

Analysing the Ability of SLAM Algorithms Fused with CNN to Acquire Deep Information

Di Wu

School of Technology, Beijing Forestry University, Beijing 100083, China

apxiao@bjfu.edu.cn

Abstract. Given that SLAM has benefited from the development of vision technology in the last decade, it has been widely popularised into people's daily life, such as drones and UAVs using sensors to acquire environmental data and construct maps to avoid obstacles and achieve path planning. However, since the extent of SLAM technology is mostly based on the texture features of the input image, SLAM is prone to misjudge the depth information when the features of the input image are not obvious. Combined with the current development of Deep Learning, the paper can consider integrating the Convolutional Neural Network (CNN), which has a good processing effect on image data, into the SLAM algorithm, so as to improve the effectiveness and accuracy of the SLAM algorithm in obtaining depth data. This paper will analyse the basic principles of SLAM and CNN, and discuss the current existence of CNN and SLAM fusion related technology.

Keywords: Convolutional Neural Network, Simultaneous localization and mapping, Deep Learning, texture features.

1. Introduction

Computer vision, as an important branch in the field of artificial intelligence, perceives external information by enabling computers to simulate the function of the human visual system [1], and plays an important role in the current big hit in the field of autonomous driving, such as Tesla's HydraNet model, which, as the current state-of-the-art computer vision system, covers almost all of the lane detection to pedestrian tracking to signal recognition, etc. road information and can predict the possible conditions on the road in time. Considering the characteristics of autonomous driving, the concept of Simultaneous localization and mapping (SLAM) in navigation and mapping can be used as a reference to satisfy the functions that need to be realised in the autonomous driving part.

SLAM is called 'real-time location-based map construction', which can help the robot to draw a map and determine the direction of travel in an unknown environment, and the problem it solves can be summarised in two words: first, what is this place; second, where am I? In order to obtain information about the external environment, data input is required through the corresponding sensors, and the current sensors used to assist SLAM are mainly classified into LiDAR and cameras [2]. Among them, when a camera is used as a sensor, the position and direction of the camera in the environment are calculated by identifying (Feature Extraction) and labelling (3D Reconstruction, as shown in Fig. 1) the feature points in the environment on the acquired image, and by using inter-frame estimation, i.e., the visual odometry method. As shown in Fig. 2.

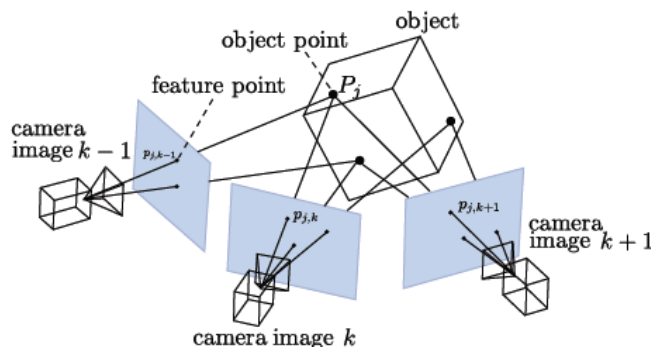


Figure 1. Move the Camera to Estimate the 3D Point Position

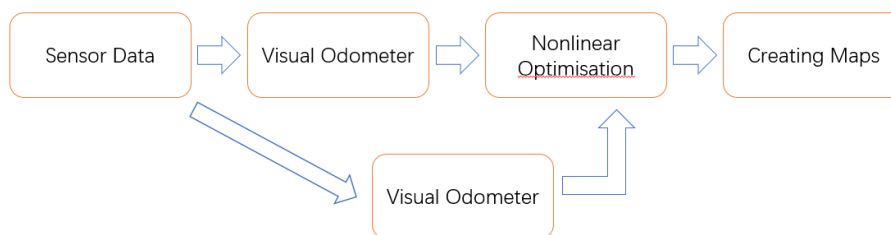


Figure 2. Visual SLAM Process

However, in reality, the measurement error of the sensor itself, the error of the system algorithm, and the complex and changing external environment all affect the accuracy of the feature points [3]. With the growth of Deep Learning in the last decade, especially the great progress of Convolutional Neural Network (CNN) in image instance segmentation, the paper can think about combining deep learning with SLAM, using neural network to learn and process the acquired image feature points, and use the effective feature points for instant positioning and map construction, to improve the reliability and robustness of SLAM algorithm.

