UnDeepVO: Monocular Visual Odometry through Unsupervised Deep Learning

UnDeepVO: Monocular Visual Odometry through Unsupervised Deep Learning

Ruihao Li[1], Sen Wang[2], Zhiqiang Long[3] and Dongbing Gu[1]

***Abstract*— We propose a novel monocular visual odometry (VO) system called UnDeepVO in this paper. UnDeepVO is able to estimate the 6-DoF pose of a monocular camera and the depth of its view by using deep neural networks. There are two salient features of the proposed UnDeepVO: one is the unsupervised deep learning scheme, and the other is the absolute scale recovery. Specifically, we train UnDeepVO by using stereo image pairs to recover the scale but test it by using consecutive monocular images. Thus, UnDeepVO is a monocular system. The loss function defined for training the networks is based on spatial and temporal dense information. A system overview is shown in Fig. 1. The experiments on KITTI dataset show our UnDeepVO achieves good performance in terms of pose accuracy.**

I. Introduction

Visual odometry (VO) enables a robot to localize itself in various environments by only using low-cost cameras. In the past few decades, model-based VO or geometric VO has been widely studied and its two paradigms, feature-based method [1]–[3] and direct method [4]–[6], have both achieved great success. However, model-based methods tend to be sensitive to camera parameters and fragile in challenging settings, e.g., featureless places, motion blurs and lighting changes.

In recent years, data-driven VO or deep learning based VO has drawn significant attention due to its potentials in learning capability and the robustness to camera parameters and challenging environments. Starting from the relocalization problem with the use of supervised learning, Kendall et al. [7] first proposed to use a Convolutional Neural Network (CNN) for 6-DoF pose regression with raw RGB images as its inputs. Li et al. [8] then extended this into a new architecture for raw RGB-D images with the advantage of facing the challenging indoor environments. Video clips were employed in [9] to capture the temporal dynamics for relocalization. Given pre-processed optical flow, a CNN based frame-to-frame VO system was reported in [10]. Wang et al. [11] then presented a Recurrent Convolutional Neural Network (RCNN) based VO method resulting in a competitive performance against model-based VO methods. Ummenhofer [12] proposed "DeMoN" which can simultaneously estimate the camera's ego-motion, image depth, surface normal and optical flow. Visual inertial odometry with deep learning was also developed in [13] and [14].

[1]Ruihao Li, Dongbing Gu are with School of Computer Science and Electronic Engineering, University of Essex, Colchester, CO4 3SQ, UK. {rlig, dgu}@essex.ac.uk

[2]Sen Wang is with Edinburgh Centre for Robotics, Heriot-Watt University, Edinburgh, EH14 4AS, UK. s.wang@hw.ac.uk

[3]Zhiqiang Long is with College of Mechatronics and Automation, National University of Defense Technology, Changsha, China.

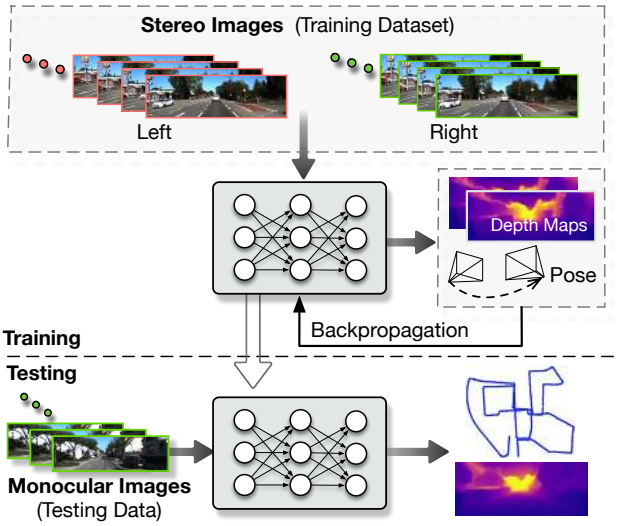


Fig. 1: System overview of the proposed UnDeepVO. After training with unlabeled stereo images, UnDeepVO can simultaneously perform visual odometry and depth estimation with monocular images. The estimated 6-DoF poses and depth maps are both scaled without the need for scale post-processing.

