CNN-SVO: Improving the Mapping in Semi-Direct Visual Odometry Using Single-Image Depth Prediction

Shing Yan Loo[1,2], Ali Jahani Amiri[1], Syamsiah Mashohor[2], Sai Hong Tang[2] and Hong Zhang[1]

***Abstract*— Reliable feature correspondence between frames is a critical step in visual odometry (VO) and visual simultaneous localization and mapping (V-SLAM) algorithms. In comparison with existing VO and V-SLAM algorithms, semi-direct visual odometry (SVO) has two main advantages that lead to state-of-the-art frame rate camera motion estimation: direct pixel correspondence and efficient implementation of probabilistic mapping method. This paper improves the SVO mapping by initializing the mean and the variance of the depth at a feature location according to the depth prediction from a single-image depth prediction network. By significantly reducing the depth uncertainty of the initialized map point (i.e., small variance centred about the depth prediction), the benefits are twofold: reliable feature correspondence between views and fast convergence to the true depth in order to create new map points. We evaluate our method with two outdoor datasets: KITTI dataset and Oxford Robotcar dataset. The experimental results indicate that the improved SVO mapping results in increased robustness and camera tracking accuracy.**

I. Introduction

Visual odometry (VO) and visual simultaneous localization and mapping (V-SLAM) have been actively researched and explored in the robotics field, including autonomous driving. As cameras become affordable and ubiquitous, being able to estimate camera poses reliably from an image sequence leads to important robotics applications, such as autonomous vehicle navigation.

Matching features between the current frame and previous frames have been one of the most important steps in solving visual odometry (VO) and visual simultaneous localization and mapping (V-SLAM). There are two main feature matching methods: indirect method and direct method. Indirect methods [1]–[3] require feature extraction, feature description, and feature matching. These methods rely on matching the intermediate features (e.g., descriptors), and they perform poorly in images with weak gradients and textureless surfaces where descriptors are failed to be matched. Direct methods [4], [5], by contrast, do not need feature description, and they operate directly on pixel intensities; therefore, any arbitrary pixels (e.g., corners, edges, or the whole image) can be sampled and matched, resulting in reliable feature matching even in images with poor texture. However, matching pixels directly requires depths of the pixels to be recovered, and such matching is defined in the direct formulation that jointly optimizes structure and motion.

[1]The authors are with Department of Computing Science, University of Alberta, Canada {lsyan,jahaniam,hzhang}@ualberta.ca

[2]The authors are with the Faculty of Engineering, Universiti Putra Malaysia, Malaysia {syamsiah,saihong}@upm.edu.my

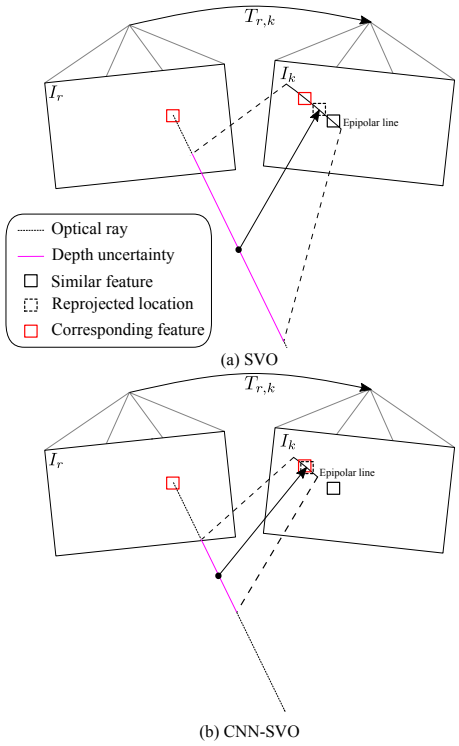


Fig. 1: Proposed map point initialization strategy. Each initialized map point has a mean depth (black dot) and an interval in which the corresponding feature should lie, as shown by the magenta line. Note that larger depth uncertainty can allow the erroneous match to happen (as illustrated in (a) where the depth filter could converge to the "similar feature" rather than the "corresponding feature"). Our improved map point initialization method (see (b)) has lower depth uncertainty for identifying the corresponding feature

Interestingly, semi-direct visual odometry (SVO) [6] is a hybrid method that combines the strength of direct and indirect methods for solving structure and motion, offering an efficient probabilistic mapping method to provide reliable map points for direct camera motion estimation. Unfortunately, one main limitation in SVO is that the map point is initialized with large depth uncertainty. Fig. 1(a) shows that the initialization of a map point with large depth uncertainty by SVO can lead to erroneous feature correspondence due to the large search range along the epipolar line.

