A Comparison of Deep Learning-based Monocular Visual Odometry Algorithms

Eunju Jeong, Jaun Lee, and Pyojin Kim

Department of Mechanical Systems Engineering, Sookmyung Women's University, Seoul, Republic of Korea
{eunju0316, jul024, pjinkim}@sookmyung.ac.kr

Abstract. Visual odometry (VO) has recently attracted significant attention, as evidenced by the increasing interest in the development of autonomous mobile robots and vehicles. Studies have traditionally focused on geometry-based VO algorithms. These algorithms exhibit robust results under a restrictive setup, such as static and well-textured scenes. However, they are not accurate in challenging environments, such as changing illumination and dynamic environments. In recent years, VO algorithms based on deep learning methods have been developed and studied to overcome these limitations. However, there remains a lack of literature that provides a thorough comparative analysis of state-of-the-art deep learning-based monocular VO algorithms in challenging environments. This paper presents a comparison of four state-of-the-art monocular VO algorithms based on deep learning (DeepVO, SfMLearner, SC-SfMLearner, and DF-VO) in environments with glass walls, illumination changes, and dynamic objects. These monocular VO algorithms are based on supervised, unsupervised, and self-supervised learning integrated with multiview geometry. Based on the results of the evaluation on a variety of datasets, we conclude that DF-VO is the most suitable algorithm for challenging real-world environments.

Keywords: Monocular Visual Odometry, Deep Learning, Challenging Environment, Service Robot

1 Introduction

The accurate positioning of robots is one of the most important tasks in the development of autonomous robots. Visual odometry (VO) allows the robot to estimate its pose incrementally using a stream of images captured by the camera on the robot.

Over the past several decades, studies have concentrated on geometry-based VO algorithms. Geometry-based VO algorithms perform feature detection, such as SIFT and feature matching, and have exhibited superior performance. However, such methods provide lower accuracy in severe illumination changes and dynamic environments. Service robots frequently encounter challenging environments in the real world.

With advancements in robotics, the environments in which autonomous service robots are deployed and operated are expanding. This expansion brings focus on deep learning-based VO algorithms to solve the limitations of geometry-based VO. Deep learning-based VO can automatically learn various indoor and outdoor environments, as well as identify features effectively.

Early deep learning-based VO methods use supervised learning [10]. The requirement of the ground truth of camera poses is a limitation of supervised methods of VO algorithms. The implementation of deep learning-based VO algorithms has been extended to unsupervised learning [1, 12]. The most recently presented methods highlight the advantages of both traditional geometry-based and deep learning-based methods by integrating geometry-based methods with deep learning to compensate for the limitations of traditional geometry-based methods and the disadvantages of deep learning-based methods, namely a lack of accuracy owing to ignoring geometric information. With the emergence of various deep

learning-based VO algorithms, it is necessary to assess and compare the existing methods to determine the most reliable VO method for application to service robots in the real world.

Furthermore, monocular VO is economical and lightweight compared to stereo VO, making it useful for various service robots. We compare four state-of-the-art deep learning-based monocular VO algorithms: DeepVO [10], SfMLearner [12], SC-SfMLearner [1], and DF-VO [11]. Our goal is to identify the VO algorithm that is applicable to service robots in real-world environments including various changing conditions.

2 Related Work

Within the existing literature, there is a lack of research that satisfies our goals. Although many VO comparisons have been conducted, they focus only on geometry-based simultaneous localization and mapping (SLAM) or visual-inertial odometry (VIO) algorithms [2, 7]. The work in [2] only compares the existing geometry-based VIO approaches and performs an analysis only on flying robots. The study of [4] compares the existing VO algorithms. However, unlike this study, [4] focuses only on traditional geometry methods such as ORB-SLAM2, rather than deep learning-based methods. Similar to our goals, [6, 9] provide a reference for autonomous service robots under challenging conditions. However, the aim of [6] is to determine an effective SLAM system, and tests are only conducted using indoor datasets, whereas we test with indoor as well as outdoor datasets. Furthermore, [9] does not test VO algorithms in challenging environments. The work in [11] compares DF-VO with SC-SfMLearner and SfMLearner, which are three of the four monocular VO algorithms that we compare, but only on the KITTI datasets.

