Unsupervised Monocular Visual-inertial Odometry Network

Peng Wei\textsuperscript{1,2*}, Guoliang Hua\textsuperscript{1*}, Weibo Huang\textsuperscript{1}, Fanyang Meng\textsuperscript{2} and Hong Liu\textsuperscript{1,2†}
\textsuperscript{1}Key Laboratory of Machine Perception, Peking University, Shenzhen Graduate School, China
\textsuperscript{2}Peng Cheng Laboratory, Shenzhen, China
\{weapon, glnhua, weibohuang, hongliu\}@pku.edu.cn, mengfy@pcl.ac.cn

Abstract

Recently, unsupervised methods for monocular visual odometry (VO), with no need for quantities of expensive labeled ground truth, have attracted much attention. However, these methods are inadequate for long-term odometry task, due to the inherent limitation of only using monocular visual data and the inability to handle the error accumulation problem. By utilizing supplemental low-cost inertial measurements, and exploiting the multi-view geometric constraint and sequential constraint, an unsupervised visual-inertial odometry framework (UnVIO) is proposed in this paper. Our method is able to predict the per-frame depth map, as well as extracting and self-adaptively fusing visual-inertial motion features from image-IMU stream to achieve long-term odometry task. A novel sliding window optimization strategy, which consists of an intra-window and an inter-window optimization, is introduced for overcoming the error accumulation and scale ambiguity problems. The intra-window optimization restrains the geometric inferences within the window through checking the photometric consistency. And the inter-window optimization checks the 3D geometric consistency and trajectory consistency among predictions of separate windows. Extensive experiments have been conducted on KITTI and Malaga datasets to demonstrate the superiority of UnVIO over other state-of-the-art VO / VIO methods. The codes are open-source\textsuperscript{1}.

1 Introduction

VO or VIO is a fundamental task that aims to track the incremental motion of the sensor and simultaneously build a map of the environment. Traditional monocular VO methods [Mur-Artal and Tardós, 2017, Geiger et al., 2011, Engel et al., 2017] utilize handcrafted features or photometric matches to calculate the trajectory from a monocular image sequence. However, these methods are impressionable to motion blur, occlusion, and textureless regions. As a complementary sensor of visual cameras, inertial measurement unit (IMU) has been widely adopted in VIO methods [Huang and Liu, 2018, Bloesch et al., 2015, Leutenegger et al., 2013, Qin et al., 2018] for its high-frequency motion measurement and relatively low cost. The use of IMU can help to increase the robustness as well as improving the accuracy.

With the development of CNN and RNN, various learning-based VO or VIO methods have been proposed. Although many supervised methods [Wang et al., 2018, Clark et al., 2017, Chen et al., 2019] have been revealed more competitive than traditional methods, the demands of a large number of labeled data, i.e., the ground truth poses acquired from high-precision devices, limit the application of the technology. Self-supervised methods [Shamwell et al., 2018, Han et al., 2019] release the pressure of collecting large quantities of ground truth, but they still require other expensive data, e.g., depth map, which degrades the flexibility. In contrast, unsupervised VO methods [Zhou et al., 2017, Bian et al., 2019] only utilize image sequences to achieve pose estimation, requiring no ground truth label nor expensive data input. However, existing unsupervised VO methods suffer from poor capability on long-term odometry task, due to the inherent limitation of only relying on visual data that may degrade in some cases. Besides, the error accumulation problem in long-term trajectory was ignored in previous methods, thus causing the mediocre result.

In this paper, an unsupervised visual-inertial odometry framework (UnVIO) is proposed. As shown in Fig. 1, by taking the consecutive images and IMU measurements as input, UnVIO is able to predict the depth map and estimate the ego-motion. In particular, a heuristic fusion module is introduced to self-adaptively fuse the visual and inertial features, enabling the model to handle data pollution. The entire framework is trained in an unsupervised end-to-end fashion, through a proposed sliding window optimization strategy. A sliding window is utilized to traverse through the sequence, where the geometric constraint and sequential constraint are exploited to optimize the geometric inferences within and among windows. The contributions can be listed as follows:

- An end-to-end unsupervised visual-inertial odometry framework (UnVIO) is proposed for estimating the ego-motion as well as predicting the depth map.

\textsuperscript{*}Equal contribution

\textsuperscript{†}Corresponding Author

\textsuperscript{1}https://github.com/Ironbrotherstyle/UnVIO
A visual-inertial feature fusion module is designed to select the most discriminative motion features for camera pose regression. The module improves the robustness to the contamination of image-IMU input.

