponding labeled ground truth. It can be seen that although they are still not smooth enough at the edges, the shapes of segments are very close to their label. And the performance of the fine-tuning is better with transformed images than original panoramic images.

4.4. Focal loss function

The original loss function used in the FCN model is the standard cross entropy loss function. The focal loss function is adding a factor that is $(1-p_t)^{\gamma}$ to it. The $(1-p_t)$ is dependent on the actual computed probability. The customized parameter γ is the most significant determination needed to be made in this step.

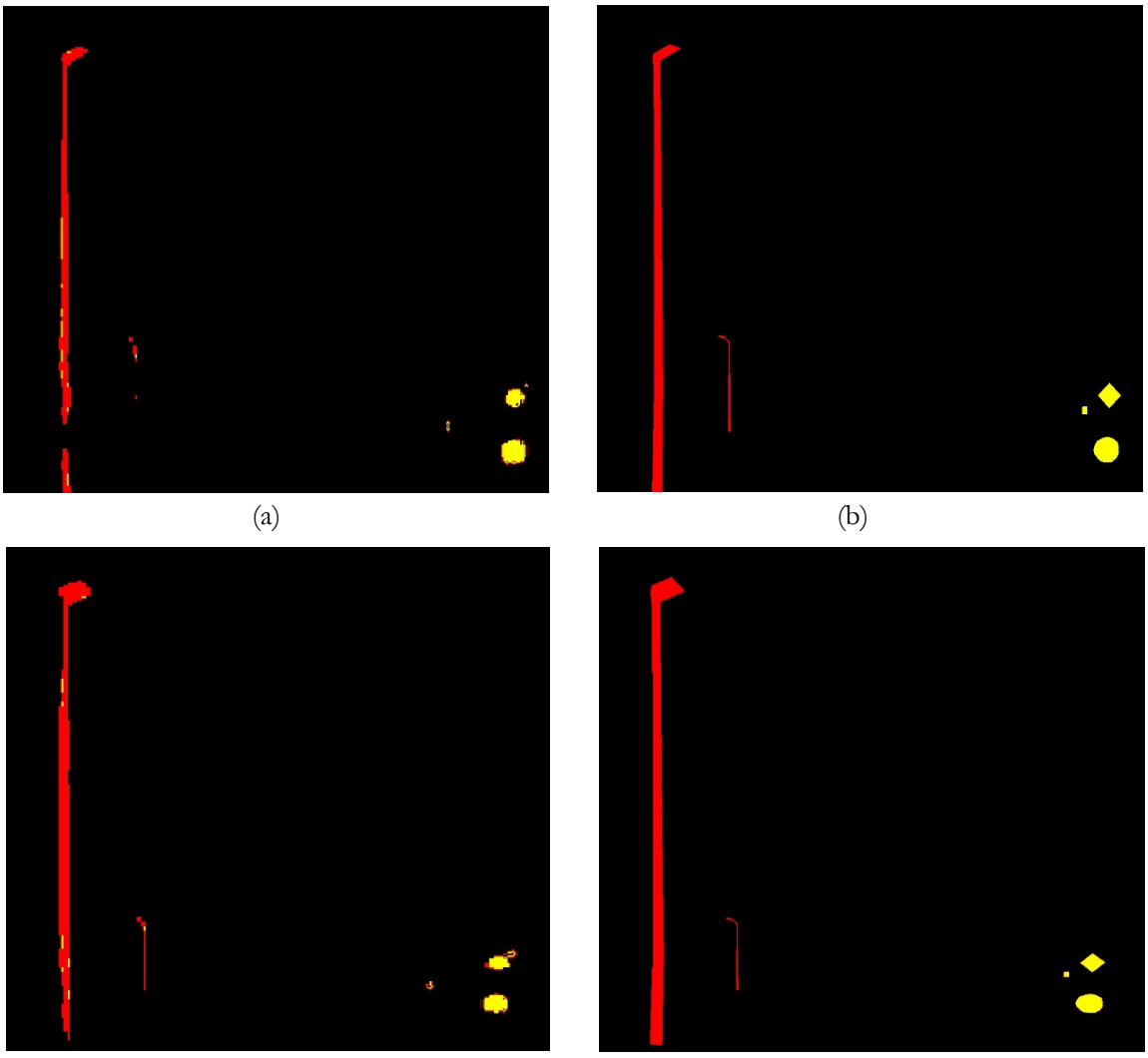
Referring to the initial paper by Lin et al., (2017), the best performance appears when setting γ value as 2. Therefore, it is also the first tried γ value of this study. But this is only the theoretically optimized value of γ, in the experiment, it could change. In consideration of the restricted hardware resources, just limited number of values can be tried for γ and merely the panoramic images will be used to select the best value of γ. When the value of γ is decided, it will be used on the transformed images.

