Challenges in Monocular Visual Odometry: Photometric Calibration, Motion Bias and Rolling Shutter Effect

Nan Yang*, Rui Wang*, Xiang Gao and Daniel Cremers

Abstract—Monocular visual odometry (VO) has seen tremendous improvements in accuracy, robustness and efficiency, and has gained exponential popularity over recent years. Nevertheless, no comprehensive evaluations have been performed to reveal the influences of the three easily overlooked, yet very influential aspects: photometric calibration, motion bias and rolling shutter effect. In this work, we evaluate these three aspects quantitatively on the state of the art of direct, feature-based and semi-direct methods, providing the community with useful practical knowledge both for better applying existing methods and developing new algorithms of VO and SLAM. Conclusions (some of which are counterintuitive) are drawn based on insightful technical and empirical analyses to all of our experiments. Possible improvements on existing methods are directed or proposed, such as a sub-pixel accuracy refinement of ORB-SLAM which boosts its performance.

I. INTRODUCTION

Modern visual simultaneous localization and mapping (SLAM) systems usually have two basic components: VO and global map optimization. While the VO component incrementally estimates camera poses and builds up a local map, small errors are accumulated and over time the estimated camera poses start to drift away from their actual positions. If a previously visited location is detected, the drift can be eliminated by global map optimization using techniques like loop closure and pose graph optimization. Although a globally consistent map can be achieved, the accumulated error cannot be totally removed from the system. In other words, the overall performance of any SLAM system is fundamentally determined by its VO accuracy. During the past few years, the VO community has seen significant progress in improving algorithm accuracy, robustness and efficiency [1]–[10]. Efforts have been made with different kinds of VO formulations, i.e., direct vs. feature-based methods, dense/semi-dense alternating optimization vs. sparse joint optimization. However, apart from these high-level diversities, it is still not clear how the performance can be influenced by some low-level aspects:

a) Photometric calibration. Pixels corresponding to the same 3D point may have different intensities across images due to camera optical vignetting, auto gain and exposure controls. Several photometric calibration techniques are proposed to approximately recover the irradiance images. While it has been proven that photometric calibration can significantly improve the performance of direct methods [9], [11], it is still unclear how it can influence other formulations such as feature-based methods or semi-direct methods [8], [10].

b) Motion bias. It has been shown in [11] that running the same sequence forward and backward can result in different VO performance. This inspires us to explore what motion pattern VO systems prefer and what causes such bias on the level of algorithm design.

c) Rolling shutter effect. Direct methods have been considered less robust against the rolling shutter effect compared to other formulations. However, when the rolling shutter effect is combined with other challenging settings, e.g., low texture scenes, it is not clear which formulation can perform better. It is also interesting to check the performance of different formulations on modern industrial level cameras, which normally have rolling shutters but with extremely fast readout speed.

These three aspects can greatly affect the performance of VO systems, yet their influences have not been systematically discussed and evaluated. For certain formulations, these aspects may even have counterintuitive effects. In this work, we perform systematic and quantitative evaluations on the three most popular formulations of VO, namely direct, feature-based and semi-direct methods. Since evaluating all existing methods is not realistic, we select the state of the art of each family, i.e., DSO [9], ORB-SLAM [6] and SVO2. Our goal is to deliver practical insights for better applying existing methods and further designing new algorithms. We draw conclusions and give insightful technical and empirical analyses to all of our experimental results. We also propose possible improvements of existing methods, e.g., a sub-pixel accuracy refined version of ORB-SLAM delivering boosted performance. Although our experiments are performed only on the three representative methods, we try our best to make the obtained knowledge generalizable to other methods.

II. RELATED WORK

In this section we briefly introduce the principles of the three VO formulations together with their respective selected representatives. Afterwards we list the datasets used for our experiments.

