Metric Monocular Localization Using Signed Distance Fields

Huaiyang Huang[1], Yuxiang Sun[1], Haoyang Ye[1] and Ming Liu[1], *Senior Member*, *IEEE*

Abstract— Metric localization plays a critical role in vision-based navigation. For overcoming the degradation of matching photometry under appearance changes, recent research resorted to introducing geometry constraints of the prior scene structure. In this paper, we present a metric localization method for the monocular camera, using the Signed Distance Field (SDF) as a global map representation. Leveraging the volumetric distance information from SDFs, we aim to relax the assumption of an accurate structure from the local Bundle Adjustment (BA) in previous methods. By tightly coupling the distance factor with temporal visual constraints, our system corrects the odometry drift and jointly optimizes global camera poses with the local structure. We validate the proposed approach on both indoor and outdoor public datasets. Compared to the state-of-the-art methods, it achieves a comparable performance with a minimal sensor configuration.

I. INTRODUCTION

Metric visual localization is a crucial building block for applications varying from surveying to autonomous navigation. It provides location information with the absolute scale for high-level applications [1], [2]. With the advancement of Simultaneous Localization and Mapping (SLAM) and Structure-from-Motion (SFM) techniques [3]–[5], localizable maps with precise structure can be built off-the-shelf, which provide fundamental prior information for global localization. However, how to localize a monocular camera metrically against a prior map with reliability remains an unsolved problem [6].

Several existing methods [7]–[9] exhibited the capability to localize camera with photometric information from Li-DAR intensity, dense visual reconstruction or sparse feature maps. However, these approaches were subject to varying illumination and appearance changes. Another solution is to match geometry [10]–[12]. Under the assumption of an accurate local reconstruction from the Bundle Adjustment (BA) or stereo matching, geometry-based or hybrid approaches generally, regard visual localization as a rigid alignment problem. For recovering the metric scale, these methods require additional sensors that provide absolute scale information are required (e.g., depth, stereo, or inertial sensors) [13]. Besides, the pre-built scene geometry can provide nontrivial prior constraints for estimating the local structure, based on which we propose to perform the motion and structure optimization in a tightly-coupled manner.

Different from previous methods, we leverage the Signed Distance Field (SDF) as a global map for an accurate and

[1]The authors are with the Robotics and Multi-Perception Laborotary, Robotics Institute, The Hong Kong University of Science and Technology. (email: hhuangat@connect.ust.hk; eeyxsun@ust.hk, sun.yuxiang@outlook.com; hy.ye@connect.ust.hk; eelium@ust.hk).

(a) An example image from EuRoC dataset.
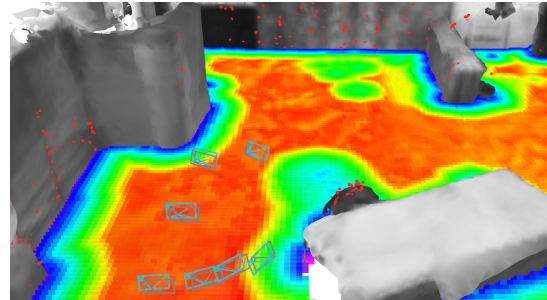
(b) Visualization of an experimental result.

Fig. 1: An example localization result (b) in an indoor environment (a). With a prior SDF (3D voxel grid), the proposed method refines the local structure (red points) and well aligns it with the scene structure (gray mesh). The camera poses (blue frustums) are estimated accurately, compared to the groudtruth (gray frustums).

robust metric localization. Several works [14]–[16] have proved that the SDF is a generic map representation for dense reconstruction, as well as for trajectory planning and pose estimation. As a SDF models an implicit scene structure, it provides direct measurements of distance along with gradients in 3D Euclidean space. Based on these properties, our system corrects the drift of visual odometry, retrieve depth prior for the landmark generation, and better optimize the local structure. Fig. 1 shows an indoor scene from EuRoC dataset [17] and the corresponding localization result. To this end, we summarize our contributions as follows:

1) A hierarchical monocular localization method using the SDF as a generic map representation.
2) A tightly-coupled approach to optimize the local structure and global poses simultaneously.
3) Various experiments on public datasets that demonstrate the effectiveness of the proposed method and the superiority over the state-of-the-art methods.

