A PILOT STUDY OF URBAN POI MAPPING USING CROWDSOURCED STREET-LEVEL IMAGERY AND DEEP LEARNING

Lanfa Liu [1,2], Baitao Zhou [1,2], Xuefeng Yi [3]*

[1] Hubei Provincial Key Laboratory for Geographical Process Analysis and Simulation, Central China Normal University, 430079 Wuhan, China - lanfa@ccnu.edu.cn, baitaozhouex@gmail.com

[2] College of Urban and Environmental Sciences, Central China Normal University, 430079 Wuhan, China

[3] School of Earth Sciences and Engineering, Hohai University, 211100 Nanjing, China - hhuyxf@sina.com

Commission IV, WG IV/4

KEY WORDS: Crowdsourced Data, Street-Level Imagery, Object Detection, Point of Interest, Deep Learning.

ABSTRACT:

Point-of-interest (POI) data contains rich semantic and spatial information, having a wide range of applications including land use, transport planning and driving navigation. However, urban POI mapping traditionally requires a lot of manpower and material resources, which only few institutions or enterprises can afford to. With the increasing amount of street-level imagery, it is possible to directly extract POI-related information from such data and automatically map the distribution of urban POIs. In the pilot study, we mainly focused on extracting POIs from billboards in street-level imagery. Firstly, the you only look once (YOLO) algorithm was considered to locate billboards in the imagery, then an optical character recognition (OCR) model was adopted to extract POI-related semantic information from the detected billboard, and finally the extracted semantic text was further processed to obtain POI results. The preliminary study shows that it is a promising way of mapping urban POIs from crowdsourced street-level data using deep learning techniques.

1. INTRODUCTION

Points of interest (POIs) refer to places that people may find useful or interesting, such as restaurants, touristic places, supermarkets (Touya et al., 2017). POI data contain rich semantic information, making it play a key role in applications of location-based services such as food delivery and driving navigation. It is possible to depict urban street landscape in the format of POI-related texts and are important complementary data for various urban studies including identifying urban building function (Lin et al., 2021). POIs can be used to geo-reference social activities, such as sending a geo-tagged tweet (Rösler and Liebig, 2013). It also plays a key role in navigation, such as providing a reference point for travel directions (Duckham et al., 2010) or being part of a personalized itinerary (Lim et al., 2015).

POIs are typically collected and maintained by large commercial companies such as Baidu, Google and AutoNavi. They are valuable asset to support many kinds of activities. However, the workload of commercial data collection is heavy and the number of required staff is often insufficient. POIs are a major part of the contributions of crowdsourcing projects such as OpenStreetMap (Yang et al., 2018). With the advancement of technology, artificial intelligence continues to develop and has achieved great success in computer vision and natural language processing, which makes it possible to automatically extract POIs from the street-level imagery using deep learning techniques, including object detection, optical symbol recognition, and semantic segmentation methods.

POI-related information is often contained in the billboard, and it is vital to locate the billboard in street-level imagery. In fact, It is possible that multiple billboards exist in a picture and the picture background is also complex, making it a challenge to delineate the position of multiple billboards. Object detection algorithms have the capability to locate objects with a bounding box in an image and identify each object. Currently, there are two strategies for object detection using deep learning techniques. The first one is two-stage object detection, which firstly generates a regional proposal from the image, and then uses a convolutional neural network model to classify the candidate region and generate the final object frame, such as SPP-Net (Purkait et al., 2017), Faster R-CNN (Ren et al., 2015), Mask R-CNN (He et al., 2017). This strategy often has higher accuracy. The second is one-stage object detection, which directly generates the category probability and position coordinate value of the object. After a single detection, the final detection result can be directly obtained, and the detection speed is faster, such as SSD (Konishi et al., 2016), YOLO (Bochkovskiy et al., 2020; Redmon et al., 2016; Redmon and Farhadi, 2018, 2017). Especially, YOLO algorithms are known as enabling end-to-end training and real-time speed while maintaining high average precision. The latest yolov5 algorithm also inherits the advantage, and has a better performance than previous ones.