A. Direct Methods

Direct methods use either all image pixels (dense) [7], all pixels with sufficiently large intensity gradient (semi-
dense) [5], or sparsely selected pixels (sparse) [9] and minimize a photometric error obtained by direct image alignment on the used pixels. Camera poses and pixel depths are estimated by minimizing the photometric error using non-linear optimization algorithms. Since much image information can be used, direct methods are very robust in low-texture scenes and can deliver relatively dense 3D reconstructions. Consequently, due to the direct image alignment formulation, direct methods are very sensitive to unmodeled artifacts such as rolling shutter effect, camera auto exposure and gain control. More crucially, the brightness constancy assumption does not always hold in practice, which drastically reduces the performance of direct methods in environments with rapid lighting change.

**Direct Sparse Odometry (DSO).** Unlike previous direct VO or SLAM algorithms which use dense or semi-dense formulations, DSO performs a novel sparse point sampling across image areas with sufficient intensity gradient. Reducing the amount of data enables real-time windowed bundle adjustment (BA) which jointly optimizes for all model parameters, including camera poses, depths, camera intrinsics and affine brightness transformation factors. The optimal parameters are obtained by minimizing the photometric error using the Gauss-Newton method, which achieves a good trade-off between speed and accuracy. Obsolete and redundant information is marginalized with the Schur complement [12], and the First Estimate Jacobians technique is involved in the non-linear optimization process [12], [13] to avoid the inconsistency and keep the observability of the system. As a direct method, DSO is fundamentally based on the brightness constancy assumption, thus the authors proposed a photometric camera calibration pipeline to recover the irradiance images [9], [11]. Performing direct image alignment on the irradiance images instead of the original images removes artifacts polluting the brightness constancy assumption, hence drastically increases the tracking accuracy [9]. An example of photometric calibration is shown in Fig. 1.

**B. Feature-based Methods**

Feature-based methods in turn extract only a sparse set of keypoints from each image and match them across multiple frames. Camera poses and feature depths are estimated by minimizing the reprojection errors between feature pairs. As modern feature descriptors are to some extent invariant to illumination and viewpoint changes, feature-based methods are more robust than direct methods to brightness inconsistencies and large viewpoint changes among consecutive frames. However, feature extraction and matching bring additional computational overhead, which highly limits the number of features that can be maintained in the system. As a result, the reconstructed 3D maps are much sparser and often cannot be used directly in practice. Moreover, in low-texture environments where not enough features can be extracted, tracking can easily get lost. Feature-based methods are also found to be more sensitive to image blurring [9].

**ORB-SLAM.** Developed by Mur-Artal et al., ORB-SLAM has become one of the most popular feature-based methods and has been widely adopted for a variety of applications. It uses ORB features [14] for all the tasks including tracking, mapping, re-localization and loop closing. To evaluate its VO performance, we disable its loop closure and global BA functionalities, and only focus on its tracking and local mapping components. To track a new frame, motion-only BA is performed on its feature matches to estimate the initial pose, which is later refined by using all the feature matches in the local map and performing the pose optimization again. A data structure called covisibility graph is used to improve system efficiency by limiting the BA to a local covisible area. Unlike in DSO, in ORB-SLAM old points and keyframes are culled out directly from the active window without marginalization.

**C. Semi-Direct Methods**

**Semi-Direct Visual Odometry (SVO).** SVO [8], [10] is so far the only semi-direct approach to our knowledge, which has been considered to be a hybrid of the two previously mentioned formulations. It extracts FAST corners and edgelets on keyframes and performs direct image alignment on those areas for initial pose estimation. Similar to DSO, SVO also uses small patches around the selected pixels to improve the accuracy and robustness. The depths of the selected pixels are estimated from multiple observations by means of a recursive Bayesian depth filter which is supposed to be fast and robust. To reduce the drift caused by incremental estimations, poses and depths are refined by BA. To do that, new patches are extracted around the selected pixels on the reference frame, and aligned to the new frame using an inverse compositional Lucas-Kanade algorithm. Based on the feature alignment, the reprojection error is computed and minimized in BA using iSAM2 [15], by which the poses and depths are refined. Due to its exceptional high efficiency...
(around 400fps on a laptop) and low cost, SVO can be easily transplanted to devices with limited computational resources, thus has gained popularity in a wide range of robotics applications.

