LARGER RECEPTIVE FIELD BASED RGB VISUAL RELOCALIZATION METHOD USING CONVOLUTIONAL NETWORK

Jiangying Qin [1], Ming Li [1,2,*], Deren Li[1], Xuan Liao[3], Jiageng Zhong [1], Hanqi Zhang[1]

[1]State Key Laboratory of Information Engineering in Surveying Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China
[2]Department of Physics, ETH Zurich, Zurich 8093, Switzerland - mingli39@ethz.ch
[3]Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hong Kong 999077, China

KEY WORDS: Visual Relocalization, Camera Relocalization, Pose Regression, Deep ConvNet, RGB Image

ABSTRACT:

Visual Relocalization is a key technology in many computer vision applications. Traditional visual relocalization is mainly achieved through geometric methods, while PoseNet introduces convolutional neural network in visual relocalization for the first time to realize real-time camera pose estimation based on a single image. Aiming at the problem of accuracy and robustness of the current PoseNet algorithm in complex environment, this paper proposes and implements a new high-precision robust camera pose calculation method (LRF-PoseNet). This method directly adjusts the size of the input image without cropping, so as to increase the receptive field of the training image. Then, the image and the corresponding pose tags are input into the improved LSTM-based PoseNet network for training, and the Adam optimizer is used to optimize the network. Finally, the trained network is used to estimate the camera pose. Experimental results on open RGB dataset show that the proposed method in this paper can obtain more accurate camera pose compared with the existing CNN-based methods.

1. INTRODUCTION

Visual relocalization plays a key role in photogrammetric computer vision, autopilot and robotics (Husain, 2019; Acharya, 2019; Wang, 2020; Pham, 2021). However, feature ambiguities have made it remain challenging. For example, the traditional geometry based visual relocation method is mainly realized by local feature matching which is based on the known 3D environment created by SfM (Sattler, 2017; Han, 2019) or SLAM (Mur-Artal, 2015; Mur-Artal, 2017). It matches the local 2D feature points extractedfrom the query image with the corresponding 3D feature points in the model to establish the corresponding relationship (Bay, 2006; Lowe, 2004; Rublee, 2011), and solves the camera posewith six degrees of freedom by PnP and other algorithms (Lepetit, 2009; Hesch, 2011). For the mismatched points that exist in the matching process, the random sampling consensus algorithm (Martin, 1981; Chum, 2005) is used to eliminate and accelerate the camera pose calculation. Geometric-based visual relocalization methods rely on correct feature matching, however, not enough matching points can be found accurately in all scenarios. Various complex situations that may exist in the actual environment, such as illumination changes, viewpoint changes, object occlusion, motion blur, and lack of texture, may affect feature matching and make it difficult to obtain accurate camera poses. On the other hand, since the matching cost increases exponentially with respect to the number of key points, the cost of matching in a large and dense feature space is very large.

In recent years, neural networks have developed rapidly and are widely used in tasks such as object recognition, image retrieval, and image classification (Eric, 2016; Melekhov, 2017; Shao, 2020; Liu, 2021). In 2015, Kendall et al. innovatively introduced convolutional neural networks to the field of image-based visual relocalization and proposed the PoseNet method (Kendall, 2015). This method uses transfer learning from large-scale classification data to directly obtain a 6-DOF camera pose from a single image in an end-to-end manner. Although PoseNet overcomes many limitations of geometry-based methods, especially reduces the dependence on rich textures and improves the robustness and efficiency of visual relocalization, its positioning accuracy still lags behind that of geometry-based methods when local features perform well. Therefore, many scholars are committed to improving the PoseNet method to improve its accuracy. (Kendall, 2016) uses bayesian convolutional neural network to estimate the uncertainty of positioning to improve the positioning accuracy of the method. (Kendall, 2017) proposes a loss function based on geometry and reprojection errors to solve the difficulty of hyperparameter training caused by the use of L_2 distance in the PoseNet loss function. (Valada, 2018) combines the geometric knowledge and semantic knowledge of the world to locate and proposes a novel "geometric consistency loss" function. (Nguyen, 2019) proposes the SP-LSTM framework based on CNN and LSTM. CNN and LSTM are used to learn the depth features and spatial dependence of images, respectively. It uses time information to improve the position accuracy. (Aragao, 2020) combines the Inception layer into the down-sampling and up-sampling symmetrical layer to solve the depth scene and 6-degree-of-freedom estimation.