Scene text recognition could be considered to extract POI scene text, which could be painted in the pictures or signboards, from POI-related objects detected on street-level imagery. As scene text tends to have different scales and shapes, including horizontal text, multi-directional text and curved text. Therefore, some effective methods are needed to achieve scene text detection. With the prediction results at the pixel-level, segmentation-based scene text detection (Husnain et al., 2019) can describe the text of various shapes. However, most segmentation-based methods require complex post-processing such as PSENet (Wang et al., 2019) and Pixel embedding (Tian et al., 2019), which result in a huge time cost in the inference procedure. To solve the problem, Differentiable Binarization

* Corresponding author: hhuyxf@sina.com

(DB) (Liao et al., 2020) was proposed, which can perform the binarization process in segmented networks. This method achieved state-of-the-art results at the time, in terms of both detection accuracy and speed. In addition, a powerful classifier is also necessary to achieve accurate recognition of POI scene text. At present, due to the great success of deep learning in the field of computer vision, many researches apply deep learning technology to Optical character Recognition (OCR), which achieves better results and higher accuracy compared with other techniques such as ANN and SVM (Bhatt and Patel, 2018). CRNN(Fu et al., 2017) is a general deep learning framework, which integrates the advantages of both Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). Not only can it achieve better results in scene text recognition, but it can also be applied to other domains such as Chinese character recognition. PaddleHub, a toolkit developed by Baidu, provides a pre-trained model that applies DB to CRNN networks, which can learn directly from text word-level or line-level annotations without requiring detailed character-level annotations to achieve text recognition. It achieves good results on the ICDAR dataset. The extracted information could be out of order or have invalid information. Natural language processing (NLP) methods should be further considered to process extracted information to obtain the POI data that meets the specification.

With the availability of crowdsourced street-level imagery, it is possible to extract POIs from these data, which has been few studied. In the present work, we mainly focused on how to generate POIs from street-level imagery. Firstly, the billboard in the imagery was detected via a YOLO algorithm. Then an optical character recognition model was considered to recognize scene text containing POI information from the billboard. Natural language processing methods will be further considered to generate accurate POI data from the text.

2. DATA AND METHODOLOGY

2.1 Data

The data provided by CCF Big Data & Computing Intelligence Contest (CCF BDCI) was considered for the pilot study of urban POI mapping. It contains 26,468 street-level images collected from cities in China, and several examples were shown in Figure 1. All images were labelled with the position of billboards and corresponding POI information. The dataset was divided into two parts for training and testing, having 20,769 and 5,699 samples, respectively.

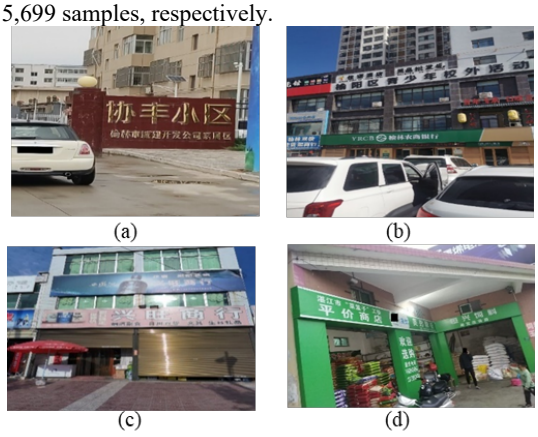

Figure 1. Examples of street-level imagery containing POI information. (a) residual area; (b) bank; (c) and (d) shops.

2.2 Methodology

This paper presents a pilot study of POI mapping using urban street-level crowdsourced data. Specifically, we first trained from scratch on YOLOv5 loaded with the yolov5x profile using 20769 training data. Then a pre-trained OCR model was adopted to perform scene text recognition and extraction on the billboard detected by YOLOv5. Finally, data cleaning was performed to obtain POI data. The overall workflow of the presented work was shown in Figure 2.

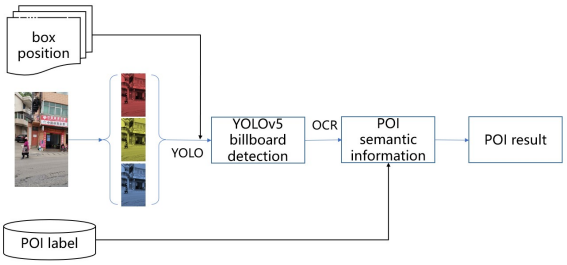

Figure 2. The workflow of urban POI mapping using the street-level imagery.

