Unsupervised Monocular Visual-inertial Odometry Network

Peng Wei[1,2*], **Guoliang Hua**[1*], **Weibo Huang**[1], **Fanyang Meng**[2] and **Hong Liu**[1,2†]
[1]Key Laboratory of Machine Perception, Peking University, Shenzhen Graduate School, China
[2]Peng Cheng Laboratory, Shenzhen, China
{weapon, glhua, 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[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.

*Equal contribution
†Corresponding Author
[1]https://github.com/Ironbrotherstyle/UnVIO

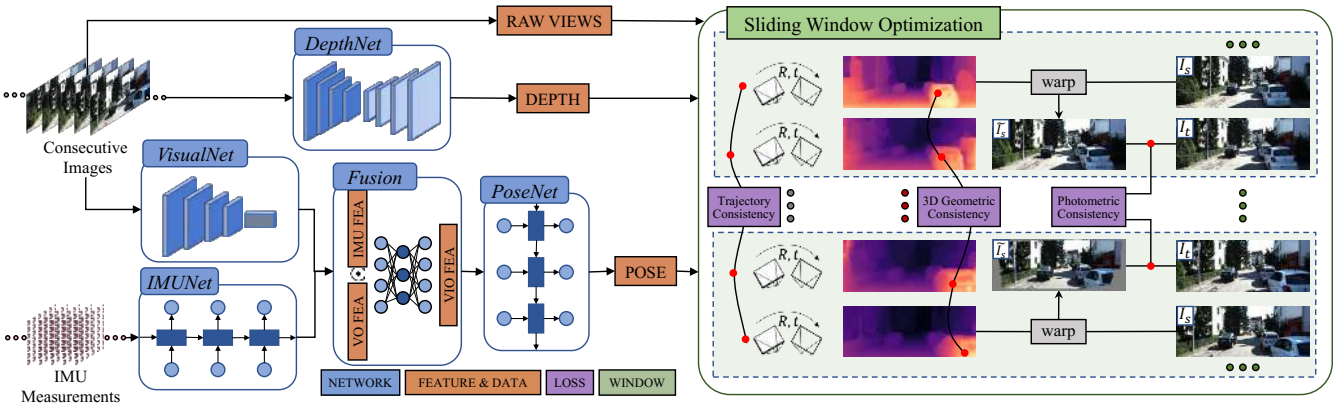


Figure 1: The pipeline of the proposed visual-inertial odometry framework (UnVIO). The *DepthNet* takes a single image as input, and outputs a dense depth map. The *PoseNet* takes the fused features from the concatenated adjacent views and contiguous IMU data to regress relative camera poses. The whole framework is trained through a sliding window optimization strategy.

- 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. VINet [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 that includes two parts: intra-window optimization and inter-window optimization (see Sec.3.3).

3.1 Depth Estimation

Given an image $I \in \mathbb{R}^{3\times H\times W}$, the *DepthNet* learns a mapping function $\mathcal{F}_D$ 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 remaps 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 reserving 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:

$$F_t^V = \mathbf{\Phi}\left(I_{t-1} \oplus I_t\right), \tag{1}$$

where $\oplus$ denotes the concatenation in the channel dimension, $\mathbf{\Phi}$ 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:

$$M = \begin{bmatrix} \alpha_{t-1}^0 & \omega_{t-1}^0 \\ \cdots & \cdots \\ \alpha_t^{n-1} & \omega_t^{n-1} \end{bmatrix} \in \mathbb{R}^{n\times 6}, \tag{2}$$

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:

$$F_i^I, \mathcal{H}^i = \mathcal{R}\left\{\left(\alpha^i, \omega^i\right); \mathcal{H}^{i-1}\right\}, \tag{3}$$

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_t^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 $F = F_t^V \oplus F_t^I$ along channel dimension will be firstly aggregated into squeezed features F', by a learned basis vector group G and a learned bias vector b:

$$F' = G \cdot F + b. \tag{4}$$

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

$$W = \sigma\left(\mathcal{F}_F\left(F'\right)\right), \tag{5}$$

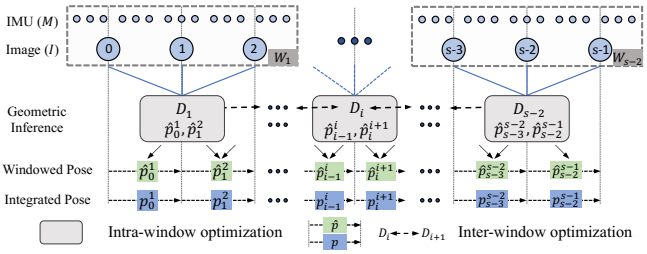


Figure 2: An illustration of the sliding window optimization, window size w is set to 3 for an instance. Photometric consistency is performed in intra-window optimization, while trajectory and 3D geometric consistency are performed in inter-window optimization.