II. RELATED WORK

A. Metric Visual Localization Against Prior Maps

A general framework of monocular visual localization is to establish a sparse feature map with local descriptors such as SIFT [18]. With the 2D-3D correspondences matching, metric poses can be estimated via a Perspective-n-Point (PnP) scheme. For instance, Lynen *et al* [8]. proposed a lightweight visual-inertial localization system for mobile devices, which demonstrated the ability of large-scale localization. In contrast to an abstract representation of sparse features, a dense reconstruction with intensity makes the direct photometric matching possible. [7], [9], [19] introduce Normalized Mutual Information (NMI) or its variant Normalized Information Distance (NID) to align a query image with the prior dense reconstruction with intensity. Wolcott *et al.* [19] proposed to search 3-DoF candidate pose hypotheses and then match their NMI with a synthetic intensity image. Stewart *et al.* [9] and Pascoe *et al.* [7] utilized a closed-form formulation to avoid the exhaustive search and scoring and estimate the 6-DoF camera pose generally. Although the NID-based similarity estimation is effective for photometric invariant localization, these methods still fail to converge in some cases due to highly non-uniform illumination.

To overcome the challenges from appearance variances, several approaches were proposed for a geometry-based or hybrid solution. In [11], Foster *et al.* presented a Monte Carlo Localization (MCL) method from correlating the height-maps originated from the alignment of the local structure with depth information. Similarly, Caselitz *et al.* [10] proposed to match the local reconstruction from visual odometry with the prior pointcloud through estimating a similarity transform. Kim *et al.* [12] proposed to directly minimize the depth residual between the stereo disparity map and depth map projected from 3D LiDAR information. For these methods, stereo or depth sensors are required to acquire depth information [11], [12], or manually initialized the scale from the groundtruth [10]. Furthermore, while a prior map is supposed to contribute to the structure optimization in local BA, current works are generally based on the assumption that the local structure is accurate. We argue that it is beneficial to introduce such factors to optimization procedures, including both the local structure and global camera poses, for a robust and drift-free visual localization.

B. SDFs for Visual Navigation

SDF was firstly introduced in computer vision as a map representation for dense reconstructions [20]. For instance, KinectFusion [21] is an accurate RGB-D reconstruction framework modelling the surroundings with volumetric Truncated Signed Distance Field (TSDF). Voxblox [22] managed to build Euclidean Signed Distance Fields (ESDF) with heterogeneous sensor types for Micro Aerial Vehicles (MAVs). As SDFs provide metric information of static obstacles, they have long been used for planning. Chomp [15] was one of the state-of-the-art gradient-based planning algorithms using SDFs.
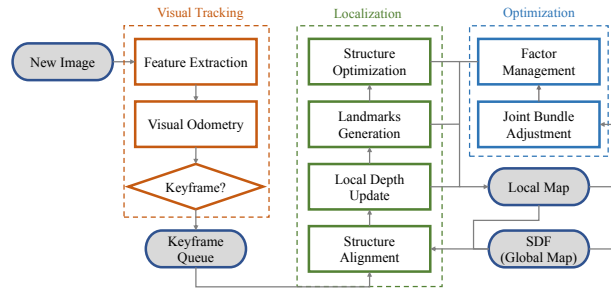

Fig. 2: Framework of the proposed system.

Besides, several methods attempted to exploit SDFs for camera pose estimations [11], [14], [16]. They either aligned the structure with SDF, or resorted to SDF for local reconstructions. Bylow *et al.* [16] proposed to a RGB-D Simultaneous Localization and Mapping (SLAM) framework. This method directly tracked camera poses by minimizing a cost function associated with the current TSDF reconstruction. Foster *et al.* [11] used the SDF as a penalty function for a better multi-view reconstruction, which served as the first step for camera localization. Existing works generally assumed an accurate structure and then tracked the camera pose by rigid structure alignment. On the contrary, we consider a tightly-coupled monocular localization problem with the SDF.

III. OVERVIEW

Fig. 2 shows the diagram of the proposed system. The proposed localization system is implemented with a *track-then-localize* scheme. Firstly, the pose of the current camera is tracked by the visual odometry front-end. With an initial local structure, we use a visual odometry framework to accomplish frame-to-frame and frame-to-model tracking. Motion and structure BA is then performed with SDF constraints to localize the camera and align the scene geometry. In the meantime, the back-end jointly optimizes the global camera poses along with the local structure.