In this paper, we propose to initialize new map points with depth prior from a single-image depth prediction neural network [7] (i.e., small variance centred about the pre-

illumination invariance.

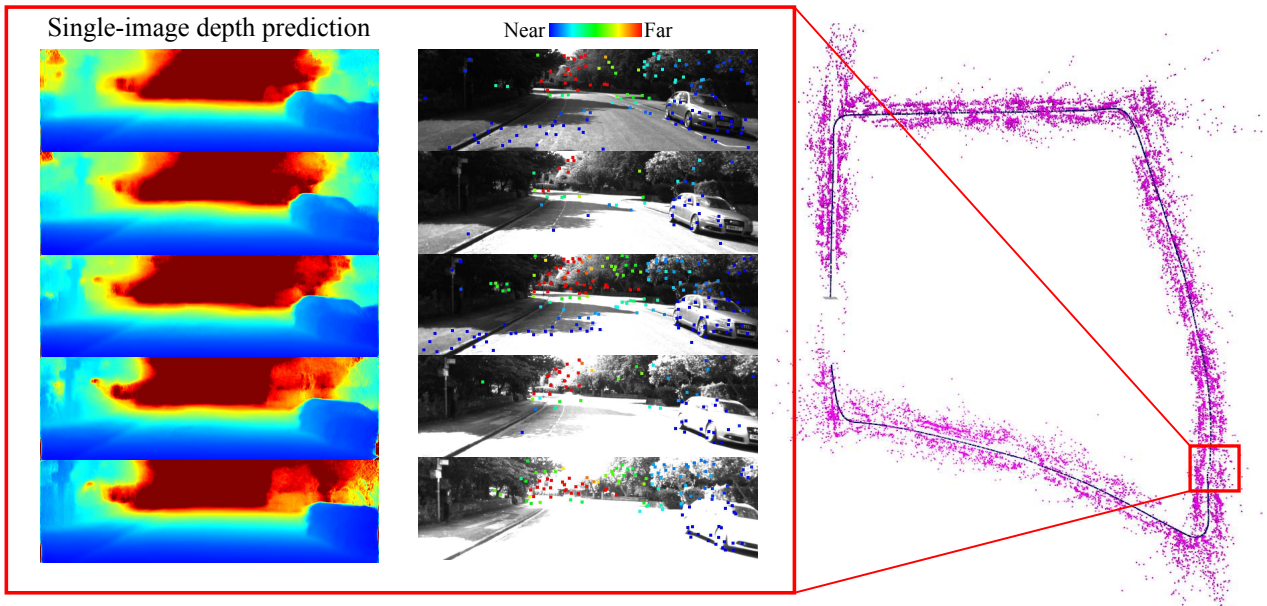


Fig. 2: CNN-SVO: Camera motion estimation in the HDR environment. (Left) The single-image depth prediction model demonstrates the illumination invariance property in estimating depth maps, and the colour-coded reprojected map points on a sample sequence of five consecutive frames show the reprojected map points to those frames for camera motion estimation (best viewed in colour). Note that CNN-SVO only predicts depth maps for the keyframes. (Right) Camera trajectory (depicted with line) and map points in magenta generated by CNN-SVO[1]

dicted depth), such that the uncertainty for identifying the corresponding features is vastly reduced (see Fig. 1(b)). Because of the robust feature correspondence and small depth uncertainty, the map point is likely to converge to its true depth quickly. Overall, the improved SVO mapping, we refer to as CNN-SVO, is able to handle challenging lighting condition, thanks to the illumination invariance property in estimating depth maps (see Fig. 2).

II. Methods

In this section, we briefly cover the working principle of SVO in Section II-A for the sake of completeness. Next, we detail our improved initialization of the map points in Section II-B.

A. Review of the SVO algorithm

First, we explain the terminology used in the rest of the paper. A feature is one of the 2D points in the image extracted from FAST corner detector [8]. To perform feature correspondence, a small image patch that is centred at the feature location is used to find its corresponding patch in the nearby view; therefore, we refer to feature correspondence as matching small image patches. A map point is a 3D point projected from a feature location whose depth is known.