where $\mathcal{F}_F$ is the decoding function of the fusion module, and σ represents the sigmoid function. The recalibrated visual-inertial features $\mathbf{F}$ 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 $\left\{\mathbf{F}_0^1, \mathbf{F}_1^2, \cdots, \mathbf{F}_{s-3}^{s-2}, \mathbf{F}_{s-2}^{s-1}\right\}$ where each item represents the visual-inertial features between two adjacent times, is fed into *PoseNet* to dig the temporal relevance:

$$\mathbf{T}_{i-1}^i, \mathcal{H}^i = \mathcal{R}\left\{\left(\mathbf{F}_{i-1}^i\right); \mathcal{H}^{i-1}\right\}, \tag{6}$$

where $\mathcal{H}^i$ is the hidden state output, and $\mathcal{R}$ is the refining function of *PoseNet*. $\mathbf{T}_{i-1}^i$ 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_{i-1}^i.

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, $\{\langle I_0, M_0\rangle, \cdots, \langle I_{s-1}, M_{s-1}\rangle\}$, 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 $\hat{p}_*$ 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 $T_{t\to s}$ available, the 3D geometric consistency can be set by:

$$D_s\left(p_s\right) K^{-1} p_s = T_{t\to s} D_t\left(p_t\right) K^{-1} p_t, \tag{7}$$

Method	Type	Metric	00	01	02	03	04	05	06	07	08	09	10	Avg (sub-t)	Avg (train)	Avg (test)
VISO-M	geo	t_{rel}	36.95	33.56	21.98	16.14	2.61	17.20	7.91	20.00	39.78	29.01	28.52	27.18	21.79	28.77
		r_{rel}	2.42	7.22	1.22	2.67	1.53	3.52	1.83	5.30	1.99	1.32	3.23	2.89	3.08	2.28
ORB-SLAM2	geo	t_{rel}	19.54	82.83	7.85	2.80	1.38	13.8	16.99	10.98	14.40	14.37	3.94	13.31	18.95	9.16
		r_{rel}	0.27	0.86	0.23	0.16	0.15	0.21	0.25	0.30	0.31	0.26	0.28	**0.26**	**0.3**	**0.27**
VINS	geo	t_{rel}	/	41.61	27.53	/	70.96	11.64	18.35	10.00	18.09	23.90	16.50	13.45	28.31	20.2
		r_{rel}	/	1.13	2.78	/	1.20	1.26	1.65	1.72	1.16	2.47	2.34	1.38	*1.56*	2.41
VIOLearner	s-sup	t_{rel}	5.62	/	4.07	/	/	3.00	/	3.60	2.93	1.51	2.04	3.84	/	1.78
		r_{rel}	3.63	/	1.48	/	/	1.40	/	2.06	1.32	0.90	1.37	1.98	/	1.14
DeepVIO	s-sup	t_{rel}	11.62	/	4.52	/	/	2.86	/	2.71	2.13	1.38	0.85	*4.77*	/	**1.12**
		r_{rel}	2.45	/	1.44	/	/	2.32	/	1.66	1.02	1.12	1.03	1.78	/	*1.08*
SfM	u-sup	t_{rel}	13.68	22.51	11.70	20.81	8.61	8.46	21.55	12.02	12.56	13.57	16.08	11.68	14.66	14.83
		r_{rel}	5.46	3.29	4.25	8.5	5.81	4.55	8.20	6.64	4.67	4.83	4.35	5.11	5.71	4.59
SC	u-sup	t_{rel}	10.03	25.78	9.07	7.52	3.24	6.23	13.56	6.45	9.92	11.52	10.44	8.34	*10.20*	10.98
		r_{rel}	3.84	1.16	2.16	2.49	0.91	1.78	2.10	2.14	1.98	3.26	4.73	2.38	2.06	4.00
Ours (NoIMU)	u-sup	t_{rel}	4.78	17.28	4.10	4.66	2.43	4.84	5.46	3.9	6.23	9.08	7.82	*4.77*	5.96	*8.45*
		r_{rel}	0.97	0.56	0.72	1.45	0.34	1.43	0.46	2.11	1.16	2.92	4.08	*1.28*	*1.02*	3.50
Ours	u-sup	t_{rel}	3.67	16.7	3.11	/	1.95	3.32	4.48	3.49	4.74	4.13	5.51	**3.67**	**5.18**	*4.82*
		r_{rel}	0.96	0.61	0.59	/	0.49	0.73	0.92	0.83	0.67	0.89	0.53	0.76	0.73	0.71

