Application of Deep Learning in Visual Odometry: A Brief Literature Review

Yijun Huang[1]
1. College of Software Engineering, Beihang University
Beijing, China
yjhuang@gmail.com

***Abstract*—Visual odometry is a technique for estimating camera egomotion based on continuous frame images and has important applications in areas such as UAV navigation and augmented reality. Traditional visual odometry mainly applies geometry-based methods, enabling near real-time applications on drones and robots. However, the classical methods have limited applications in challenging cases due to problems such as sensitivity to scene illumination and difficulty in detecting dynamic environments. With the booming development of deep learning in recent years, related techniques combined with visual odometry have emerged as a viable complement, but the current review still focuses on the traditional methods. Therefore, in this paper, we review the classical development of visual odometry to highlight the progress of deep learning incorporating or improving VO traditional methods in recent years and discuss possible current issues and trends.**

Keywords— Deep Learning, Visual Odometry,

I. INTRODUCTION

Visual odometry is a technique that estimates the camera pose without a priori knowledge by detecting the motion of the surrounding environment. In contrast to VSLAM, which focuses on global consistency, VO focuses mainly on the consistency of local trajectories. The term was coined by D. Nister because vision-based localization is similar to wheel odometry in that it incrementally estimates the motion of a vehicle by integrating the number of turns of its wheels over time[1].

Since it was first proposed by D. Nister et al.[2] Since 204, visual odometry has played an important role in many aspects, such as augmented and visual reality, Mars rover exploration, autonomous driving, and navigation of drones. However, these clear methods need another angle of enhancement due to the poor robustness of traditional geometry-based methods in the presence of large differences in illumination and drastic changes in environmental dynamics. With the advancement and development of neural networks and deep learning, the traditional way of using geometric methods in visual mapping has been aided by sophisticated, implicit but more effective deep learning efforts due to their superior performance in vision-related tasks. Considering the cost of training and storing models for these networks, integrating AI-VSLAM with UAVs is a challenge.

In this paper, we will review the outstanding contributions of traditional methods in VO and free up more space for an in-depth discussion of highly promising deep learning methods in the VO domain. In contrast to SLAM, in which we are only concerned with the local consistency of trajectories, local maps are used to obtain more accurate estimates of local trajectories (e.g., in bundle adjustment), while SLAM is concerned with the consistency of global maps. This paper is organized as follows. Section II shows some relevant reviews and surveys. Section III provides a brief review of the main previous contributions of DeepVO. Section IV outlines the mainstream ways of combining deep learning and VO. Section V discusses the current potential and challenges of using deep learning as a recognizer at a relatively low cost. The last section summarizes our survey and evaluation.

II. RELATED WORKS

Because of the close connection between visual odometry and SLAM technology, VO is often part of SLAM reviews, and there are few reviews specifically on visual odometry compared to SLAM technology. Early classic reviews on VO include a series of tutorials presented by D. Scaramuzza et al[3]. Also, in presenting the development of VO, researchers focused mainly on traditional approaches. Even after 2016, when deep learning has become very popular, reviews and reviews are still scarce, despite the progress made in recent years in visual odometry methods incorporating deep learning.

MoShan et al[4]. introduced the application of VIO in MAV; Chen et al[5]. outlined the main applications of VO in SLAM mainly by traditional VO techniques, and also spent a subsection on the integration of deep learning and SLAM in general; in addition, the development of deep learning in SLAM was also discussed in the review by Li et al[6]. However, they mainly introduce the possible directions of deep learning in SLAM such as semantic SLAM, and do not describe the specific potential and methods of deep learning for VO applications; Lai et al[7]. provide a more systematic review of the combination of VSLAM and deep learning, summarize the advantages of the current use of deep learning to deal with SLAM problems, and list the advantages of VSLAM in recent years by applying Li et al[8]. provide a detailed overview of the evolution of VSLAM, but they focus on depth estimation and semantic map building, without a specific collation of VO. Wang et al[9]. give a comprehensive introduction to the methods and applications of deep learning in VO, and list the current problems in the field.

III. GEOMETRY-BASED APPROACHES

The traditional implementation methods of VO mainly include the feature point method, direct method, and hybrid semi-direct tracking method.