SVO [6] contains two threads running in parallel: tracking thread and mapping thread. In the tracking thread, the camera pose of a new frame is obtained by minimizing the photometric residuals between the reference image patches (from which the map points are back-projected) and the image patches that are centred at the reprojected locations in the new frame. The optimization steps for obtaining camera pose through the inverse compositional formulation [9] can be found in [6]. Concurrently, the mapping thread creates new map points using two processes: initialization of new map points with large depth uncertainty and update of depth uncertainty of the map points with *depth-filters*; consequently, a new map point is inserted in the map if the depth uncertainty of the map point is small.

Given the camera poses of two frames one of which is the keyframe, the depth of a feature can be obtained using the following two steps: finding the feature correspondence along the epipolar line in the non-keyframe, and then recover the depth via triangulation. Since the occurrence of outlier matching is inevitable, a *depth-filter* is modeled as a two-dimensional distribution [10], [11]: the first dimension describes the probability distribution of the depth, and the second dimension models the inlier probability. Therefore, given a set of depth measurements, a *depth-filter* approximates the mean depth and the variance (the first dimension) of the feature and separates the outliers from the inliers (the second dimension). The depth uncertainty (i.e., approximated variance) of the feature is updated when there is a new depth measurement, and the *depth-filter* is considered to have converged if the updated depth uncertainty is small. Then, the converged *depth-filters* that contain the true depths are used to create new map points by back-projecting the points at those feature locations according to their true depth. In this paper, we are focusing on improving the mapping in SVO [6]; therefore, we assume that the poses of the images

[1] https://github.com/yan99033/CNN-SVO

can be successfully recovered in the tracking thread.

B. Improved initialization of map points in SVO mapping

The effective implementation of *depth-filter* in SVO mapping and the use of direct matching of pixels has enabled SVO to achieve high frame rate camera motion estimation. However, SVO mapping initializes new map points in a reference keyframe with large uncertainty and their mean depths are set to the average scene depth in the reference frame. While such an initialization strategy is reasonable for the scene with one dominant plane—e.g., the floor plane—where the dominant depth information exists, the large depth uncertainty has limited the capability of the mapping to determine the true depths of the map points for the scene in which the depths of the map points vary considerably. Particularly, large depth uncertainty introduces two problems: possible erroneous feature correspondence along the epipolar line in the nearby frames and a high number of depth measurements to converge to the true depth.

With single-image depth prediction as the prior knowledge of the scene geometry, our CNN-SVO able to obtain a much better estimate of the mean and a smaller initial variance of the *depth-filter* than SVO to allow it to converge to the true depth of the map point. Fig. 3 illustrates the CNN-SVO pipeline, in which we add the CNN depth estimation module (marked in green) to provide strong depth priors in the map points initialization process when a keyframe is selected—the initialization of *depth-filters*.

Given a set of triangulated depth measurements, the goal of using *depth-filter* is to separate the good measurements from the bad measurements: good measurements are normally distributed around the true depth, and bad measurements are uniformly distributed within an interval $[\rho_i^{\min}, \rho_i^{\max}]$. Specifically given a set of triangulated inverse depth measurements $\rho_i^1, \rho_i^2, \cdots, \rho_i^N$ that correspond to the same feature, the measurement ρ_i^n is modeled in SVO using a *Gaussian + Uniform* mixture model:

$$p(\rho_i^n|\rho_i,\gamma_i) = \gamma_i \mathcal{N}(\rho_i^n|\rho_i,\tau_i^2) + (1-\gamma_i)\mathcal{U}(\rho_i^n|\rho_i^{\min},\rho_i^{\max}) \tag{1}$$

where ρ_i is the true inverse depth, τ_i^2 the variance of the inverse depth, and γ_i the inlier ratio. Assuming the inverse depth measurements $\rho_i^1, \rho_i^2, \cdots, \rho_i^N$ are independent, [10] shows that the approximation of the true inverse depth posterior can be computed incrementally by the product of a Gaussian distribution for the depth and a Beta distribution for the inlier ratio:

$$q(\rho_i,\gamma_i|a_n,b_n,\mu_n,\sigma_n^2) = Beta(\gamma_i|a_n,b_n)\mathcal{N}(\rho_i|\mu_n,\sigma_n^2) \tag{2}$$

where a_n and b_n are the parameters in the Beta distribution, and μ_n and σ_n^2 the mean and variance of the Gaussian depth estimate. The incremental Bayesian update step for a_n, b_n, μ_n, and σ_n^2 is described in detail in [10], [11]. Once σ_n^2 is lower than a threshold, the *depth-filter* is converged to the true depth.

