Robotics and Autonomous Systems 132 (2020) 103597

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

journal homepage: www.elsevier.com/locate/robot

A Robust Stereo Feature-aided Semi-direct SLAM System

Xiangrui Zhao [a,1], Lina Liu [a,1], Renjie Zheng [c], Wenlong Ye [c], Yong Liu [a,b,*]

[a] Institute of Cyber-Systems and Control, Zhejiang University, Hangzhou, China
[b] State Key Laboratory of Industrial Control Technology, Zhejiang University, Hangzhou, China
[c] Alibaba Group Inc., Hangzhou, China

ARTICLE INFO

Article history:
Available online 7 July 2020

Keywords:
Visual SLAM
Hybrid method
Image brightness rectification

ABSTRACT

In autonomous driving, many intelligent perception technologies have been put in use. However, visual SLAM still has problems with robustness, which limits its application, although it has been developed for a long time. We propose a feature-aided semi-direct approach to combine the direct and indirect methods in visual SLAM to allow robust localization under various situations, including large-baseline motion, textureless environment, and great illumination changes. In our approach, we first calculate inter-frame pose estimation by feature matching. Then we use the direct alignment and a multi-scale pyramid, which employs the previous coarse estimation as a priori, to obtain a more precise result. To get more accurate photometric parameters, we combine the online photometric calibration method with visual odometry. Furthermore, we replace the Shi–Tomasi corner with the ORB feature, which is more robust to illumination. For extreme brightness change, we employ the dark channel prior to weaken the halation and maintain the consistency of the image. To evaluate our approach, we build a full stereo visual SLAM system. Experiments on the publicly available dataset and our mobile robot dataset indicate that our approach improves the accuracy and robustness of the SLAM system.

1. Introduction

Perception and localization in the unknown environment are the foundation of autonomous driving. In an outdoor open space, the car can use the Global Positioning System(GPS) to obtain an accurate pose estimation. The localization does not drift over time and location and is available in most scenarios. However, in the mountains of the wild and the urban environment with many tall buildings, GPS signals are often obscured, and the localization accuracy drops sharply, which dramatically increases the difficulty of navigation and vehicle control.

The most commonly used solution in autonomous driving is employing a 3D LiDAR with pre-build high-precision maps. LiDAR can perform accurate distance measurement with only a few centimeters of error. It can easily achieve high localization accuracy even in GPS-denied environments. Although the consumer-level LiDAR continues to progress currently and its price gradually decreases, there is only one manufacturer of mass-production automotive-grade LiDAR, which means that LiDAR is still far from practical applications.

Vision is an essential way for autonomous vehicles to perceive the environment. It relies on small-sized and inexpensive cameras only and provides sufficient information. In driver assistance systems, the vision has been used for a long time, such as lane detection and car distance measurement. In recent years, visual SLAM technology that provides maps and localization has made remarkable achievements. Its localization accuracy has reached a practical level. Besides, the visual SLAM system needs fewer changes to the vehicle when it is deployed, and can even be made as a plug and play module in the future, which has great potential for application.

However, the current technical immaturity severely restricts the practical performance of visual SLAM, especially in textureless environment, large-baseline motion ,and various illumination. General speaking, the feature-based (indirect) method [1–4] has been regarded as a more suitable approach for the cases with large baseline, fast motion and varied illumination compared with direct method [5–10], as the feature points used in the indirect methods are more robust to scale, rotation, and illumination than the gradient points used in the direct method. The main drawback of the indirect method is that it is less adaptable to the textureless environment where the feature point detection often fails, while the direct method can perform much better, as it employs pixel-to-pixel match based on the hypothesis of constant brightness. The map generated by the indirect method is much sparser than the direct method. Thus it is a hot research trend to fuse the

* Corresponding author at: Institute of Cyber-Systems and Control, Zhejiang University, Hangzhou, China.
E-mail address: yongliu@iipc.zju.edu.cn (Y. Liu).
[1] Xiangrui Zhao and Lina Liu contribute equally to this work.

https://doi.org/10.1016/j.robot.2020.103597
0921-8890/

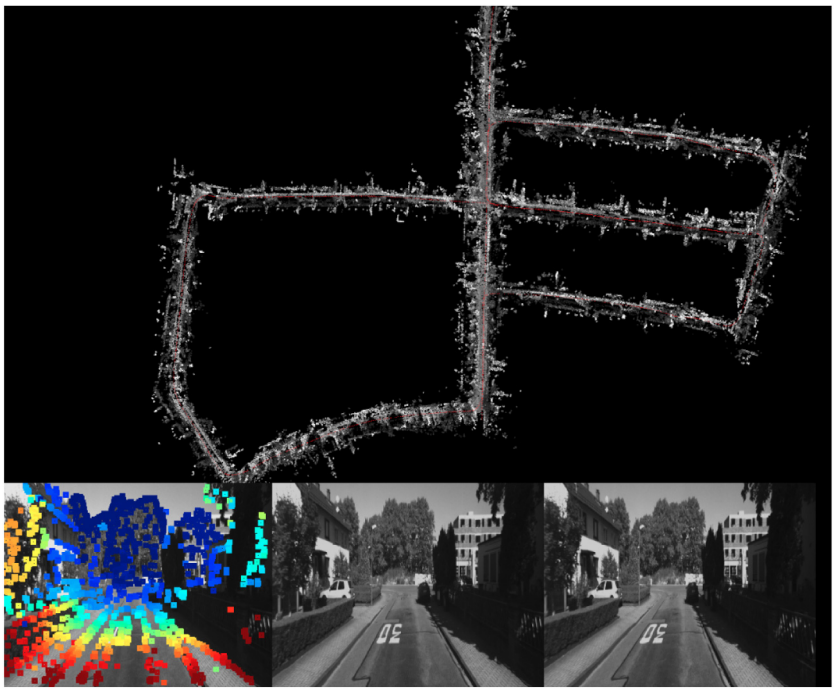


Fig. 1. Trajectory and point cloud on KITTI sequence 05.

advantages of both the direct and the indirect methods in the SLAM system.

In this paper, we propose a novel approach to fuse both the direct and the indirect methods together through a hybrid pyramid, which enables our approach robust to large-baseline motion, textureless environment, and illumination change (see Fig. 1). The main contributions of our work are given as follows:

- We perform online photometric calibration to get more precise photometric parameters used in the direct method. By replacing the Shi–Tomasi corner with ORB feature point, we improve the accuracy of calibration and its robustness to illumination change.
- We design a hybrid pyramid to fuse direct and indirect methods for improving estimation accuracy and robustness.
- We build a full stereo semi-direct visual SLAM system with loop-closure detection, pose graph optimization, and relocalization module to evaluate our approach. Experiments show that our method has superior accuracy and robustness compared with the state-of-the-art ORB-SLAM2[4].
- In terms of extreme brightness change, we employ the dark channel prior to remove halation and maintain the consistency of image pixels, which avoids direct method failure.

2. Related work

The feature-based frontend has long been considered as the mainstream method of visual odometry, which makes full use of the robustness of feature points to illumination and scale. It first selects feature points from the image that are more robust to illumination and scale. These points generally do not change much between frames and can be used for data association between frames. Then, the descriptors of these feature points are extracted for matching. After the feature matching between the two frames is obtained, the pose transformation between the two frames can be calculated through the eight-point method, PnP, or reprojection error minimization. MonoSLAM [1] is the first real-time monocular SLAM system. It extracts few feature points in the frontend and uses extended kalman filter as backend. PTAM [2] is a milestone in the history of visual SLAM. It has an optimization-based backend that applies bundle adjustment to optimize camera poses and key points, which further improves the accuracy of pose estimation. ORB-SLAM [3,4] is one of the most advanced and robust SLAM systems. It employs bag-of-words [11] to perform loop closure, making it suitable for wide range motion.

Feature-based indirect methods rely on feature points. The extraction and description of features usually cost much time. And in a textureless environment where has little features, few feature points will lead to a decrease in accuracy and pose estimation failure. Recently, some direct methods have appeared, aiming to use the image to estimate poses directly, skip the key point selection and matching steps, and do not perform data association between frames. The direct method was first used on RGBD cameras in DTAM [5]. It does not extract sparse feature points for each frame but directly aligns each pixel of the image and uses an inverse depth filter. LSD-SLAM [7] is the first large-scale monocular direct SLAM. Compared with DTAM, it only selects pixels with a high gradient to perform multi-scale pyramid alignment and tracking. The direct method is greatly affected by photometric since it estimates poses on the raw image. Jacob proposes photometric calibration [12] and a more robust and accurate direct sparse odometry [13], which significantly improved the practicability and accuracy of the direct method.