D. Datasets

We use the following datasets for our experiments which cover a variety of real-world settings, e.g., indoor/outdoor, texture/textureless, global/rolling shutters.

The TUM Mono VO Dataset [11] contains 50 sequences captured by using a global shutter camera with two different lenses. Camera response function, dense attenuation factors and exposure time of each image are provided for photometric calibration.

The EuRoC MAC Dataset [16] contains 11 sequences recorded by global shutter cameras mounted on a drone. It covers three different indoor environments. Some of the sequences are quite challenging as they have extremely unstable motion and strong brightness change.

The ICL-NUIM Dataset [17] has been extended by Kerl et al. [18] to provide both simulated rolling shutter and global shutter sequences of the same indoor environment, which we use to compare the performance differences of the selected VO methods on images with rolling/global shutters.

The Cityscapes Dataset [19] provides a long street view sequence captured by industrial rolling shutter cameras in Frankfurt. We use it to evaluate how the selected methods work against realistic rolling shutter effect.

III. Evaluation

A. Photometric Calibration

In the first experiment, we evaluate how photometric calibration can influence the performances of the chosen direct, feature-based and semi-direct methods, focusing more on analyzing its impacts on formulations other than direct method. We use the 50 original sequences from the TUM Mono VO Dataset and their corresponding ones after photometric calibration, i.e., with the nonlinear camera response function $G$ and pixel-wise vignetting factors $V$ calibrated. Example images with/without photometric calibration can be found in Fig. 1. Each method runs 10 times on each of these 100 sequences. The accumulative histogram of the alignment error $e_{align}$, rotation drift $e_r$ (in degree) and scale drift $e_s$ are calculated [11] and shown in Fig. 3. It is worth noting that the exposure times $t$ into the formulation of ORB-SLAM and SVO (not open-sourced) is not straightforward, therefore we do not use them for all three methods. However, for reference, we also show the results of DSO with all calibration information used, i.e., $G$, $V$ and $t$.

As can be seen in Fig. 3, with $G$ and $V$ calibrated, the performance of DSO increases significantly. In addition, using $t$ can further boost the tracking accuracy. However, interestingly, photometric calibration reduces the overall performance of ORB-SLAM and SVO. We give our explanations in the following.

DSO. As direct methods are fundamentally built up on the assumption that a 3D point should have the same intensity across different images, it is straightforward to conclude that removing or reducing brightness inconsistencies among images can improve their performances. More evidence can also be found in [9].

ORB-SLAM. To better understand the performance declines of ORB-SLAM in Fig. 3, we further show its performance on each sequence in Fig. 2, where the differences of the alignment errors with/without photometric calibration $e_{align}^{PC} - e_{align}^{nPC}$ are shown in the left and middle, alignment errors of all runs are shown in the right. As can be seen there, both the tracking robustness and accuracy decrease after photometric calibration: ORB-SLAM fails on 6 sequences and generally performs worse on the other sequences. It is worth noting that the performance decline is not consistent over sequences. To understand the inconsistency, we need to relook at the inverse camera response function $G^{-1}$ shown in Fig. 1a. This nonlinear function can be roughly divided into three linear parts with pixel values $I$ belonging to $[0, 90)$, $[90, 190)$ and $[190, 255]$. Due to the different slopes, applying $G^{-1}$ compresses intensities in $[0, 90)$ and stretches the ones in $[190, 255]$. In other words, it reduces contrast of dark areas while increases it for bright areas.