2.2.1 YOLO: YOLO is a popular object detection algorithm because of its speed and accuracy. YOLOv5 was released by Ultralytics in May 2020, which inherited the advantages of YOLOv4, namely adding SPP-Net (Figure. 3) and putting forward new data enhancement methods (Wu et al., 2021). With the use of SPP-Net in the YOLOv5 framework, YOLOv5 is able to use sparse feature synthesis to augment training data and improve generalization performance to unseen data. The data loader can perform three kinds of data enhancements, including scaling, color space adjustment, and mosaic enhancement. Moreover, the anchor mechanism of Faster R-CNN is utilized to strengthen the ability of the YOLOv5 algorithm to small target detection in the image through a multi-scale mechanism in the process of image detection. In addition, it provides the YOLOv5 algorithm with high adaptability to different sizes of images. Due to the varying size of billboards in crowdsourced images, YOLOv5 was adopted to perform the task of billboard detection on the street-level imagery. Figure 4 illustrates the network structure of the YOLO algorithm.

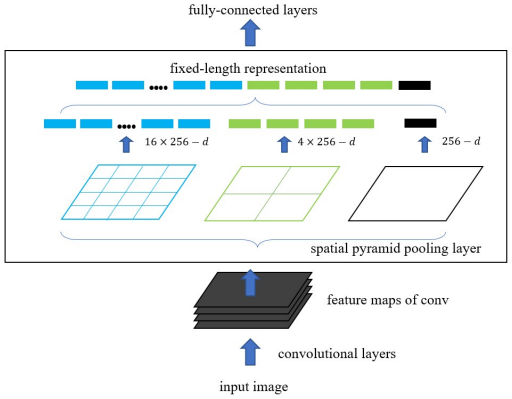

Figure 3. the SPP-net algorithm.

The network structure of YOLOv5 is mainly divided into three parts. The first part in the YOLOv5 architecture uses cross stage partial network (CSPNet) (Wang et al., 2020) as its backbone to extract the features from the input image. With the use of CSPNet architecture, the number of model parameters and floating-point operation per second (FLOPS) are reduced, as gradient change information in large-scale backbones is integrated into the feature map. By doing that, it not only

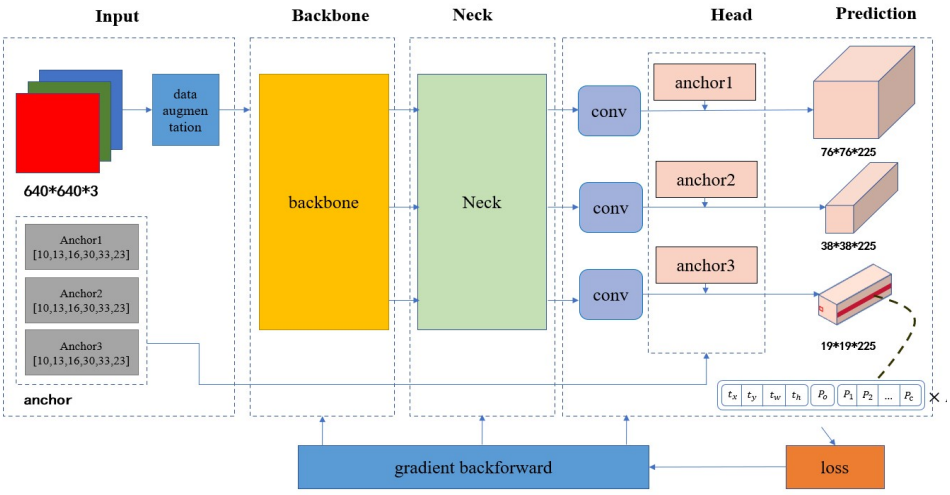

Figure 4. The network structure of the YOLO algorithm.

ensures the improvement in the inference speed and accuracy, but also reduces the overall model size.