However, all the above mentioned methods require the ground truth of camera poses or depth images for conducting the supervised training. Currently obtaining ground truth datasets in practice is typically difficult and expensive, and the amount of existing labeled datasets for supervised training is still limited. These limitations suggest us to look for various unsupervised learning VO schemes, and consequently we can train them with easily collected unlabeled datasets and apply them to localization scenarios.

VO related unsupervised deep learning research mainly focuses on depth estimation, inspired by the image wrap technique "spatial transformer" [15]. Built upon it, Garg et al. [16] proposed a novel unsupervised depth estimation method by using the left-right photometric constraint of stereo image pairs. This method was further improved in [17] by wrapping the left and right images across each other. In this way, the accuracy of depth prediction was improved by penalizing both left and right photometric losses. Instead of using stereo image pairs, Zhou et al. [18] proposed to use consecutive monocular images to train and estimate both ego-motion and depth, but the system cannot recover the scale from learning monocular images. Nevertheless, these unsupervised

learning schemes have brought deep learning technologies and VO methods closer and showed great potential in many applications.

In this paper, we propose UnDeepVO, a novel monocular VO system based on unsupervised deep learning scheme (see Fig. 1). Our main contributions are as follows:

- We demonstrate a monocular VO system with recovered absolute scale, and we achieve this in an unsupervised manner by harnessing both spatial and temporal geometric constraints.
- Not only estimated pose but also estimated dense depth map are generated with absolute scales thanks to the use of stereo image pairs during training.
- We evaluate our VO system using KITTI dataset, and the results show UnDeepVO achieves good performance in pose estimation for monocular cameras.

Since UnDeepVO only requires stereo imagery for training without the need of labeled datasets, it is possible to train it with an extremely large number of unlabeled datasets to continuously improve its performance.

The rest of this paper is organized as follows. Section II introduces the architecture of our proposed system. Section III describes different types of losses used to facilitate the unsupervised training of our system. Section IV presents experimental results. Finally, conclusion is drawn in Section V.

II. System Overview

Our system is composed of a pose estimator and a depth estimator, as shown in Fig. 2. Both estimators take consecutive monocular images as inputs, and produce scaled 6-DoF pose and depth as outputs, respectively.

For the pose estimator, it is a VGG-based [19] CNN architecture. It takes two consecutive monocular images as input and predicts the 6-DoF transformation between them. Since rotation (represented by Euler angles) has high nonlinearity, it is usually difficult to train compared with translation. For supervised training, a popular solution is to give a bigger weight to the rotational loss [7] as a way of normalization. In order to better train the rotation with unsupervised learning, we decouple the translation and the rotation with two separate groups of fully-connected layers after the last convolutional layer. This enables us to introduce a weight normalizing the rotation and the translation predictions for better performance. The specific architecture of the pose estimator is shown in Fig. 2.

The depth estimator is mainly based on an encoder-decoder architecture to generate dense depth maps. Different from other depth estimation methods [17], [18] which produce disparity images (inverse of the depth) from the network, the depth estimator of UnDeepVO is designed to directly predict depth maps. This is because training trails report that the whole system is easier to converge when training in this way.

For most monocular VO methods, a predefined scale has to be applied. One feature of our UnDeepVO is to recover absolute scale of 6-DoF pose and depth, it is credited to