Since the purpose of deep learning is to represent the target at multiple levels, thus constructing multi-layer networks. And in order to represent the abstract semantics of the data, multiple layers of high-level feature information are needed so that the trained model obtains better feature robustness [4]; and CNN, as a member of the field of Deep Learning, is a kind of deep neural network with a convolution structure, which is better adapted to the structure of the image compared to other neural networks, facilitating the extraction and allocation of features, and improving the reliability and robustness of SLAM algorithm. It is easy to extract and assign features, and the weights are shared to reduce the training parameters set in the network simplifies the network structure and makes the neural network more adaptable [5].

Therefore, this paper will explore the use of CNN, which are currently more effective in processing image feature points, to improve the ability of SLAM to collect image feature points by segmenting the input image data by instances, and at the same time, optimise the noise information between the feature points, so that the SLAM can focus on the calculation of effective and accurate feature points, and enhance the connection between the visual images (inter-frames), which will reduce the loss of arithmetic power and improve the operation efficiency to ensure that the algorithm can operate in a more efficient way. This reduces arithmetic power loss and improves the operational efficiency of the algorithm, thus ensuring that the algorithm operates in a more efficient manner. Improve the operational efficiency of the algorithm, thus ensuring that the mechanical equipment can understand and update the environmental information in a timely manner [6].

2. Convolutional Neural Network

Deep learning is a branch of machine learning, introduced by Hinton et al. in 2006, which has achieved great success in many fields as it has far surpassed previous related techniques in language and image recognition [7].

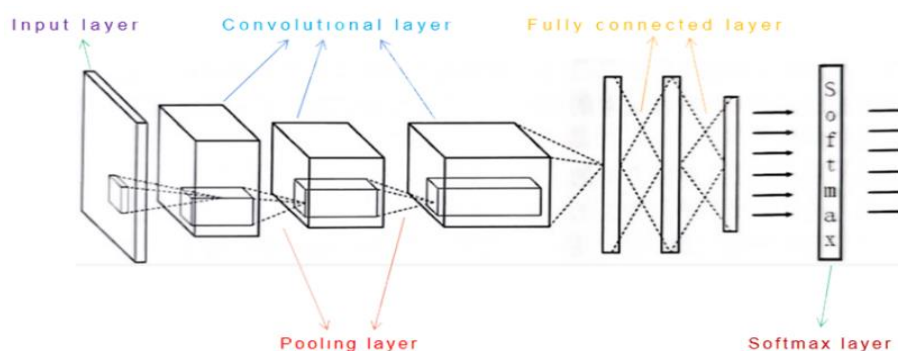


Figure 3. Basic Structure of CNN

The CNN in deep learning has a good effect on image data and can produce satisfactory results according to the needs. The CNN architecture mainly consists of two nonlinear convolutional layers which can be obtained by training, two fixed subsampling layers, and one fully connected layer, while the number of hidden layers is usually at least 5 layers [8]. As shown in Fig. 3.

3. Techniques for fusing SLAM and CNN

This section will be divided into three parts. In the first part, the paper will recommend the basic principles of SLAM to acquire feature points and compute them. In the second part, the paper will explore the principles and effects of ORB-SLAM2. Finally, the paper analyses the basic principle of SLAM algorithm fused with CNN and the effect of acquiring depth information.

3.1. Principle of SLAM's computation of image feature points

Features are used as a mathematical expression of image data, and feature points are pixel points that maintain their relevant characteristics even after the position of the camera acquiring the image has changed. In order to accurately describe a feature point, Key-points are usually used to describe the coordinates, size and direction of the feature point in the image [9]; and Descriptors are used to describe the surrounding information of the pixel point corresponding to the feature point. Theoretically, it is considered that to identify whether it is the same feature point or not, it is only necessary to observe whether the distances of the corresponding Descriptors of two feature points are the same in the vector space [10].

The steps for extracting feature points are as follows:

- (1) Calculate the principal directions of the feature points to facilitate the processing of the descriptors and add rotationally invariant properties to them;
- (2) Describe the surrounding image region of the feature point acquired in the first step.

As shown in Fig. 4, a pyramid model for scaling the image and extracting pixels is built to better describe the rotation and scale of feature points in the image. In this, each layer of the pyramid model will scale the image equally.