As features like ORB generally work better on images with higher contrast (more evenly distributed intensity histogram), we assume the performance declines of ORB-SLAM are mainly caused by dark frames. To verify this we perform ORB feature matching before and after photometric calibration on four sample image pairs, where two of them are with dark images and the other two are with bright ones. The numbers of feature matches and image histograms are shown in Fig. 4. It can be seen there, after photometric calibration the numbers of ORB feature matches decrease on the dark image pairs and increase on the bright ones. Although sometimes the drop of the number of matched...
features may not seem crucial (as in the second column of Fig. 4), the effect can be accumulated over multiples frames. When the system projects all features within the local map to the newest frame to search for their correspondences, the number of matches can decrease drastically as shown in Fig. 5. As a result, only few features from the newest frame will be considered as inliers and added into the system, which is the main reason for the tracking failures in Fig. 2. This conclusion can be generalized to other feature-based methods if the used features are fundamentally based on image gradients. To apply these methods, we suggest to enable the camera gamma correction and auto exposure when possible.

**SVO.** In Fig. 3, photometric calibration also reduces the performance of SVO, but the performance declines are overall smaller than those of ORB-SLAM. Recall that SVO extracts FAST corners and edgelets and uses image patches around them as features for feature matching. The photometric calibration used in our experiments can reduce both the numbers and the qualities of the extracted corners and edgelets on dark images, and can also influence the following matching of image features. This results in the performance declines in Fig. 3. On the other hand, SVO performs direct image alignment for initial pose estimation,
to which photometric calibration is supposed to be beneficial. We believe this is the reason for the reduced performance declines compared to those of ORB-SLAM.

### B. Motion Bias

The term motion bias here refers to the difference of VO performance caused by running the same sequence forward and backward. As shown in the top two subfigures of Fig. 6, experiments in [9], [11] demonstrate that DSO does not suffer from such bias, while ORB-SLAM performs better when running backward. While this issue was raised there, no analysis, conclusion or possible remedies were given, which we are going to address in this section. To get a more thorough understanding of the influence of motion bias, we first perform the same experiment for SVO and show the result at the bottom of Fig. 6. Surprisingly, SVO does not perform very well on this dataset and it fails for all backward runs. Since SVO 2.0 is not open-sourced by far, we cannot analyze the reason but can only suspect that this is due to some implementation issues. We exclude SVO from the remaining experiments on the TUM Mono VO Dataset in this section.

Both for direct and feature-based methods, triangulation is a necessary step for estimating depths of newly observed 3D points. According to [20], the images of a 3D point in two frames can be mapped by (assuming there is no relative rotation between the two frames or the relative rotation is very small)

\[
x' = x + Kt/z \iff z = Kt/(x' - x),
\]

where \(x\) and \(x'\) are the point images in homogeneous coordinates, \(K\) the intrinsic camera matrix, \(t\) the relative translation between the two frames, \(z\) the point depth in the first frame. This means that better depth estimation is achieved with larger disparity between an image pair. When the camera is moving forward in a relatively open area, new points will emerge from the image center and have relatively small motions among consecutive frames. This pattern of optical flow introduces poorly initialized depths into the system. On the contrary, when moving backward, points close to the camera come into the field of view with large parallaxes, thus their depths are better initialized. We claim this is the main reason for the improved performance of ORB-SLAM when running backward.

To verify this, we check the sequences on which ORB-SLAM performs much better running backward and show them in the first 5 subfigures in Fig. 7. It can seen that all of them fulfill our description above. We also check the two counter examples, namely sequence 31 and 44, on which ORB-SLAM performs better running forward. At the end part of sequence 31 there is a large amount of high frequency textures (leaves) as shown in the last image in Fig. 7, which makes ORB-SLAM not able to initialize or fail directly after the initialization. In sequence 44, interestingly, the camera is moving most of the time backward, which in fact verifies our conclusion.

To further support our analysis, we run ORB-SLAM on the EuRoC MAV Dataset, where the sequences are captured in a relatively closed indoor environment and the camera motion is rather diverse without any clear pattern. We assume ORB-SLAM should deliver similar results running forward and backward. The result is shown in Fig. 8 and it verifies our assumption.

The analysis above does not explain why DSO performs consistently running forward and backward. We claim here the performance gain of DSO mainly comes from its implementation and feature-based or semi-direct methods can be improved taking into consideration the issues discussed in the following.