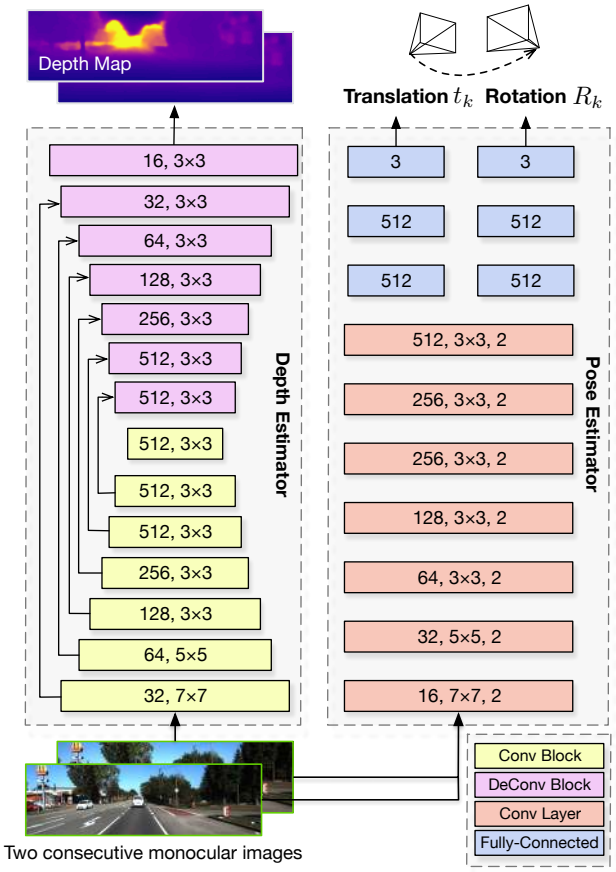


Fig. 2: Architecture of our UnDeepVO.

our training scheme shown in Fig. 3. During training, we feed both left images and right images into the networks, and get 6-DoF poses and depths of left sequences and right sequences, respectively. We then use the input stereo images, estimated depth images and 6-DoF poses to construct the loss function considering the geometry of a stereo rig.

As shown in Fig. 3, we utilize both spatial and temporal geometric consistencies of a stereo image sequence to formulate the loss function. The red points in one image all have the corresponding ones in another. Spatial geometric consistency represents the geometric projective constraint between the corresponding points in left-right image pairs, while temporal geometric consistency represents the geometric projective constraint between the corresponding points in two consecutive monocular images (more details in section IV). By using these constraints to construct the loss function and minimizing them all together, the UnDeepVO learns to estimate scaled 6-DoF poses and depth maps in an end-to-end unsupervised manner.

III. Objective Losses for Unsupervised Training

UnDeepVO is trained with losses through backpropagation. Since the losses are built on geometric constraints rather than labeled data, UnDeepVO is trained in an unsupervised manner. Its total loss includes spatial image losses and

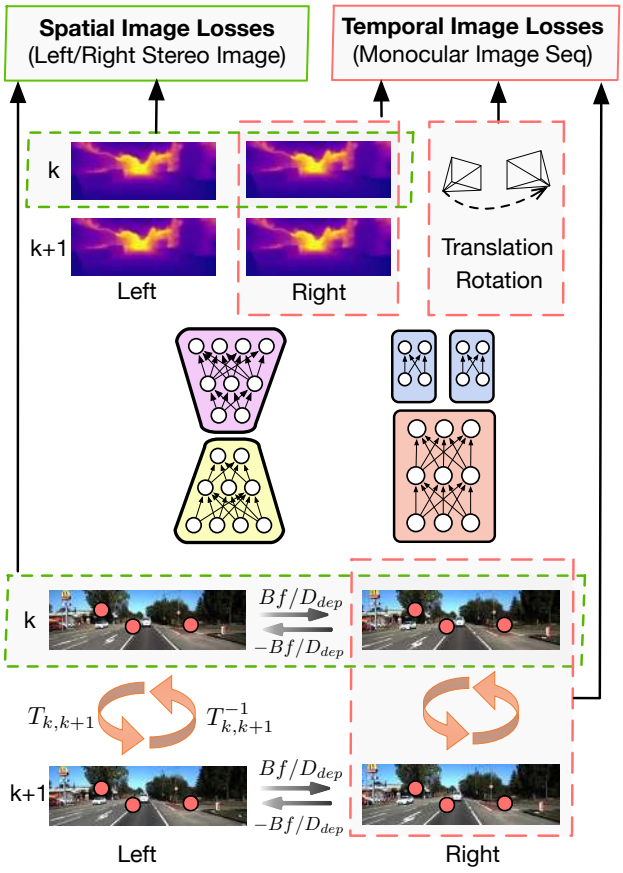


Fig. 3: Training scheme of UnDeepVO. The pose and depth estimators take stereo images as inputs to estimate 6-DoF poses and depth maps, respectively. The total loss including spatial losses and temporal losses can then be calculated based on raw RGB images, estimated depth maps and poses.

temporal image losses, as shown in Fig. 3. The spatial image losses drive the network to recover scaled depth maps by using stereo image pairs, while the temporal image losses are designed to minimize the errors on camera motion by using two consecutive monocular images.