In recent years, the SLAM systems combining direct and indirect methods have become popular. The SVO [14] is a semi-direct approach. It extracts feature points at the frontend but does not calculate descriptors. It directly uses the optical flow to perform feature matching and pose estimation between frames. However, it still uses the traditional way of minimizing reprojection error for backend optimization. Although SVO takes less time than other methods, it is designed for small computing platforms and mostly is used for tracking with down-view cameras. Krombach [15,16] uses the pose obtained by the feature-based method as the initial value of the direct method. It can increase the stability of the direct visual odometry. However, the accuracy needs to be improved since it does not perform photometric calibration. Kim [17] proposes using partitioned photometric estimation,

modifies the photometric error function, and improves the partial illumination problem of the direct method. Younes [18] has presented a combination of the direct and indirect method as feature-based direct monocular odometry. They present a VO method that is based on DSO but uses feature-based tracking when optimization. The problem is that when the direct method is going to fail, it has begun to produce large calculation errors. Lee [19] proposes a loose couple approach of ORB-SLAM2 and DSO to improve localization accuracy. However, its frontend and backend are almost independent, which cannot share estimation information to improve the pose precision.

3. Online photometric calibration based on feature points

Bergmann [20] presents to extract the Shi–Tomasi corner points by using the gain-robust KLT algorithm and select good candidate points for tracking. However, in some datasets and real-world scenarios, there are larger exposure changes and non-contiguous frames. The gain-robust KLT algorithm is still strongly affected. Therefore, we propose to use ORB feature points to track. The ORB feature point is to extract the BRIEF descriptor for each FAST corner point, using KNN matching, cross filter, ratio test, and RANSAC verification to obtain feature matches in two images. Since we use the BRIEF descriptor, the matches obtained by this method are not affected by illumination. Perfect match pairs can be obtained even in the case of big illumination changes and non-contiguous frames.

After getting points P tracked between images, the energy function is given by

$$E = \sum_{p\in P}\sum_{i\in F_p} w_i^p \left\| \underbrace{O_i^p - f\left(e_i V\left(x_i^p\right) L^p\right)}_{r(f,V,e_i,L^p)} \right\|_h \tag{1}$$

where point $p \in P$ is visible in frame F_p, w_i^P is a weight factor for residual r, O_i^p is the pixel intensity of p in image i, e_i is the exposure time, L_p is the radiance of p and x_i^p is the spatial location of the projection of p onto image i. We use the Huber norm $\|\cdot\|_h$ for robust estimation, parametrized by $h \in \mathbb{R}$.

When getting a new frame, it is rectified by the existing response and vignette function. The exposure time of the new frame can be calculated by the weighted least squares.

$$E = \sum_{i=1}^{M}\sum_{p\in R_i} w_i^p \left(\frac{f^{-1}\left(O_i^p\right)}{V\left(x_i^p\right)} - e_i L^p\right)^2 \tag{2}$$

where P_i is the set of scene points visible in the i'th image and f^{-1} is the inverse of the response function. Each residual is only dependent on the exposure time of its frame and the radiance of its scene point. So that the exposure time can be calculated efficiently.

4. Feature-aided semi-direct method

Direct visual odometry does not contain feature extraction and matching, which is not robust to large-baseline motion. Therefore, we add the feature-based method into it and propose a hybrid pyramid.

4.1. Feature points selection

Feature points selection is the same as LDSO [21].

- We first extract corner points by using a dynamic grid gradient threshold. The maximum quantity of corner points is 2000, so that we can get more points with different gradients in the image.
- Further selection is applied to the obtained corner points using the Shi–Tomasi point selection method to select points with higher repeatability.
- Calculate the BRIEF descriptor on the points selected in the second step and convert them to bag-of-words vectors.
- Repeat the above steps in the grid with different sizes.

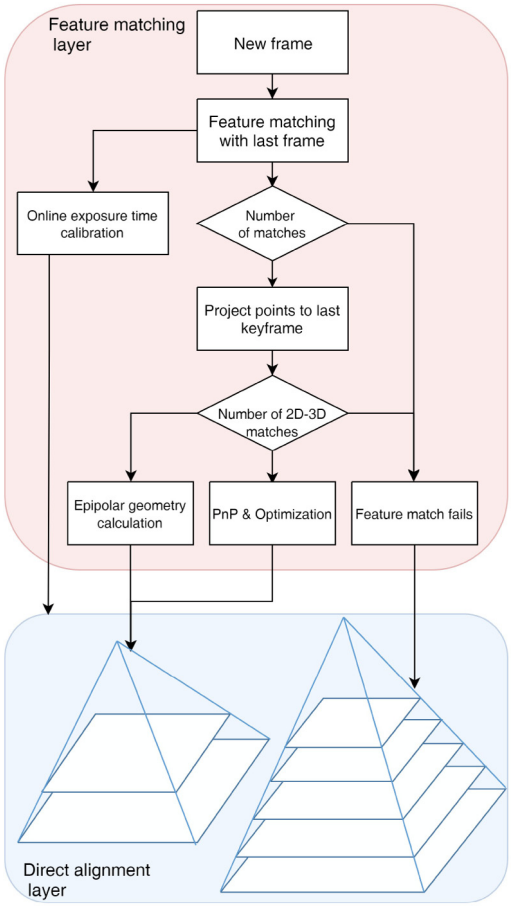


Fig. 2. Pipeline of hybrid pyramid.

4.2. Hybrid pyramid

The top layer of the image pyramid in DSO is used for rough calculation. Then the upper layer's pose estimation is passed to the next layer, thereby achieving coarse-to-fine calculation, getting more accurate pose estimation. We can find from the experiment that the top layer can only get the coarse pose estimation, and the last layer obtains the most precise result.

In DSO, it uses a constant motion model to get prior. It is possible to avoid being unable to iterate to the global optimal due to the massive displacement under some circumstances. However, in the case of large-baseline motion, such as high-speed vehicles, the error of pose estimation is still big. In this section, we propose a hybrid pyramid to improve its robustness (see Fig. 2).

4.2.1. Feature matching layer

When a new frame arrives, the feature points of the current frame are matched with the previous one. 2D–2D matches are obtained. If the quantity of matches is greater than N_1, the process goes to the next step. Otherwise, the feature matching fails due to the environment containing repetitive textures or too few features. Then the process goes to the direct alignment layer.

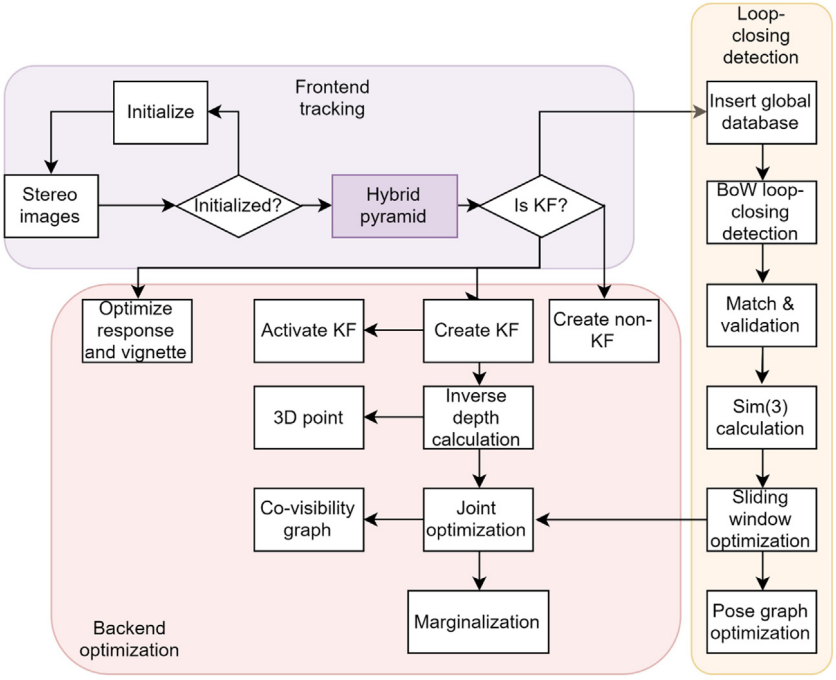


Fig. 3. Pipeline of the whole SLAM system.