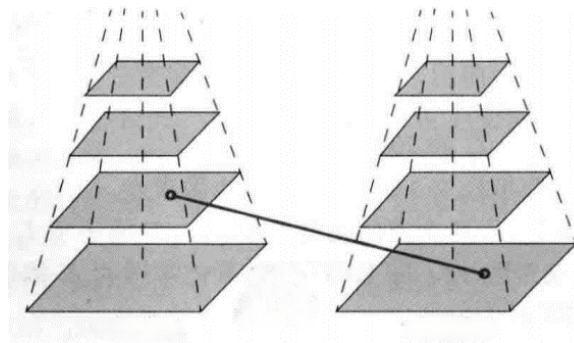


Figure 4. Matching of Images with Different Scaling Magnifications by Pyramid Modelling [9]

The rotation of features is achieved using the greyscale prime method, combined with the properties of the pyramid model. The steps are as follows:

- (1) Let there exist an image block in a given image with moments:

$$m_{pq} = \sum_{x,y \in B} x^p y^q I(x,y) \quad (1)$$

$$p, q = \{0,1\} \quad (2)$$

- (2) The centre of mass (centre of weights) of the image block is found through the moments:

$$C = \left(\frac{m_{10}}{m_{00}}, \frac{m_{01}}{m_{00}} \right) \quad (3)$$

(3) To better orient the feature points, the geometric centre O of the image block is connected to the centre of mass C of the whole image to form the vector:

$$\theta = \arctan(m_{01}/m_{10}) \quad (4)$$

The above method makes the Key-points have a scale and rotation description, which improves the robustness of the feature points across different images. Finally the feature points between the two frames are paired and the coordinates of these feature points are taken to calculate the depth [11].

3.2. Fundamentals and effects of LSD-SLAM

LSD-SLAM is a monocular SLAM algorithm based on the direct method, which can not only construct large-scale environment maps, but also environment maps similar to the actual environment. The algorithm can both be aligned based on the direct image, which leads to a highly accurate estimation of the pose; its 3D environment can also be inserted into keyframes in the form of a pose map, which leads to a real-time reconstruction with a semi-dense depth map that exists in a link [12]. And by employing an explicit scale-drift-aware formulation, the method can be made to operate on challenging sequences, where drastically varying scene scales are included [13].

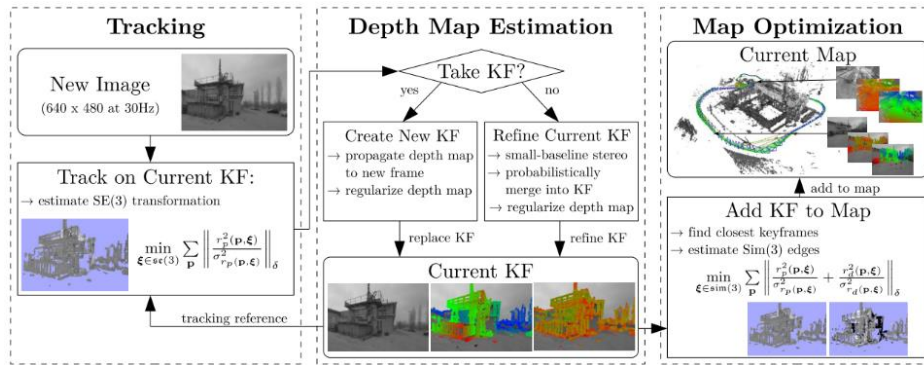


Figure 5. LSD-SLAM Overall Framework [12]

As shown in Fig. 5, three of the core components of LSD-SLAM are:

(1) Image Position Tracking: The main task of the tracking part is to estimate the corresponding camera position based on successive images, and it estimates the camera position corresponding to the current image based on the current key frame. The initial values needed to calculate the camera pose can be used as the camera pose data corresponding to the previous image frame [14].

(2) Depth estimation: firstly, according to the change of the estimated camera position, it determines whether to use the image of the frame to create a keyframe or not, when the distance of the camera from the camera at the moment corresponding to the currently used keyframe is greater than a certain threshold, then it uses the image of the frame to create a new keyframe, and obtains a new depth map. If the distance is too short, the depth estimation result of that frame image is Kalman fused to the currently used keyframe. Whether or not new keyframes are created, depth estimation is required and depth regularisation is performed to screen out outliers [15].