A. Spatial Image Losses of a Stereo Image Pair

The spatial losses are constructed from the geometric constraints between left and right stereo images. It is composed of left-right photometric consistency loss, disparity consistency loss and pose consistency loss. UnDeepVO relies on these losses to recover the absolute scale for the monocular VO.

1) Photometric Consistency Loss: The left-right projective photometric error is used as photometric consistency loss to train the network. Specifically, for the overlapped area between two stereo images, every pixel in one image can find its correspondence in the other with horizontal distance D_p [16]. Assume $p_l(u_l, v_l)$ is a pixel in the left image and $p_r(u_r, v_r)$ is its corresponding pixel in the right image. Then, we have the spatial constraint $u_l = u_r$ and $v_l = v_r + D_p$. The distance D_p can be calculated by

$$D_p = Bf/D_{dep} \tag{1}$$

where B is the baseline of the stereo camera, f is the focal length and D_{dep} is the depth value of the corresponding pixel. We can construct a D_p map with the depths estimated from the network to define the constraints between the left and right images. With this spatial constraint and the calculated D_p map, we could synthesize one image from the other through "spatial transformer" [15]. The combination of an L1 norm and an SSIM term [20] is used to construct the left-right photometric consistency loss:

$$L^l_{pho} = \lambda_s L^{SSIM}(I_l, I'_l) + (1-\lambda_s) L^{l_1}(I_l, I'_l) \tag{2}$$

$$L^r_{pho} = \lambda_s L^{SSIM}(I_r, I'_r) + (1-\lambda_s) L^{l_1}(I_r, I'_r) \tag{3}$$

where I_l, I_r are the original left and right images respectively, I'_l is the synthesized left image from the original right image I_r, and I'_r is the synthesized right image from the original left image I_l, L^{SSIM} is the operation defined in [21] with a weight λ_s, and L^{l_1} is the L1 norm operation.

2) Disparity Consistency Loss: Similarly, the left and right disparity maps (inverse of depth) are also constrained by D_p. The disparity map UnDeepVO used is

$$D_{dis} = D_p \times I_W \tag{4}$$

where I_W is the width of original image size. Denote the left and right disparity maps as D^l_{dis} and D^r_{dis}, respectively. We can use D_p to synthesize $D^{l'}_{dis}, D^{r'}_{dis}$ from D^r_{dis}, D^l_{dis}. Then, the disparity consistency losses are

$$L^l_{dis} = L^{l_1}(D^l_{dis}, D^{l'}_{dis}) \tag{5}$$

$$L^r_{dis} = L^{l_1}(D^r_{dis}, D^{r'}_{dis}) \tag{6}$$

3) Pose Consistency Loss: We use both left and right image sequences to predict the 6-DoF transformation of the camera separately during training. Ideally, these two predicted transformations should be basically identical. Therefore, we can penalize the difference between them by

$$L_{pos} = \lambda_p L^{l_1}(\mathbf{x}'_l, \mathbf{x}'_r) + \lambda_o L^{l_1}(\varphi'_l, \varphi'_r) \tag{7}$$

where λ_p is the left-right position consistency weight, λ_o is the left-right orientation consistency weight, and $[\mathbf{x}'_l, \varphi'_l]$ and $[\mathbf{x}'_r, \varphi'_r]$ are the predicted poses from the left and right image sequences, respectively.

B. Temporal Image Losses of Consecutive Monocular Images

Temporal loss is defined according to the geometric constraints between two consecutive monocular images. It is an essential part to recover the 6-DoF motion of camera. It comprises photometric consistency loss and 3D geometric registration loss.

1) Photometric Consistency Loss: The photometric loss is computed from two consecutive monocular images. Similar to DTAM [4], in order to estimate 6-DoF transformation, the projective photometric error is employed as the loss to minimize. Denote I_k, I_{k+1} as the kth and $(k+1)$th image frame, respectively, and $p_k(u_k, v_k)$ as one pixel in I_k, and $p_{k+1}(u_{k+1}, v_{k+1})$ as the corresponding pixel in I_{k+1}. Then, we can derive p_{k+1} from p_k through

$$p_{k+1} = KT_{k,k+1}D_{dep}K^{-1}p_k \tag{8}$$

where K is the camera intrinsics matrix, D_{dep} is the depth value of the pixel in the kth frame, $T_{k,k+1}$ is the camera coordinate transformation matrix from the kth frame to the $(k+1)$th frame. We can synthesize I'_k and I'_{k+1} from I_{k+1} and I_k by using estimated poses and "spatial transformer" [15]. Therefore, the photometric losses between the monocular image sequence are