A. Feature Point Method

The feature point method is a representative method and an early attempt of VO. It can be compared to the principal component of an image, and the feature point method extracts the sparse representative information of the image and estimates the adjacent key frame motion of the whole image based on the overall motion of the feature points. Classical feature point extraction methods include the early Harris corner point[11], FAST corner point[12], and so on. These classical corner point recognition algorithms were proposed earlier and are not stable enough in the case of large image changes. Scale Invariant Feature Transform (SIFT)[13] is a classic algorithm that is robust to illumination, scale, and rotation. However, with the recognition effect comes a huge amount of computation, which makes it difficult for SLAM to meet its real-time requirements. To improve the speed of the algorithm operation, H. Bay et al. proposed Speeded Up Robust Features (SURF)[14] to reduce the computational effort in the SIFT integration process. However, both of these methods still require significant computational costs, and executing the computation in real-time may be challenging.

To reconcile accuracy, robustness, and computational effort, Rublee et al[15]. improved the BRIEF descriptor proposed by M. Calonder et al[16]. and addressed the direction invariance of FAST by proposing the oriented FAST and rotated BRIEF (ORB) descriptor. In 2015, based on Klein et al's Parallel Tracking and Mapping (PTAM)[17], R. Mur-Artal et al.[18] proposed a landmark solution: a robust and accurate real-time ORB-SLAM system. They later improved on this system with ORB-SLAM2[19], which further supported calibrated binocular and RGB-D cameras, and C. Campos et al. went on to introduce ORB3[20], which enriched the sub-maps to improve robustness and further incorporated IMU to enrich the calibration data.

B. Direct Method

The feature point method is clear and straightforward, but there are still some problems. For example, even if the ORB speed is already quite fast, it still takes about 20ms[21], and if we want to do a 30-frame real-time SLAM, then we need each frame to be around 33ms on average. Thus, most of the time overhead is spent on feature point extraction. In addition, the image itself is discarded when feature points are used. Although feature points can reflect the image in a sense, an image has after all millions of pixels, and feature points are often only a few hundred, which may be difficult to reflect potentially useful image information in some cases. Meanwhile, in some occasions where feature points are not too significant, such as along the direction of a wall or an empty corridor, it is difficult to identify the camera movement by feature points alone. All these may pose related problems. That is, there is a certain relationship among them.

Therefore, in some cases, the direct method may be more appropriate. In contrast to the feature point method, the direct method does not require a one-to-one match, and the projection is considered successful as long as the previous points have reasonable projection residuals in the current image: success depends mainly on the judgment of the depth of the map points and the camera pose, not on what the image looks like locally. The direct method saves a lot of time in feature extraction and matching is easily portable to embedded systems and can be integrated with IMUs, of which the LK optical flow technique[22] is a well-known approximate example. Since its introduction, the optical flow method has been continuously developed[23]. Direct methods like this seem to directly use image pixel grayscale information and geometric information to construct error functions by graphical optimization to minimize the cost function and thus obtain the best camera pose. In practice, Engel et al. proposed the large-scale direct monocular simultaneous localization and mapping (LSD-SLAM) algorithm[24] and applied it to a stereo camera, combining temporal and static stereo in a direct, real-time SLAM approach[25]. realistic conditions with some robustness considering also illumination variations. After this, Engel et al. further proposed DSO-SLAM[26]. however, its parameters in the code need to be adjusted to adapt to the new scene requirements each time the scene is changed in a dynamic environment, and there are problems such as scale drift in practical application scenarios.

IV. Deep Learning Approaches

However, although geometry-based SLAM has been able to achieve CPU real-time in classical scenes, traditional, geometry-based SLAM methods still have some problems: for the feature point method, identifying feature points may encounter some difficulties in the case of insignificant features. In addition, additional arithmetic power is required to extract features, and these computations account for most of the entire VO process, and these feature points are discarded soon after matching, resulting in a large degree of waste; for the direct optical flow method, the assumption of constant features such as overall illumination of the rigid scene is required, and these are difficult to implement in scenes such as outdoors. Therefore, with the development of deep learning and its great advantages shown in visual recognition, many VOs incorporating deep learning have been proposed. Convolutional neural networks, a network structure, were the first to come into view due to their dominant level in object recognition and detection problems.