(3) Map Optimisation: This section performs two tasks: the pose optimisation and the point cloud map optimisation functions. The bitmap optimisation here differs from the first part (tracking) in that the tracking part is a bitmap estimation for each image frame, which is evaluated using the 6-degree-of-freedom transformation matrix $se(3)$ for depth estimation, whereas the bitmap in the third (present) part is based on the bitmap estimation between keyframes, which is evaluated using the similarity matrix $sim(3)$. When a keyframe becomes a reference frame and its corresponding depth map is no longer optimised, the depth map of that frame is fused to the global map. Loopback detection and scale estimation, which is evaluated using the similarity transformation matrix for two keyframes in close proximity, still uses the direct image matching method [16].

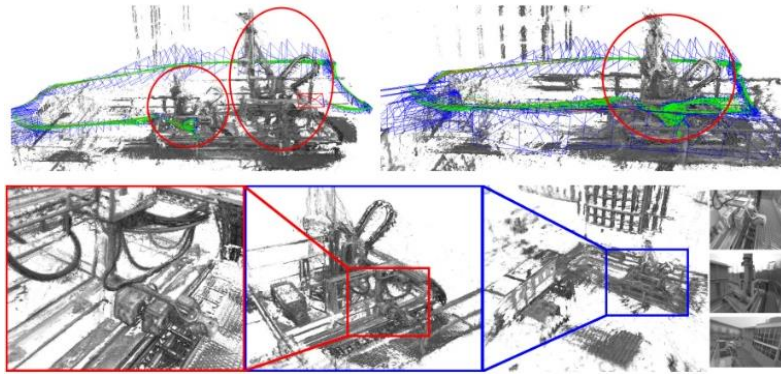


Figure 6. Effect of LSD-SLAM Algorithm [12]

As shown in Fig. 6, the LSD-SLAM algorithm is mainly used to maintain the global environment map and track the pixels of the global environment map, which contains the keyframe pose maps and associated probabilistic semi-dense depth maps, whereas the existing direct methods are purely measured by visual odometers, which are significantly different. And LSD-SLAM contains two important innovations: (1) a direct approach that explicitly merges pixels and detects scale drift to align two keyframes on $\text{sim}(3)$, and (2) a novel probabilistic approach that incorporates noise on the predicted depth map into the tracking. Depth maps are mainly represented as point clouds, which allow semi-dense or even highly accurate 3D world reconstruction of the environment.

3.3. Principles of SLAM algorithms combined with CNN

CNN as a deep learning model is commonly used for iterative processing of images and videos. Because CNNs can extract features from images and learn them, they gain a greater ability to process images and continuous data than traditional neural networks. Convolutional operations are the core feature of CNNs, which perform sliding window computation on an image through a filter (also known as a convolutional kernel) and a pooling layer (maximum pooling) to extract features from the image. Convolutional operations reduce the number of weights and thus the computational effort of the model, while also being able to preserve the spatial structure information of the image. On the other hand, to ensure accuracy i.e. without changing the feature map dimensions, the pooling layer reduces the computational effort and improves the robustness of the model [17].

The following three main directions for deep learning combined with SLAM are currently on the market:

- 1) Replacing one or several modules in traditional SLAM with deep learning methods;
- 2) Adding semantic information to traditional SLAM;
- 3) End-to-End SLAM

The exploration of this paper focuses on reference [6], which combines CNN with SLAM by improving the matching accuracy of effective pixel feature points of images between frames in visual ranging, so that when the sensor didn't create a new keyframe, it predicts the related depth map by CNN, and with the help of semantic labels provided by CNN, so that the semantic information in the SLAM is associated with each element of the three-dimensional global model [18]. The structure of the algorithm is shown in Fig. 7.