$$L^k_{pho} = \lambda_s L^{SSIM}(I_k, I'_k) + (1-\lambda_s)L^{l_1}(I_k, I'_k) \tag{9}$$

$$L^{k+1}_{pho} = \lambda_s L^{SSIM}(I_{k+1}, I'_{k+1}) + (1-\lambda_s)L^{l_1}(I_{k+1}, I'_{k+1}) \tag{10}$$

2) 3D Geometric Registration Loss: 3D geometric registration loss is to estimate the transformation by aligning two point clouds, similar to the Iterative Closest Point (ICP) technique. Assume $P_k(x,y,z)$ is a point in the kth camera coordination. It can then be transformed to the $(k+1)$th camera coordination as $P'_k(x,y,z)$ by using $T_{k,k+1}$. Similarly, points in the $(k+1)$th frame can be transformed to kth frame. Then, the 3D geometric registration losses between two monocular images are

$$L^k_{geo} = L^{l_1}(P_k, P'_k) \tag{11}$$

$$L^{k+1}_{geo} = L^{l_1}(P_{k+1}, P'_{k+1}) \tag{12}$$

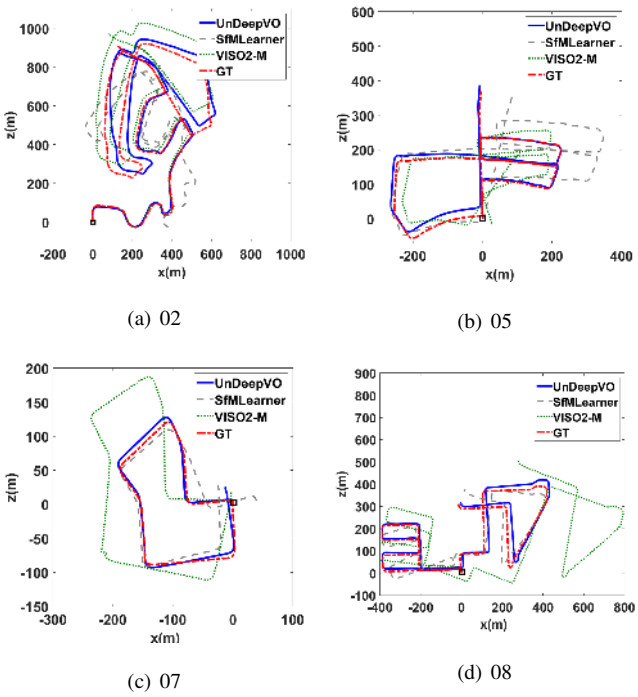
In summary, the final loss function of our system combines the previous spatial and temporal losses together. The left-right photometric consistency loss has been used in [16] and [17] to estimate depth map. [18] introduced the photometric loss of a monocular image sequence for ego-motion and depth estimation. However, to the best of our knowledge, UnDeepVO is the first to recover both scaled camera poses and depth maps by benefiting all these losses together with the 3D geometric registration and pose consistency losses.

IV. Experimental Evaluation

In this section, we evaluated the proposed UnDeepVO system.[1] The network models were implemented with the TensorFlow framework and trained with NVIDIA Tesla P100 GPUs. For testing, we used a laptop equipped with NVIDIA GeForce GTX 980M and Intel Core i7 2.7GHz CPU. The GPU memory needed for pose estimation is less than 400MB with 40Hz real-time performance.

Adam optimizer was employed to train the network for up to 20-30 epochs with parameter $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The learning rate started from 0.001 and decreased by half for every 1/5 of total iterations. The sequence length of images feeding to the pose estimator was 2. The size of image input to the networks was 416×128. We also resized the output images to a higher resolution to compute the losses and fine-tuned the networks in the end. Different kinds of data augmentation methods were used to enhance the performance and mitigate possible overfitting, such as image color augmentation, rotational data augmentation and left-right pose estimation augmentation. Specifically, we randomly selected 20% images for color augmentation with random brightness in range [0.9, 1.1], random gamma in range [0.9, 1.1] and random color shifts in range [0.9, 1.1]. For rotational data augmentation, we increased the proportion of rotational data to achieve better performance in rotation estimation. Pose estimation consistency of left-right images was also used for left-right pose estimation augmentation.