After the feature extraction part, the data is augmented with a Path Aggregation Network (PANet) (Liu et al., 2018). PANet adopts an improved Feature Pyramid Network (FPN) structure, which shortens the information path between the bottom-level and top-level features. At the same time, the adaptive feature pool method is used to aggregate the feature grid and each feature layer, so that the useful information in each feature layer is directly propagated to subsequent sub-networks. Since underlying features are effectively utilized and propagated between layers, the accuracy of small object detection is improved, which also makes the model perform well on unseen data.

The last part is the detection part, which utilizes three different sized YOLO layers of the feature map to achieve multi-scale prediction, enabling the model to efficiently handle small, medium and large objects.

The YOLO algorithm divides the image involved in detection into n × n grids, among which each grid has different detection tasks. The whole network structure is composed of two full connection layers and 24 convolution layers. After the full connection layer, the tensor of n × n × (B × 5 + C) is output, in which B represents the number of predicted targets in each grid, and C denotes the number of categories. The final detection result can be obtained by regressing the detection box position and judging the category probability of the tensor data. The YOLO-based method performed not well on small targets as there tends to be multiple targets in the same grid without detailed grid division.

2.2.2 Pre-trained OCR model: For the sequence recognition problem, the most suited neural networks are RNN (Medsker and Jain, 1999), while for an image-based problem, the most suited are CNN. To recognize POI-related text in the billboard, it would be better to combine CNNs and RNNs. DB is a popular segmentation-based framework for detecting arbitrary-shape scene text (Liao et al., 2020). Optimized along with a DB module, a segmentation network can adaptively set the thresholds for binarization, which not only simplifies the post-processing but also enhances the performance of text detection. DB and CRNN were integrated as the solution for OCR and an end-to-end pretrained model was provided by PaddleHub, which was trained on ICDAR 2015 dataset. The overall framework of the pretrained model is shown in Figure 5. It consists of three components, including the convolutional layers, the recurrent layers, and a transcription layer, from bottom to top.

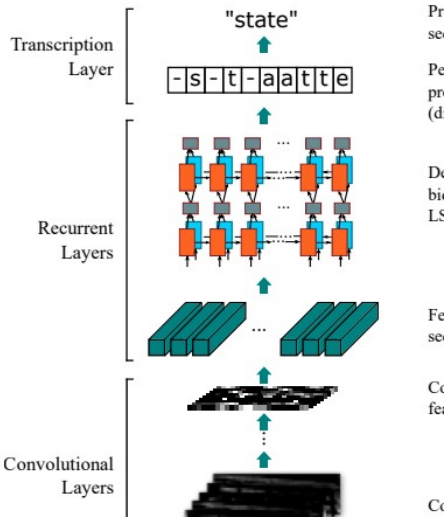

Figure 5. The network architecture of convolutional recurrent neural network for recognizing scene text. The architecture consists of three parts: 1) convolutional layers, which extract a feature sequence from the input image; 2) recurrent layers, which predict a label distribution for each frame; 3) transcription layer, which translates the per-frame predictions into the final label sequence(Shi et al., 2016).

2.3 Evaluation

All experiments were performed on the PyTorch framework. The computation was done with a workstation operated under Ubuntu 18.04 operation system. The workstation is equipped with 16-core Intel Xeon 5218@2.3GHz processor, 256 GB of RAM and NVIDIA RTX3090 GPU. The initial learning rate is 0.01, the learning rate decay value is $5e^{-4}$. The YOLOv5 was training from scratch with the yolov5x network structure and activating three parameters of rectangular training, multi-scale training and single data class. The model trained with a total of 175 million parameters. The evaluation metric of Intersection Over Union (IOU) was adopted in the experiment. YOLOv5 defines the loss function as follows:

Figure 6. the results of billboard detection on test data using YOLOv5.

$$IOU = \left|\frac{A \cap B}{A \cup B}\right| \quad (1)$$

$$GIOU = IOU - \left|\frac{R/(A \cup B)}{R}\right| \quad (2)$$

$$Loss_{GIOU} = 1 - GIOU \quad (3)$$

A represents the billboard area predicted by the model, and B represents the ground truth. R represents the smallest convex closed box containing A and B. The sample with IOU greater than a certain value is denoted as positive sample, while samples with IOU less than a certain value are labeled as negative samples. C represents the number of categories detected by YOLOv5. The precision rate (P), recall rate (R), and mean Average Precision (mAP) were considered to evaluate the model performance.