- **a) Depth representation.** Instead of using depth directly like ORB-SLAM, DSO uses an inverse depth parametrization that affects the validity range of linearization and can better cope with distant features [21]. We claim that the distant points, which are poorly initialized from the image center, have less impact on DSO than on ORB-SLAM.

- **b) Point sampling strategy.** DSO samples points evenly across the entire image, which can be beneficial to avoid selecting many points from locations that only give poor initializations (e.g., image center).

---

\(1\) We already use the settings recommended by one coauthor of SVO 2.0.
TABLE I: Results of original ORB-SLAM and our refined ORB-SLAM. We run both methods 10 times on each of the selected sequences from Fig. 7. The data of original ORB-SLAM is obtained from [11].

<table>
<thead>
<tr>
<th>Seq.</th>
<th>Original Refined</th>
<th>Original Refined</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>12.29</td>
<td>3.06</td>
</tr>
<tr>
<td>23</td>
<td>10.33</td>
<td>5.52</td>
</tr>
<tr>
<td>29</td>
<td>21.84</td>
<td>10.52</td>
</tr>
<tr>
<td>46</td>
<td>27.18</td>
<td>14.89</td>
</tr>
<tr>
<td>47</td>
<td>20.57</td>
<td>10.85</td>
</tr>
<tr>
<td>mean</td>
<td>18.44</td>
<td>8.97</td>
</tr>
</tbody>
</table>

**c) Point management.** In ORB-SLAM, features extracted from a new frame will be added into the system, if they can match those features that are already in the system but haven’t been matched before. If all these features gather together at the central image area, they will be added with inaccurate depth estimations. In contrary, DSO only samples candidate points from the new frame but does not add them to the system immediately. The depth estimations of these points keep being refined (outliers are removed) before they are activated and added. Moreover, points are only selected to be activated if they can keep the uniform spatial distribution of all activated points. All these strategies prevent problematic points from being added into the system.

**d) Discretization artifacts.** In direct methods, the depth of a newly observed point is initialized by searching for its correspondence in the reference frame along the epipolar line using sub-pixel accuracy. In feature-based methods, however, a new feature is extracted and matched to a previously observed feature with both of them at discretized image locations. Thus feature-based methods suffer more from pixel discretization artifact, especially when matching those distant features emerging from the image center running forward. To verify our analysis, we first perform the experiment in Fig. 9 where we run DSO and ORB-SLAM forward and backward on sequences sampled to different resolutions. The performance of DSO drops a little on low resolution sequences, but overall it is robust to such artifact. In contrast, the performance gaps of ORB-SLAM between running forward and backward increase significantly with reduced resolutions (thus severer discretization artifact).

In our second experiment, we adopt a sparse optical flow algorithm to refine the feature matching step of ORB-SLAM to achieve sub-pixel precision. We use the iterative Lucas-Kanade method implemented in OpenCV and run the refined ORB-SLAM on the first 5 sequences shown in Fig. 7. The result is shown in Table I. ORB-SLAM performs similarly running backward as before but much better (more than 50% on average) running forward, which supports our analysis. For reference we also show the results on all the sequences in Fig 10.

**C. Rolling Shutter Effect**

In the first experiment of this section, we run DSO, ORB-SLAM and SVO 10 times on each of the 4 Living Room sequences of the ICL-NUIM Dataset, as well as on their simulated rolling shutter correspondences where the rolling shutter effect is relatively strong. The results are shown in Fig. 11a. All three methods are influenced by rolling shutter effect, yet the performance declines of DSO and SVO are apparently larger than the one of ORB-SLAM. This result verifies that feature-based methods are more robust to the rolling shutter effect than direct methods. To show the influence of the rolling shutter effect on direct methods, we show examples of the reconstructed scene by DSO in Fig. 11b and Fig. 11c. It can be seen that the delivered reconstruction has very large scale drift on the rolling shutter sequence (the big structure in the background of Fig. 11c is the drifted reconstruction of the painting in the foreground).