Hence, each *depth-filter* is initialized with the following parameters: the mean of the inverse depth μ_n, the variance of the inverse depth σ_n^2, and the inlier ratio a_n and b_n. Table I compares the initialization of the parameters between SVO and CNN-SVO. The key difference is that CNN-SVO initializes the mean and the variance of the feature using learned scene depth instead of using the average and minimum scene depths in the reference keyframe. We empirically found that setting the depth variance to $\frac{1}{(6d_{\text{CNN}})^2}$ provides adequate room for noisy depth prediction to converge; we will be losing the absolute scale if the depth variance is large (e.g., replacing 6 with a higher number) by allowing more uncertainty in the measurement. Based on the initialized μ_n and σ_n^2, a depth interval $[\rho_i^{\min}, \rho_i^{\max}]$ can be defined by

$$\rho_i^{\min} = \mu_n + \sqrt{\sigma_n^2} \tag{3}$$

$$\rho_i^{\max} = \begin{cases} 0.00000001, & \text{if } \mu_n - \sqrt{\sigma_n^2} < 0 \quad (4a) \\ \mu_n - \sqrt{\sigma_n^2}, & \text{otherwise} \quad (4b) \end{cases}$$

so that the corresponding feature can be found in the limited search range along the epipolar line in the nearby view (see Fig. 1). By obtaining strong depth prior from the single-image depth prediction network, the benefits are twofold: smaller uncertainty in identifying feature correspondence and faster map point convergence, as illustrated in Fig. 4.

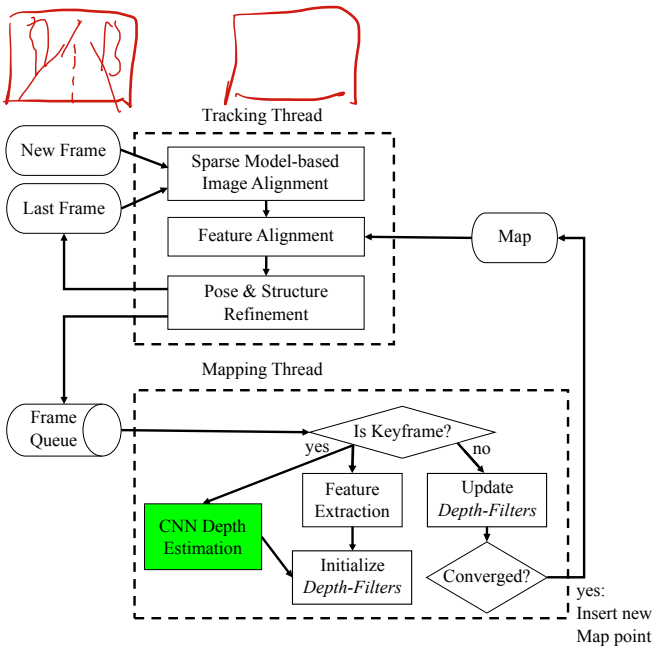


Fig. 3: The CNN-SVO pipeline. Our work augments the SVO pipeline [6] with the CNN depth estimation module (marked in green) to improve the mapping in SVO

III. Evaluation

We compare our method against the state-of-the-art direct and indirect methods, namely direct sparse odometry (DSO) [5], semi-direct visual odometry (SVO) [6], and ORB-SLAM without loop closure [12]. We use the absolute trajectory error (ATE) as the performance metric that has been used in the aforementioned papers. In addition, we indicate with 'X' for methods that are unable to complete the sequence due to lost tracking in the middle of the sequence (see Section III-A).