A sliding window optimization strategy, consisting of an intra-window optimization and an inter-window optimization, is proposed for unsupervised VIO to tackle the error accumulation and scale ambiguity problems.

2 Related Work

Traditional methods. Traditional visual odometry methods can be divided into two categories, feature-based methods and direct methods. ORB-SLAM2 [Mur-Artal and Tardós, 2017] is a classical feature-based method which extracts hand-crafted features and utilizes bundle adjustment to estimate ego-motion in real-time. DSO [Engel et al., 2017] is a sparse direct method that performs the epipolar matching to achieve camera tracking, based on the assumption of photometric consistency.

In order to increase the robustness and improve the performance, researchers exploit IMU measurements as supplemental information, hence extending VO methods to VIO methods [Bloesch et al., 2015, Huang and Liu, 2018]. OKVIS [Leutenegger et al., 2013] is a tightly-coupled method which optimizes the reprojection error and IMU error at the same time. VINS-Mono [Qin et al., 2018] fuses preintegrated IMU measurements with visual feature observations to achieve accurate pose estimation.

Supervised/Self-Supervised learning methods. By harnessing the deep convolutional and recurrent neural networks, Wang et al. [Wang et al., 2018] designed a supervised architecture to estimate camera pose from the monocular image sequence. VNNet [Clark et al., 2017] firstly tackled VIO in a supervised manner. Chen et al. [Chen et al., 2019] exploited two masking strategies for visual-inertial sensor fusion.

Some self-supervised learning methods were proposed for releasing the pressure of collecting quantities of ground truth labels for supervised learning. Shamwell et al. [Shamwell et al., 2018] presented VIOLearner that carried out online error correction in multiple scales to refine the pose estimation. But VIOLearner required the depth map as input, thus limiting its applicability on other scenes without supplying depth data. DeepVIO [Han et al., 2019] is a self-supervised VIO method that uses 3D geometric constraint as supervision. However, DeepVIO needs an awesome pretrained stereo network PSMNet [Chang and Chen, 2018] to provide the accurate and dense disparity map for training.

Unsupervised learning methods. Zhou et al. [Zhou et al., 2017] proposed an unsupervised framework of pose estimation and depth prediction. The framework can be trained by only using image sequences. Shen et al. [Shen et al., 2019] proposed a matching loss constrained by epipolar geometry and improved the odometry performance. In addition to the photometric matching loss, PatchGAN [Vankadari et al., 2019] adopted the generative adversarial approach to promote the depth prediction and pose estimation results. Different from these unsupervised VO methods, we propose an unsupervised VIO method that significantly improves the odometry performance. In contrast to self-supervised VIO methods, ours requires no extra expensive data as input, but achieves competitive results through the sliding window optimization strategy and visual-inertial fusion module.

3 Unsupervised Visual-inertial Odometry

An overview of the unsupervised visual-inertial odometry framework is shown in Fig. 1. The DepthNet learns a mapping from single RGB image to depth map (see Sec.3.1). By taking the raw monocular image sequence and IMU measurements as input, the visual-inertial odometry networks estimate the ego-motion (see Sec.3.2). The whole framework is trained in a sliding window optimization strategy and visual-inertial fusion module.

3.1 Depth Estimation

Given an image \( I \in \mathbb{R}^{3 \times H \times W} \), the DepthNet learns a mapping function \( F_D \) that infers the scene depth of per pixel,
i.e., $D = \mathcal{F}_D(I)$. The DepthNet is designed based on an encoder-decoder architecture where the encoder part maps the RGB image into a high dimensional feature space, and the decoder maps these features into depth values. The pretrained ResNet18 [He et al., 2016] is adopted as the encoder and the skip-connection is exploited between the encoder and decoder for retaining structure details. For the decoder, the nearest neighboring upsampling operation followed by a Conv layer is used to expand the resolution. Exponential linear units are appended after each Conv layer as recommended in [Godard et al., 2017].

3.2 Visual-inertial Odometry

Two parallel networks are designed to extract visual features and inertial features, followed by a visual-inertial fusion module to select the most efficient features. Then, the PoseNet takes the fused temporal visual-inertial features as input to regress 6 DOF poses.