Although SVO performs feature matching followed by BA for refining structures and poses, it uses direct image alignment on corners and edgelets to provide initial poses for BA. It is also worth mentioning, as can be seen in Fig. 11a, that the overall performances of DSO and SVO on this dataset significantly transcend the one of ORB-SLAM under both global and rolling shutter settings. The main reason is that the scenes in the dataset are indoor environments with low textured structures such as walls, floors and doors and thus are very challenging for feature-based methods. As a result, the selected direct and semi-direct methods
outperform the feature-based method even on rolling shutter images.

While the results above coincide with our intuition, it sometimes can be misleading. One may easily draw the conclusion that on sequences with enough texture and captured using rolling shutter cameras, feature-based methods should be preferable than direct or semi-direct methods. This is not always the case. It is worth noting that the rolling shutter effect in the extended ICL-NUIM Dataset is artificially simulated. On modern industrial level cameras, pixel read-out speeds are usually extremely fast such that the rolling shutter effect is to some extent neglectable for many applications. In the second experiment, we aim at comparisons on images with such realistic rolling shutter effect. As there is no such dataset that provides both real global shutter and rolling shutter sequences, we only compare the VO accuracies on realistic rolling shutter sequences. For this purpose, we use the Frankfurt sequence of the Cityscapes Dataset and split it into smaller segments (each with around 6000 frames). To our surprise, ORB-SLAM always fails on the selected segments: whenever the camera rotates strongly at street corners or large occlusion occurs due to moving vehicles. We thus suspect the failures are not caused by the rolling shutter effect. In Fig. 12 we show the estimated camera trajectories of DSO and SVO. While SVO is able to run on the entire selected segments, it suffers from severe scale drifts. In contrast DSO can still deliver satisfying results. Without a proper dataset, it is difficult to analyze the

---

Fig. 9: Performance differences of ORB-SLAM (top) and DSO (down) on the TUM Mono VO Dataset due to motion bias at different image resolutions. While DSO delivers similar results under different settings, ORB-SLAM performs consistently better when running backward and the performance gaps increase with reduced image resolutions.

Fig. 10: Performance comparison between the original ORB-SLAM and our refined version on the full TUM Mono VO Dataset. With sub-pixel accuracy refinement of feature matching, ORB-SLAM performs better running forward and similarly running backward.

Fig. 11: Results on the extended ICL-NUIM Dataset with original global shutter setting and the simulated rolling shutter setting.
When implemented properly, even direct methods can still be robust (to some extent, of course) to the rolling shutter effect.

**IV. Conclusions**

We present a thorough evaluation for state-of-the-art direct, semi-direct and feature-based methods on photometric calibration, motion bias and the rolling shutter effect, with the aim to provide practical inputs to the community for better applying existing methods and developing new VO and SLAM algorithms. Our main conclusions are:

1. With photometric calibration, the performance of direct methods gets improved significantly, while for semi-direct and feature-based methods, it depends on the used feature, the camera response function and the overall brightness of the scene. When the features are computed based on image gradients, we suggest to enable the camera gamma correction and auto exposure. For direct methods, when photometric calibration information is not available, online calibration methods should be developed.

2. Compared to direct methods, feature-based methods have a relatively larger performance bias when running forward and backward. While this may be mainly an implementation issue, possible reasons are discussed: depth representation, point selection and management, discretization artifact. When adopting existing feature-based methods for applications like autonomous driving, more effort should be taken to address the motion bias.

3. Direct and semi-direct methods are more sensitive to the rolling shutter effect. But when the rolling shutter effect is not so strong, or the environment is low textured, the performance might depend more on the specific implementation. When implemented properly, even direct methods can still deliver satisfying results. Besides, a specific dataset is needed for getting a better understanding on the rolling shutter effect.

4. The used feature-based methods are more sensitive to pixel discretization artifact. When possible, images with higher resolutions are preferred. Moreover, sub-pixel accuracy refinement on feature extraction and matching can boost their performance.

5. Based on the current implementations, the monocular version of DSO is generally more robust than those of SVO and ORB-SLAM under different settings.

**References**