In order to combine the advantages of the geometric-based visual relocation method and the deep learning-based visual relocation method, scholars have proposed a visual relocation method based on fusion of geometry and deep learning, that is, to estimate the camera pose by combining geometry methods and deep learning methods. The deep learning part is used to learn and predict the 3D position of the pixel in world coordinates and geometric methods infer the camera pose from these correspondences. (Guzman, 2014) try to use hybrid methods for visual relocation, but their main limitation is that they require the use of RGB-D images for training and testing (Cavallari, 2017). (Brachmann, 2016) optimizes this limitation and proposes to use only automatic context random forest from RGB images for positioning. (Meng, 2016) realizesimage localization by using regression forest to estimate the initial camera pose, then queries the nearest neighbor key frame image, and optimizes the initial pose by sparse feature matching between the camera input image and the nearest key frame. (Brachmann, 2017) uses the architecture of VGG style to predict 3Dcoordinates and proposes a distinguishable RANSAC to learn the matching function that

* Corresponding author

optimizes the quality of pose. Although these methods improve positioning accuracy, they require many predictions about the scene coordinates, which causes RANSAC to spend more and more time to infer the best camera pose.

Based on this, this paper proposes a high-precision camera pose estimation method LRF-PoseNet. This method mainly improves PoseNet from the following three aspects. First, by improving the cropping method of the input image to obtain a larger receptive field. Secondly, the Adam optimizer is used to optimize the network. Finally, the LSTM structure is introduced into the PoseNet neural network to perform structural dimensionality reduction on the fully connected layer and select the most useful feature correlation for the task of camera pose estimation. Experiments show that the method proposed in this paper has better accuracy and stronger robustness.

2. METHODOLOGY

The experimental images used in this paper are automatically generated sample annotations (i.e. camera poses) in advance by SfM. During image preprocessing, in order to obtain a fixed size image, this paper proposes to directly resize the experimental image to the corresponding size without cropping. Then, the obtained images and corresponding annotations are put into the high-precision positioning network based on LSTM proposed in this paper for training. The network introduces LSTM structure on the basis of PoseNet network structure to perform structural dimension reduction on the full connection layer and select the most useful feature correlation for pose estimation task. In addition, this paper chooses the Adam optimizer to optimize the network to train the most suitable parameters. Figure 1 is the architecture of this paper.

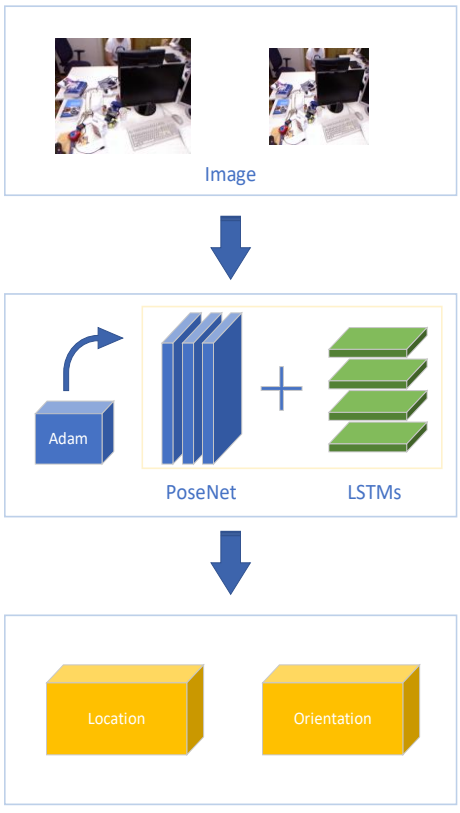


Figure 1. Architecture of the proposed method.

2.1 Image Processing for Larger Receptive Fields

PoseNet is based on the GoogleNet network, and one of its disadvantages is that it has strict restrictions on the network structure. Among them, the size of the RGB image input to the network must be 224*224 pixels specified by GoogleNet (Seifi, 2019). However, the actual RGB image participating in the training may not be the specified size. To solve this problem, the processing method of PoseNet is to scale the size of the image's minimum side to 256 pixels according to the aspect ratio of the original image, and then crop the 224*224 window in the middle of the scaled image as the training image. As shown in Figure 2, 2a is the original image, 2b is the image whose height is scaled to 256 pixels according to the aspect ratio, and 2c is the image with 224*224 pixels in the center of 2b. One disadvantage of this processing is that the image information outside the cropping window will be lost and cannot be added to the network for training. However, the lost part may also contain key information to assist positioning, thereby affecting the accuracy of positioning.