Table 1: Comparison of odometry performance with existing geometry-based (geo), self-supervised (s-sup), and unsupervised (u-sup) VO or VIO approaches on KITTI odometry dataset. The best, second-best, and third-best results of t_{rel} and r_{rel} are respectively highlighted in **bold**, underline and *italic*. / indicates that the data could not be acquired or the method fails on this sequence.

where K is the camera intrinsic matrix. Also, Equ. (7) can be converted to indicate 2D reprojection constraint:

$$p_s \sim K T_{t\to s} D_t\left(p_t\right) K^{-1} p_t, \tag{8}$$

where $\sim$ means 'equal in the homogeneous coordinate. According to Equ. (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 $\tilde{I}_s$ and I_t.

Intra-window optimization. The image-IMU stream $\{\langle I_i, M_i\rangle, \cdots, \langle I_{i+w-1}, M_{i+w-1}\rangle\}$ 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 $\langle \hat{p}_i^{i+1}, \cdots, \hat{p}_{i+w-2}^{i+w-1}\rangle$ between adjacent frames within the window, are checked by photometric consistency:

$$L_{\text{photo}} = \sum_{\langle s,t\rangle} \left(\lambda_1 \cdot \left|\tilde{I}_s - I_t\right| + \lambda_2 \cdot \text{SSIM}\left(\tilde{I}_s, I_t\right)\right), \tag{9}$$

where $\langle s,t\rangle$ 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]:

$$L_{\text{smooth}} = \sum_{i,j}\left(|\partial_x d_{i,j}|\, e^{-|\partial_x I_{i,j}|} + |\partial_y d_{i,j}|\, e^{-|\partial_y I_{i,j}|}\right). \tag{10}$$

The intra-window optimization loss can be summarized as:

$$L_{\text{intra}} = \alpha_1 \cdot L_{\text{photo}} + \alpha_2 \cdot L_{\text{smooth}}, \tag{11}$$

where α_1 and α_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., $\{\mathbf{T}_i^{i+1}, \cdots, \mathbf{T}_{i+w-2}^{i+w-1}\} \to \langle \hat{p}_i^{i+1}, \cdots, \hat{p}_{i+w-2}^{i+w-1}\rangle$. 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. $\{\mathbf{T}_0^1, \cdots, \mathbf{T}_{s-2}^{s-1}\} \to \langle p_0^1, \cdots, p_{s-2}^{s-1}\rangle$. The estimated camera poses $\hat{p}_*$ that aggregated from the windowed estimation and the corresponding p_* that estimated from the integrated sequential information are checked for the trajectory consistency by:

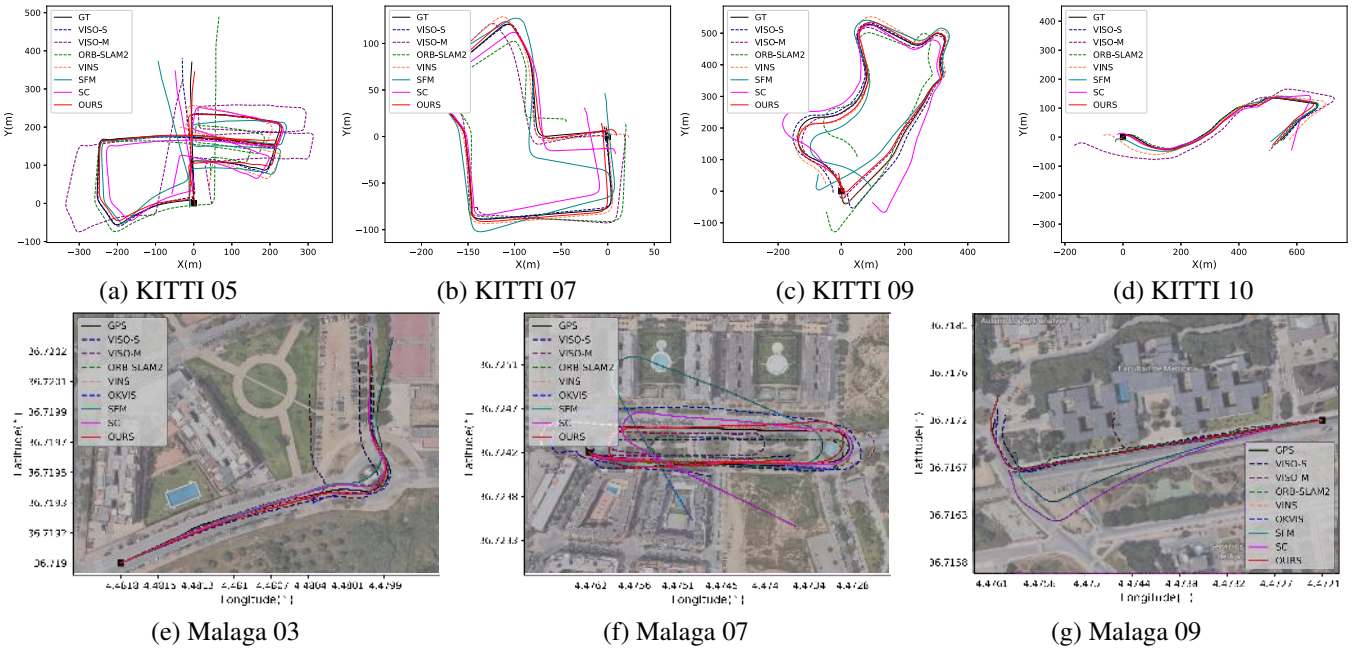
$$L_{\text{pose}} = \sum_i \left|\hat{p}_i^{i+1} - p_i^{i+1}\right|. \tag{12}$$

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 Equ. (7) to check the 3D geometric consistency. The 3D geometric consistency loss L_{3D} for inter-window optimization is defined as:

$$L_{3D} = \sum_i \frac{\left|\tilde{D}_i - T_{i+1\to i} D_{i+1}\right|}{\tilde{D}_i + T_{i+1\to i} D_{i+1}}, \tag{13}$$

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

$$L_{\text{inter}} = \alpha_3 \cdot L_{\text{pose}} + \alpha_4 \cdot L_{3D}. \tag{14}$$



(a) KITTI 05 (b) KITTI 07 (c) KITTI 09 (d) KITTI 10

(e) Malaga 03 (f) Malaga 07 (g) Malaga 09

Figure 3: Trajectory estimation on KITTI and Malaga dataset. (a), (b) are KITTI 05, 07 in the training set. (c), (d) are KITTI 09, 10 that are used for testing. (e), (f), (g) are the test trajectories of Malaga 03, 07, 09 overlaid on Google Map (GPS is served as reference instead). Best viewed in the colored electronic version.

To summarize, the loss function of sliding window optimization can be written in:

$$L_{\text{final}} = L_{\text{intra}} + L_{\text{inter}} . \tag{15}$$

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 832×256, 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), ORB-SLAM2 [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 t_{rel} is the average translational RMSE drift (%) on length of 100m-800m, and r_{rel} is the average rotational RMSE drift (°/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