A. Review of Supervised Deep Learning

In this context, Kendall et al[27]. proposed PoseNet capable of generating six degrees of freedom of a camera directly from a single RGB input image and was the first implementation of camera pose estimation. Since CNN extracts more powerful features than conventional feature detectors, the system can achieve high accuracy even under certain extreme conditions, such as strong illumination and blurred images. Later the authors improved PoseNet and also proposed improvements based on Bayesian analysis[28], which improved the accuracy of relocation; another direction of improvement by the authors was to improve the performance of PoseNet by using multi-view geometry as a source of training data[29]. Li et al[30]. extended PoseNet to accommodate color and depth inputs from RGB-D cameras using a dual-stream convolutional neural network, which showed robust performance in the face of challenging situations, becoming the first work to solve the deep CNN-based indoor relocation problem using RGB-D cameras. Wang et al[31]. proposed DeepVO, a recurrent convolutional neural

network (RCNN)-based VO approach that is competitive with model-based VO approaches, as a notable advance.

B. Review of unsupervised or self-supervised Deep Learning

All these above are the applications of supervised learning methods in VO: supervised learning methods tend to obtain better pose estimation results. However, SLAM is a niche area where it is often difficult and expensive to obtain real ground truth datasets in practice. It is difficult to build datasets suitable for large supervised learning and to label ground truth, while the number of available labeled datasets for supervised training is still limited. Li et al[32]. proposed a new monocular visual ranging (VO) system called UnDeepVO, which can achieve recovery of absolute scale. As an unsupervised approach, compared to DeepVO, UnDeepVO can be trained using a large number of unlabeled datasets to continuously improve its performance. Ummenhofer proposed DeMoN[33], which can simultaneously estimate camera self-motion, image depth, surface normals, and optical flow. Compared to popular single-image depth networks, DeMoN learns the concept of matching and thus can be better generalized to structures not seen during training. both DeMoN and UnDeepVO use stereo images to train the network to eliminate the important scale ambiguity problem in monocular V and are the first network models to use unsupervised learning methods to estimate the depth and pose of continuous images. The GANVO[34] proposed by Almalioglu et al. In contrast to traditional VO methods, pose and depth estimation does not require strict parameter tuning while being able to address the problem that traditional depth estimators based on autoencoder decoders tend to generate overly smooth images. However, unsupervised methods suffer from the drawback of insufficient supervisory information, so self-supervised algorithms by adding known image features as supervisory signals are also widely proposed. The D3VO self-supervised monocular depth estimation network proposed by Yang et al. tightly combines predicted depth, pose, and uncertainty into a direct visual ranging approach to enhance front-end tracking and back-end nonlinear optimization .e k can be analyzability, etc.

C. Methods Comparison

As shown in Table 1, due to the rapid development of deep learning in VO applications in recent years, this paper collates the progress of the main network models according to the characteristics of different models. The collation criteria include five main dimensions to evaluate: the use of network structure, the type of training (supervised or not), the test dataset used, whether it is an end-to-end model, and the type of camera applied to the model.

TABLE I. CHARACTERISTICS OF THE MAJOR DEEP LEARNING VISUAL ODOMETRY MODELS

Name	Structure	Type	Benchmark	E2E	Sensor
PoseNet	CNN	Supervised	Cambridge Landmarks, 7 Scenes dataset[35]	Y	Mono
DeepVO	RCNN	Supervised	KITTI[36] Benchmark	Y	Mono
D3VO	DeepThingNet	Self-supervised	KITTI & EuRoC[37]	N	Mono
UndeepVO	RCNN	Unsupervised	KITTI Benchmark	Y	Mono
GANVO	GAN	Unsupervised	KITTI & Cityscapes[38]	Y	Mono
DeMoN	Bootstrap Net	Supervised	SUN3D[39] & MVS[40]	Y	Mono

From the above table for the summary of influential models in recent years, it can be seen that due to the complexity and specificity of VO, the applied network structure has changed from the early CNN ruling the situation to the present blossoming; in addition, the number of available large-scale datasets in VO or SLAM is still limited due to the development of VO datasets suitable for deep learning slightly lagging behind the development of network models.

As a result, unsupervised and self-supervised approaches have been emerging since 2017 and can obtain stronger generalization while maintaining accuracy. In addition, since neural networks can be likened to a black box, most applications have adopted an end-to-end model, i.e., replacing the process from feature extraction to camera pose estimation in traditional VO methods; in terms of the cameras used, thanks to the advantage of deep learning in reducing the estimated absolute depth, researchers in most application scenarios favor monocular cameras to reduce the cost and improve the generalization capability, researchers have favored monocular cameras in most applications to reduce costs and improve generalization capabilities.