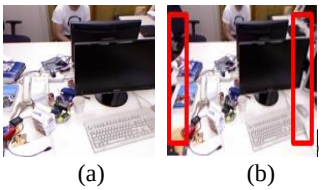
2a 2b 2c

Figure 2. Schematic diagram of PoseNet image preprocessing.

In order to solve the problem of image information loss in PoseNet, this paper proposes to use the entire field of view of the image, that is, only need to scale the input image to 224×224 pixels, as shown in Figure 3. Figure 4 shows the difference between the receptive field of the input image of PoseNet and the network of this paper. There is significant texture information in the red box area in the figure, but the PoseNet network discards this information. Using the direct zoom method proposed in this paper will result in different aspect ratios, but considering that the new aspect ratio changes are consistent for all images in the dataset, losing the original aspect ratio will not have any impact on network performance. In addition, although this zoom method will reduce the resolution of the image, the receptive field is more important than the image resolution, because the pooling layer in the network can smooth the high-frequency details of the high-resolution image. Therefore, using the direct zoom method will get higher positioning accuracy. The experimental results in this paper also prove this.

(a) (b)

Figure 3. Schematic diagram of LRF-PoseNet image preprocessing. (a–b) correspond to before and after processed.



(a) (b)

Figure 4. Field of view comparison between PoseNet and LRF-PoseNet. (a–b) correspond to different methods.

2.2 Network Structure

GoogleNet is a new deep learning framework proposed by Christian Szegedy in 2014. It was originally designed for object classification and detection. It uses an average pooling layer after the convolutional layer to collect the information of each feature channel in the entire image. PoseNet uses a fully connected layer (FC) after average pooling layer to learn the correlation between features. But the regression pose is not accurateafter the high-dimensional output of the FC layer. Specifically, compared with the amount of available training data, the dimensionality of 2048D image embedding through fully connected layers is usually relatively large. Therefore, the linear
pose regressor has multiple degrees of freedom, and resulting over-fitting, which is tend to cause inaccurate predictions on test images that are different from training images. Reducing the dimensionality of FC can reduce this adverse effect. But studies have shown that the use of LSTM memory block networks for dimensionality reduction is more effective (Walch, 2017). Compared with PoseNet's application of dropout to avoid overfitting, the estimation of this method is more accurate, which proves the rationality of using LSTMs in this paper. However, directly inputting the 2048D vector into the LSTM does not work well. This is because even if the LSTM storage unit can rememberthe features far away, the length of 2048 is too long for the LSTM. In order to solve this problem, this paper resizes the vector to a 32×64 matrix and introduces four LSTMs in the up, down, left and right directions. Then these four outputs are connected and passed to the fully connected layer. It simulates the structured dimensionality reduction function and improves the accuracy of pose calculation. The network structure is shown in Figure 5, where the blue part represents the module inherited from PoseNet, and the yellow part represents the improved module of the network.

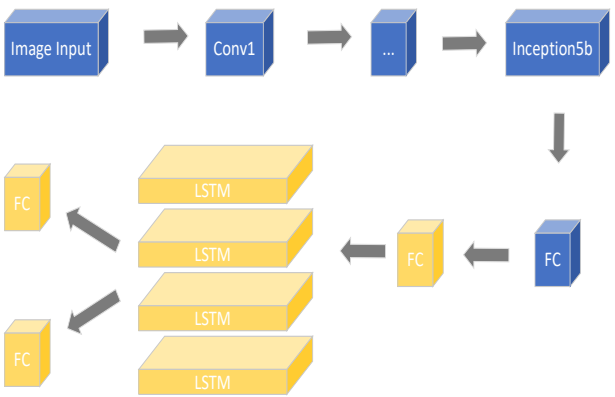


Figure 5. Our image positioning network.