**Visual feature extraction.** Two adjacent frames $I_{t-1}$, $I_t$ from the image sequence are concatenated along the channel dimension as the input of *VisualNet*. The architecture of *VisualNet* is made up of the first 7 Conv layers of *FlowNet* [Dosovitskiy et al., 2015] and a global average pooling. The process of visual motion feature extraction can be formulated as:

$$F^V_i = \Phi (I_{t-1} \oplus I_t),$$

where $\oplus$ denotes the concatenation in the channel dimension, $\Phi$ is the feature extraction function of *VisualNet*.

**Inertial feature extraction.** IMU measures the linear acceleration and angular velocity of the embedded body at a faster rate than the visual measurement. The sampled raw IMU measurements from time $t - 1$ to $t$ are arrayed in the following form:

$$M = \begin{bmatrix} \alpha^0_{t-1} & \omega^0_{t-1} \\ \vdots & \vdots \\ \alpha^n_{t-1} & \omega^n_{t-1} \end{bmatrix} \in \mathbb{R}^{n \times 6},$$

where $\alpha, \omega \in \mathbb{R}^3$ are the linear acceleration and angular velocity respectively, $n$ is the number of IMU samples. The sequential IMU measurements are then sent into a two-layer LSTM [Hochreiter and Schmidhuber, 1997] to get the inertial motion features:

$$F^I_i, \mathcal{H}^i = \mathcal{R} \{ (\alpha^i, \omega^i) ; \mathcal{H}^{i-1} \},$$

where $\mathcal{R}$ represents the recurrent function of IMUNet, $\mathcal{H}^i$ is the hidden state. By this way, sequential IMU measurements are integrated into the final inertial motion feature $F^I_i$.

**Visual-inertial feature fusion.** A straightforward but effective fusion strategy is designed to fuse visual features and inertial features. The concatenated visual-inertial features $F = F^V_i \oplus F^I_i$ along channel dimension will be firstly aggregated into squeezed features $F'$, by a learned basis vector group $G$ and a learned bias vector $b$:

$$F' = G \cdot F + b.$$  

Then, $F'$ are decoded to a weight vector $W$ that indicates the importances of channel-wise visual and inertial features:

$$W = \sigma (\mathcal{F}_F (F')),$$

where $\mathcal{F}_F$ is the decoding function of the fusion module, and $\sigma$ represents the sigmoid function. The recalibrated visual-inertial features $F$ are obtained through the Hadamard product of $W$ and $F$: $\mathbf{F} = F \odot W$.

**Pose estimation.** Given an image-IMU stream, the motion feature set $\{F^V_i, F^I_{i-1}, \ldots, F^V_{i-w}, F^I_{i-w} \}$ where each item represents the visual-inertial features between two adjacent times, is fed into PoseNet to dig the temporal relevance:

$$\mathbf{T}_{i-1}, \mathcal{H}^i = \mathcal{R} \{ (F^i_{i-1}) ; \mathcal{H}^{i-1} \},$$

where $\mathcal{H}^i$ is the hidden state output, and $\mathcal{R}$ is the refining function of PoseNet. $\mathbf{T}_{i-1}$ is the refined motion features between frame $i - 1$ and frame $i$, which is subsequently sent to a linear layer to obtain 6 DOF camera pose $p_t$.

3.3 Sliding Window Optimization

The key supervision of the unsupervised visual-inertial odometry framework comes from the multi-view geometric constraint and sequential constraint. Given a sequence of visual-inertial measurements at different times, $\{\{I_0, M_0\}, \ldots, \{I_{s-1}, M_{s-1}\}\}$, a sliding window traverses through the sequence, with consistency check leveraged to optimize the geometric inferences (depth and camera pose). An example of sliding window optimization with step size 1 and window size 3 is shown in Fig. 2. In each window $W$, the depth $D$ and camera poses $\mathbf{p}$ are independently predicted from the windowed visual-inertial measurements through our framework. The photometric consistency check is utilized to achieve individual intra-window optimization. To handle the error accumulation and scale ambiguity problems, additional inter-window optimization is designed to constrain predictions of different windows through checking 3D geometric consistency and trajectory consistency.