Most importantly, because the four algorithms (DeepVO [10], SfMLearner [12], SC-SfMLearner [1], and DF-VO [11]) have not been intensively compared in challenging environments that are commonly encountered in the real world, it is not easy to establish which monocular VO algorithm is suitable for service robots in real-world environments. We evaluate the algorithms using accuracy metrics such as the absolute trajectory error (ATE) and relative position error (RPE) [8]. We provide a comprehensive assessment of the publicly available state-of-the-art deep learning-based monocular VO algorithms by comparing them on the KITTI datasets [3], an outdoor urban environment, and author-collected real-world challenging datasets.

3 Deep Learning-Based Monocular VO Algorithms

Table 1
Classification of four monocular VO algorithms for comparison

	Learning / Algorithm	Supervised	Unsupervised	Self-supervised
Deep learning-based	DeepVO [10]	✓		
	SfMLearner [12]		✓	
	SC-SfMLearner [1]		✓	
Deep learning and geometry-based	DF-VO [11]			✓

3.1 DeepVO

DeepVO [10] is the first end-to-end method for monocular VO through deep learning. DeepVO uses a supervised training method that requires a ground-truth 6-DoF camera pose to train the network. DeepVO can achieve simultaneous representation learning and sequential modeling of monocular VO by combining convolutional neural networks (CNNs) with recurrent neural networks (RNNs). The CNNs capture different geometric features and patterns of different images, while the RNNs model the camera motion

from an image sequence. Sequential dependence and dynamic scenes of an image sequence, which humans cannot easily model, are automatically learned by the RNNs. DeepVO does not rely on any modules of the conventional VO methods, including camera calibration for pose estimation, and it does not require careful adjustment of the VO system parameters.

3.2 SfMLearner

SfMLearner [12] is one of the first deep learning-based monocular VO algorithms using unsupervised learning. The algorithm outputs the relative pose of the camera movement and depth of the input image. Although many subsequent VO algorithms based on unsupervised learning use stereo datasets to train their models, SfMLearner uses monocular datasets. The SfMLearner algorithm uses view synthesis during the training. A synthesized target view can be created from the depth, pose, and visibility of a nearby view in an input image. Using the warped source and target views, SfMLearner formulates a view synthesis objective as supervision. To improve the algorithm performance, SfMLearner trains an “explainability” prediction network. The network outputs an “explainability” mask based on where the network believes that the direct view synthesis will be accordingly modeled for each target pixel. This “explainability” mask is used as a weight for the view synthesis objective.

3.3 SC-SfMLearner

SC-SfMLearner [1] is another deep learning-based VO algorithm that uses unsupervised learning. The algorithm outputs the relative pose estimation and depth estimation. The structure of the algorithm is similar to that of SfMLearner, as the training is supervised by the photometric loss between the actual and synthesized images. To overcome the limitations of SfMLearner, such as dynamic scenes and occlusions, SC-SfMLearner incorporates the method of the self-discovered weight mask M, which differs from the "explainability" mask of SfMLearner. Without explicitly separating inconsistent scene structures, the self-discovered mask assigns low weights to inconsistent pixels and high weights to consistent pixels in a scene. To address scale inconsistency issues, SC-SfMLearner includes the use of geometry consistency loss, which minimizes the difference between two consecutive predicted depth maps that are related by the relative camera pose. Although this only directly affects the depth estimation, the tight link between the depth estimation and the pose estimation method enables scale-consistent pose prediction results to be achieved.