2.3 Network Optimizer

PoseNet uses Stochastic gradient descent (SGD) to optimize the network. SGD is a very common optimization algorithm in neural network model training (Zinkevich, 2011). It is based on the gradient descent algorithm. The basic idea of the gradient descent algorithm is to obtain the partial derivative of each hyperparameter, then the current gradient can be obtained, and then the parameter is updated in the opposite direction of the gradient, and iteratively updated in this way. Then the global optimal solution of the hyperparameter can be obtained. However, the SGD algorithm has two shortcomings in dealing with practical problems. The first is that it is difficult to choose an appropriate learning rate so it uses the same learning rate for all parameters. But in practical applications, for sparse data or features, we may want to update faster, and for features that do not appear frequently, we hope that it can be updated slower to reduce training costs. The SGD algorithm cannot satisfy this point. Second, because the SGD algorithm uses the gradient descent of a random sample as the average gradient descent of the overall sample, this also makes it easy for SGD to converge to a local optimum, and in some cases may be trapped in a saddle point. Adam method is a method to obtain better performance by calculating the adaptive learning rate of each parameter. It adds first-order momentum and second-order momentum on the basis of SGD. The first-order momentum is shown in equation (1), where β_1 is a hyperparameter, often taking an empirical value of 0.9. The first-order momentum is the exponential moving average of the gradient direction at each time, approximately equal to the average of the sum of the gradient vectors at the latest $1/1-\beta_1$ time.

$$m_t = \beta_1 \cdot m_{t-1} + (1-\beta_1) \cdot g_t \quad (1)$$

In other words, the descending direction at time t is determined not only by the gradient direction of the current point, but also by the descending direction accumulated before. It avoids training problems caused by extreme current gradients.

The second-order momentum solves the problem of learning rate, and the historical update frequency is measured by the second-order momentum: the sum of the squares of all gradient values so far in this dimension. The second-order momentum is shown in equation (2), where β_2 is a hyperparameter.

$$V_t = \beta_2 \cdot V_{t-1} + (1-\beta_2) \cdot g_t^2 \quad (2)$$

For parameters that are frequently updated, since we have accumulated a lot of knowledge about it, we do not want it to be affected too much by a single sample, so we hope that the learning rate will be lower; For parameters that are updated occasionally, because there is too little information to understand, we hope to learn more from each occasional sample, that is, the learning rate is higher. This improvement can effectively solve the problems of low training accuracy and high cost caused by the consistent learning rate of all parameters of SGD.

3. EXPERIMENTS

3.1 Experimental Data and Computing Environment

In order to verify the effectiveness of the LRF-PoseNet method proposed in this paper, this paper conductes multiple sets of experiments, and compares the experimental results with the experimental results of the PoseNet open source code. This paper uses Pytorch to program the proposed new method. In the experiment, the processor used is Intel(R) Core (TM) i7-8750H, the memory is 8GB, the GPU is GeForce GTX 1060, and the batch size of 75 is used to finetune the network. Set the initial learning rate for 500 epochs to 0.0005. For the Adam optimization algorithm, set β_1 =0.9, β_2 =0.999. The image data used in this paper comes from the public databases CoRBS dataset and TUM RGB-D benchmark dataset. The CoRBS data has a total of 563 training images and 48 test images (Wasenmüller, 2016). The TUM data has a total of 1197 training images and 325 test images, respectively.

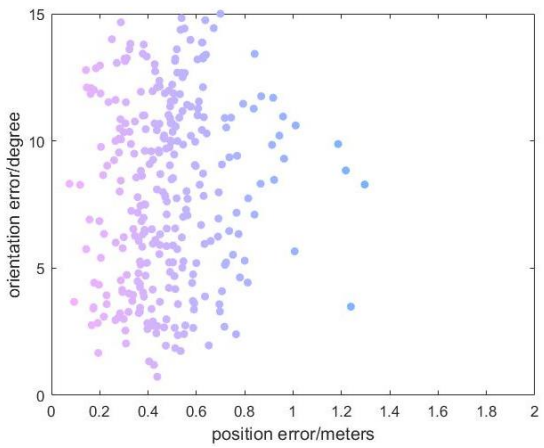
3.2 Experimental Results and Analysis

In order to show the experimental results of different methods fairly and objectively, this paper provides the average position accuracy and average orientation accuracy results of the two datasets using two methods, as shown in Table 1. The average position and orientation accuracy of the original PoseNet of the two sets of data are 0.48m, 5.08° and 0.35m, 4.16°, respectively. The average accuracy of the LRF-PoseNet method proposed in this paper is significantly higher than that of PoseNet. The position accuracy and orientation accuracy of CoRBS data and TUM data are increased by 0.44m, 2.64° and 0.14m, 0.54°, respectively, and the error is reduced by about 91%, 52% and 40%, 13%.