TABLE I: A comparison between SVO and CNN-SVO in the intialization of parameters. The parameters are defined by some prior knowledge of the scene, where d_{avg} is the average scene depth in the reference keyframe, d_{CNN} the depth prediction from the single-image depth prediction network, and d_{min} the minimum scene depth in the reference keyframe

	SVO	CNN-SVO
μ_n	$\frac{1}{d_{\text{avg}}}$	$\frac{1}{d_{\text{CNN}}}$
σ_n^2	$\frac{1}{(6d_{\text{min}})^2}$	$\frac{1}{(6d_{\text{CNN}})^2}$

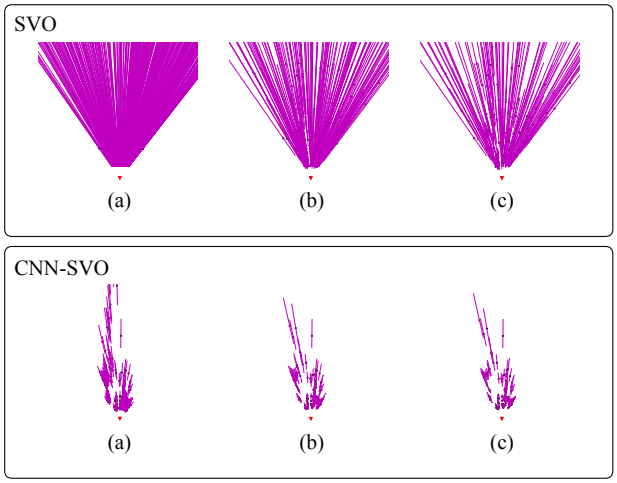


Fig. 4: The improved mapping strategy is able to provide faster convergence of the map points. The length of the magenta line represents the depth uncertainty. (a) initialization of the depth filters where SVO uses a large interval to model the uncertainty of each initial map point whereas CNN-SVO uses a short interval; (b) depth estimates of the map points by the *depth-filters* after three updates; (c) depth estimates of the map points by the *depth-filters* after five updates

To provide depth prediction in the initialization of map points in CNN-SVO, we adopt the Resnet50 variant of the encoder-decoder architecture from [7] that has already been trained on Cityscape dataset. Next, we fine-tune the network on stereo images in KITTI raw data excluding KITTI Odometry Sequence 00-10 using original settings in [7] for 50 epochs. To produce consistent structural information, even on overexposed or underexposed images, the brightness of the images has been randomly adjusted throughout the training, creating the effect of illumination variation. This consideration is useful for a neural network to handle high dynamic range (HDR) environments (see Fig. 2).

To design the system with real-time capability, we resize the images to 512×256 for depth map inference, and then we resize the depth map back to original shape for VO processing. While two separate threads have been designed to handle mapping and tracking, GPU is used to provide the depth maps for the keyframes. The hardware is an Intel i7 processor[2] with NVidia GeForce GTX Titan X graphics card.

To scale the depth prediction for other datasets, the scaled depth d_{current} can be obtained by the inferred depth d_{trained} multiplied by the ratio of current focal length f_{current} to trained focal length f_{trained}, that is:

$$d_{\text{current}} = \frac{f_{\text{current}}}{f_{\text{trained}}} d_{\text{trained}} \tag{5}$$

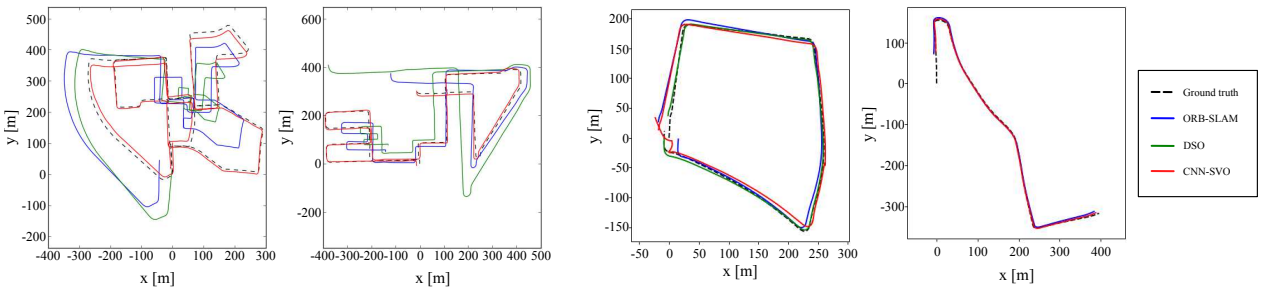
We use eleven KITTI Odometry sequences and nine Oxford Robotcar sequences for performance benchmarking. As for the images, we use the left camera from KITTI binocular stereo setup and the centre camera of the Bumblebee XB3 trinocular stereo setup from Oxford Robotcar. Both of the image streams are captured using global shutter cameras. Note that the ground truth poses from Oxford Robotcar dataset are not reliable for evaluation [13], because of the poor and inconsistent GPS signals; we still use the ground truth for both quantitative and qualitative evaluation purposes. The frame rates are 10 frames per second (FPS) and 16 FPS for KITTI and Oxford Robotcar, respectively. To maintain the same aspect ratio that is used by the network input, the images in the Oxford Robotcar dataset have been cropped to 1248x376 throughout the evaluation process. We skip the first 200 frames for all the Oxford Robotcar sequences because of the extremely overexposed images at the beginning of the sequences. Since the network has not been trained on Oxford Robotcar dataset, we analyze the scale of the odometry relative to absolute scale for both datasets (see Section III-C).