$$P = \frac{TP}{TP + FP} \quad (4)$$

$$R = \frac{TP}{TP + FN} \quad (5)$$

$$mAP = \frac{\sum_i^C \int_0^1 P_i(r)dr}{C} \quad (6)$$

3. RESULTS AND DISCUSSION

Figure 6 shows several detected billboards in street-level imagery using the YOLOv5 model. When there is a large inclination or a large number of shadow in the picture, the confidence of detected billboards is lower than that of the inclination or the picture with less shadow (Figure. 6). There is only one billboard label on each image in the CCF BDCI dataset, while YOLOv5 can locate multiple billboards on the street-level image (high confidence is for correctly detected billboards, low for additionally detected billboards). This indicates that YOLOv5 has great potential in locating POI-relate billboards. YOLOv5 provides multiple models with different size that can be used in different situation, and model ensemble inference often achieves better results than single model inference. The next phase of our work will consider to further fine-tune parameters of the YOLOv5 model and adopt ensemble modelling approach to further improve the performance.

Due to a street-level imagery often contains multiple billboards in parallel or clustered together, The bounding box tends to include parts of other billboards in the image or to contain billboards that are similar to nearby billboards in color, which can result in low precision (Figure. 7). This result is mainly caused by the tilt angle of the image being too large. Introducing some image quality evaluation indicators to control the data quality of street-level images may be helpful for obtaining higher accuracy results.

Figure 7. Example of test images with labelled (left) and predicted (right) bounding boxes.

The performance of YOLOv5 model on the proposed validation and test data are depicted in Table 1. Figure 8 shows the box loss and performance metric curves obtained during training of YOLOv5 model. The detection of billboard achieved an accuracy of 0.81 with mAP at 0.6 IOU threshold value.

Table 1: Performance of YOLOv5 model on the CCF BDCI dataset. In the table, N = the number of images.

	N	Precision	Recall	mAP@0.6
validation	2077	0.73	0.739	0.81

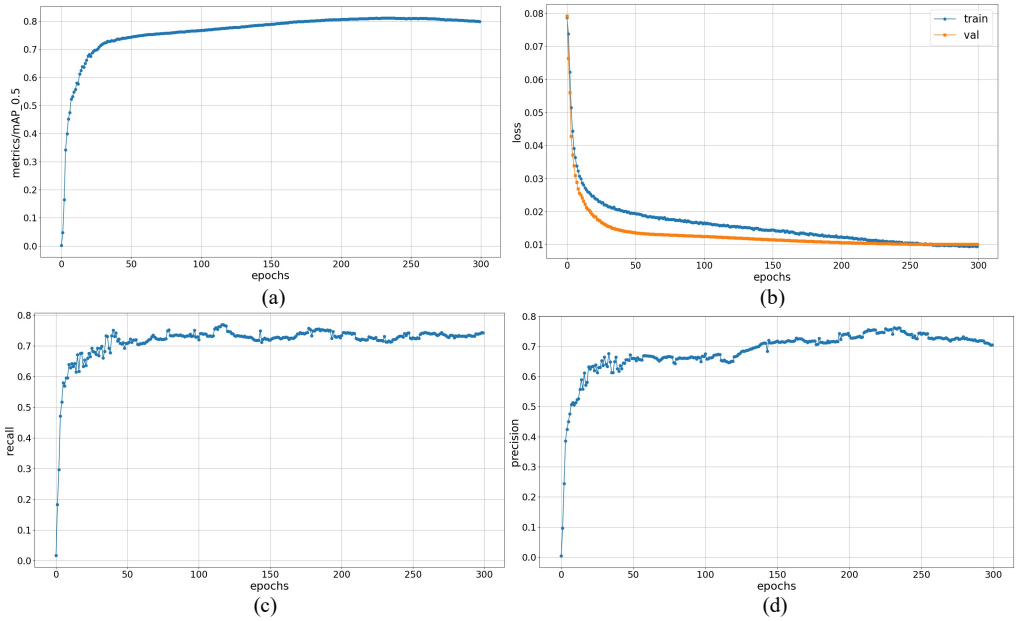

Figure 8. The precision, recall, mAP and class box loss plots during training epochs. (a) mAP at training stage. (b) box loss at training and validation stage. (c) and (d) precision and recall values during training of YOLOv5.