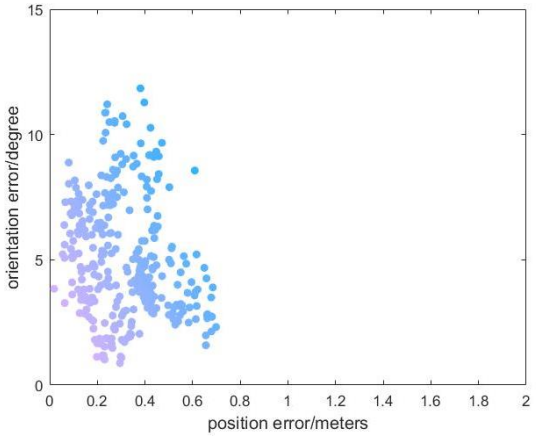
		PoseNet	LRF-PoseNet
CoRBS	Position(m)	0.48	0.04
	Orientation(°)	5.08	2.44
TUM	Position(m)	0.35	0.21
	Orientation(°)	4.16	3.62

Table 1. Comparison table of average accuracy.

Figure 6 is a scatter plot of the position and orientation errors of the query image after using two methods on the TUM data. The horizontal axis is the position error, in meters, and the vertical axis is the direction error, in degrees. Figure 6a is the result of PoseNet running, and Figure 6b is the result of LRF-PoseNet running. It can be seen from the figure that the error point of the improved method in this paper is closer to the origin point, that is, the position error and the orientation error are smaller, which shows that the method in this paper has significantly improved the accuracy of the PoseNet method. Especially for some points difficult in positioning, the PoseNet method has more points with larger positioning error, and its position error and orientation error can reach about 1.5m and 15 degrees, respectively. For LRF-PoseNet, the maximum position error and orientation error are about 0.7m and 12 degrees respectively, and the number of such points is very small. That is the position accuracy and orientation accuracy are greatly improved compared with PoseNet.



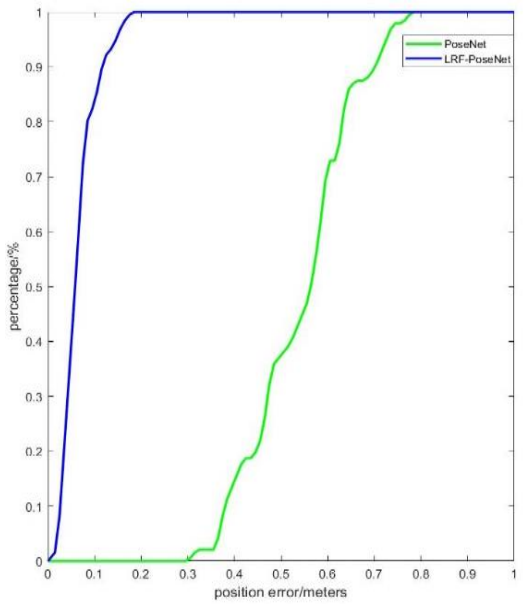
(a)



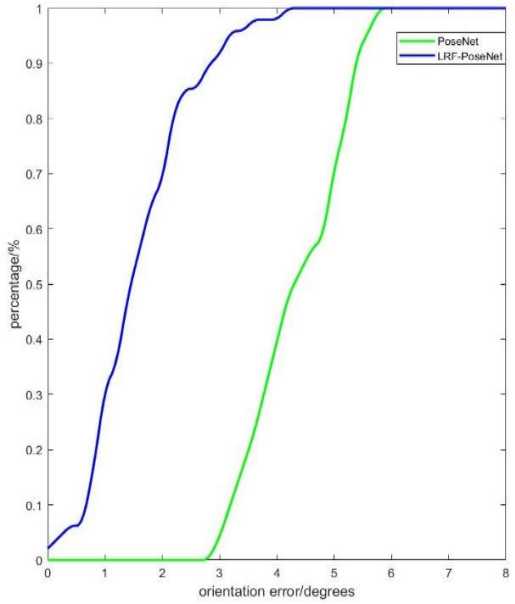
(b)

Figure 6. Scatter plot of errors. (a–b) correspond to position error and orientation error.