We set the maximum and the minimum number of tracked features in a frame to 200 and 100, respectively. Regarding the *depth-filter*, we modify SVO to use 5 previous keyframes to increase the number of measurements in the *depth-filters*. We also enable bundle adjustment during the evaluation process.

A. Accuracy evaluation

The ATEs of KITTI dataset and Oxford Robotcar dataset are collected with a median of 5 runs, and they are shown in Table II and Table III, respectively. Our system is able to track all the sequences except for KITTI Sequence 01, because of failure to match features accurately in the scene with repetitive structure. We also demonstrate that our competitors fail to track most of the Oxford Robotcar sequences, which contain severely overexposed images. While ORB-SLAM is able to track features in consistent lighting conditions, the vanished textural information in overexposed and underexposed images has resulted in failure to match feature in these HDR environments. The main reason of tracking failure in DSO is its inability of affine brightness modeling to handle severe brightness change in the sequences, and the problem has also been reported in stereo DSO [14]. Scale drift is also noticeable in the trajectories on KITTI dataset produced by DSO and ORB-SLAM (without loop closure). SVO is designed to perform well in a planar scene; therefore, it fails to identify corresponding features effectively in the outdoor

[2]Intel i7-4790K, 4 cores, 4.0GHz, 32GB RAM



(a) (b)

Fig. 5: Qualitative comparison of camera trajectories produced by ORB-SLAM (without loop closure), DSO, and CNN-SVO. (a) KITTI Sequence 00 and 08; (b) Oxford Robotcar Sequence 2014-05-06-12-54-54 and 2014-06-25-16-22-15. SVO is not included in this figure because it is not able to complete the trajectory due to tracking and mapping failures

TABLE II: Absolute keyframe trajectory RMSE (in metre) on KITTI dataset

Sequence	SVO	CNN-SVO	DSO	ORB-SLAM (w/o loop closure)
00	X	**17.5269**	113.1838	77.9502
01	X	X	X	X
02	X	50.5119	116.8108	**41.0064**
03	X	3.4588	1.3943	**1.0182**
04	58.3970	2.4414	**0.422**	0.9302
05	X	**8.1513**	47.4605	40.3542
06	X	**11.5091**	55.6173	52.2282
07	X	**6.5141**	16.7192	16.546
08	X	**10.9755**	111.0832	51.6215
09	X	**10.6873**	52.2251	58.1742
10	X	**4.8354**	11.090	18.4765

TABLE III: Absolute keyframe trajectory RMSE (in metre) on Oxford Robotcar dataset