V. FURTHER DISCUSSIONS

Since 2015, deep learning has been increasingly integrated with VO technology. Rich applications have sprung up based on traditional feature extraction or optical flow computation. However, the current research still has some possible problems. The next section will discuss the challenges faced and possible directions for development.

A. Challenges and Difficulties

Dynamic scenes and dynamic objects in the scenes. VO assumes that the environment is static to integrate the egomotion of the camera, however, the scenes in which VO is performed are likely to encounter dynamic objects, such as pedestrians, animals, etc.; in this regard, the illumination may change more drastically, for example, the illumination may not be uniformly distributed in outdoor environments.

Deep learning is poorly interpretable, and because the training set cannot contain all scenes, the visual odometry trained by deep learning is often limited to certain specific scenes and performs poorly in some unfamiliar scenes.

Insufficient training data. In addition to the KITTI benchmark and Cityscapes datasets mentioned above, the mainstream VO datasets available include the RobotCar dataset[41], which contains different weather and scenery in the same location and was collected using a car driven in Oxford for a year. Also available is the previously mentioned EuRoc MAV dataset, which is a dataset that can be used for VO and VSLAM by collecting data through MAV. All of these datasets can be used for self-motion estimation, and the Cityscapes and KITTI datasets can also be used to complete scene segmentation. Although these training sets are relatively rich, they tend to be

limited to certain scenes, for which overfitting may lead to a decrease in the model's generalization ability and struggle to perform in some unfamiliar environments, which is exactly where VO tends to run.

B. Prospects and Directions

Use semantic predictive feedback to reduce the interference of dynamic objects in the scene. Semantic segmentation of the scene is performed during image processing and the results of the semantic segmentation are used as a correction to modify the operation of the VO. For example, Barnes et al[42]. distinguish static and dynamic parts of the scene by integrating a per-pixel ad-hoc mask in the VO to determine unreliable regions in the image.

Unsupervised learning does not require much hard-to-label ground truth, and thus semi-supervised, self-supervised, or unsupervised methods can be used to reduce the requirement for training set labeling when the dataset is not fully developed.

Fuse additional information such as IMU into the network structure for loss processing, while giving each other synchronization feedback, that is, in the direction of Deep VIO.

For the problem of poor model prediction improvement due to poor interpretability of deep learning, the degree of overfitting can be reduced by methods in artificial intelligence. Yang et al[43]. introduce the Bayesian distribution of weight factors to improve the generalization ability of network models in the prediction process and improve certain robustness for translation and rotation; they can also provide improved network structures to enhance the robustness for scenarios that do not appear in the training set and generalization.

VI. CONCLUSIONS

In this paper, we review the classical methods of VO, and on this basis, we make a brief arrangement and summary of the fusion and application of deep learning in VO in recent years. From the summary, we can see that compared with the traditional methods, the deep learning methods have good results in the case of very sparse or insignificant features. As deep learning continues to develop recognition capabilities in various visual tasks, research attention is increasingly turning toward deep learning and VO fusion. Despite the current shortcomings compared to clear solutions with geometry, they have shown great potential in various areas of VO applications.

ACKNOWLEDGMENT

This work was inspired by *A brief survey of visual odometry for micro aerial vehicles* by Ben M. Chen. In addition, This thesis owes sincere gratitude to my supervisor, Hongguang Li, for his guidance and expectations all along.

REFERENCES

[1] Scaramuzza D, Fraundorfer F, Pollefeys M. Closing the loop in appearance-guided omnidirectional visual odometry by using vocabulary trees. Robot Auton Syst. 2010;58(6):820–827. doi: 10.1016/j.robot.2010.02.013.

[2] D. Nister， O. Naroditsky, and J. Bergen， "Visual odometry"， Proc. Int. Conf. Computer Vision and Pattern Recognition， pp. 652-659， 2004.

[3] D. Scaramuzza and F. Fraundorfer, "Visual odometry. part i: The rst 30 years and fundamentals", IEEE Robot. Autom. Mag, vol. 18, pp. 8092, 2011.

[4] Mo Shan et al., "A brief survey of visual odometry for micro aerial vehicles," IECON 2016 - 42nd Annual Conference of the IEEE Industrial Electronics Society, 2016, pp. 6049-6054, doi: 10.1109/IECON.2016.7793198.

[5] Y. Chen， Y. Zhou， Q. Lv and K. K. Deveerasetty， "A Review of V-SLAM"， 2018 IEEE International Conference on Information and Automation （ ICIA ）， 2018， pp. 603-608， doi： 10.1109/ICInfA.2018.8812387.