3.4 DF-VO

DF-VO [11] is a monocular VO algorithm that integrates traditional multi-geometry-based methods with deep predictions. DF-VO incorporates multi-view geometry and deep learning to overcome the limitations and to highlight the advantages of both the deep learning and geometry methods. DF-VO uses self-supervised learning to train and fine-tune the deep networks and does not require ground-truth data. DF-VO uses the optical flow and single-view depth prediction from the deep networks as an intermediate output to establish 2D-2D/3D-3D correspondences for camera pose estimation. Moreover, monocular VO exhibits a scale-drift issue in which errors accumulate. However, DF-VO is capable of providing scale-consistent predictions even in long sequences using well-trained deep neural networks. Depth models with consistent scales are used for scale recovery, which mitigates the scale-drift issue in most monocular VO/SLAM systems.

4 Experiment and Analysis

We implement four state-of-the-art deep learning-based VO methods on a desktop computer using an Intel Core i7-10700 CPU operating at 2.90 GHz with 8 GB of RAM. We use a PC equipped with a graphics card (NVIDIA GeForce RTX 2070). We test and evaluate four monocular VO algorithms, namely DeepVO, SfMLearner, SC-SfMLearner, and DF-VO with PyTorch using the source code provided by the authors on the GitHub website on Ubuntu 18.04 64-bit OS.

To compare the performance of the deep learning-based VO algorithms, we test and evaluate the algorithms on largely two datasets: the KITTI datasets and author-collected datasets with an iPhone 12 Pro Max.

The purpose of this comparison is to determine which VO algorithm performs the best in challenging real-world environments. We discuss the results in two main categories: general outdoor urban environments and challenging indoor/outdoor environments. The results from the KITTI datasets represent the performance in outdoor urban environments, whereas the results from the author-collected datasets represent environments with challenging real-world scenarios. Monocular VO algorithms tend to have scale-inconsistency issues. We aligned the scales of each resulting pose trajectory to the ground truth. We compare the results of the experiments with frequently used accuracy metrics: the absolute trajectory error (ATE) and relative pose error (RPE). The ATE compares the absolute distance between the estimated pose and ground-truth pose trajectory, making it more suitable for evaluating visual SLAM algorithms. The RPE is an error metric that captures the local accuracy of the trajectory over a fixed time interval Δ. This corresponds to the drift of the trajectory, which makes it a particularly useful accuracy metric for VO systems [8]. We discuss the results of our evaluation, focusing on the RPE, particularly the translation component. Smaller ATE and RPE values indicate higher pose accuracy. In Tables 2 and 3, the bold font represents the best results in terms of ATE and RPE, respectively, whereas the underlined font represents the second-best results.