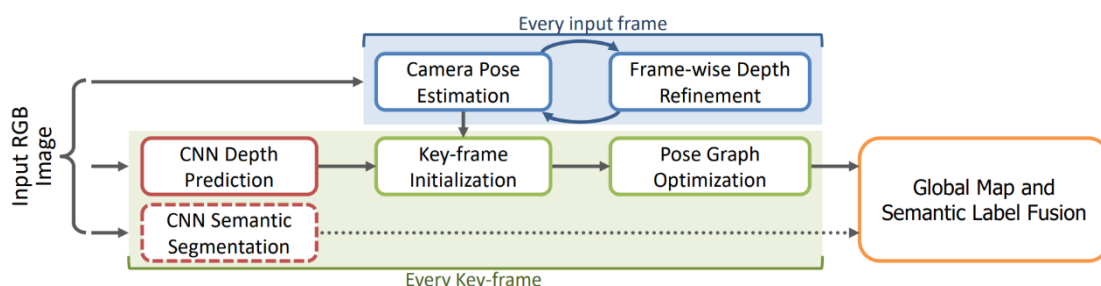


Figure 7. CNN-SLAM Overview

A keyframe-based SLAM paradigm is employed, in particular using the direct semi-dense approach in [4] as a baseline. In this approach, visually distinct subsets of frames are collected as keyframes, whose pose is globally refined based on a pose map optimisation. Also, to estimate the pose information of the camera, by estimating the scale of change between each input frame and the nearest keyframe; and using the CNN to predict only the keyframes in the depth map to ensure that the algorithm has a high frame frequency. In particular, if the pose information in a keyframe differs from the currently estimated pose information, a new keyframe is created in the current image frame and the CNN is allowed to estimate the depth information for the current keyframe [19]. In addition, to construct uncertainty row maps, the pixel confidence in each depth estimate can be measured. Whereas the pixels of the camera used for SLAM do not coincide with the pixels of the camera that captured the dataset, a specific depth map normalisation procedure can be proposed in this regard, showing the robustness to different camera parameters. Another convolutional neural network is used to estimate the semantic segmentation of the input frames, i.e., used in the context of semantic tag fusion. Finally, in order to optimise the relative pose information in all directions, pose maps can be created on key frames [20]. As shown in Fig. 8.

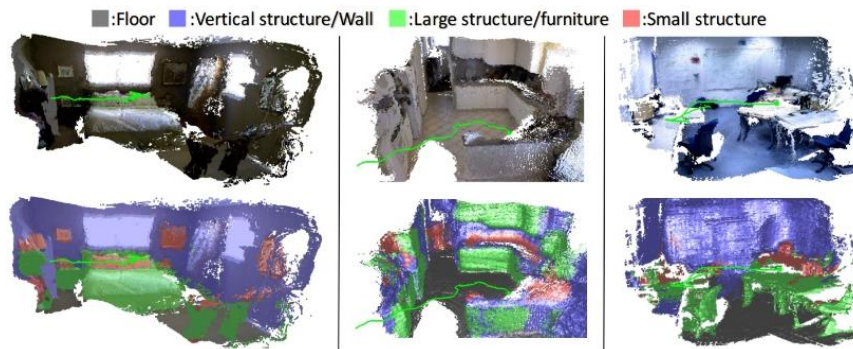


Figure 8. Effect of CNN-SLAM Algorithm [6]

To refine the depth image so that the keyframes are combined with the CNN's estimates, small baseline stereo matching can be used and the depth map can be minimised to enhance the colour consistency between the keyframes and the relevant input frames. Whereas, in order to provide a more accurate prediction region, the operation will be performed mainly in the image pixel regions where gradients are present, i.e., the refinement of the depth values in this area.

4. Discussion

In this paper, the LSD-SLAM algorithm and the CNN-SLAM algorithm are explored by reading the literature [6], and it can be seen that relying on CNN to predict the depth information allows for the existence of a stereo baseline between consecutive frames of an image, making it possible to reconstruct a three-dimensional scene with a camera that is purely rotationally in motion, whereas the results of the LSD-SLAM suffer from a noticeable noise due to the insufficient stereo baseline required for the estimation of the depth information. As shown in Fig. 9.

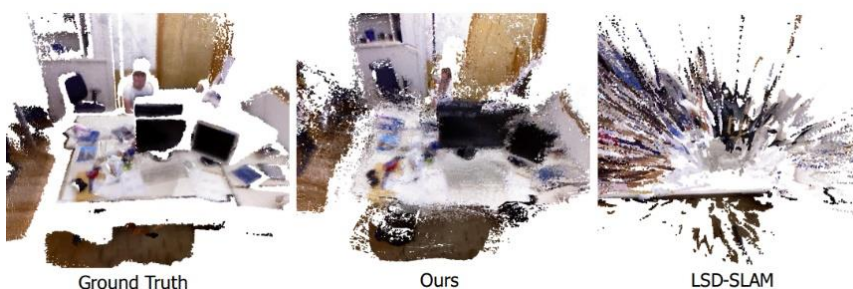


Figure 9. Comparison of CNN-SLAM and LSD-SLAM [6]

5. Conclusion

In this paper, after reviewing papers on combining CNNs with visual SLAM, the paper has learnt that combining SLAM with neural networks has a very good compatibility, and that the depth predictions it generates can help to solve the problems prevalent in the industry, and can address the inherent limitations of monocular reconstruction, i.e. the algorithms allow for higher accuracy, better real-time performance, and a wider range of applications. However, the use of CNNs means that a large amount of high-quality image data will be required to optimise the algorithm in the future, but this data is difficult to obtain and label; the more samples there are, the more the CNN needs to undergo a large amount of training and inference, which costs more; and at the same time, the CNN's lack of robustness to external factors such as noise, changes in illumination, and changes in viewpoints, which may affect the performance of the map localisation and SLAM technology performance.