[1] Video:https://www.youtube.com/watch?v=5RdjO93wJqo&t



(a) 02 (b) 05 (c) 07 (d) 08

Fig. 4: Trajectories of Sequence 02, 05, 07 and 08. Results of SfMLearner [18] are post-processed with 7-DoF alignment to ground truth since it cannot recover the scale. UnDeepVO and SfMLearner use images with size 416×128. Images used by VISO2-M are 1242×376.

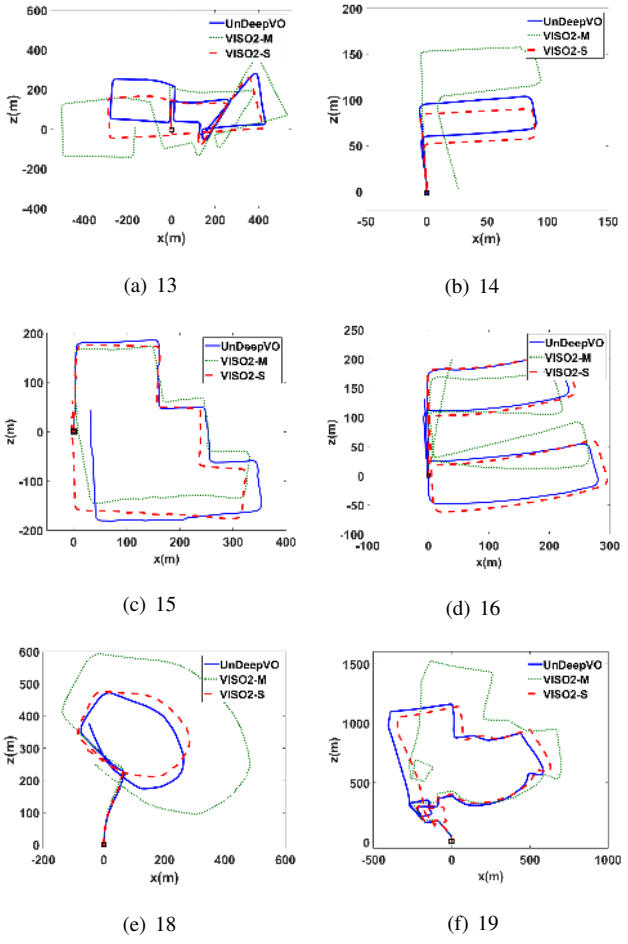
A. Trajectory Evaluation

We adopted the popular KITTI Odometry Dataset [22] to evaluate the proposed UnDeepVO system, and compared the results with SfMLearner [18], monocular VISO2-M and ORB-SLAM-M (without loop closure). In order to implement fair qualitative and quantitative comparison, we used the same training data as in SfMLearner [18] (sequences: 00-08). The trajectories produced by different methods are shown in Fig. 4, the comparison here shows the goodness of the network fit and is meaningful for structure-from-motion

TABLE I: VO results with our proposed UnDeepVO. All the methods listed in the table did not use loop closure. Note that monocular VISO2-M and ORB-SLAM-M (without loop closure) did not work with image size 416 × 128, the results were obtained with image size 1242×376. 7-DoF (6-DoF + scale) alignment with the ground-truth is applied for SfMLearner and ORB-SLAM-M.

Seq.	UnDeepVO (416×128)		SfMLearner [18] (416×128)		VISO2-M (1242×376)		ORB-SLAM-M (1242×376)	
	$t_{rel}(\%)$	$r_{rel}(^\circ)$	$t_{rel}(\%)$	$r_{rel}(^\circ)$	$t_{rel}(\%)$	$r_{rel}(^\circ)$	$t_{rel}(\%)$	$r_{rel}(^\circ)$
00	**4.14**	**1.92**	65.27	6.23	18.24	2.69	25.29	7.37
02	5.58	2.44	57.59	4.09	**4.37**	**1.18**	×	×
05	**3.40**	**1.50**	16.76	4.06	19.22	3.54	26.01	10.62
07	**3.15**	**2.48**	17.52	5.38	23.61	4.11	24.53	10.83
08	**4.08**	**1.79**	24.02	3.05	24.18	2.47	32.40	12.13
mean	**4.07**	**2.02**	36.23	4.56	17.93	2.80	27.05	10.23

- t_{rel}: average translational RMSE drift (%) on length of 100m-800m.
- r_{rel}: average rotational RMSE drift (°/100m) on length of 100m-800m.