**Prior knowledge of multi-view geometry.** When a camera moves in a scene, objects that can be seen in adjacent views form the geometric constraint. Denote $I_s, I_t$ are two adjacent frames of source view and target view respectively, and $p_s, p_t$ are two pixel points that correspond to the same 3D map point of the scene. With the depth maps $D_s, D_t$ and the ego-motion transform matrix $T_{t \rightarrow s}$ available, the 3D geometric consistency can be set by:

$$D_s (p_s) K^{-1} p_s = T_{t \rightarrow s} D_t (p_t) K^{-1} p_t,$$
where $K$ is the camera intrinsic matrix. Also, Eq. (7) can be converted to indicate 2D reprojection constraint:

$$p_s \sim KT_{t \rightarrow s}D_t(p_t)K^{-1}p_t,$$

where $\sim$ means 'equal in the homogeneous coordinate. According to Eq. (8), a sample grid can be generated and used to warp $I_s$ into synthesized target-view image $\tilde{I}_s$, through a bilinear sampling. The photometric consistency check is defined by the appearance similarity between $I_s$ and $I_t$.

**Intra-window optimization.** The image-IMU stream $\{(I_s, M_t), \ldots, (I_{s+w-1}, M_{t+w-1})\}$ of each sliding window is utilized for geometric inference generation and intra-window optimization. The middle frame of the window is taken as the target view, while others are source views. The predicted depth map $D$ of the target view and estimated camera poses $\langle \tilde{p}^{i+1}_t, \ldots, \tilde{p}^{i+w-2}_t \rangle$ between adjacent frames within the window, are checked by photometric consistency:

$$L_{\text{photo}} = \sum_{(s, t)} \left( \lambda_1 \cdot \tilde{I}_s - I_t + \lambda_2 \cdot \text{SSIM}(\tilde{I}_s, I_t) \right),$$

where $(s, t)$ denotes all the source-target pairs. SSIM [Wang et al., 2004] represents the structural similarity metric. Additional smoothness loss is also adopted for alleviating the shortage of photometric consistency on the textureless regions as recommended in [Shen et al., 2019]:

$$L_{\text{smooth}} = \sum_{s, t} \left| \partial_x d_s \langle p_s, I_t \rangle \rangle + \partial_y d_s \langle p_s, I_t \rangle \rangle \right|.$$

The intra-window optimization loss can be summarized as:

$$L_{\text{intra}} = \alpha_1 \cdot L_{\text{photo}} + \alpha_2 \cdot L_{\text{smooth}},$$

where $\alpha_1$ and $\alpha_2$ are weighting factors.

**Inter-window optimization.** It is prone to fall into a local optimum by only relying on the optimization within windowed frames, due to the lack of sequential constraint that may cause the universal scale ambiguity and the accumulated error problems in monocular odometry. We consider the inter-window optimization, including trajectory consistency check and 3D geometric consistency check.

Partial information of the sequence is exploited to estimate the windowed ego-motion in intra-window optimization, i.e., $\{T_{i+1}^i, \ldots, T_{i+w-1}^i\} \rightarrow \langle \tilde{p}^{i+1}_t, \ldots, \tilde{p}^{i+w-2}_t \rangle$. To take the inter-window relevance into account, the integrated information is also exploited to estimate the camera poses for the entire sequence, i.e., $\{T_{i}^1, \ldots, T_{i+w-1}^i\} \rightarrow \langle p^{1}_s, \ldots, p^{i+w-1}_s \rangle$. The estimated camera poses $\tilde{p}_s$ that aggregated from the windowed estimation and the corresponding $p_s$ that estimated from the integrated sequential information are checked for the trajectory consistency by:

$$L_{\text{pose}} = \sum_i \left| \tilde{p}^{i+1}_t - p^{i+1}_t \right|.$$

With the sliding window traverses through the sequence, the middle-frame depth map that determines the scale of each window is predicted. Therefore, to ensure the uniform scale of contiguous windows, we project the depth map into 3D point clouds and then perform the 3D transform based on Eq. (7) to check the 3D geometric consistency. The 3D geometric consistency loss $L_{3D}$ for inter-window optimization is defined as:

$$L_{3D} = \sum_i \left| \tilde{D}_t - T_{i+1}^{i+1}D_{i+1}^{i+1} \right| / \tilde{D}_t + T_{i+1}^{i+1}D_{i+1}^{i+1},$$

where $\tilde{D}_t$ is the warped depth map from $D_t$. The loss function of inter-window optimization is then concluded as:

$$L_{\text{inter}} = \alpha_3 \cdot L_{\text{pose}} + \alpha_4 \cdot L_{3D}.$$
To summarize, the loss function of sliding window optimization can be written in:

\[ L_{\text{final}} = L_{\text{intra}} + L_{\text{inter}}. \]  

(15)

4 Experiments

In this section, both quantitative and qualitative results compared with traditional and learning-based VO/VIO methods are presented. The ablation study is employed to demonstrate the effectiveness of each component of our method.