If feature matching succeeds, the matched points of the previous frame are projected onto the last keyframe. We select the points whose depth is greater than 0 and within 100 times of the baseline.

Finally, we choose the calculation method based on the quantity of 2D–3D matches N:

- If $N < N_1$, the feature matching fails. Then it goes to direct alignment layer.
- If $N_1 < N < N_2$, we use the 2D–2D epipolar geometry to calculate pose estimation.
- If $N_2 < N$, we use the PnP algorithm to calculate pose estimation.

N_1, N_2 are fixed parameters.

With the feature matches, we can simultaneously calculate the exposure time for photometric rectification in the direct alignment layer.

4.2.2. Direct alignment layer

According to the results of the feature matching layer, we get the following situations:

- When the feature matching fails, we use a constant motion model to give the prior.
- When the 2D–2D pose estimation succeeds, the rotation of prior is obtained by the eight-point method, and the translation is obtained by the constant motion model.
- When the 2D–3D pose estimation succeeds, the prior is given by the feature matching and optimization directly.

The quantity of image pyramid layers is also different when performing the direct alignment. When the prior can be obtained by feature matching, we only use a two-layer pyramid. If the feature matching fails, we still use a five-layer pyramid, which ensures that the direct alignment can be optimized to the global optimal result under different conditions, and reduces computational cost for real-time performance.

4.2.3. Hybrid residual function

We use a hybrid residual function to improve the accuracy of the final optimization since we can get the reprojection error of the feature matching and the direct alignment photometric error. The total residual function is given by:

$$E(\xi) = w_{rep} \left\| E_{rep} \right\|^2 + w_{photo} \left\| E_{photo} \right\|^2 \tag{3}$$

where w_{rep} and w_{photo} are weights of the two errors. We can rewrite the cost function as:

$$\begin{aligned} E(\xi) &= \begin{bmatrix} e_{rep} \\ e_{photo} \end{bmatrix}^T \begin{bmatrix} w_{rep} & \\ & w_{photo} \end{bmatrix} \begin{bmatrix} e_{rep} \\ e_{photo} \end{bmatrix} \\ &= \boldsymbol{e}^T \mathbf{W} \boldsymbol{e} \end{aligned} \tag{4}$$

where $\mathbf{e}$ is joint residual, $\mathbf{W}$ is the information matrix and ξ is the camera's pose.

5. Stereo feature-aided semi-direct SLAM system

For further evaluation of our approach, we implement a full stereo visual SLAM system. As is shown in Fig. 3, the frontend tracking module is based on our hybrid pyramid. The whole system resembles ORB-SLAM2, which consists of three parallel threads: frontend tracking thread, backend sliding window optimization thread and loop-closure detection thread.

- Frontend tracking thread. It includes monocular image distortion rectification, stereo images epipolar rectification, and feature extraction. Then it uses the hybrid pyramid to track and finally determine whether it is a keyframe.
- Backend sliding window optimization thread. It first performs a stereo scale estimation, recovers the inverse depth of 3D points in the current frame, and then minimizes the hybrid residual to optimize the points and poses in the sliding window. Finally, it judges whether the frames and points in the sliding window need to be marginalized and perform marginalization.
- Loop-closure detection and optimization thread. It first uses the bag-of-words model to perform loop-closure detection.

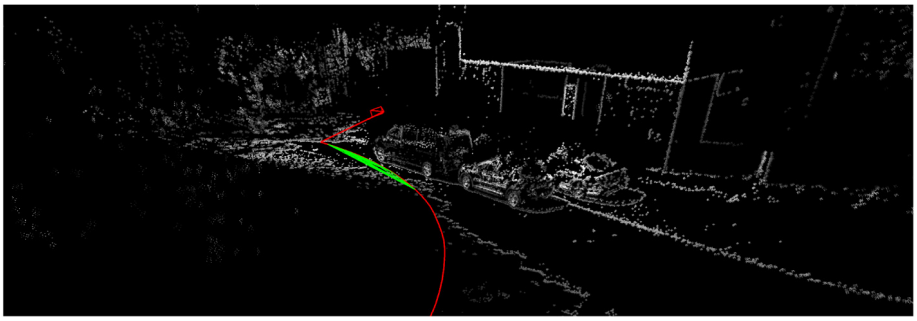


Fig. 4. Here is an example from [10]. Despite the photometric rectification the direct method still fails due to the extreme brightness change.

Fig. 5. Some images in Cityscapes Dataset [22] have severe brightness change due to the illumination and camera auto exposure and gain control.

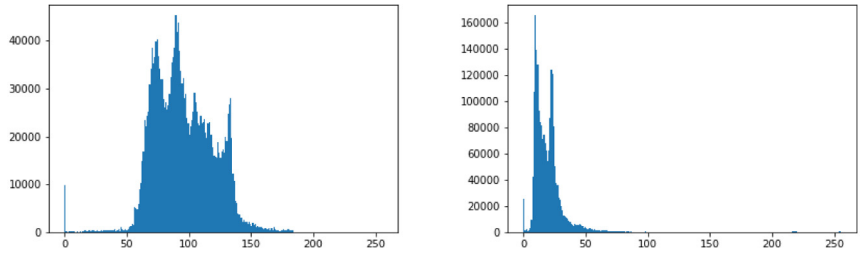


Fig. 6. The dark channel histogram of two images in the second row of Fig. 5.

When detecting a closed loop, it projects the verified loop-closure frame to the current sliding window and calculates the Sim(3) transformation. Finally, it performs optimization through the pose graph to correct the pose error and global map points.

6. Image brightness rectification using dark channel prior

The online photometric calibration that we use here models exposure time, response function, and vignette, which are parameters of the camera. When the light source change that is causing halation or local and global brightness change, modeling camera parameters only does not help as well, leading to direct alignment failure in [10]. As is shown in Fig. 5, it is common in autonomous vehicles. When a car is going through the shadow in an avenue, the brightness changes frequently, and the halation will severely damage the image.

Building on our previous work [23], we consider the pixel value as a combination of original scene radiance and atmospheric light. More specifically, inspired by [24], we model the image as follows:

$$\mathbf{I}(x) = \mathbf{J}(x)t(x) + \mathbf{A}(1 - t(x)) \tag{5}$$

where $\mathbf{I}$ is the acquired pixel intensity, $\mathbf{J}$ is the scene radiance, $\mathbf{A}$ is the global atmospheric light and t is the transmissivity. To remove halation and get $\mathbf{J}$, we need to recover $\mathbf{A}$ and t from $\mathbf{I}$. We transform Eq. (5) to:

$$\frac{I^c(x)}{A^c} = t(x)\frac{J^c(x)}{A^c} + 1 - t(x) \tag{6}$$

where the superscript c represents each channel of the color image.

First, we assume that the transmissivity $t(x)$ in a window is constant and define it as $\tilde{t}(x)$. With a given $\mathbf{A}$, we calculate the minimum on both sides of Eq. (6):

$$\min_{y\in\Omega(x)}(\min_c \frac{I^c(y)}{A^c}) = \tilde{t}(x)\min_{y\in\Omega(x)}(\min_c \frac{J^c(y)}{A^c}) + 1 - \tilde{t}(x) \tag{7}$$

and then:

$$\tilde{t}(x)(1 - \min_{y\in\Omega(x)}(\min_{c}\frac{J^c(y)}{A^c})) = 1 - \min_{y\in\Omega(x)}(\min_{c}\frac{I^c(y)}{A^c}) \tag{8}$$

According to [24], the dark channel prior is defined as:

$$J^{dark}(x) = \min_{y\in\Omega(x)}(\min_{c\in r,g,b} J^c(y)) = 0 \tag{9}$$

where J^c represents each channel of the color image and $\Omega(x)$ is a window with center x. We calculated the histogram of the dark channel image with and without halation in Fig. 5, which shows that it approximately meets the prior.