(a) 13 (b) 14 (c) 15 (d) 16 (e) 18 (f) 19

Fig. 5: Trajectories of KITTI dataset with our UnDeepVO. No ground truth of poses is available for these sequences. Trajectories with both monocular VISO2-M and stereo VISO2-S are plotted. Our UnDeepVO works well on these sequences and is comparable to VISO2-S.

problem. Note that all the methods took monocular images for testing, and we post-process the scales for SfMLearner and ORB-SLAM-M as they cannot recover the scale of pose and depth. VISO2-M employed the fixed camera height for scale recovery. For ORB-SLAM-M, we disabled the local mapping and loop closure in order to perform VO only for comparison. The KITTI Odometry Dataset only provides the ground-truth of 6-DoF poses for Sequence 00-10. As shown in Fig. 4, the trajectories of UnDeepVO are qualitatively closest to the ground truth among all the methods. For sequences 11-21, there is no ground-truth available, and the trajectories of our method and VISO2-M are given in Fig. 5. The results of stereo VISO2-S (image resolution 1242 × 376) are provided for reference. As shown in the figure, our system's performance is comparable to that of VISO2-S.

The detailed results (shown in Fig. 4) are listed in Table I for quantitative evaluation. We use the standard evaluation method provided along with KITTI dataset. Average translational root-mean-square error (RMSE) drift (%) and average rotational RMSE drift (°/100m) on length of 100m-800m are adopted. Since SfMLearner and ORB-SLAM-M cannot recover the scale of 6-DoF poses, we aligned their poses to the ground-truth with 6-DoF and scale (7-DoF). For monocular VISO2-M and ORB-SLAM without loop closure, they can not work with our input settings (image resolution 416 × 128), so we provide the results of both system with high resolution 1242 × 376. All the methods here did not use any loop closure technology. As shown in Table I, our method achieves good pose estimation performance among the monocular methods even with low resolution images and without the scale post-processing.

B. Depth Estimation Evaluation

Our system can also produce the scaled depth map by using the depth estimator. Fig. 6 shows some raw RGB images and their corresponding depth images estimated from our system. As shown in Fig. 6, the different depths of cars and trees are explicitly estimated, even the depth of trunks is predicted successfully. The detailed depth estimation results are listed in Table II. As shown in the table, our method outperforms the supervised one [23] and the unsupervised one without scale [18], but performs not as good as [17]. This could be caused by a few reasons. First, we only used parts of KITTI dataset (KITTI odometry dataset) for training while all other methods use full KITTI dataset to train their networks. Second, [17] used higher resolution (512 × 256) input and a different net (ResNet-based architecture). Third,

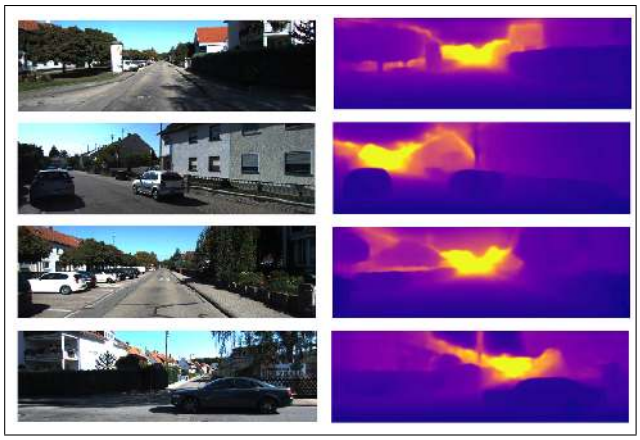


Fig. 6: Depth images produced by our depth estimator. The left column are raw RGB images, and the right column are the corresponding depth images estimated.

TABLE II: Depth estimation results on KITTI using the split of Eigen et al. [23].