The key idea of our method is to optimize the local structure and global poses simultaneously, utilizing the volumetric SDF as an global constraint. The SDF, denoted as $\phi : \mathbb{R}^3 \to \mathbb{R}$, implies an isosurface $\phi(\mathbf{p}) = 0$ representing the pre-built world model. For any point $\mathbf{p}$ in the 3D Euclidean space, SDF provides a distance measurement $\phi(\mathbf{p})$ to the nearest isosurface, along with a smooth gradient $\nabla\phi(\mathbf{p})$. Based on these properties, the structure could be aligned and refined with this implicit model for global localization.

Throughout this paper, we denote camera poses as matrices of the special Euclidean group $\mathbf{T}_k \in \mathbf{SE}(3)$, which transform a 3D point $\mathbf{p}_i$ from the world coordinate to the camera coordinate by $\mathbf{p}_i^c = \mathbf{T}_k \cdot \mathbf{p}_i$. The update of a camera pose with an incremental twist $\boldsymbol{\xi} \in \mathfrak{se}(3)$ using a left-multiplicative formulation $\oplus : \mathfrak{se}(3) \times \mathbf{SE}(3) \to \mathbf{SE}(3)$, denoted as:

$$\boldsymbol{\xi} \oplus \mathbf{T}_k := \exp(\boldsymbol{\xi}^\wedge) \cdot \mathbf{T}_k. \tag{1}$$

The corresponding pixel of a 3D point $\mathbf{p}_i$ observed in the k-th image, is denoted as $\mathbf{u}_{i,k} = (u, v)^T \in \mathcal{I}$. And we use
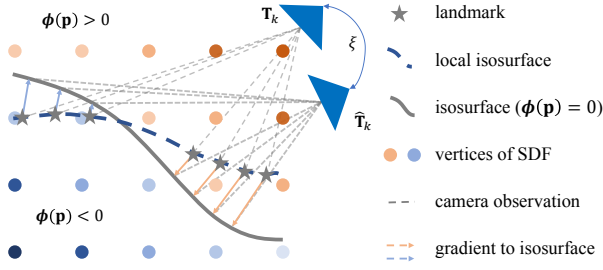
Fig. 3: Illustration for pose refinement. The implicit structure is inferred from local sparse features. With the smooth gradient at each landmark pointing to global isosurface, we align the local structure with map and the camera pose is corrected seamlessly.

$\pi : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ to denote the camera projection model, which transform a point to the image coordinate by:

$$\mathbf{u_{i,k}} = \pi(\mathbf{p_i^c}). \tag{2}$$

IV. METHODOLOGY

A. Visual Tracking

To demonstrate the capability of tightly coupling with existing visual odometry algorithms, here we adopt built-in modules of ORB SLAM2 [23] for a tracking front-end. The tracking module extracts FAST corners and computes their ORB descriptors [24], which are tracked with the previous frame via matching the descriptors in a local window. The camera pose is then estimated via a PnP scheme. After the visual tracking, all the essential information is delivered to the following modules for localization, including the features extracted, the pose of the current frame and 3D correspondences that are tracked successfully.

We separate the local structure into two sets according to their observation sources. We denote $\mathcal{P} = \{\mathbf{p}_i\}$ with $\mathbf{p}_i \in \mathbb{R}^3$ as the set of all 3D features belonging to the local structure. We divide $\mathcal{P}$ into two separate sets as: $\mathcal{P} = \mathcal{M} \cup \mathcal{N}$, where $\mathcal{M}$ represents for points has a constraint with the pre-built map and $\mathcal{N}$ represents for points only constrained by multi-view stereo.

B. Pose Refinement

To eliminate the inevitable drift of visual odometry over time, we refine the pose of current keyframe with map constraints. Fig. 3 illustrates this localization process. Similar to [16], for every $\mathbf{p}_i \in \mathcal{M}_k$, we define an energy function measures the fitness of local structure with the map, which is given by:

$$e_i^{sdf} = \phi\left(\mathbf{p}'\right) = \phi\left(\left(\boldsymbol{\xi} \oplus \mathbf{T}_k\right)\mathbf{p}_i\right), \tag{3}$$