The other two value of γ mentioned in Figure 3.8 will be used to experiment, for they have been chosen as representative γ value in the initial paper. Then the prediction results are compared using mean IoU as shown in Table 4.1.

	γ = 1	γ = 2	γ = 5
Mean IoU	29.43%	34.64%	29.20%

Table 4.1 Mean IoU of panoramic images with different γ.

It can be seen from Table 4.1 that sets γ as 2 will get the best results. Therefore, the FCN model with focal loss function that parameter γ is 2 will be used to train both panoramic images and transformed images. The prediction results can be seen in Figure 4.5.



(a) (b)

(c) (d)

Figure 4.5 Predictions of FCN model with focal loss function. (a) (b) The prediction and labeled ground truth of panoramic images. (c) (d) The prediction and labeled ground truth of transformed images.

The Figure 4.5 presents the predictions and the corresponding labeled ground truth of both panoramic images and transformed images, showing that the objects are identified but with rough edges and the shapes of traffic signs are not precise enough. And still, the transformed images perform better than the panoramic images. It can be seen from the Figure 4.5 that there is a light pole in the left bottom and in the prediction of panoramic image (a) it just has a small part of the pole while in the prediction of transformed image (b) it almost predicts the whole light pole.

5. RESULTS

The three results from pre-trained FCN model, fine-tuning and FCN model with focal loss function are evaluated in aspects of accuracy and IoU. Accuracy is given in Table5.1, Table5.2, and Table5.3 respectively.

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

Table 5.1 Accuracy of the results from the 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 5.2 Accuracy of the results from fine-tuning.

Class	Panoramic Images	Transformed Images
Light Poles	94.36%	94.67%
Traffic Sign	87.58%	85.39%
Average	90.97%	90.03%

Table 5.3 Accuracy of the results from the FCN model with focal loss function.

It can be seen from the first two accuracy tables that the transformed images are predicted a bit more accurate than panoramic images. For light poles class, panoramic images have higher accuracy than transformed images, vice versa in the traffic sign class. In the last accuracy table, the average accuracy of panoramic images is slightly higher than it of transformed images, mostly due to the difference in the accuracy of traffic signs. Overall, fine-tuning the FCN model instead of using it to predict directly improve the semantic segmentation results a lot. However, the focal loss function does not play a role in the accuracy evaluation. There are no increases in average accuracy.
Below presents the IoU in Table 5.4, Table 5.5 and Table 5.6.

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

Table 5.4 IoU of the results from the 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 5.5 IoU of the results from fine-tuning.

Class	Panoramic Images	Transformed Images
Light Poles	36.98%	42.00%
Traffic Sign	32.30%	33.02%
Mean	34.64%	37.51%

Table 5.6 IoU of the results from the FCN model with focal loss function.

It has more changes in IoU for it is a very sensitive metric. The mean IoU of the pre-trained model is only several percentages 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. When focal loss function is added, the mean IoU is improved by 5% and 1% compared to the mean IoU form fine-tuning. Although the IoU of light poles in panoramic images is decreased a little, the IoU of traffic signs is increased up to 12%. For transformed images, IoU of traffic signs is almost the same as it from fine-tuning while IoU of light poles has been raised by 2%. 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.
The whole image predictions of direct prediction from pre-trained FCN model, fine-tuning and FCN model with focal loss function are shown in Figure 5.1 and Figure 5.2 with the related original images.