4.1 KITTI Dataset

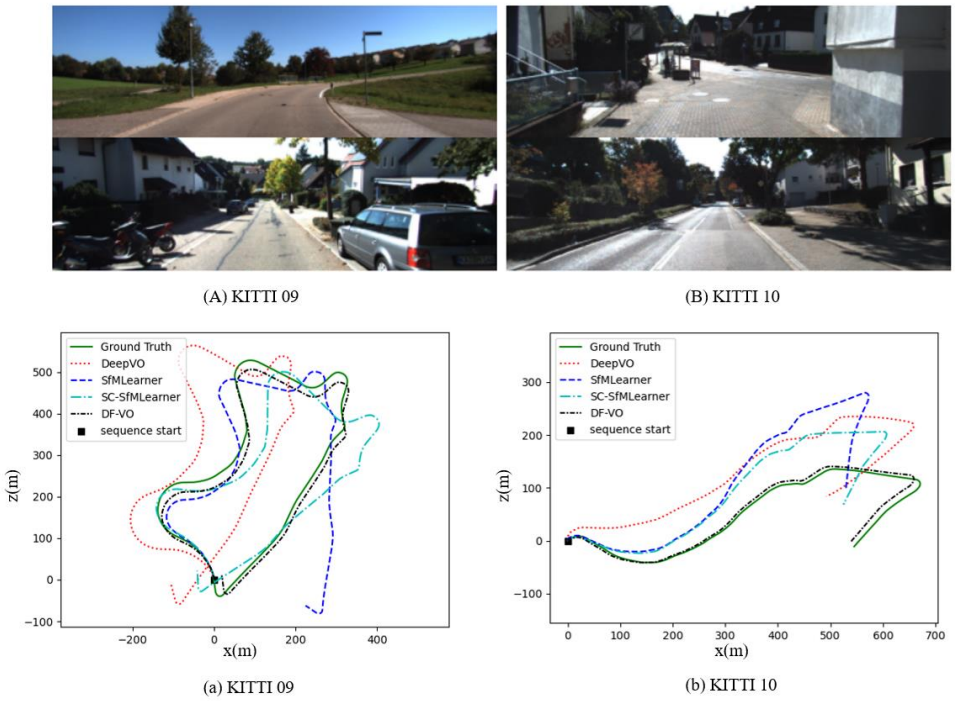

Fig. 1 Example images and trajectories of DeepVO, SfMLearner, SC-SfMLearner, and DF-VO in sequences 09 (left) and 10 (right) from KITTI odometry benchmark.

The KITTI dataset is a well-known and widely acknowledged dataset for training and testing VO algorithms. We use this dataset to compare the performances of the algorithms in urban outdoor environments. The dataset was recorded using a vehicle equipped with modern autonomous driving sensors [3]. We employ KITTI sequences 09 and 10, as illustrated in Figs. 1 (A) and (B), to compare the positional estimation. We specifically select sequence 09 owing to its scene diversity, which ranges from wide-open roads to tight residential spaces. Furthermore, the sequence 09 dataset enables testing how effectively each algorithm detects loop closure. We select sequence 10 because it represents outdoor environments in which autonomous service robots would generally operate.

Table 2
Results of DeepVO, SfMLeaner, SC-SfMLearner, and DF-VO on KITTI dataset sequences 09 and 10

Algorithm	Error	KITTI dataset sequence		Avg
		09	10	
DeepVO [10]	ATE (m)	30.70	22.76	26.73
	RPE (m/s)	0.831	1.002	0.917
	Runtime (s)	299.886	223.522	261.704
SfMLearner [12]	ATE (m)	61.69	**4.38**	33.04
	RPE (m/s)	0.215	0.069	0.142
	Runtime (s)	173.502	135.301	154.402
SC-SfMLearner [1]	ATE (m)	22.75	12.00	17.38
	RPE (m/s)	0.190	0.115	0.153
	Runtime (s)	66.022	58.946	62.484
DF-VO [11]	ATE (m)	**7.91**	**4.38**	**6.15**
	RPE (m/s)	**0.093**	**0.048**	**0.071**
	Runtime (s)	278.091	207.790	242.941

We use existing pretrained VO models that are provided by the authors on the GitHub website. All four deep learning-based monocular VO algorithms (DeepVO, SfMLearner, SC-SfMLearner, and DF-VO) were pretrained with several KITTI datasets. DF-VO, SC-SfMLearner, and SfMLearner were trained using sequences 00 to 08. We measure the runtime of the four algorithms for predicting the camera pose.

For sequence 09, DF-VO exhibits the best performance in pose accuracy in terms of the ATE, which is significantly different from that of the other three algorithms. DF-VO also shows the best result in terms of the RPE, with SC-SfMLearner being the second best. Fig. 1 presents the trajectory results of the KITTI datasets for sequences 09 and 10. All of the monocular VO algorithms are aligned with the starting point of the ground truth. The starting and ending points of the ground-truth trajectory of sequence 09 are the same. In terms of the trajectory estimation, we evaluate how effectively each algorithm detects the loop closing. As illustrated in Fig. 1, DF-VO provides the best detection of the loop closure.

For sequence 10, DF-VO and SfMLearner both exhibit the best results in terms of the ATE. In the case of the RPE, SC-SfMLearner shows the best performance after DF-VO. As we focus more on the RPE than the ATE, DF-VO has the best accuracy in both KITTI datasets, namely sequences 09 and 10. The algorithm that exhibits the second-best results is SfMLearner.

Although DF-VO was not trained with sequences 09 and 10, unlike the other three algorithms, it uses geometric information. The results demonstrate an accurate trajectory, and small ATE and RPE values. Thus, DF-VO demonstrates the highest effectiveness and robustness among the four deep learning-based VO approaches that best suit service robots in large-scale urban outdoor environments with roads and numerous buildings.