Methods	Dataset	Scale	Error metric			
			Abs Rel	Sq Rel	RMSE	RMSE log
Eigen [23]	K (raw)	✓	0.214	1.605	6.563	0.292
MonoDepth [17]	K (raw)	✓	0.148	1.344	5.927	0.247
SfMLearner [18]	K (raw)	×	0.208	1.768	6.856	0.283
UnDeepVO	K (odo)	✓	0.183	1.73	6.57	0.268

the temporal image sequence loss we used could introduce some noise (such as moving objects) for depth estimation.

V. Conclusions

In this paper, we presented UnDeepVO, a novel monocular VO system with unsupervised deep learning. The system makes use of spatial losses and temporal losses between stereo image sequences for unsupervised training. During testing, the proposed system can perform the pose estimation and dense depth map estimation with monocular images. Our system recovers the scale during the training stage, which distincts itself from other model based or learning based monocular VO methods. In general, unsupervised learning based VO methods have the potential to improve their performance with the increasing size of training datasets. In the next step, we will investigate how to train the UnDeepVO with large amount of datasets to improve its performance, such as robustness to image blurs, camera parameters, or illumination changes. In the future, we also plan to extend our system to a visual SLAM system to reduce the drift. Developing an unsupervised DeepVO system with stereo cameras or RGB-D cameras is also in consideration.

ACKNOWLEDGMENT

The authors would like to thank Robin Dowling for his support in experiments. The first author has been financially supported by scholarship from China Scholarship Council. This work was supported in part by EPSRC Robotics and Artificial Intelligence ORCA Hub (grant No. EP/R026173/1).

References

[1] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "MonoSLAM: Real-time single camera SLAM," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 1052–1067, 2007.

[2] G. Klein and D. Murray, "Parallel tracking and mapping for small AR workspaces," in *Mixed and Augmented Reality, 2007. ISMAR 2007. 6th IEEE and ACM International Symposium on*. IEEE, 2007, pp. 225–234.

[3] R. Mur-Artal, J. Montiel, and J. D. Tardos, "ORB-SLAM: a versatile and accurate monocular SLAM system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[4] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison, "DTAM: Dense tracking and mapping in real-time," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2011, pp. 2320–2327.

[5] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-scale direct monocular SLAM," in *European Conference on Computer Vision (ECCV)*. Springer, 2014, pp. 834–849.

[6] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[7] A. Kendall, M. Grimes, and R. Cipolla, "PoseNet: A convolutional network for real-time 6-DOF camera relocalization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 2938–2946.

[8] 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.

[9] R. Clark, S. Wang, A. Markham, N. Trigoni, and H. Wen, "VidLoc: 6-DoF video-clip relocalization," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[10] G. Costante, M. Mancini, P. Valigi, and T. A. Ciarfuglia, "Exploring representation learning with CNNs for frame-to-frame ego-motion estimation," *IEEE robotics and automation letters*, vol. 1, no. 1, pp. 18–25, 2016.

[11] S. Wang, R. Clark, H. Wen, and N. Trigoni, "DeepVO: Towards end-to-end visual odometry with deep recurrent convolutional neural networks," in *Robotics and Automation (ICRA), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2043–2050.

[12] B. Ummenhofer, H. Zhou, J. Uhrig, N. Mayer, E. Ilg, A. Dosovitskiy, and T. Brox, "DeMoN: Depth and motion network for learning monocular stereo," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[13] R. Clark, S. Wang, H. Wen, A. Markham, and N. Trigoni, "VINet: Visual-Inertial odometry as a sequence-to-sequence learning problem." in *AAAI*, 2017, pp. 3995–4001.

[14] S. Pillai and J. J. Leonard, "Towards visual ego-motion learning in robots," *arXiv preprint arXiv:1705.10279*, 2017.

[15] M. Jaderberg, K. Simonyan, A. Zisserman, *et al.*, "Spatial transformer networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 2017–2025.

[16] R. Garg, G. Carneiro, and I. Reid, "Unsupervised CNN for single view depth estimation: Geometry to the rescue," in *European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 740–756.

[17] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[18] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*, 2015.

[20] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Is L2 a good loss function for neural networks for image processing?" *ArXiv e-prints*, vol. 1511, 2015.

[21] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[22] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[23] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Advances in neural information processing systems*, 2014, pp. 2366–2374.