Figure 7 as the cumulative error histogram of the CoRBS test set shows the positioning performance of the two methods from a more quantitative and intuitive perspective. Figure 7a shows the position error, and Figure 7b shows the orientation error. Generally speaking, compared with PoseNet algorithm, the method proposed in this paper is more competitive. From the point of view of position error, PoseNet has no points with a positioning error within 0.2m, while the corresponding point of LRF-PoseNet reaches 79%. The positioning accuracy is greatly improved compared with the original PoseNet. For the orientation error, the percentage of the orientation error within 3 degrees of the two methods is 4% and 93% respectively. The percentage of LRF-PoseNet's orientation error within 3 degrees is 89% higher than that of PoseNet. When the orientation error is 3.5 degrees, the corresponding percentages of the two methods are 16% and 100% respectively, that is, the orientation error of all points of LRF-PoseNet is within 3.5 degrees. At this time, there are still many points with orientation accuracy greater than 3.5 degrees.



(a)



(b)

Figure 7. Cumulative histogram of errors. (a–b) correspond to position error and orientation error.

4. CONCLUSION

This paper changes the PoseNet image preprocessing method to obtain a larger receptive field, proposes a new neural network based on LSTM and uses the Adam optimizer to optimize the network, which improves the accuracy and robustness of the original PoseNet visual relocalization. Through the statistics and comparison of the experimental results, it is proved that the LRF-PoseNet method proposed in this paper has a significant improvement in the performance of PoseNet. On the one hand, the method proposed in this paper is more accurate than PoseNet. On the other hand, the method in this paper is more robust to complex environment and can achieve high-precision image positioning more effectively.

ACKNOWLEDGEMENTS

This research was funded by the National Natural Science Foundation of China (NSFC), grant number 41901407, the College Students' Innovative Entrepreneurial Training Plan Program, grant number S2020634016. Meanwhile, we thank the editors and reviewers for their valuable comments.

REFERENCES

Husain S S., Bober M., 2019. REMAP: Multi-layer entropy-guided pooling of dense CNN features for image retrieval. *IEEE Transactions on Image Processing*, 28(10), pp. 5201-5213.

Acharya D., Ramezani M., Khoshelham K., Winter S., 2019. BIM-Tracker: A model-based visual tracking approach for indoor localisation using a 3D building model. *ISPRS Journal of Photogrammetry and Remote Sensing*,150(4), pp. 157–171.

Wang J. Wang P. Dai D.et al., 2020. Regression Forest Based RGB-D Visual Relocalization Using Coarse-to-Fine Strategy. *IEEE Robotics and Automation Letters*, 5(3), pp. 4431-4438.

Pham T., Seto W., Dafty s., et al., 2021. Rover Relocalization for Mars Sample Return by Virtual Template Synthesis and Matching. *IEEE Robotics and Automation Letters*, 6(2), pp. 4009-4016.

Sattler T., Leibe B., Kobbelt L, 2017. Efficient & effective prioritized matching for large-scale image-based localization. *IEEE Trans. Pattern Anal. Mach. Intell.* 39 (9), pp. 1744–1756.

Han X.F., Laga H., Bennamoun M., 2019. Image-based 3D Object Reconstruction: State-of-the-Art and Trends in the Deep Learning Era. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*

Mur-Artal R., Montiel J.M.M., Tardos J.D., 2015. Orb-slam: A versatile and accurate monocular slam system. *IEEE Transactions on Robotics*, 31 (5), pp. 1147–1163.

Mur-Artal R., Tardos J D., 2017. ORB-SLAM2: an Open-Source SLAM System for Monocular, Stereo and RGB-D Cameras. *IEEE Transactions on Robotics*, 33(5), pp. 1255-1262.

Bay H., 2006. SURF: Speeded Up Robust Features. *Computer Vision & Image Understanding*, 110(3), pp. 404-417.

Lowe D. G, 2004. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2), pp.91–110.

Rublee E., Rabaud V., Konolige K., and Bradski G, 2011. ORB: An efficient alternative to SIFT or SURF. *International Conference on Computer Vision*, pp. 2564-2571.