References

- [1] Chang Q L, Hung K W, Jiang J M. Deep learning based image Super-resolution for nonlinear lens distortions. *Neurocomputing*, 2018, 275: 969-982.
- [2] Yuan Z K, Cheng K, Tang J H, et al. RGB-D DSO: Direct Sparse Odometry With RGB-D Cameras for Indoor Scenes. *Ieee Transactions on Multimedia*, 2022, 24: 4092-4101.
- [3] Li L, Yang M. Point Cloud Registration Based on Direct Deep Features With Applications in Intelligent Vehicles. *Ieee Transactions on Intelligent Transportation Systems*, 2022, 23: 13346-13357.
- [4] Chen M Q, Yu L J, Zhi C, et al. Improved faster R-CNN for fabric defect detection based on Gabor filter with Genetic Algorithm optimization. *Computers in Industry*, 2022, 134.
- [5] He K M, Gkioxari G, Dollar P, et al. Mask R-CNN. *Ieee Transactions on Pattern Analysis and Machine Intelligence*, 2020, 42: 386-397.
- [6] Keisuke T, Federico T, Iro L, Nassir N. CNN-SLAM: Real-time dense monocular SLAM with learned depth prediction. *Computer Vision and Pattern Recognition*, 2017.
- [7] Egrioglu E. Median-Pi artificial neural network for forecasting. *Neural computing & applications*, 2019, 31.
- [8] Pang B, Nijkamp E, Wu Y N. Deep Learning with TensorFlow: A Review. *Journal of Educational and Behavioral Statistics*, 2020, 45: 227-248.
- [9] Shiozaki T, Dissanayake G. Eliminating Scale Drift in Monocular SLAM Using Depth from Defocus. *Ieee Robotics and Automation Letters*, 2018, 3: 581-587.
- [10] Zhou D F, Dai Y C, Li H D. Ground-Plane-Based Absolute Scale Estimation for Monocular Visual Odometry. *Ieee Transactions on Intelligent Transportation Systems*, 2020, 21: 791-802.
- [11] Zhang Z, Huang Y H, Bao S, et al. Panoramic annular image restoration algorithm by prediction based on the lens design characteristics. *Applied Optics*, 2023, 62: 518-527.
- [12] Engel J, Schps T, and Cremers D. LSD-SLAM: LargeScale Direct Monocular SLAM. In *European Conference on Computer Vision (ECCV)*, 2014.
- [13] Zhang Y J, Zhao J X, Han H Y. A 3D Machine Vision-Enabled Intelligent Robot Architecture. *Mobile Information Systems*, 2021.
- [14] Zhou Y, Yan F H, Zhou Z. Handling pure camera rotation in semi-dense monocular SLAM. *Visual Computer*, 2019, 35: 123-132.
- [15] Niu Z Y, Zhong G Q, Yu H. A review on the attention mechanism of deep learning. *Neurocomputing*, 2021, 452: 48-62.
- [16] Lu J, Wang Z, Hua B W, et al. Automatic point cloud registration algorithm based on the feature histogram of local surface. *Plos One*, 2020, 15.
- [17] Cui Y, Yu Z K, Han J C, et al. Dual-Triple Attention Network for Hyperspectral Image Classification Using Limited Training Samples. *Ieee Geoscience and Remote Sensing Letters*, 2022, 19.

- [18] Chen S Y, Zhang Z X, Zhong R F, et al. A Dense Feature Pyramid Network-Based Deep Learning Model for Road Marking Instance Segmentation Using MLS Point Clouds. *Ieee Transactions on Geoscience and Remote Sensing*, 2021, 59: 784-800.
- [19] Sakpere W, Adeyeye-oshin M, Mlitwa N B W. A state-of-the-art survey of indoor positioning and navigation systems and technologies. *South African Computer Journal*, 2017, 29: 145-197.
- [20] Wang P, Shen X, Lin Z, Cohen S, Price B, and Yuille A L. Towards unified depth and semantic prediction from a single image. In *Proc. Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 2800–2809, 2015.