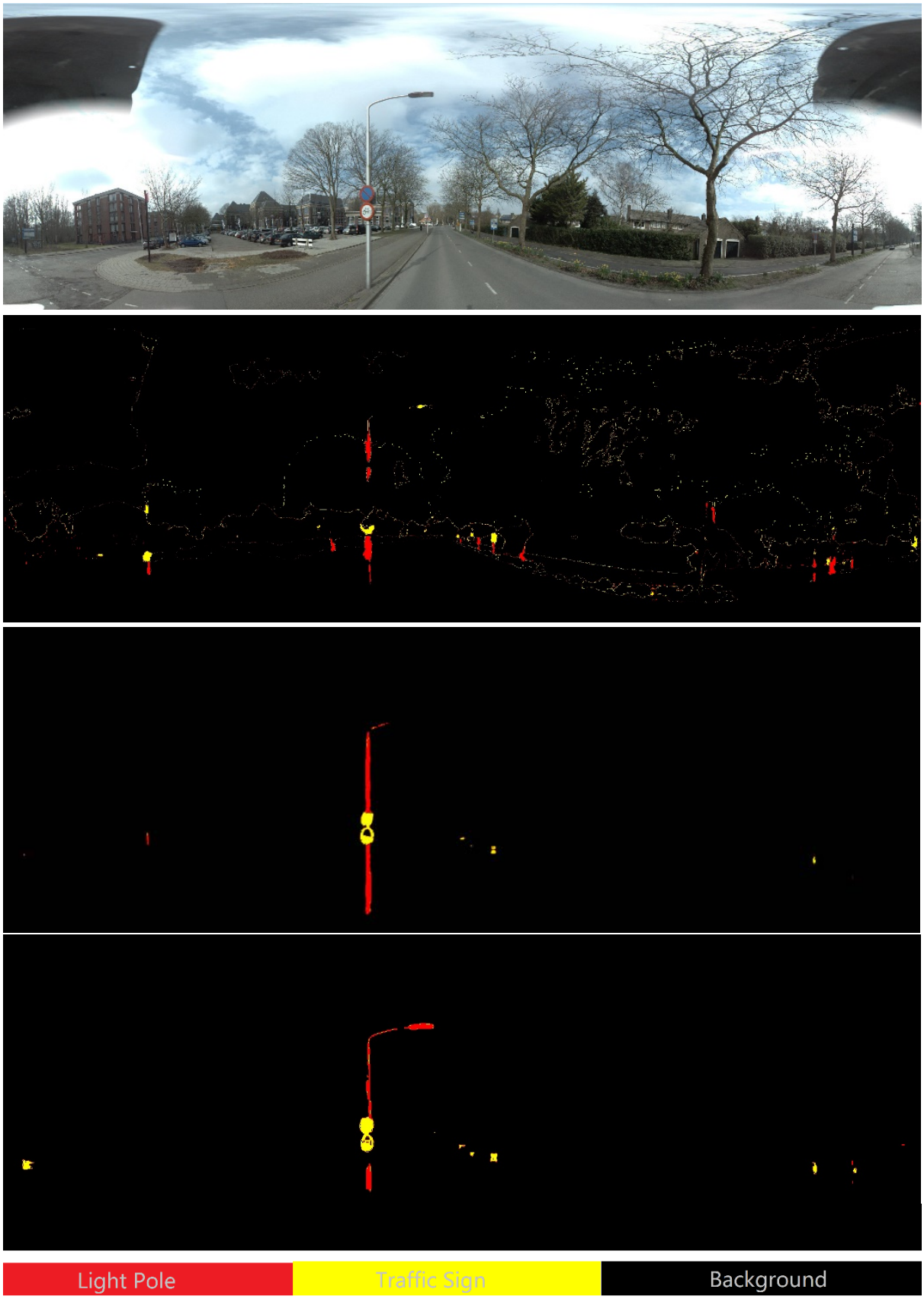


Figure 5.1 The panoramic image and its whole predicting images. The first prediction is directly from pre-trained FCN model. The second prediction is from fine-tuning. The third prediction is from FCN model with focal loss function.