[6] A. Li， X. Ruan， J. Huang， X. Zhu and F. Wang， "Review of vision-based simultaneous Localization and Mapping，" 2019 IEEE 3rd Information Technology， Networking， Electronic and Automation Control Conference （ ITNEC ）， 2019， pp. 117-123， doi： 10.1109/ITNEC.2019.8729285.

[7] D. Lai， Y. Zhang 和 C. Li， "A Survey of Deep Learning Application in Dynamic Visual SLAM"， 2020 International Conference on Big Data & Artificial Intelligence & Software Engineering （ICBASE），2020，pp. 279-283，doi：10.1109/ICBASE51474.2020.00065.

[8] Li, R., Wang, S. & Gu, D. Ongoing Evolution of Visual SLAM from Geometry to Deep Learning: Challenges and Opportunities. Cogn Comput 10, 875–889 (2018). https://doi.org/10.1007/s12559-018-9591-8

[9] Wang, S. Ma, J. Chen, F. Ren and J. Lu, "Approaches, Challenges, and Applications for Deep Visual Odometry: Toward Complicated and Emerging Areas," in IEEE Transactions on Cognitive and Developmental Systems, vol. 14, no. 1, pp. 35-49, March 2022, doi: 10.1109/TCDS.2020.3038898.

[10] He, M., Zhu, C., Huang, Q. et al. A review of monocular visual odometry. Vis Comput 36, 1053–1065 (2020). https://doi.org/10.1007/s00371-019-01714-6

[11] C. Harris and M. Stephens, "A combined corner and edge detector", Proc. Alvey Vis. Conf., vol. 15, no. 50, pp. 5244, 1988.

[12] Rosten, E., Drummond, T.: Machine learning for high-speed corner detection. In: European Conference on Computer Vision, pp. 430–443. Springer, Berlin (2006)

[13] D. G. Lowe, "Distinctive image features from scale-invariant keypoints", International Journal of Computer Vision (IJCV), vol. 60, no. 2, pp. 91-110, 2004.

[14] H. Bay, A. Ess, T. Tuytelaars and L. V. Gool, "SURF: Speeded up robust features", Computer Vision and Image Understanding (CVIU), vol. 110, no. 3, pp. 346-359, 2008.

[15] Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: ORB: an efficient alternative to SIFT or SURF. In: 2011 IEEE international conference on computer vision (ICCV), pp. 2564–2571. IEEE (2011)

[16] M. Calonder, V. Lepetit, C. Strecha, and P. Fua. Brief: Binary robust independent elementary features. In In European Conference on Computer Vision, 2010. 1, 2, 3, 5

[17] Klein, G., Murray, D.: Parallel tracking and mapping for small AR workspaces. In: 6th IEEE and ACM International Symposium on Mixed and Augmented Reality, 2007 (ISMAR 2007), pp. 225–234. IEEE (2007)

[18] R. Mur-Artal, J. M. M. Montiel and J. D. Tardós, "ORB-SLAM: A Versatile and Accurate Monocular SLAM System," in IEEE Transactions on Robotics, vol. 31, no. 5, pp. 1147-1163, Oct. 2015, doi: 10.1109/TRO.2015.2463671.

[19] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras," in IEEE Transactions on Robotics, vol. 33, no. 5, pp. 1255-1262, Oct. 2017, doi: 10.1109/TRO.2017.2705103.

[20] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. M. Montiel and J. D. Tardós, "ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual–Inertial, and Multimap SLAM," in IEEE Transactions on Robotics, vol. 37, no. 6, pp. 1874-1890, Dec. 2021, doi: 10.1109/TRO.2021.3075644.

[21] F. Endres, J. Hess, N. Engelhard, J. Sturm, D. Cremers and W. Burgard, "An evaluation of the RGB-D SLAM system," 2012 IEEE International Conference on Robotics and Automation, 2012, pp. 1691-1696, doi: 10.1109/ICRA.2012.6225199.

[22] B.D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision", Proc. DARPA Image Understanding Workshop, pp. 121-130, 1981.

[23] Baker, S., Matthews, I.: Lucas-Kanade 20 years on: a unifying framework. Int. J. Comput. Vis. 56(3), 221–255 (2004)

[24] Engel, J., Schöps, T., Cremers, D.: LSD-SLAM: large-scale direct monocular SLAM. In: European Conference on Computer Vision, pp. 834–849. Springer, Cham (2014)

[25] J. Engel, J. Stückler and D. Cremers, "Large-scale direct SLAM with stereo cameras," 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2015, pp. 1935-1942, doi: 10.1109/IROS.2015.7353631.