With the pre-trained OCR model, POI-related text was successfully recognized. Several examples were shown in Figure. 9 and the text were in Chinese. For some complicated cases, the extracted text is disordered resulting text recognition errors. The results was further processed to obtain accurate POI data, such as "华源超市" for Figure.9 (b) and "重庆移动忠县分公司三分局" for Figure.9 (c). In addition, the result was seriously affected by the quality of street-level data, such as the light condition, the shooting angle and the shadow. Due to the diversity of language types and characters in billboards, the DB+CRNN pre-training model was not enough to recognize all characters. it is better to consider more effective network frameworks for the task such as Mask R-CNN+LSTM+Attention framework.

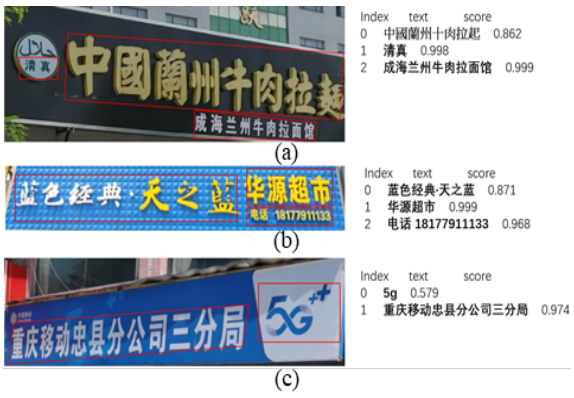

Figure 9. Examples of POI-related text recognition using the pre-training model (text in Chinese). (a) restaurant; (b) convenience shop; (c) mobile shop.

4. CONCLUSION

In this paper, a deep learning-based approach was presented for urban POI mapping using street-level imagery. YOLOv5 and OCR models were adopted to detect the position of POI-related billboard and extract POI scene text, and finally the results were further cleaned to obtain POI data. From the pilot study, YOLOv5 performs well in billboard detection on street-level image data. When the IOU is 0.5, the mAP value for validation data is 0.80; when the IOU is 0.6, the mAP value is increased to 0.81. The pre-trained OCR model provided by PaddleHub can recognize numbers, English, traditional and simplified Chinese characters. However, text recognition errors are prone to occur, which directly reduced the accuracy of POI mapping. The results indicated a scene text recognition model that performs well in different orientation, font types, and languages was expected for the task.

However, it should be pointed out that it is a challenging task to automatically map urban POIs directly from street-level imagery. It encounters many problems for scene text recognition that should be solved such as font types, typography, multilingual, and lighting and blur problems. Extracting the POI information on the billboard is difficult, as it not only need to recognize the text, but also need to deal with the invalid information. In future work, POI-related attribute determination and structured processing of the recognized text should be performed to improve the accuracy and precision of POI mapping.

ACKNOWLEDGMENTS

This research was funded by the National Natural Science Foundation of China (Grants No. 42101070).

REFERENCES

Bhatt, P.P., Patel, I., 2018. Optical Character Recognition Using Deep Learning – a Technical Review. *National Journal of System and Information Technology*. 11, 55–66.

Bochkovskiy, A., Wang, C.-Y., Liao, H.-Y.M., 2020. Yolov4: Optimal Speed And Accuracy Of Object Detection. *arXiv preprint arXiv:*2004.10934.

Duckham, M., Winter, S., Robinson, M., 2010. Including Landmarks In Routing Instructions. *Journal of Location Based Services*. 4, 28–52.

Fu, X., Ch'Ng, E., Aickelin, U., See, S., 2017. CRNN: A Joint Neural Network For Redundancy Detection. *2017 IEEE Int. Conf. Smart Comput. SMARTCOMP 2017*. pp. 1-8.

He, K., Gkioxari, G., Dollár, P., Girshick, R., Doll, P., Girshick, R., Ai, F., 2017. Mask R-Cnn. *Proceedings of the IEEE International Conference on Computer Vision*. pp. 2961–2969.