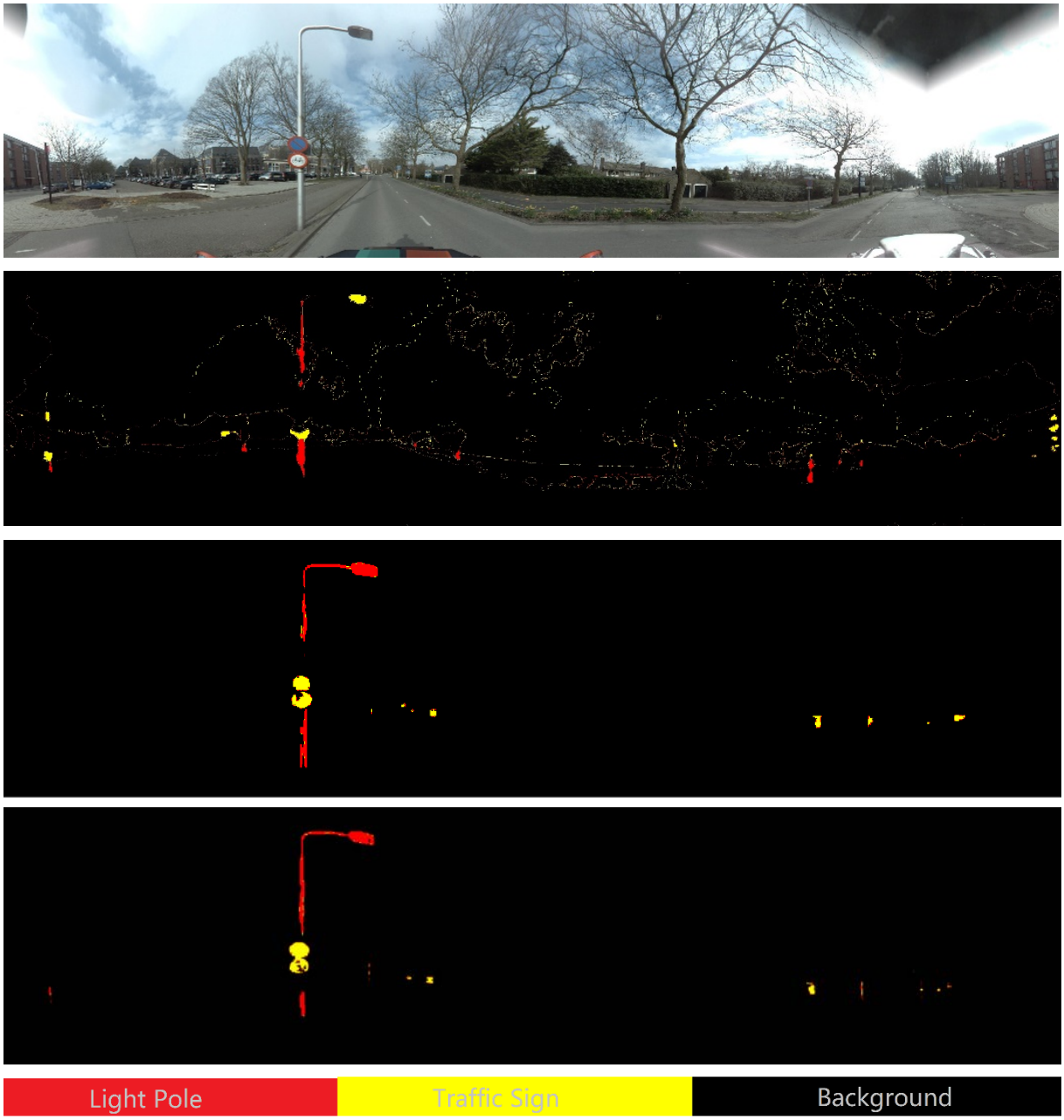


Figure 5.2 The transformed image and its whole predicting images. The first prediction is directly from pre-trained FCN model. The second prediction is from fine-tuning. The third prediction is from FCN model with focal loss function.

6. DISCUSSIONS

Although the workflow of this study has been successfully completed and the three sets of expected results have been achieved, there remain some limitations in the research.

Firstly, the semantic information is missing in the transformation process. Generally, projecting the smaller field of the panoramic image into the perspective image, the better the projection performance. Therefore, in the beginning, several projecting ways have been tried such as 60-degree field of view for each projection. It was found that the smaller field of view leads to more pieces of projected images to be stitched. And because of the different projecting direction, there may be a clear disposition in the seam line. In order to keep both semantic information and projection performance, the 90-degree field of view for projections from four directions is the best choice. Although the disposition problem remains as shown in Figure 6.1, most of the images are keeping complete semantic information of the target objects (light poles and traffic signs).

Figure 6.1 An example of disposition in the seam line.

The second limitation lies in that dataset is not big enough. Although image cropping and data augmentation have made a large increase in the number of images, it is small for fine-tuning. If enrich the dataset by adding newly captured images or using other data augmentation methods, the prediction results of fine-tuning may be improved further.

The parameter setting of the focal loss function is also not perfect. For every independent experiment, the optimized value of the parameter may change a lot, and it needs to attempt patiently. Under the limited time and financial condition, only three values have been tried as the parameter. Therefore, the results got from the last step of this study is relatively good and there maybe exist a more appropriate value for this parameter.