Then we substitute Eqs. (9)–(8) and get the estimation of transmissivity $\tilde{t}(x)$:

$$\tilde{t}(x) = 1 - \min_{y\in\Omega(x)}(\min_{c}\frac{I^c(y)}{A^c}) \tag{10}$$

As is shown in Fig. 6, not all of the dark channel in no halation image is zero. So we introduce a constant parameter ω into Eq. (10) to compensate:

$$\tilde{t}(x) = \omega(1 - \min_{y\in\Omega(x)}(\min_{c}\frac{I^c(y)}{A^c})) \tag{11}$$

The above inferences are with the assumption that **A** is known. In the haze removal methods, the atmospheric light **A** is estimated from the most non-transparent pixels. Tan [25] used the most intense pixel as the atmospheric light. But in practice, the brightest pixel does not always represent light. Sometimes it could be on a white building. As is shown in Fig. 6, the dark channel of the image can approximate the atmospheric light as well. It is more robust than just taking the brightest pixel. We can use the dark channel to estimate the atmospheric light.

- Pick the top 1% brightest pixels in the dark channel.
- Take the average intensity of those pixels in raw image as the atmospheric light.

Then we can give the recovery equation:

$$\mathbf{J}(x) = \frac{\mathbf{I}(x) - \mathbf{A}}{max((t(x), t_0))} + \mathbf{A} \tag{12}$$

where t_0 is a threshold that prevent the image to become too white when transmissivity $t(x)$ is extremely small (see Fig. 4).

7. Evaluation

In this section, we evaluate our algorithms on KITTI Dataset [26], self-collected Mobile Robot Dataset [27] and Cityscapes Dataset [22].

The KITTI Dataset [26] is a popular public odometry dataset. It contains data from various scenarios: rural, urban, and highway. The data collection sensors include 64-ring LiDAR, GPS/INS, two color, and two grayscale cameras.

The Mobile Robot Dataset [27] is collected on a skid-steering robot. Sensors are shown in Fig. 7, including a 16-ring LiDAR, an IMU, and two grayscale cameras. The LiDAR and cameras are sampled at 10 Hz, and the IMU is sampled at 200 Hz. All the sensors are well-synchronized with uniform timestamps. We do not have the most commonly used RTK-GPS as ground truth since there are many tall buildings and trees in our campus so that the GPS is not reliable. Instead, we use LiDAR to generate ground truth pose with a pre-built high-precision point cloud map. We collect five sequences of data in our campus. Table 1 illustrates the specifics of the dataset. The sequence 1, 2, 3, 5 runs under the same scene A and the sequence 4 runs under scene B. Datasets are collected at different times under different weather conditions, such as daytime, cloudy days, and even rainy days. As we can see from the table, great changes have taken place in different sequences, such as illumination, dynamic objects, and weather conditions, which is enough to verify the robustness of our approach.

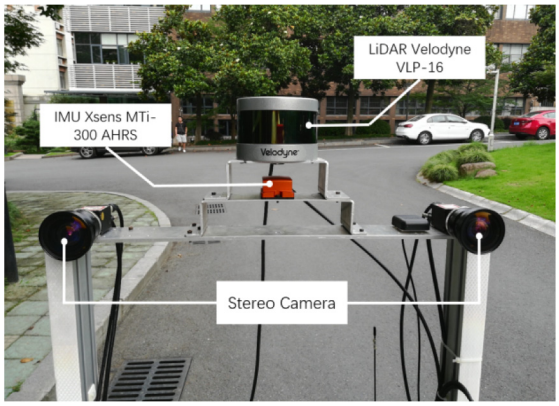


Fig. 7. Data collection platform of mobile robot dataset.

The Cityscapes Dataset [22] focuses on semantic understanding of urban street scenes. The images are collected by a stereo camera with a 22 cm baseline. The whole sequence is divided into many small parts since it is not an odometry dataset. Besides, it only provides low-frequency consumer-level GPS data, which makes it hard to evaluate trajectory accuracy. But in the validation sequence frankfurt, that are some typical scenes where the extreme brightness changes cause the direct method to fail. So we only use it to test image brightness rectification.

7.1. Online photometric calibration

We first test our photometric calibration on Monocular Visual Odometry Dataset [12] since it provides exposure time of each frame. Fig. 8 shows the estimation result and ground truth in Sequence 47. The red line is the exposure time estimation of our algorithm, the blue line is the exposure time of the KLT tracker based photometric calibration, and the black line is ground truth. It can be seen from the figure that the estimated exposure time of our algorithm is closer to the ground truth. The KLT tracker based calibration algorithm shows a large error in some cases with strong exposure, indicating that our approach has higher precision and robustness.

To evaluate its performance in autonomous driving, we test on KITTI [26] dataset. But this dataset does not provide exposure time. We only give the rectified images in Fig. 9. Some parts are too intense, and the pixels have been overexposed in the raw images, especially at the top of the house in the distance. It is difficult to see the edge. When using the photometric calibration algorithm based on the KLT tracker, the photometric correction is wrong, and the entire image becomes overexposed. Our approach of using the ORB tracker can reduce the image exposure so that the overexposure is well corrected, and the texture of the whole image is more precise.

7.2. Hybrid pyramid tracking

We use the KITTI [26] dataset to test the pose estimation between frames in the frontend. Most of the data is captured on a moving car in high-speed environments. The motion baseline between two frames is large, which can verify the improvement of accuracy by using the hybrid pyramid algorithm proposed in this paper. We perform frontend tracking only to show the advantages of our approach. The specific steps are as follows:

Table 1
Mobile robot dataset.

Num	Date	Time	Scene	Illumination	Weather	Dynamic scene	Direction	length (m)
1	6.1	9:18	A	Daytime	Sunny	A little	Counterclockwise	341.990
2	6.1	9:32	A	Daytime	Sunny	A little	Clockwise	325.425
3	6.5	18:16	A	Dusk	After raining	Abundant	Counterclockwise	361.765
4	6.5	18:38	B	Dusk	After raining	Abundant	Clockwise	777.568
5	6.12	17:48	A	Dusk	Cloudy	Several	Counterclockwise	406.877

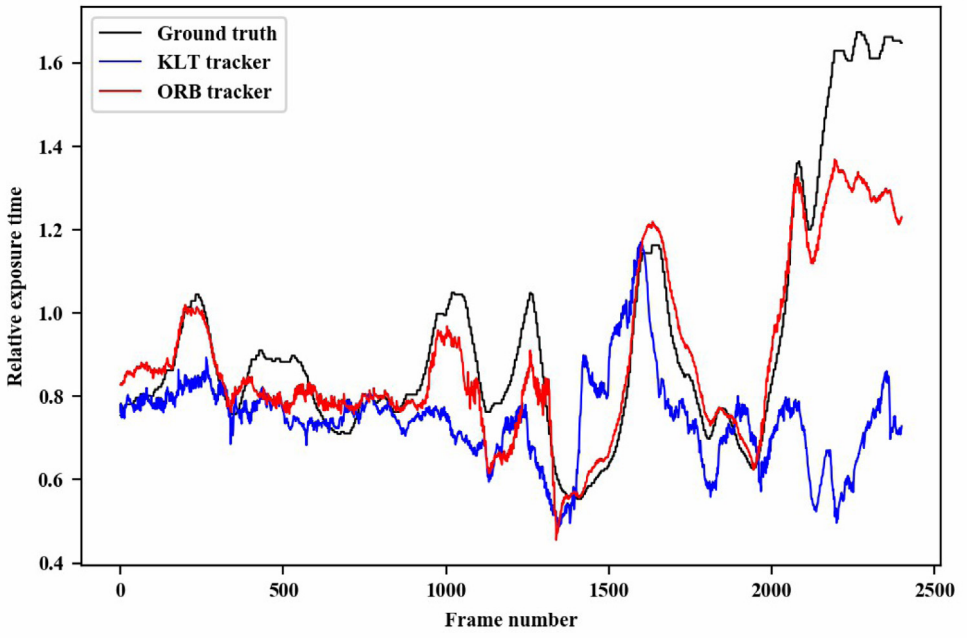


Fig. 8. Exposure time in Sequence 47.

(a) Raw image (b) KLT tracker (c) ORB tracker

Fig. 9. Photometric rectification on KITTI Sequence 00.