4.2 Author-Collected Dataset

Fig. 2 Example images in sequences 00 to 06 of author-collected dataset from iPhone 12 Pro Max

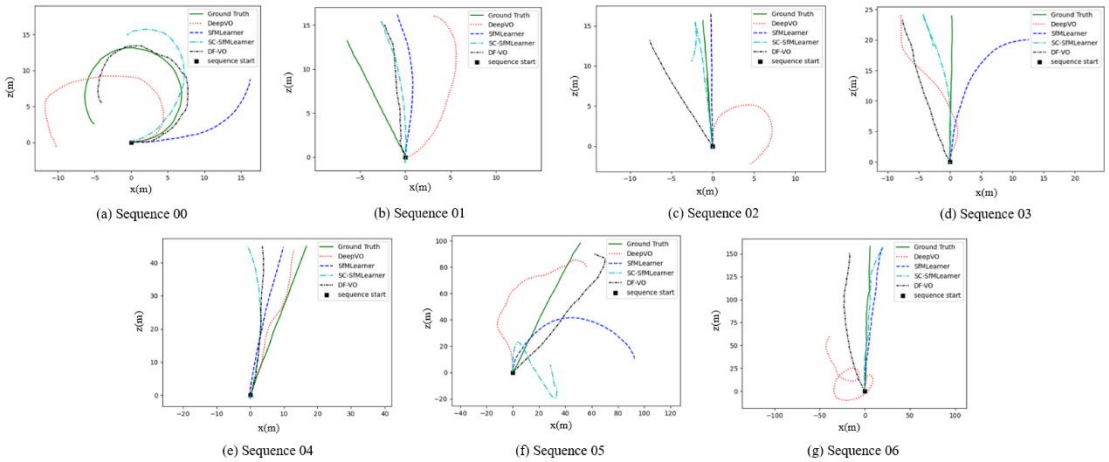

Fig. 3 Estimated trajectories with DeepVO, SfMLearner, SC-SfMLearner, and DF-VO in sequences 00 to 06 from author-collected dataset

As there is a lack of existing datasets for challenging environments, such as glass walls, dynamic objects, and illumination changes, we acquire consecutive RGB images and 6-DoF camera poses using an iPhone 12 Pro Max with ARKit VIO. We consider the recorded 6-DoF camera poses using ARKit as the ground-truth trajectory of the camera. We acquire continuous images of each dataset, as illustrated in Figs. 2 (A) to (F), at 30 Hz from the video captured using the iPhone 12 Pro Max, and we resize the obtained RGB images to 1226 × 370 for testing.

Table 3
Results of DeepVO, SfMLeaner, SC-SfMLearner, and DF-VO on sequences 00 to 06 from author-collected dataset

Algorithm	Error	Glass		Illumination		Dynamic			Avg
		00	01	02	03	04	05	06	
DeepVO [10]	ATE (m)	1.80	0.95	3.01	1.69	0.83	10.79	30.67	6.99
	RPE (m/s)	0.028	0.033	0.036	0.024	0.071	0.042	0.083	0.045
SfMLearner [12]	ATE (m)	4.03	**0.52**	**1.35**	1.87	**0.59**	11.70	**2.45**	3.22
	RPE (m/s)	**0.026**	0.022	**0.023**	**0.016**	0.060	0.040	**0.015**	0.029
SC-SfMLearner [1]	ATE (m)	2.63	0.60	1.62	1.35	2.35	25.32	7.89	5.97
	RPE (m/s)	0.027	0.026	0.029	0.018	0.058	0.030	0.026	0.031
DF-VO [11]	ATE (m)	**1.66**	0.55	1.53	**0.71**	1.24	**4.71**	4.99	**2.20**
	RPE (m/s)	0.034	**0.016**	0.026	0.021	**0.022**	**0.021**	0.022	**0.023**

4.2.1 Glass Wall

The inconsistency of the phase is a characteristic of glass walls that makes positional tracking difficult. We collect two datasets to compare the VO algorithms in this challenging environment. Fig. 2 (A) depicts a dataset with a camera navigating a circular track around glass walls with high transparency. Fig. 2 (B) presents a dataset of the camera navigating in a straight line next to a glass wall with high reflectivity.