7. CONCLUSIONS

7.1. Conclusions

In this study, we research the approaches to identify the light poles and traffic signs from the panoramic images. We not only explore how to project the panoramic images into normal perspective images but also attempt the different values for the parameter of the FCN model with focal loss to find the appropriate one. Implement semantic segmentation on the panoramic images and transformed images and compare their predictions from the pre-trained FCN model, fine-tuning and FCN model with focal loss.

There are obviously big improvements that the IoU raised by up to 26% and 33% 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. The IoU of predictions from FCN model with focal loss has increased by 5% and 1% in the panoramic images and transformed images.

Overall, panoramic images have worse predicting results than transformed images in the pre-trained model, fine-tuning and FCN model with focal loss. 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 applying deep learning model on panoramas, the images need to be pre-processed or the used network needs to be adjusted.

We have implemented 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.

7.2. Answers to research questions

1. How many classes needed for the semantic segmentation?
   There are three classes for the semantic segmentation. Two of them are the target objects of this research, which are street furniture consisting of light poles and traffic signs. The third class is the background including all other objects.

2. How can the panoramic images be used in an FCN model, which means how to deal with the panoramic characteristics in the FCN model?
   The panoramic images are transformed into normal perspective images from four directions, therefore, the panoramic characteristics are eliminated in the transformed images. Then the two datasets are both used as training dataset for FCN model.

3. Which pre-trained FCN model is the most appropriate for this study?
   The pre-trained FCN model used in this study is based on VGG-16 network and it is trained on the cityscapes dataset, for this dataset comprised by many street view images which the visual angels are highly similar to the used panoramic images.

4. How to fine-tune the classifier with part of the panorama images?
   Half of the panoramic images are used to fine-tune the FCN model. In order to get better fine-tuning results, data pre-processing methods including image cropping and data augmentation that is contrast enhancement are implemented. Taking advantage of the limited number of images to enlarge the training dataset and it effectively improves training performance.

5. How to set the appropriate parameters for fine-tuning?
   Most of the parameters like learning rate use the default value. The modified parameter including batch size and epoch is determined by testing. Gradually increasing the value of batch size until out of memory, find that the biggest value for it is 8. And set a relatively big epoch 100, observing the loss, find that after about 50 epoch the loss is the minimum.

6. How to decide the optimal parameters of the focal loss function?
   At first, three candidate value for γ is chosen from the initial paper of focal loss function. Then used them to train the FCN model with panoramic images and produce predictions. Comparing the mean IoU of the predictions and the one with the biggest mean IoU is decided to be the final choice.

7. Which method is chosen to evaluate the results?
   Intersection over Union (IoU) and accuracy are chosen to evaluate the results.

8. Are the results improved by finetuning? If so, how much does it improve?
   Yes. The mean IoU of results from fine-tuning has been improved by 26% and 33% for panoramic images and transformed images compared with it directly from the pre-trained model.

9. Are the results improved by adding focal loss function? If so, how much does it improve?
   Yes. The mean IoU of results from FCN model with focal loss function has been improved by 5% and 1% for panoramic images and transformed images compared with it from fine-tuning.

LIST OF REFERENCES

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.

Argyros, A. A., Bekris, K. E., & Orphanoudakis, S. C. (2001). Robot homing based on corner tracking in a sequence of panoramic images. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001* (Vol. 2, p. II-3-II-10). IEEE Comput. Soc.

Badrinarayanan, V., Kendall, A., & Cipolla, R. (2015). SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(12), 2481–2495.

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 *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 9905 LNCS, pp. 714–731). Springer Verlag.

Bhongale, K., & Gore, S. (2017). Design of robot navigation monitoring system using image feature analysis and omnidirectional camera images. In *2017 2nd International Conference for Convergence in Technology, I2CT 2017* (Vol. 2017–Janua, pp. 405–409). IEEE.

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, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & 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), 834–848.

Chen, X., Ma, H., Wan, J., Li, B., & Xia, T. (2016). Multi-View 3D Object Detection Network for Autonomous Driving. *Cvpr2017*, 1907–1915.

Coors, B., Condurache, A. P., & Geiger, A. (2018). SphereNet : Learning Spherical Representations for Detection and Classification in Omnidirectional Images. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*(Vol. 11213 LNCS, pp. 525–541). Springer Verlag.

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. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) Workshops* (Vol. 3).

Creusen, I. M., Hazelhoff, L., & de With, P. H. N. (2012). A semi-automatic traffic sign detection, classification, and positioning system, *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.

Floros, G., & Leibe, B. (2012). Joint 2D-3D temporally consistent semantic segmentation of street scenes. In *2012 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2823–2830). IEEE.