[26] J Engel, V Koltun, D Cremers et al., "Direct sparse odometry[J]", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 3, pp. 611, 2018.

[27] A. Kendall, M. Grimes and R. Cipolla, "PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization," 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2938-2946, doi: 10.1109/ICCV.2015.336.

[28] A. Kendall and R. Cipolla, "Modelling uncertainty in deep learning for camera relocalization", Proc. IEEE Int. Conf. Robot. Autom. (ICRA), pp. 4762-4769, May 2016.

[29] A. Kendall and R. Cipolla, "Geometric loss functions for camera pose regression with deep learning", Proc. 30th IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 6555-6564, 2017.

[30] R. Li, Q. Liu, J. Gui, D. Gu and H. Hu, "Indoor relocalization in challenging environments with dual-stream convolutional neural networks", IEEE Transactions on Automation Science and Engineering, 2017.

[31] S. Wang， R. Clark， H. Wen 和 N. Trigoni， “DeepVO：Towards end-to-end visual odometertry with deep recurrent convolutional neural networks ”， Robotics and Automation （ ICRA ） 2017 IEEE International Conference on， pp. 2043-2050， 2017.

[32] R. Li, S. Wang, Z. Long and D. Gu, "UnDeepVO: Monocular Visual Odometry Through Unsupervised Deep Learning," 2018 IEEE International Conference on Robotics and Automation (ICRA), 2018, pp. 7286-7291, doi: 10.1109/ICRA.2018.8461251.

[33] B. Ummenhofer， H. Zhou， J. Uhrig， N. Mayer， E. Ilg， A. Dosovitskiy， et al.， “DeMoN： Depth and Motion Network for learning monocular stereo ”， Conference on Computer Vision and Pattern Recognition （CVPR）， 2017.

[34] Y. Almalioglu, M. R. U. Saputra, P. P. B. d. Gusmão, A. Markham and N. Trigoni, "GANVO: Unsupervised Deep Monocular Visual Odometry and Depth Estimation with Generative Adversarial Networks," 2019 International Conference on Robotics and Automation (ICRA), 2019, pp. 5474-5480, doi: 10.1109/ICRA.2019.8793512.

[35] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi and A. Fitzgibbon, "Scene coordinate regression forests for camera relocalization in RGB-D images", Computer Vision and Pattern Recognition (CVPR) 2013 IEEE Conference on, pp. 2930-2937, 2013.

[36] Andreas Geiger, Philip Lenz and Raquel Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite", Conference on Computer Vision and Pattern Recognition (CVPR), 2012.

[37] Michael Burri, Janosch Nikolic, Pascal Gohl, Thomas Schneider, Joern Rehder, Sammy Omari, et al., "The EuRoC micro aerial vehicle datasets", The International Journal of Robotics Research, 2016.

[38] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, et al., "The cityscapes dataset for semantic urban scene understanding", Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3213-3223, 2016.

[39] J. Xiao， A. Owens and A. Torralba， “SUN3D： A Database of Big Spaces Reconstructed Using SfM and Object Labels ”， IEEE International Conference on Computer Vision （ICCV）， pp. 1625-1632， Dec. 2013.

[40] S. Fuhrmann， F. Langguth and M. Goesele， “Mve-a multiview reconstruction environment ”， Proceedings of the Eurographics Workshop on Graphics and Cultural Heritage （GCH）， vol. 6， pp. 8， 2014.

[41] Maddern W, Pascoe G, Linegar C, Newman P. 1 Year, 1000km: the Oxford robotCar dataset. The International Journal of Robotics Research (IJRR) 2017;36(1):3–15.

[42] D. Barnes, W. Maddern, G. Pascoe and I. Posner, "Driven to distraction: Self-supervised distractor learning for robust monocular visual odometry in urban environments", Proc. IEEE Int. Conf. Robot. Autom. (ICRA), pp. 1894-1900, 2018.

[43] X. Yang, X. Li, Y. Guan, J. Song and R. Wang, "Overfitting reduction of pose estimation for deep learning visual odometry," in China Communications, vol. 17, no. 6, pp. 196-210, June 2020, doi: 10.23919/JCC.2020.06.016.