Sequence	SVO	CNN-SVO	DSO	ORB-SLAM (w/o loop closure)
2014-05-06-12-54-54	X	8.657	**4.708**	10.6596
2014-05-06-13-09-52	X	**9.1947**	X	X
2014-05-06-13-14-58	X	**10.1865**	X	X
2014-05-06-13-17-51	X	**8.26**	X	X
2014-05-14-13-46-12	X	**13.7513**	X	X
2014-05-14-13-50-20	X	**32.4199**	X	X
2014-05-14-13-53-47	X	**6.3017**	X	X
2014-05-14-13-59-05	X	6.1515	**2.4532**	X
2014-06-25-16-22-15	X	**3.703**	X	6.558

scene, where depths of the features can vary considerably. We attribute the robust tracking of CNN-SVO to its ability to match features in consecutive frames with additional depth information, even when the images are overexposed or underexposed (see Fig. 2). The qualitative comparison of the camera trajectories can be found in Fig. 5 for KITTI dataset and Robotcar dataset, respectively. In Fig. 5 (b), an S-like curve is produced by CNN-SVO near the end of trajectory in Sequence 2014-05-06-12-54-54, which is caused by a moving car in front of the camera. Since the network has not been trained on Oxford Robotcar sequences, the experimental results suggest generalization ability to the structurally similar scene.

B. Runtime evaluation

Local BA (about 29 ms) and single-image depth prediction (about 37 ms) have been the most demanding processes in the pipeline, but both processes are only required when new keyframes are created. Despite the computational demand, we experimentally found that CNN-SVO runs faster at 16 FPS with Oxford Robotcar dataset than 10 FPS with KITTI dataset. This is due to the close distance between frames in high frame rate sequence, and hence lesser keyframes are selected relative to the total number of frames from the sequence. For this reason, real-time computation can be achieved.

C. Scale evaluation

Since the network is trained on rectified stereo images with known baseline, we examine the scale of the odometry based on predicted depth from the network. Table IV (a) shows that the scale of the odometry is close to absolute scale in KITTI dataset because the training images are mostly from KITTI dataset. For Oxford Robotcar dataset, we scale the depth predictions using Eq. 5, and the scale of the VO is between 0.9 and 0.97 (see Table IV (b)). We offer two possible explanations for the inconsistent odometry scale. First, as mentioned in the Oxford Robotcar dataset documentation, the provided ground truth poses are not accurate, and the reasons are as follows: inconsistent GPS signals and scale drift in the large-scale map (see Section III in [13]). Second, the single-image depth prediction network has not been trained on the images in the Oxford Robotcar dataset, so the recovery of absolute scale cannot be guaranteed.

TABLE IV: Scale relative to absolute scale in VO output from CNN-SVO

(a) KITTI Dataset

Sequence	Scale
Sequence 00	0.9296
Sequence 01	X
Sequence 02	0.921
Sequence 03	1.0811
Sequence 04	1.1876
Sequence 05	0.9837
Sequence 06	0.9602
Sequence 07	1.0246
Sequence 08	1.0014
Sequence 09	1.043
Sequence 10	1.0512

(b) Oxford Robotcar Dataset

Sequence	Scale
2014-05-06-12-54-54	0.8953
2014-05-06-13-09-52	0.9321
2014-05-06-13-14-58	0.9172
2014-05-06-13-17-51	0.9399
2014-05-14-13-46-12	0.9103
2014-05-14-13-50-20	0.9737
2014-05-14-13-53-47	0.9427
2014-05-14-13-59-05	0.9473
2014-06-25-16-22-15	0.9236

IV. Conclusion

In this paper, we have improved SVO mapping, called CNN-SVO, by initializing the map points with low uncertainty and the mean depth obtained from a single-image depth prediction neural network. The proposed method has two main advantages: (1) features can be matched effectively by limiting the search range along the epipolar line in nearby views, assuming the camera poses are known, and (2) the map points are initialized with lower depth uncertainty, therefore they are able to converge to their true depths faster. With the combination of single-image depth prediction and implementation of *depth-filters*, CNN-SVO can perform mapping and estimate camera motion reliably. Thanks to the illumination invariance property in the single-image depth prediction network, depth maps produced from overexposed or underexposed images can still be used to facilitate feature correspondence between views, overcoming a key limitation of the original SVO.

Nevertheless, there are still shortcomings we are planning to address in the future. First, the threshold of the map point uncertainty is increased to allow map points with larger uncertainty to be inserted for camera motion tracking. This is due to the limited observations of the corresponding features that can be found in the nearby frame as a consequence of limited frame rate. Hence, the increase in uncertainty threshold implicitly assumes accurate depth prediction from the single-image depth prediction network. Second, although the network is able to produce depth maps from overexposed images, it still could not produce useful depth map with blank image—i.e., completely overexposed image. Because blank images rarely occur in an extended period of time, we estimate the pose of the blank images with constant velocity model until new features can be extracted. Then local BA is applied to jointly correct the map points and camera poses. This problem can be mitigated using exposure compensation algorithm [15]. Lastly, we facilitate feature matching by limiting the search space of the corresponding feature along the epipolar line in nearby frames. This feature matching strategy does increase the tolerance of illumination change, but it does not solve the inherent problem of photometric constancy assumption in direct methods. Thus, incorporation of additional photometric calibration [16] can further improve the feature matching performance.