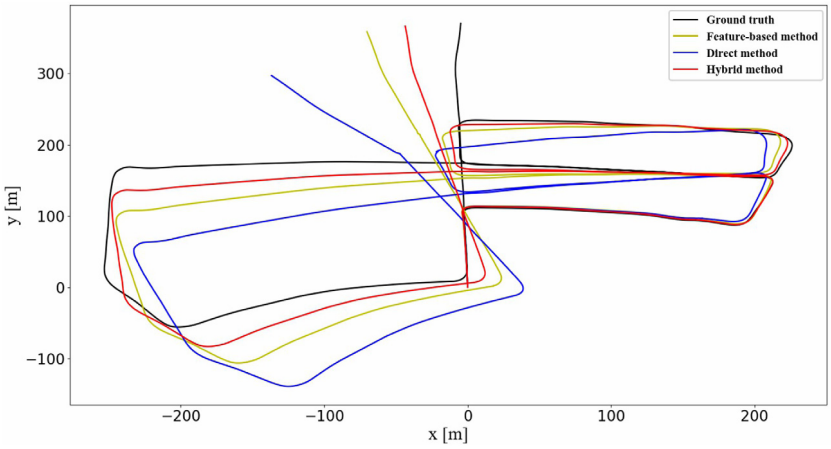


Fig. 10. Trajectory comparison in Sequence 05.

Table 2
Average running time.

Method	Time
Direct method	30.3 ms
Feature-based method	24.2 ms
Hybrid method	35.7 ms

- Extract the FAST corner points of current image.
- Use the SGBM algorithm to estimate the depth map of current frame, get the depth of each extracted corner point.
- Use 2D points of current frame and 3D points of previous frame to perform pose estimation by different methods.
- Estimate poses between frames and draw trajectories for comparison.

In this experiment, the direct method uses a 5-layer pyramid. The parameters of the feature-based method are the same as ORB-SLAM2. In the hybrid pyramid, we extract 300 ORB feature points in the first layer to match and use two layers to perform the direct alignment. Other parameters are the same as the direct method.

In Fig. 10, the yellow line is the trajectory of the feature-based method, the blue line is the direct method, and the red line is our hybrid method. The black line is ground truth in Sequence 05. We can find that the direct method is less accurate because of the great movement. After adding feature matching in the upper layer to obtain the initial value, the accuracy of direct alignment can exceed the feature matching method. The experiment shows that under the same conditions without other optimizations, the pose calculated by the hybrid pyramid proposed in this paper is more accurate. It improves the robustness of the direct method to large baseline motion.

The processor of the computing platform in the experiment is Intel i7-4790@3.60 GHz. Table 2 shows the time consumption of the three methods. The hybrid pyramid method proposed in this paper takes more time than others. But it still meets the requirements of real-time.

7.3. Stereo feature-aided semi-direct SLAM system

We test the trajectory accuracy on public KITTI Dataset [26] and Mobile Robot Dataset to verify the accuracy and robustness of the whole system. We compare the results with the state-of-the-art ORB-SLAM2 system.

First, we test the visual odometry and disable the loop-closure detection thread. We run the feature-based method (ORB-SLAM2), direct method, and our method on 11 sequences. Each algorithm is run six times and averaged. Table 3 shows the average ATE and RMSE of the feature-based method, direct method, and hybrid method. We can see that the accuracy of

Table 3
Visual odometry trajectory error without loop-closure.

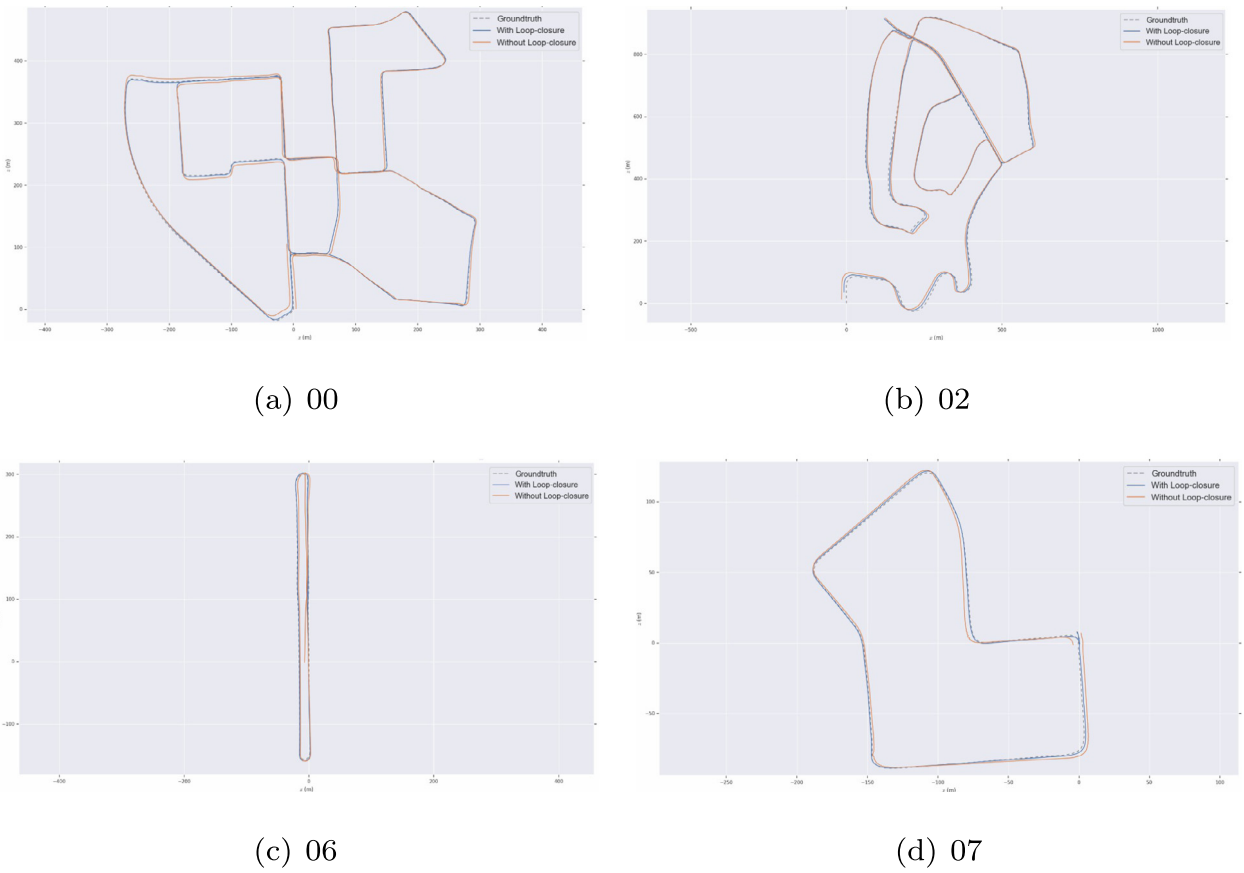
	Hybrid method		Direct method		ORB-SLAM2	
	Mean (m)	RMSE (m)	Mean (m)	RMSE (m)	Mean (m)	RMSE (m)
Seq00	**3.9753**	**4.6433**	12.0296	13.3774	4.2218	4.7353
Seq01	**8.4216**	**10.9853**	46.5368	57.9955	10.9960	11.6609
Seq02	**7.2184**	**8.6793**	10.8594	12.7483	8.8580	10.2612
Seq03	0.9342	1.0717	8.3263	9.3555	**0.6867**	**0.8008**
Seq04	0.4295	0.4659	2.9640	3.3881	**0.3089**	**0.3469**
Seq05	2.1815	2.4141	7.4165	8.0278	**1.8962**	**2.0399**
Seq06	2.7914	3.0485	7.4993	8.3491	**2.1110**	**2.3154**
Seq07	**1.0849**	**1.4193**	X	X	1.3127	1.4960
Seq08	**2.7613**	**3.3839**	23.6450	25.6290	3.4244	3.8558
Seq09	**2.8014**	**3.4086**	22.0984	24.2153	2.8966	3.5395
Seq10	**0.6620**	**0.7185**	2.2024	2.6776	1.1124	1.2423

Table 4
ATE of our approach on KITTI with and without Loop-closure.

	With loop-closure		Without loop-closure	
	Mean (m)	RMSE (m)	Mean (m)	RMSE (m)
Seq00	**1.0297**	**1.1185**	3.9753	4.6433
Seq02	**3.6286**	**4.1144**	7.2184	8.6793
Seq05	**0.8628**	**0.9859**	2.1815	2.4141
Seq06	**1.2887**	**1.3956**	2.7914	3.0485
Seq07	**0.8231**	**0.8979**	1.0849	1.4193

Table 5
ATE on KITTI with loop-closure.