Husnain, M., Missen, M.M.S., Mumtaz, S., Jhanidr, M.Z., Coustaty, M., Luqman, M.M., Ogier, J.M., Choi, G.S., 2019. Recognition Of Urdu Handwritten Characters Using Convolutional Neural Network. *Applied Sciences*. 9(13), 2758.

Konishi, Y., Hanzawa, Y., Kawade, M., Hashimoto, M., 2016. Fast 6D Pose Estimation From A Monocular Image Using Hierarchical Pose Trees. *European Conference on Computer Vision*. pp. 398–413.

Liao, M., Wan, Z., Yao, C., Chen, K., Bai, X., 2020. Real-Time Scene Text Detection With Differentiable Binarization. *Proceedings of the AAAI Conference on Artificial Intelligence*. pp. 11474–11481.

Lim, K.H., Chan, J., Leckie, C., Karunasekera, S., 2015. Personalized Tour Recommendation Based On User Interests And Points Of Interest Visit Durations. *Twenty-Fourth International Joint Conference on Artificial Intelligence*. pp. 1778–1784.

Lin, A., Sun, X., Wu, H., Luo, W., Wang, D., Zhong, D., Wang, Z., Zhao, L., Zhu, J., 2021. Identifying Urban Building Function By Integrating Remote Sensing Imagery And POI Data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. 14, 8864–8875.

Liu, S., Qi, L., Qin, H., Shi, J., Jia, J., 2018. Path Aggregation Network For Instance Segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 8759–8768.

Medsker, L., Jain, L.C., 1999. Recurrent Neural Networks: Design And Applications. *CRC press*.

Purkait, P., Zhao, C., Zach, C., 2017. SPP-Net: Deep Absolute Pose Regression With Synthetic Views. *arXiv Preprint arXiv:*1712.03452.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You Only Look Once: Unified, Real-Time Object Detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 779–788.

Redmon, J., Farhadi, A., 2018. Yolov3: An Incremental Improvement. *arXiv Preprint arXiv:*1804.02767. pp. 1–6.

Redmon, J., Farhadi, A., 2017. YOLO9000: Better, Faster, Stronger. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 7263–7271.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster R-Cnn: Towards Real-Time Object Detection With Region Proposal Networks. *Advances in neural information processing systems*. 28, 1–9.

Rösler, R., Liebig, T., 2013. Using Data From Location Based Social Networks For Urban Activity Clustering. *Geographic Information Science at the Heart of Europe.* Springer, pp. 55–72.

Shi, B., Bai, X., Yao, C., 2016. An End-To-End Trainable Neural Network For Image-Based Sequence Recognition And Its Application To Scene Text Recognition. *IEEE transactions on pattern analysis and machine intelligence*. 39, 2298–2304.

Tian, Z., Shu, M., Lyu, P., Li, R., Zhou, C., Shen, X., Jia, J., 2019. Learning Shape-Aware Embedding For Scene Text Detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 4234–4243.

Touya, G., Antoniou, V., Olteanu-Raimond, A.-M., Van Damme, M.-D., 2017. Assessing Crowdsourced POI Quality: Combining Methods Based On Reference Data, History, And Spatial Relations. *ISPRS International Journal of Geo-Information.* 6(3), 80.

Wang, C.Y., Mark Liao, H.Y., Wu, Y.H., Chen, P.Y., Hsieh, J.W., Yeh, I.H., 2020. CSPNet: A New Backbone That Can Enhance Learning Capability Of CNN. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. pp. 390–391.

Wang, W., Xie, E., Li, X., Hou, W., Lu, T., Yu, G., Shao, S., 2019. Shape Robust Text Detection With Progressive Scale Expansion Network. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2019-June, 9328–9337.

Wu, W., Id, H.L., Li, L., Long, Y., Wang, X., Wang, Z., Li, J., Chang, Y., 2021. Application Of Local Fully Convolutional Neural Network Combined With YOLO V5 Algorithm In Small Target Detection Of Remote Sensing Image. *PloS one*. 16(10), e0259283.

Yang, S., Shen, J., Konečný, M., Wang, Y., Štampach, R., 2018. Study On The Spatial Heterogeneity Of The POI Quality In Openstreetmap. *Proceedings of the 7th International Conference on Cartography and GIS, Sozopol, Bulgaria*. pp. 18–23.