DF-VO performs the best in terms of ATE in a glass wall environment with high transparency (sequence 00), followed by DeepVO. In terms of the RPE measurement, SfMLearner exhibits the best performance, followed by SC-SfMLearner. In the environment of glass walls with high reflectivity (sequence 01), SfMLearner performs the best, followed by DF-VO. Regarding the RPE, DF-VO yields the best results, followed by SfMLearner.

DF-VO achieves the most accurate motion estimation results in terms of the ATE in glass sequences 00 and 01, whereas SfMLearner performs the best with regard to the RPE. As we focus more on the RPE than the ATE, we can conclude that the SfMLearner is the algorithm that best fits environments with glass walls. The reason that DF-VO yield high accuracy is that it recognizes features of other non-glass objects, such as columns, in sequences 00 and 01.

4.2.2 Illumination Change

The classic geometry-based VO is accurate and reliable in controlled environments, such as well-textured environments with no illumination changes. However, the accuracy tends to be low in real-world environments with illumination changes [9, 11]. We acquire two sets of illumination change datasets, as illustrated in Fig. 2, to compare the performance of VO only for illumination change differences. Fig. 2 (C) depicts the indoors with a low amount of illumination, whereas Fig. 2 (D) shows a high amount of illumination in the same place as the first dataset (Fig. 2 (C)). The two datasets differ only in the amount of illumination. Both datasets represent static environments with no moving people or objects. Outdoor datasets are not strictly comparable only in terms of illumination differences because various illumination changes occur within a short time, such as shadows and weather changes, and numerous challenging environments include aspects such as dynamic objects. We acquire static indoor datasets in which we can control the illumination changes.

The algorithm that exhibits the smallest ATE and RPE differences between the dark and bright datasets (sequences 02 and 03) is the robust VO method under light-changing environments. SC-SfMLearner exhibits the smallest ATE difference between the dark and bright environments, with higher accuracy than SfMLearner. This is because, compared to SfMLearner, SC-SfMLearner was designed to be robust in illumination-changing environments. DF-VO is the algorithm with the smallest difference in the RPE results. As we focus more on the RPE than the ATE, we can conclude that DF-VO is the most robust algorithm in an illumination-changing environment.

4.2.3 Dynamic Objects

The majority of VO algorithms are based on the strong assumption that the surrounding environment is static. Furthermore, moving objects and people may distort the pose estimation. Moving objects in the scene may alter the measurements of the photo-consistency errors between consecutive images of datasets.

We acquire three datasets, 04 to 06, which are presented in Figs. 2 (E), (F), and (G). These datasets depict various dynamic objects in the real world, such as walking humans and driving cars. Among the three sequences, the first dataset, namely sequence 04 (Fig. 2 (E)), has the shortest sequence of 99.40 m. The dynamic surroundings mainly consist of moving cars in the driveway. Sequence 06 (Fig. 2 (G)) has the longest sequence length of 161.18 m.

Sequence 04 is the shortest of the dynamic datasets 04 to 06. In sequence 04, SfMLearner performs the best in terms of the ATE. In terms of the RPE, DF-VO provides the best results.

Sequence 05 has a sequence length of 111.64 m. DF-VO yields the smallest ATE value in this sequence, exhibiting the best performance among the four algorithms. Many repetitive patterns appeared in sequence 05, such as sidewalk blocks. DF-VO can identify the repetitive features using geometric knowledge. In terms of the RPE, DF-VO still shows the best performance. However, SC-SfMLearner also performs exceptionally well.