References

[1] H. Lim, J. Lim, and H. J. Kim, "Real-Time 6-DOF Monocular Visual SLAM in a Large-Scale Environment," in *Proc. IEEE International Conference on Robotics and Automation (ICRA'14)*. Hong Kong, China: IEEE, May 2014, pp. 1532–1539.

[2] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: A Versatile and Accurate Monocular SLAM System," *IEEE Trans. Robot.*, vol. 31, no. 5, pp. 1147–1163, 2015.

[3] S. Song, M. Chandraker, and C. C. Guest, "Parallel, Real-Time Monocular Visual Odometry," in *Proc. IEEE International Conference on Robotics and Automation (ICRA'13)*. Karlsruhe, Germany: IEEE, May 2013, pp. 4698–4705.

[4] J. Engel, T. Schps, and D. Cremers, "LSD-SLAM: Large-scale Direct Monocular SLAM," in *Proc. European Conference on Computer Vision (ECCV'14)*. Zurich, Switzerland: Springer, Sept. 2014, pp. 834–849.

[5] J. Engel, V. Koltun, and D. Cremers, "Direct Sparse Odometry," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 3, pp. 611–625, 2018.

[6] C. Forster, M. Pizzoli, and D. Scaramuzza, "SVO: Fast Semi-Direct Monocular Visual Odometry," in *Proc. IEEE International Conference on Robotics and Automation (ICRA'14)*. Hong Kong, China: IEEE, May 2014, pp. 15–22.

[7] C. Godard, O. M. Aodha, and G. J. Brostow, "Unsupervised Monocular Depth Sstimation with Left-Right Consistency," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*. Honolulu, Hawaii: IEEE, July 2017.

[8] E. Rosten, R. Porter, and T. Drummond, "Faster and better: A machine learning approach to corner detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 1, pp. 105–119, 2010.

[9] S. Baker and I. Matthews, "Lucas-kanade 20 years on: A unifying framework," *Int. J. Comput. Vis.*, vol. 56, no. 3, pp. 221–255, 2004.

[10] C. Forster, Z. Zhang, M. Gassner, M. Werlberger, and D. Scaramuzza, "SVO: Semi-Direct Visual Odometry for Monocular and Multicamera Systems," *IEEE Trans. Robot.*, vol. 33, no. 2, pp. 249–265, 2017.

[11] G. Vogiatzis and C. Hernndez, "Video-based, real-time multi-view stereo," *Image Vis. Comput.*, vol. 29, no. 7, pp. 434 – 441, 2011.

[12] R. Mur-Artal and J. D. Tardos, "ORB-SLAM2: an Open-Source SLAM System for Monocular, Stereo and RGB-D Cameras," *IEEE Trans. Robot.*, vol. 33, no. 5, pp. 1255–1262, 2016.

[13] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000 km: The Oxford RobotCar Dataset," *Int. J. Robotics Res.*, vol. 36, no. 1, pp. 3–15, 2017.

[14] R. Wang, M. Schwŕer, and D. Cremers, "Stereo DSO: Large-Scale Direct Sparse Visual Odometry with Stereo Cameras," in *Proc. IEEE Internation Conference on Computer Vision (ICCV'17)*, Venice, Italy, Oct. 2017.

[15] Z. Zhang, C. Forster, and D. Scaramuzza, "Active exposure control for robust visual odometry in hdr environments," in *Proc. IEEE International Conference on Robotics and Automation (ICRA'17)*. Marina Bay Sands, Singapore: IEEE, May 2017, pp. 3894–3901.

[16] P. Bergmann, R. Wang, and D. Cremers, "Online Photometric Calibration of Auto Exposure Video for Realtime Visual Odometry and SLAM," *IEEE Robot. Autom. Lett.*, vol. 3, no. 2, pp. 627–634, 2018.