4.1 Datasets

KITTI Dataset. KITTI dataset [Geiger et al., 2012] serves as a prevalent driving dataset, with stereo images at 10Hz, IMU data at 100Hz, accurate pose and laser scan. Seqs 00-10 of the odometry partition are used, except for 03 where IMU data is not acquirable. Seqs 00-08 excluding 03 are adopted for training and 09-10 are utilized for testing.

Malaga Dataset. Malaga [Blanco-Claraco et al., 2014] is an outdoor dataset. Stereo images at 20Hz, IMU measurements at 100Hz and GPS are provided. In our implementation, rectified left images are downsampled to 10Hz. Seqs 01, 02, 04, 05, 06, 08 are adopted for training and Seqs 03, 07, 09 are used for qualitatively evaluating since no ground truth pose is offered.

4.2 Training Details

All the models are implemented by using the Pytorch framework on a computer equipped with an Nvidia GeForce GTX1080 Ti GPU. Adam optimizer with learning rate \(10^{-4}\), \(\beta_1 = 0.9, \beta_2 = 0.999\) is utilized. Images for training on both datasets are resized to \(832 \times 256\), meanwhile, the IMU samples \(n\) is set to 11. The training process converges after about 100000 iterations with a batch size of 4. Besides, the length of training sequence \(s\) and window size \(w\) are 5 and 3 respectively in our experiment. The weights for loss functions are empirically given as: \(\alpha_1 = 1, \alpha_2 = 0.1, \alpha_3 = 0.1, \alpha_4 = 0.1, \lambda_1 = 0.15, \lambda_2 = 0.85\).

4.3 Odometry Evaluation

The evaluation of odometry is carried out among traditional VO methods VISO-M [Geiger et al., 2011] (monocular version), VISO-S [Geiger et al., 2011] (stereo version), ORBSLAM2 [Mur-Artal and Tardós, 2017], VIO methods VINS-Mono [Qin et al., 2018], OKVIS [Leutenegger et al., 2013], self-supervised VIO methods VIOlearner [Shamwell et al., 2018], DeepVIO [Han et al., 2019], and unsupervised VO methods SFM [Zhou et al., 2017], SC [Bian et al., 2019]. All the monocular methods need to be evaluated after making 7 DOF (6 DOF + scale) alignment with ground truth, apart from VINS and OKVIS that can recover the scale. Notably, we implemented the above open-source methods except for VIOlearner and DeepVIO to get the odometry results. KITTI benchmark [Geiger et al., 2013] is utilized as the evaluation criterion, where \(L_{\text{rel}}\) is the average translational RMSE drift (\%) on length of 100m-800m, and \(R_{\text{rel}}\) is the average rotational RMSE drift (\(^\circ\)/100m) on length of 100m-800m.

The quantitative results of odometry evaluation on KITTI dataset are summarized in Table 1. Seqs 00, 02, 05, 07, 08 of the training set are selected for evaluation in [Shamwell et al., 2018], therefore, we calculate the average errors of these...
Figure 4: Qualitative comparison of depth estimation among SfMLearner, SC, and the proposed UnVIO on KITTI and Malaga dataset. It is clear that the proposed method predicts depth maps with more details and sharper edges compared with competitors.

Table 2: The robustness test of VIO on four settings: camera-IMU calibration error (Mis-c), unsynchronization (Unsyn), IMU disturbance (IMU-D), and camera degradation (Cam-D).

<table>
<thead>
<tr>
<th>Seq</th>
<th>M</th>
<th>Mis-c: 10°</th>
<th>Unsyn: 20ms</th>
<th>IMU-D: 30%</th>
<th>Cam-D: 30%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ours</td>
<td>4.13</td>
<td>4.60</td>
<td>16.05</td>
<td>5.01</td>
</tr>
<tr>
<td></td>
<td>VINS</td>
<td>34.53</td>
<td>3.60</td>
<td>2.73</td>
<td>29.15</td>
</tr>
<tr>
<td>09</td>
<td>Ours</td>
<td>5.51</td>
<td>5.10</td>
<td>8.33</td>
<td>9.02</td>
</tr>
<tr>
<td></td>
<td>VINS</td>
<td>27.76</td>
<td>2.41</td>
<td>22.13</td>
<td>3.60</td>
</tr>
</tbody>
</table>

Table 3: Comparison of quantitative depth results on KITTI 09, 10. SfMLearner and state-of-the-art SC are used as a reference. The best of each metric is highlighted in bold.