In terms of both the ATE and RPE, SfMLearner yields the best results in sequence 06. The second-best algorithm is DF-VO. Sequence 06 is the longest compared to sequences 04 and 05. It is also the most dynamic, with various people walking by. Although SC-SfMLearner is designed to be more robust to dynamic objects than SfMLearner, our experiments demonstrate the opposite results. This is because the extent of the dynamic surroundings in our dataset is different from that used in [1]. We believe that the extent of the dynamic surroundings in our dataset better fits the outdoor environment of the real world that service robots would encounter.

Overall, DF-VO performs the best in sequences 04 to 06, followed by SfMLearner, in terms of both ATE and RPE. We can conclude that DF-VO is the monocular VO algorithm that best suits dynamic outdoor environments, and it is the most suitable algorithm for autonomous service robots that are used in indoor/outdoor real-world environments.

5 Conclusions

We compare state-of-the-art deep learning-based VO algorithms to demonstrate a visual odometry algorithm that exhibits the most accurate performance in challenging environments for indoor/outdoor autonomous service robots. We evaluate the performance of four VO algorithms using KITTI datasets for outdoor urban environments. Furthermore, we evaluate author-collected indoor and outdoor datasets that best represent challenging environments that autonomous service robots could generally encounter. These environments include glass walls, illumination changes, and environments with dynamic moving objects. We compare the performance of the deep learning-based algorithms in each dataset and analyze their effectiveness in overcoming these challenges. We assess the VO algorithms based on the positional accuracy, graded by well-known accuracy metrics, namely the ATE and RPE. Through these analyses, we identify DF-VO as the best deep learning-based monocular VO method that provides the most robust results even in challenging environments and explore the limitations of each algorithm. The results and conclusions presented in this paper will provide insight for research on expanding the types of environments in which autonomous robots can be utilized.

Acknowledgement

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2021R1F1A1061397).

6 References

1. J. Bian, Z. Li, N. Wang, H. Zhan, C. Shen, M. Cheng, I. Reid (2019) Unsupervised scale-consistency depth and ego-motion learning from monocular video. Neural Information Processing Systems (NeurIPS)
2. J. Delmerico, D. Scaramuzza (2018) A Benchmark Comparison of Monocular Visual-Inertial Odometry Algorithms for Flying Robots. IEEE International Conference on Robotics and Automation (ICRA), pp. 2502-2509
3. A. Geiger, P. Lenz, C. Stiller, R. Urtasun (2013) Vision meets robotics: The KITTI dataset. International Journal of Robotics Research (IJRR)
4. M. He, C. Zhu, Q. Huang, B. Ren (2019) A review of monocular visual odometry. The Visual Computer 36(2): 1053-1065
5. A. Kasar (2019) Benchmarking and comparing popular visual SLAM algorithms. Asian Journal For Convergence In Technology (AJCT) ISSN -2350-1146
6. T. Lee, C. Kim, D. D. Cho (2019) A monocular vision sensor-based efficient SLAM method for indoor service robots. IEEE Transactions on Industrial Electronics, 66(1): 318-328
7. A. Merzlyakov, S. Macenski (2021) A Comparison of Modern General-Purpose Visual SLAM Approaches. IEEE/RSJ International Conference on Intelligent Robots and Systems
8. J. Sturm, N. Engelhard, F. Endres, W. Burgard, D. Cremers (2012) A benchmark for the evaluation of RGB-D SLAM systems. IEEE/RSJ International Conference on Intelligent Robots and Systems pp. 573-380
9. K. Wang, S. Ma, J. Chen, F. Ren (2020) Approaches, challenges and applications for deep visual odometry: Toward to complicated and emerging areas. IEEE Transactions on Cognitive and Developmental Systems, doi:10.1109/TCDS.2020.3038898
10. S. Wang, R. Clark, H. Wen, N. Trigoni (2017) Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. 2017 IEEE International Conference on Robotics and Automation (ICRA)
11. H. Zhan, C. S. Weerasekera, J. Bian, R. Garg, I. Reid (2020) Visual Odometry Revisited: What Should Be Learnt? IEEE International Conference on Robotics and Automation (ICRA) pp. 4203-4210
12. T. Zhou, M. Brown, N. Snavely, D. Lowe (2017) Unsupervised learning of depth and ego-motion from video. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)