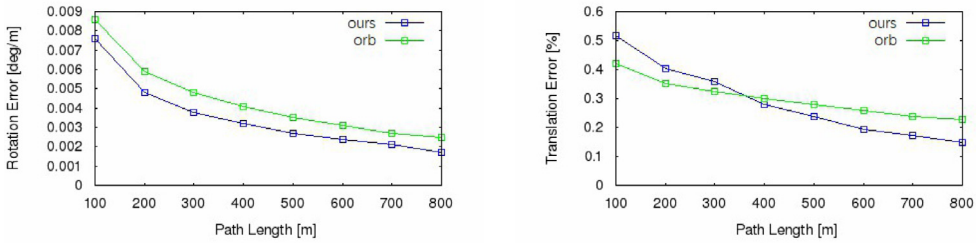
	Hybrid method		ORB-SLAM2	
	Mean (m)	RMSE (m)	Mean (m)	RMSE (m)
Seq00	**1.0297**	**1.1185**	1.1358	1.2681
Seq01	**8.4216**	**10.9853**	10.9960	11.6609
Seq02	**3.6286**	**4.1144**	5.8315	6.9868
Seq03	0.9342	1.0717	**0.6867**	**0.8008**
Seq04	0.4295	0.4659	**0.3089**	**0.3469**
Seq05	0.8628	0.9859	**0.7318**	**0.8135**
Seq06	1.2887	1.3956	**0.8863**	**0.9068**
Seq07	**0.8231**	**0.8979**	1.3127	1.4959
Seq08	**2.7613**	**3.3839**	3.4244	3.8558
Seq09	**2.8014**	**3.4086**	2.8966	3.5395
Seq10	**0.6620**	**0.7185**	1.1124	1.2423



(a) 00 (b) 02

(c) 06 (d) 07

Fig. 11. KITTI loop-closure test.



(a) Rotation Error (b) Translation Error

Fig. 12. Average RPE on KITTI Dataset.

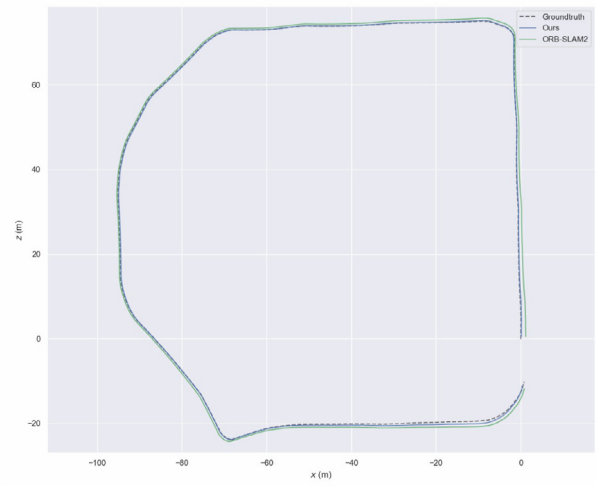
Table 6
ATE on mobile robot dataset.

	Hybrid method		ORB-SLAM2	
	Mean (m)	RMSE (m)	Mean (m)	RMSE (m)
Seq1	**0.3586**	**0.4370**	0.7555	0.8619
Seq2	**0.4206**	**0.4597**	1.1723	1.2915
Seq3	**0.6231**	**0.7524**	1.2454	1.3366
Seq4	**0.9056**	**1.0150**	2.2877	2.4919
Seq5	**0.3700**	**0.4876**	0.6108	0.6943

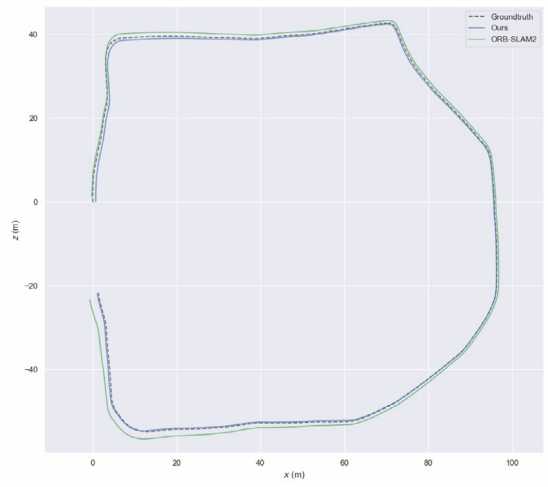
the direct method is low, most of the root mean square errors are above 10 m, and sequence 07 even fails, mainly due to the large-baseline motion and low frame rate of the KITTI dataset. The direct method does not have good adaptability, and the accuracy of the hybrid method is much higher than the direct one. Compared with the feature-based method, our approach is comparable to the most advanced ORB-SLAM2, and the accuracy on the sequence 00,01,02,07,08,09,10 even exceeds it.

Then, we evaluate the loop-closure module using the sequence 00,02,05,06, and 07. All five sequences have closed-loop paths. We compare the trajectory accuracy of the full SLAM system with and without loop-closure to verify the impact of the loop-closure detection and optimization module on the localization accuracy. In this experiment, the frontend tracking thread and the backend sliding window optimization thread use the same parameters and settings. Table 4 lists the ATE after optimization with and without loop-closure detection. Fig. 11 shows the trajectory estimation and the ground truth. We can find that after the loop-closure detection and optimization, the root mean square error is much smaller, which shows that the module can continuously correct error for a long time, thereby obtaining higher localization accuracy.

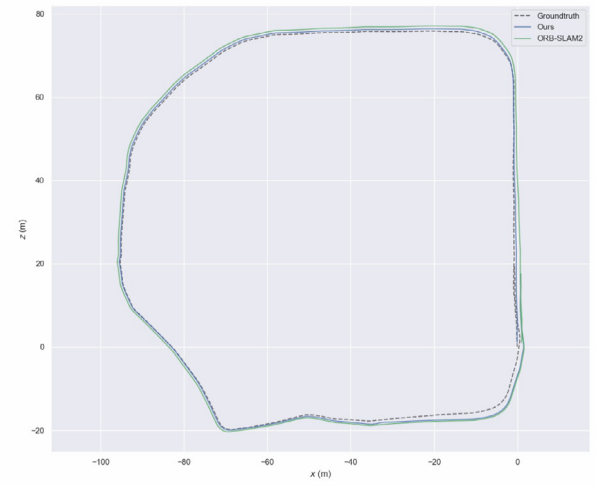
Finally, we test the full SLAM system compared with ORB-SLAM2 on KITTI Dataset [26] and Mobile Robot Dataset. In Fig. 12,



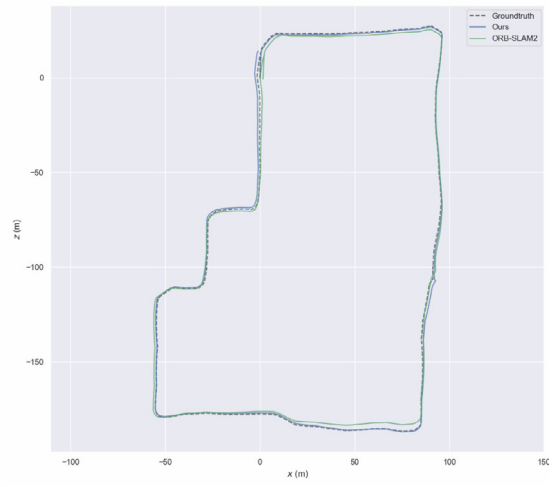
(a) Seq1



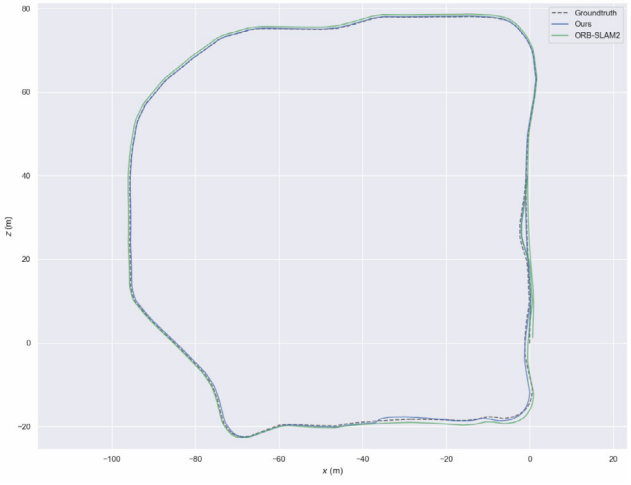
(b) Seq2



(c) Seq3



(d) Seq4



(e) Seq5

Fig. 13. Trajectories on mobile robot dataset.