Lepetit V., Moreno-Noguer F., Fua P., 2009. EPnP: An Accurate O(n) Solution to the PnP Problem. *International Journal of Computer Vision*, 81(2), pp. 155-166.

Hesch J.A., Roumeliotis S.I., 2011. A Direct Least-Squares (DLS) method for PnP. *International Conference on Computer Vision*, pp. 383-390.

Martin, A., Fischler, et al, 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6), pp. 381-395.

Chum O., Matas J., 2005. Matching with PROSAC " Progressive Sample Consensus. *IEEE Computer Society Conference on Computer Vision & Pattern Recognition*, pp. 220-226.

Eric Brachmann, Frank Michel, Alexander Krull, Michael Ying Yang, Stefan Gumhold, and Carsten Rother, 2016. Uncertainty-driven 6D pose estimation of objects and scenes from a single RGB image. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3364–3372.

Melekhov I., Ylioinas J., Kannala J., et al., 2017. Image-based Localization using Hourglass Networks. *IEEE International Conference on Computer Vision Workshops*, pp.107-115.

Z. Shao, W. Zhou, X. Deng, M. Zhang and Q. Cheng, 2020. Multilabel Remote Sensing Image Retrieval Based on Fully Convolutional Network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13, pp. 318-328.

Liu, X. Han, Z. Liu Y. and Zwicker M., 2021. Fine-Grained 3D Shape Classification With Hierarchical Part-View Attention. *IEEE Transactions on Image Processing*, 30, pp. 1744-1758.

Kendall A., Grimes M., Cipolla R., 2015. Posenet: A convolutional network for real-time 6-dof camera relocalization. *IEEE international conference on computer vision*, pp. 2938-2946.

Kendall A., Cipolla R., 2016. Modelling uncertainty in deep learning for camera relocalization. *IEEE international conference on Robotics and Automation*, pp. 4762-4769.

Kendall A., Cipolla R., 2017. Geometric Loss Functions for Camera Pose Regression with Deep Learning. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6555-6564.

Valada A., Radwan N., Burgard W., 2018. Incorporating semantic and geometric priors in deep pose regression. *Workshop on Learning and Inference in Robotics: Integrating Structure, Priors and Models at Robotics: Science and Systems*, 1(3), pp. 1-3.

Nguyen A., Do TT, Caldwell DG, et al, 2019. Real-Time 6DOF Pose Relocalization for Event Cameras with Stacked Spatial LSTM Networks. *IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

Aragao D., Nascimento T. and Mondini A., 2020. SpaceYNet: A Novel Approach to Pose and Depth-Scene Regression Simultaneously. *International Conference on Systems, Signals and Image Processing*, pp. 217-222.

Guzman-Rivera A., Kohli P., Glocker B., Shotton J., Sharp T., Fitzgibbon A., Izadi S., 2014. Multi-output learning for camera relocalization. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1114–1121.

Cavallari T., Golodetz S., Lord N.A., Valentin J., Di Stefano L., Torr P.H., 2017. On-the-fly adaptation of regression forests for online camera relocalisation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4457–4466.

Brachmann E., Michel F., Krull A., Ying Yang M., Gumhold S., et al., 2016. Uncertainty-driven 6d pose estimation of objects and scenes from a single rgb image. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3364–3372.

Meng L., Chen J., Tung F., Little J.J., de Silva C.W., 2016. Exploiting random rgb and sparse features for camera pose estimation. *BMVC*.

Brachmann E., Krull A., Nowozin S., Shotton J., Michel F., Gumhold S., Rother C., 2017. Dsac - differentiable ransac for camera localization. *The IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6684-6692.

Zinkevich M., Weimer M., Smola A J., et al, 2011. Parallelized Stochastic Gradient Descent. *Conference on Neural Information Processing Systems*, pp. 1–4.

Seifi S., Tuytelaars T., 2019. How to Improve CNN-Based 6-DoF Camera Pose Estimation. *International Conference on Computer Vision Workshop*.

Wasenmüller O., Meyer M., Stricker D., 2016. Corbs: Comprehensive rgb-d benchmark for slam using kinect v2. *Applications of Computer Vision*, pp. 1–7.

Walch F., Hazirbas C., Leal-Taixe L., et al, 2017. Image-based localization using LSTMs for structured feature correlation. *IEEE International Conference on Computer Vision*, pp. 627–637.