<table>
<thead>
<tr>
<th>Method</th>
<th>Error metric</th>
<th>Accuracy metric (δ)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Abs Rel</td>
<td>Sq Rel</td>
</tr>
<tr>
<td>SfM</td>
<td>0.3272</td>
<td>3.1131</td>
</tr>
<tr>
<td>SC</td>
<td>0.1629</td>
<td>0.9644</td>
</tr>
<tr>
<td>Ours</td>
<td>0.1322</td>
<td>0.73005</td>
</tr>
</tbody>
</table>

Table 4: The ablation study of components on VO results. ’IMU’, ’SW’, ’Fusion’ mean IMU input, sliding window optimization and fusion strategy.

<table>
<thead>
<tr>
<th>Method</th>
<th>IMU</th>
<th>SW</th>
<th>Fusion</th>
<th>Seq 09</th>
<th>Seq 10</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>10.35</td>
<td>4.67</td>
<td>10.58</td>
</tr>
<tr>
<td>Ours</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>5.63</td>
<td>1.10</td>
<td>6.39</td>
</tr>
<tr>
<td>Ours</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>5.36</td>
<td>1.19</td>
<td>5.74</td>
</tr>
</tbody>
</table>

4.4 VIO Robustness Evaluation

Four settings that simulate sensor data collapse due to physical and thermal changes in the VIO system are conducted to test the robustness of UnVIO. Specifically, Mis-c:10° indicates adding 10° to the rotation matrix of camera-IMU extrinsic parameters. Unsyn:20ms means randomly adding 20ms to the IMU time-stamp. IMU-D:30% represents adding white noise to the accelerometer data and random walk noise to the gyroscope measurements at a rate of 30%. Cam-D:30% means adding blur, partial occlusion or full occlusion on the input images with a probability of 30%. Table 2 details the performance of VINS and the proposed method trained on KITTI on the abovementioned conditions. The proposed method has no performance deterioration on Mis-c because the extrinsic parameters are not required in our method. However, VINS seems to be confused and holds large errors as VINS relies on delicate calibration. On Unsyn, our method shows the advantages of handling the unsynchronized image-
IMU stream than VINS. Even when the input data is polluted, i.e. IMU-D and Cam-D, the proposed method still achieves better performance, showing better robustness to polluted input data.

4.5 Depth Evaluation

Since depth and pose estimation are coupled tasks, we test the performance of DepthNet following the odometry split. Table 3 gives quantified comparison with SfM, SC, and the proposed UnVIO on KITTI dataset. The results show that our method achieves the best performance in all metrics. Fig. 4 sketches the predicted depth maps. The first row lists the monocular inputs of KITTI and Malaga dataset, and the second to the last rows are the predicted depth maps corresponding to each input of the three methods. Intuitively, our results retain more details of the edges and contours of the predicted objects, e.g., the road sign and distant cars.

4.6 Ablation Study

Ablation study has been done to demonstrate the effectiveness of each proposed component, as shown in Table 4. By extending VIO methods to unsupervised VIO methods, the performance gains remarkable improvements via taking both image and IMU data as input. It can also be concluded that the proposed sliding window optimization is able to promote the translational and rotational performance. This is because the sliding window optimization strategy not only considers the intra-window photometric consistency, but also focuses on inter-window 3D geometric consistency and trajectory consistency to handle the widespread problems in monocular odometry. By self-adaptively fusing the visual-inertial features through the visual-inertial fusion module, the performance is further improved.

5 Conclusions

An unsupervised visual-inertial odometry framework (UnVIO) which only utilizes the monocular image-IMU stream for training and testing, is proposed in this paper. A visual-inertial feature fusion module is introduced to enable UnVIO to show robustness to polluted data. Besides, a novel sliding window optimization strategy with the advantages of overcoming scale ambiguity and error accumulation is proposed. Experimental results show that our method not only outperforms other unsupervised methods and traditional methods but also performs competitively with self-supervised VIO methods that need extra expensive depth data.

Acknowledgments

This work is supported by National Natural Science Foundation of China (No.U1613209, 61906103), National Natural Science Foundation of Shenzhen (No.JCYJ20190808182209321)

References


[Huang and Liu, 2018] W. Huang and H. Liu. Online initialization and automatic camera-imu extrinsic calibration for monocular