Fig. 14. Direct method fails due to light interference.

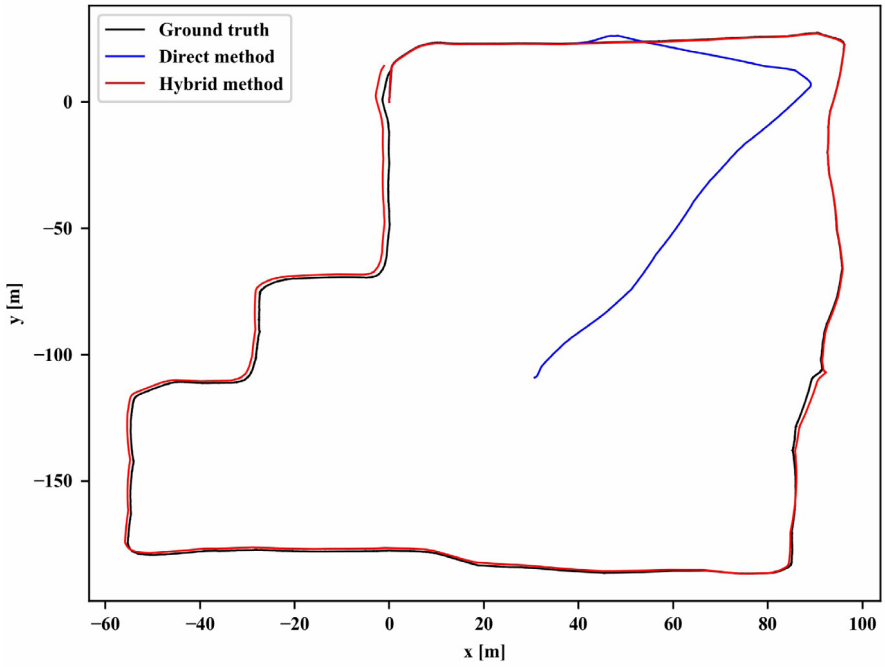


Fig. 15. The result of direct tracking failure.

we draw the average RPE of the 11 sequences, that is, the curves of translation error and rotation error relative to the path, respectively, where the blue curve is our approach and green curve is ORB-SLAM2. As is shown in the figure, our approach is better than ORB-SLAM2 in terms of rotation. The error in short distance ORB-SLAM2 is smaller, and in long-distance, our approach is better. The reason is that the high-speed scene leads to a reduction in the translation accuracy of our approach. Because the pose estimation accuracy still partially depends on the photometric error. Although we use feature matching to provide the prior, it is still slightly worse than the feature-based method under the large-baseline motion. It can also be seen in Table 5 that the accuracy of the sequence 03,04,05,06 is marginally lower than that of the ORB-SLAM2. The main reason is that there are more short-distance high-speed movements in these four sequences, resulting in a reduction in translation accuracy, which shows that our approach needs to be improved for high-speed scenarios.

Fig. 13 and Table 6 gives the trajectories and error on Mobile Robot Dataset compared with ORB-SLAM2. In this dataset, the robot moves slowly and does not have a large baseline movement, which is conducive to obtaining more accurate calculation results in direct alignment. At the same time, the outdoor scene is rich in texture and can continuously track stable feature points. Even in the case of illumination changes, the pose prior can be obtained through the feature matching to ensure that the direct alignment layer is not affected. Therefore, our approach outperforms ORB-SLAM2 in all sequences.

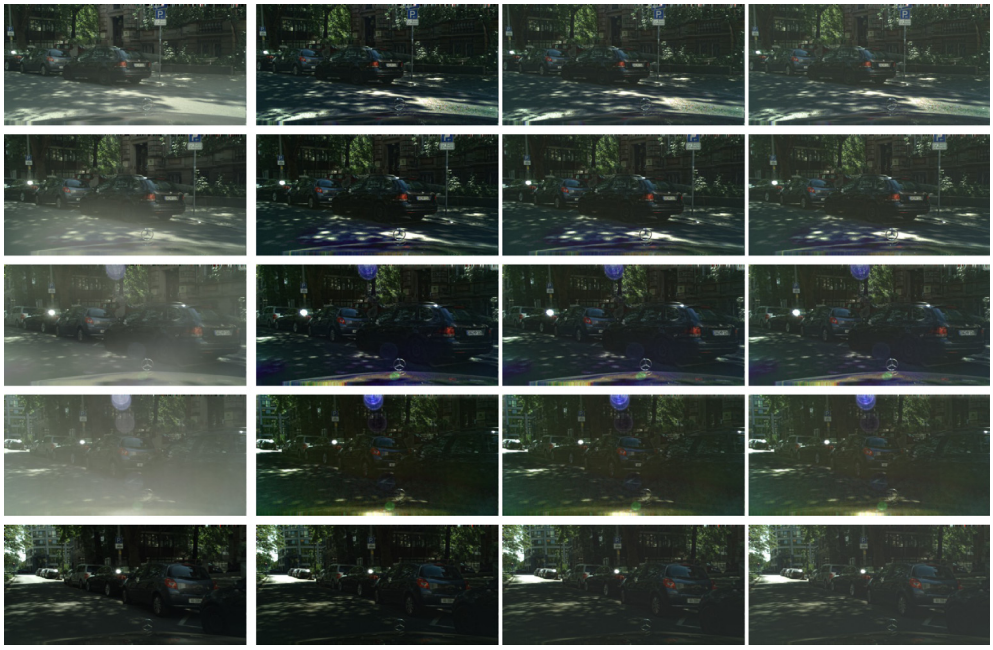
As is shown in Fig. 14, the data collected by the mobile robot has a large light interference. There are many cases of excessive darkness or overexposure. Some of the pedestrians and cars are overexposed because of the sunlight. In Fig. 15, the red line is the trajectory of the hybrid algorithm proposed in this paper, and the blue line is the trajectory of the original direct method. Our system can still work under light interference, while the direct method gets the wrong results. It shows that the hybrid pyramid algorithm proposed in this paper has higher illumination robustness than the original direct method.

7.4. Image brightness rectification

We test on validation sequence frankfurt of Cityscapes Dataset [22], which contains extreme brightness changes that cause direct method failure. In Fig. 16, we sample five images uniformly in the sequence. The first and the last images do not have halation. But there are some high-brightness areas in the dark channel due to the strong reflection of the ground. The other three images have halation, and its intensity increases gradually. The details of the image are less affected so that the feature-based method works well. But the whole image becomes brighter, which brings

(a) Raw image (b) Dark channel (c) Transmissivity (d) Rectification

Fig. 16. The raw image, dark channel, transmissivity and rectification on Cityscapes Dataset.



(a) Raw image (b) ω=1.05 (c) ω=1.15 (d) ω=1.25

Fig. 17. Comparison with different compensation parameter ω.

a drastic jump in the photometric error of the direct method. From the perspective of the dark channel and transmissivity, the dark channel is bright, and transmissivity is small where there is halation. That is, the proportion of scene radiance in the image pixel is less than the atmospheric light, according to Eq. (5). After rectification, the brightness of the non-halation image slightly decreased, but the consistency of the image sequence has significantly improved. There is no longer jump in the photometric error, which guarantees the direct method works well.

In Fig. 17, we test with different compensation parameter w. The images in the fourth column are a little bit brighter than that in the second column, which means that the compensation parameter w can control the rectification level. If the images are too dark, we can increase w.

For further evaluation, we test our SLAM system on the rectified sequence. Fig. 18 shows the trajectory and the point cloud of the part where illumination changes extremely. It proves that our brightness rectification is helpful for the SLAM system.