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.

Greenhalgh, J., & Mirmehdi, M. (2012). Real-Time Detection and Recognition of Road Traffic Signs. *IEEE Transactions on Intelligent Transportation Systems*, *13*(4), 1498–1506.

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.

Ikeuchi, K., Sakauchi, M., Kawasaki, H., & Sato, I. (2004). Constructing Virtual Cities by Using Panoramic Images. *International Journal of Computer Vision*, *58*(3), 237–247.

Jianxiong Xiao, & Long Quan. (2009). Multiple view semantic segmentation for street view images. In *2009 IEEE 12th International Conference on Computer Vision* (pp. 686–693). IEEE.

Khan, J. F., Bhuiyan, S. M. A., & Adhami, R. R. (2011). Image Segmentation and Shape Analysis for Road-Sign Detection. *IEEE Transactions on Intelligent Transportation Systems*, *12*(1), 83–96.

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).
Labrosse, F. (2007). Short and long-range visual navigation using warped panoramic images. *Robotics and Autonomous Systems*, *55*(9), 675–684.
Lin, G., Milan, A., Shen, C., & Reid, I. (2016). RefineNet: Multi-Path Refinement Networks for High-Resolution Semantic Segmentation. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017* (Vol. 2017-January, pp. 5168–5177). Institute of Electrical and Electronics Engineers
Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollar, P. (2017). Focal Loss for Dense Object Detection. In *2017 IEEE International Conference on Computer Vision (ICCV)* (pp. 2999–3007). IEEE.
Liu, W., Wen, Y., Yu, Z. & Yang, M.. (2016). Large-Margin Softmax Loss for Convolutional Neural Networks. *Proceedings of The 33rd International Conference on Machine Learning, in PMLR* 48:507-516.
Long, J., Shelhamer, E., & Darrell, T. (2015a). Fully Convolutional Networks for Semantic Segmentation. *Computer Vision and Pattern Recognition (CVPR)*, 3431–3440.
Lyu, Y., Vosselman, G., Xia, G., Yilmaz, A., & Yang, M. Y. (2018). The UAVid Dataset for Video Semantic Segmentation. *Arxiv.org*, *2018*, 1-9.
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.
Mičušík, B., Martinec, D., & Pajdla, T. (2004). 3D Metric Reconstruction from Uncalibrated Omnidirectional Images. *Asian Conference on Computer Vision*, (02), 28–30.
Pajdla, T., & Hlaváč, V. (1999). Zero phase representation of panoramic images for image based localization. In *Computer Analysis of Images and Patterns* (pp. 838–838). Springer, Berlin, Heidelberg.
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.
Paszke, A., Chaurasia, A., Kim, S., & Culurciello, E. (2016). ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation. *arXiv:1606.02147*.
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.
Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 9351, pp. 234–241). Springer, Cham.
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.
Schwarz, M., & Behnke, S. (2017). Data-efficient Deep Learning for RGB-D Object Perception in Cluttered Bin Picking. *Warehouse Picking Automation Workshop (WPAW), IEEE International Conference on Robotics and Automation (ICRA)*, (May), 2–4.
Shrivastava, A., Gupta, A., & Girshick, R. (2016). Training Region-based Object Detectors with Online Hard Example Mining. *arXiv:1604.03540 [cs.CV]*.
Song, M., Watanabe, H., & Hara, J. (2018). Robust 3D reconstruction with omni-directional camera based on structure from motion. In *2018 International Workshop on Advanced Image Technology, IWAIT 2018* (pp. 1–4). IEEE.
Sturm, P. (2000). A method for 3D reconstruction of piecewise planar objects from single panoramic images. In *Proceedings - IEEE Workshop on Omnidirectional Vision, OMNIVIS 2000* (pp. 119–126).
Su, Y.-C., & Grauman, K. (2017). Learning Spherical Convolution for Fast Features from 360 Imagery. *NIPS*.
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.

Wen, Y., Zhang, K., Li, Z., & Qiao, Y. (2016). A discriminative feature learning approach for deep face recognition. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 9911 LNCS, pp. 499–515). Springer, Cham.

Yang, M., Liao, W., Li, X., & Rosenhahn, B. (2018). Vehicle Detection in Aerial Images. In *IEEE International Conference on Image Processing (ICIP), 2018*.

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 *Proceedings - 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.

Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2016). Pyramid Scene Parsing Network. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2881-2890.*