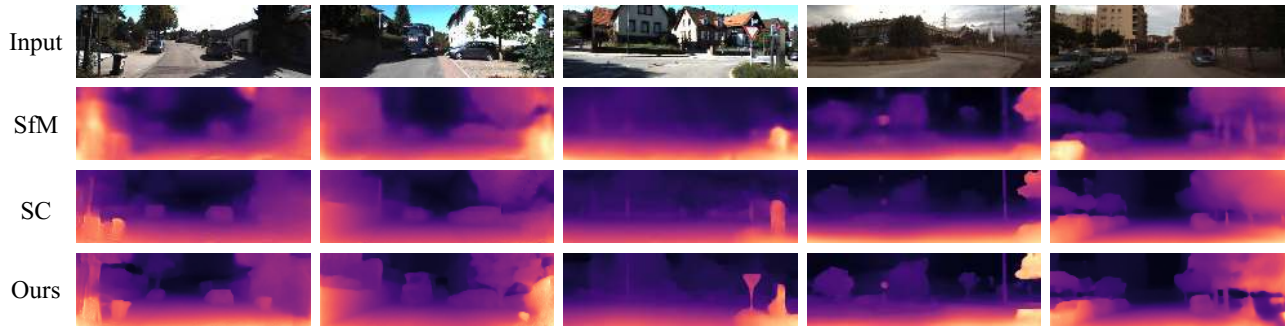


Figure 4: Qualitative comparison of depth estimation among SfMLearner, SC, and the proposed UnVIO on KITTI and Malaga dataset. It is clear that the proposed method predicts depth maps with more details and sharper edges compared with competitors.

Seq	M	Mis-c: 10°		Unsyn: 20ms		IMU-D: 30%		Cam-D: 30%	
		t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}
09	Ours	4.13	0.89	4.60	0.93	16.05	5.01	6.54	1.48
	VINS	34.53	3.60	28.06	2.73	29.15	4.08	30.22	2.87
10	Ours	5.51	0.53	5.10	0.63	8.33	2.09	9.02	1.63
	VINS	27.76	2.41	22.13	3.50	28.31	3.60	19.31	2.65

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).

Method	Error metric			Accuracy metric (δ)		
	Abs Rel	Sq Rel	RMSE	< 1.25	$< 1.25^2$	$< 1.25^3$
SfM	0.3272	3.1131	9.5216	0.4232	0.7010	0.8476
SC	0.1629	0.9644	4.9129	0.7760	0.9315	0.9773
Ours	**0.1322**	**0.73005**	**4.2443**	**0.8324**	**0.9509**	**0.9821**

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**.

Method	IMU	SW	Fusion	Seq 09		Seq 10		Avg	
				t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}
Ours				10.35	3.67	10.58	6.26	10.46	4.96
Ours	✓			5.63	1.10	6.39	0.88	6.01	0.99
Ours	✓	✓		5.36	1.19	5.74	0.54	5.55	0.87
Ours	✓	✓	✓	4.13	0.89	5.51	0.53	4.82	0.71

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

sequences in column Avg(sub-t). The results of Avg(sub-t) show that our method outperforms other self-supervised VIO methods in both t_{rel} and r_{rel} metrics, although without using extra depth data. The average results of all the training set except Seq 03 whose IMU data is not available are used for the complete comparison on the training set (see column Avg(train)). Our method clearly performs better in t_{rel} than the unsupervised VO methods and traditional methods that may hold larger accumulated errors. Additionally, average errors on the test set (see column Avg(test)) are provided. It can be observed that the proposed method significantly improves the translational performance compared with the unsupervised VO methods and traditional methods on unseen scenes, validating the superiority of UnVIO. Compared with self-supervised VIO methods, UnVIO also achieves competitive results with lower rotational error r_{rel} on the test set. The reason that other self-supervised VIO methods gain better t_{rel} may be that the extra depth data can provide more determinate information for geometric inference. Besides, our vision-only method, i.e, Ours (NoIMU), performs better than other unsupervised VO methods, which indicates the predominance of our framework when implemented without IMU data.

Fig.3 illustrates the trajectories generated by various methods on KITTI and Malaga dataset. The proposed UnVIO can predict more accurate trajectories than other learning-based methods on KITTI and is superior to traditional monocular VO methods. They may hold large drift due to scale ambiguity and accumulated error. It is obvious that our method outperforms reference methods on Malaga dataset, where VINS and OKVIS are likely to fail at the begining part of the trajectories because of long-time initialization.

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 VO 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

[Bian *et al.*, 2019] J. Bian, Z. Li, N. Wang, H. Zhan, C. Shen, M. Cheng, and I. Reid. Unsupervised scale-consistent depth and ego-motion learning from monocular video. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 35–45, 2019.

[Blanco-Claraco *et al.*, 2014] J. L. Blanco-Claraco, F. Á. Moreno-Dueñas, and J. González-Jiménez. The málaga urban dataset: High-rate stereo and lidar in a realistic urban scenario. *The International Journal of Robotics Research (IJRR)*, 33(2):207–214, 2014.