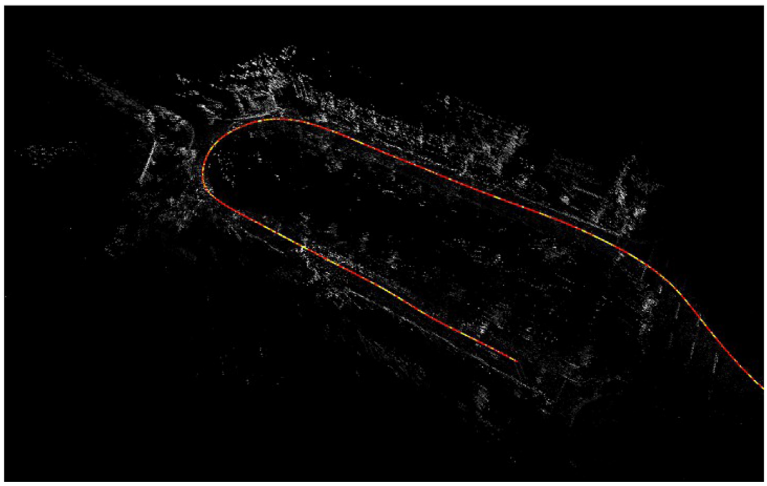


Fig. 18. Part of trajectory and point cloud on Cityscapes Dataset [22], sequence frankfurt.

8. Conclusion

In this paper, we propose a robust stereo feature-aided semi-direct SLAM system for robust pose estimation under challenging environments. We also perform online photometric calibration to obtain better photometric parameters and apply it into visual odometry. Furthermore, we extend the stereo visual SLAM system based on our hybrid pyramid frontend for evaluation. For extreme brightness change, we employ the dark channel prior to rectify and maintain the consistency of image pixels. The photometric rectification test, hybrid pyramid tracking test, image brightness rectification test, and SLAM test on KITTI [26] and mobile robot dataset indicate that our method is accurate and robust in challenging real-world scenarios.

In future work, we plan to integrate the inertial measurement unit to deal with stronger inter-frame motion and combine the image brightness rectification and the photometric calibration to eliminate regional illumination changes.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant 61836015 and the National Key Research and Development Program of China under Grant 2018AAA0101503.

References

[1] A.J. Davison, I.D. Reid, N.D. Molton, O. Stasse, MonoSLAM: Real-time single camera SLAM, IEEE Trans. Pattern Anal. Mach. Intell. 29 (6) (2007) 1052–1067.

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

[3] R. Mur-Artal, J.M.M. Montiel, J.D. Tardós, ORB-SLAM: A versatile and accurate monocular SLAM system, IEEE Trans. Robot. 31 (5) (2015) 1147–1163.

[4] R. Mur-Artal, J.D. Tardós, ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras, IEEE Trans. Robot. 33 (5) (2017) 1255–1262.

[5] R.A. Newcombe, S.J. Lovegrove, A.J. Davison, DTAM: Dense tracking and mapping in real-time, in: 2011 International Conference on Computer Vision, 2011, pp. 2320–2327.

[6] J. Engel, T. Schöps, D. Cremers, LSD-SLAM: Large-scale direct monocular SLAM, in: D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), Computer Vision – ECCV 2014, Springer International Publishing, Cham, 2014, pp. 834–849.

[7] J. Engel, J. Stückler, D. Cremers, Large-scale direct SLAM with stereo cameras, in: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2015, pp. 1935–1942.

[8] D. Caruso, J. Engel, D. Cremers, Large-scale direct SLAM for omnidirectional cameras, in: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2015, pp. 141–148.

[9] V. Usenko, J. Engel, J. Stückler, D. Cremers, Direct visual-inertial odometry with stereo cameras, in: 2016 IEEE International Conference on Robotics and Automation, ICRA, 2016, pp. 1885–1892.

[10] R. Wang, M. Schwörer, D. Cremers, Stereo DSO: Large-scale direct sparse visual odometry with stereo cameras, in: 2017 IEEE International Conference on Computer Vision, ICCV, 2017, pp. 3923–3931.

[11] D. Galvez-López, J.D. Tardos, Bags of binary words for fast place recognition in image sequences, IEEE Trans. Robot. 28 (5) (2012) 1188–1197.

[12] J. Engel, V.C. Usenko, D. Cremers, A photometrically calibrated benchmark for monocular visual odometry, 2016, CoRR, abs/1607.02555, http://arxiv.org/abs/1607.02555.

[13] J. Engel, V. Koltun, D. Cremers, Direct sparse odometry, IEEE Trans. Pattern Anal. Mach. Intell. 40 (3) (2018) 611–625.

[14] J. Engel, J. Sturm, D. Cremers, Semi-dense visual odometry for a monocular camera, in: 2013 IEEE International Conference on Computer Vision, 2013, pp. 1449–1456.

[15] N. Krombach, D. Droeschel, S. Behnke, Combining feature-based and direct methods for semi-dense real-time stereo visual odometry, in: W. Chen, K. Hosoda, E. Menegatti, M. Shimizu, H. Wang (Eds.), Intelligent Autonomous Systems, Vol. 14, Springer International Publishing, Cham, 2017, pp. 855–868.

[16] N. Krombach, D. Droeschel, S. Houben, S. Behnke, Feature-based visual odometry prior for real-time semi-dense stereo SLAM, Robot. Auton. Syst. 109 (2018) 38–58.

[17] P. Kim, H. Lee, H.J. Kim, Autonomous flight with robust visual odometry under dynamic lighting conditions, Auton. Robots 43 (6) (2019) 1605–1622.

[18] G. Younes, D.C. Asmar, J.S. Zelek, FDMO: Feature assisted direct monocular odometry, in: A. Trémeau, G.M. Farinella, J. Braz (Eds.), Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISIGRAPP 2019, Vol. 5, VISAPP, Prague, Czech Republic, February 25–27, 2019, SciTePress, 2019, pp. 737–747.

[19] S.H. Lee, J. Civera, Loosely-coupled semi-direct monocular SLAM, IEEE Robot. Autom. Lett. 4 (2) (2019) 399–406.

[20] P. Bergmann, R. Wang, D. Cremers, Online photometric calibration of auto exposure video for realtime visual odometry and SLAM, IEEE Robot. Autom. Lett. 3 (2) (2018) 627–634.

[21] X. Gao, R. Wang, N. Demmel, D. Cremers, LDSO: Direct sparse odometry with loop closure, in: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2018, pp. 2198–2204.

[22] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset for semantic urban scene understanding, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 3213–3223.

[23] X. Zhao, R. Zheng, W. Ye, Y. Liu, A robust stereo semi-direct SLAM system based on hybrid pyramid, in: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2019, pp. 5376–5382.
[24] K. He, J. Sun, X. Tang, Single image haze removal using dark channel prior, IEEE Trans. Pattern Anal. Mach. Intell. 33 (12) (2011) 2341–2353.
[25] R.T. Tan, Visibility in bad weather from a single image, in: 2008 IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.
[26] A. Geiger, P. Lenz, R. Urtasun, Are we ready for autonomous driving? The KITTI vision benchmark suite, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 3354–3361.
[27] X. Zuo, W. Ye, Y. Yang, R. Zheng, T. Vidal-Calleja, G. Huang, Y. Liu, Multimodal localization: Stereo over LiDAR map, J. Field Robotics (2020).

Xiangrui Zhao received his B.S. degree in automation from Huazhong University of Science and Technology in 2018. He is currently a Ph.D. candidate of the Institute of Cyber Systems and Control, Department of Control Science and Engineering, Zhejiang University. His latest research interests include robotics vision and SLAM systems.

Lina Liu received her B.S. degree in automation from Zhejiang University in 2018. She is currently working toward the M.S. degree in control science and engineering at Zhejiang University, Hangzhou, China. Her research interests include computer vision and deep learning.

Renjie Zheng received his B.S. degree in communications engineering from Zhejiang University of Technology in 2016 and M.S. degree in control engineering from Zhejiang University in 2019. He is currently with Alibaba Group, Hangzhou, China. His latest research interests include reinforce learning and SLAM systems.

Wenlong Ye received his B.S. degree in automation in 2017 and M.S. degree in control engineering from Zhejiang University in 2020. He is currently with Alibaba Group, Hangzhou, China. His latest research interests include visual inertial odometry and multiple sensor fusion.

Yong Liu received the B.S. degree in computer science and engineering and the Ph.D degree in computer science from Zhejiang University, Zhejiang, China, in 2001 and 2007, respectively. He is currently a professor of Institute of Cyber-Systems and Control at Zhejiang University. His main research interests include: intelligent robot systems, robot perception and vision, deep learning, big data analysis, and multi-sensor fusion. He has published over 30 research papers on machine learning, computer vision, information fusion, and robotics.