[Bloesch *et al.*, 2015] M. Bloesch, S. Omari, M. Hutter, and R. Siegwart. Robust visual inertial odometry using a direct ekf-based approach. In *IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 298–304, 2015.

[Chang and Chen, 2018] J. Chang and Y. Chen. Pyramid stereo matching network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5410–5418, 2018.

[Chen *et al.*, 2019] C. Chen, S. Rosa, Y. Miao, C. X. Lu, W. Wu, A. Markham, and N. Trigoni. Selective sensor fusion for neural visual-inertial odometry. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10542–10551, 2019.

[Clark *et al.*, 2017] R. Clark, S. Wang, H. Wen, A. Markham, and N. Trigoni. Vinet: Visual-inertial odometry as a sequence-to-sequence learning problem. In *Thirty-First AAAI Conference on Artificial Intelligence (AAAI)*, pages 3995–4001, 2017.

[Dosovitskiy *et al.*, 2015] A. Dosovitskiy, P. Fischer, E. Ilg, P. Husser, C. Hazirbas, V. Golkov, P. v. d. Smagt, D. Cremers, and T. Brox. Flownet: Learning optical flow with convolutional networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2758–2766, 2015.

[Engel *et al.*, 2017] J. Engel, V. Koltun, and D. Cremers. Direct sparse odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(3):611–625, 2017.

[Geiger *et al.*, 2011] A. Geiger, J. Ziegler, and C. Stiller. Stereoscan: Dense 3d reconstruction in real-time. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 963–968, 2011.

[Geiger *et al.*, 2012] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361, 2012.

[Geiger *et al.*, 2013] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research (IJRR)*, 32(11):1231–1237, 2013.

[Godard *et al.*, 2017] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 270–279, 2017.

[Han *et al.*, 2019] L. Han, Y. Lin, G. Du, and S. Lian. Deepvio: Self-supervised deep learning of monocular visual inertial odometry using 3d geometric constraints. *arXiv preprint arXiv:1906.11435*, 2019.

[He *et al.*, 2016] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 770–778, 2016.

[Hochreiter and Schmidhuber, 1997] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

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

visual-inertial SLAM. In *International Conference on Robotics and Automation (ICRA)*, pages 5182–5189, 2018.

[Leutenegger *et al.*, 2013] S. Leutenegger, P. T. Furgale, V. Rabaud, M. Chli, K. Konolige, and R. Siegwart. Keyframe-based visual-inertial slam using nonlinear optimization. In *Proceedings of Robotis Science and Systems (RSS)*, 2013.

[Mur-Artal and Tardós, 2017] R. Mur-Artal and J. D. Tardós. ORB-SLAM2: an open-source SLAM system for monocular, stereo and RGB-D cameras. *IEEE Transactions on Robotics (TRO)*, 33(5):1255–1262, 2017.

[Qin *et al.*, 2018] T. Qin, P. Li, and S. Shen. Vins-mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Transactions on Robotics (TRO)*, 34(4):1004–1020, 2018.

[Shamwell *et al.*, 2018] E. J. Shamwell, S. Leung, and W. D. Nothwang. Vision-aided absolute trajectory estimation using an unsupervised deep network with online error correction. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2524–2531, 2018.

[Shen *et al.*, 2019] T. Shen, Z. Luo, L. Zhou, H. Deng, R. Zhang, T. Fang, and L. Quan. Beyond photometric loss for self-supervised ego-motion estimation. In *International Conference on Robotics and Automation (ICRA)*, pages 6359–6365, 2019.

[Vankadari *et al.*, 2019] M. B. Vankadari, S. Kumar, A. Majumder, and K. Das. Unsupervised learning of monocular depth and ego-motion using conditional patchgans. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5677–5684, 2019.

[Wang *et al.*, 2004] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing (TIP)*, 13(4):600–612, 2004.

[Wang *et al.*, 2018] S. Wang, R. Clark, H. Wen, and N. Trigoni. End-to-end, sequence-to-sequence probabilistic visual odometry through deep neural networks. *The International Journal of Robotics Research (IJRR)*, 37(4-5):513–542, 2018.

[Zhou *et al.*, 2017] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1851–1858, 2017.