where $\mathbf{p}'$ is the position of $\mathbf{p}$ in camera coordinate. $\mathcal{M}_k \subset \mathcal{M}$ is the features tracked in the current frame with SDF constraints. $\mathbf{T}_k \in \mathbf{SE}(3)$ is initially estimated by the tracking module. As the optimal solution is generally not far from the initial guess, we use an incremental update policy $\hat{\mathbf{T}}_k \leftarrow \boldsymbol{\xi} \oplus \mathbf{T}_k$ to correct the trivial drift. The update is parametrized
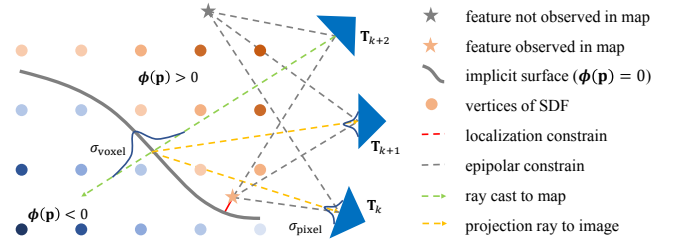

Fig. 4: Illustration for landmark generation and depth update. If we find zero-crossing voxel (green ray), the feature is matched via epipolar-line search (orange ray) and the depth will be initialized according to the variances (blue curves). If not, we triangulate the 3D position (gray feature) by temporal matching only (gray rays). The structure is then optimized combining visual (gray dashed line) and SDF factors (red line).

as a twist $\boldsymbol{\xi} \in \mathfrak{se}(3)$ and is set to be zero at the start of the pose optimization.

The optimal pose is solved via minimizing the full energy function over all the points constrained by the SDF, formulated as:

$$\hat{\boldsymbol{\xi}} = \arg\min_{\boldsymbol{\xi}} \sum_{\mathbf{p}_i \in \mathcal{M}_k} \|\phi\left(\left(\boldsymbol{\xi} \oplus \mathbf{T}_k\right)\mathbf{p}_i\right)\|_\gamma, \tag{4}$$

where $\|\|_\gamma$ denotes robust Huber norm for convergence regardless of outliers. And the twist is updated by:

$$\boldsymbol{\xi} \leftarrow \log\left(\delta\boldsymbol{\xi} \oplus \mathbf{T}(\boldsymbol{\xi})\right). \tag{5}$$

We use Levenberg-Marquardt algorithm to solve $\delta\boldsymbol{\xi}$ iteratively. The system is represented as:

$$\delta\boldsymbol{\xi} = \mathbf{H}^{-1}\mathbf{b}, \tag{6}$$

$$\mathbf{H} = \mathbf{J}^T\boldsymbol{\Omega}^{-1}\mathbf{J} + \beta\mathbf{I} \quad \text{and} \quad \mathbf{b} = -\mathbf{J}^T\boldsymbol{\Omega}^{-1}\mathbf{r}, \tag{7}$$

where the $\boldsymbol{\Omega}$ is the covariance matrix calculated from the pre-defined weight, described in Section. IV-C. And

$$\begin{aligned} \mathbf{J}_i &= \mathbf{J}_{\text{sdf}} \cdot \mathbf{J}_{\text{pose}} \\ &= \frac{\partial\phi\left(\mathbf{p}'\right)}{\partial\mathbf{p}'} \cdot \frac{\partial\left(\left(\boldsymbol{\xi} \oplus \mathbf{T}_k\right)\mathbf{p}_i\right)}{\partial\boldsymbol{\xi}}, \end{aligned} \tag{8}$$

where $\mathbf{J}_{\text{sdf}} = \nabla\phi\left(\mathbf{p}'\right)$ is the three-dimensional gradient indicating the direction of the nearest isosurface at global position $\mathbf{p}'$. We use trilinear interpolation between neighbourhood vertices to calculated an approximate gradient for each $\mathbf{p}'$ at every optimization step.

C. Landmark Generation and Depth Update

New landmarks are generated in two steps. The first step is to lookup the zero-crossing voxel along the optical ray. For an efficient query, we define a visual frustum according to camera intrinsics. The voxels are stored in a hash-table to support fast loopup [22].

As shown in Fig. 4, if the ray is cast successfully, the depth $d_\mathbf{p}$ of current feature is computed by trilinear interpolation and is assumed with a variance of one pixel. By searching

along the epipolar line, temporal matching is then performed with all the keyframes that have overlap in camera view. If ray-casting fails, we directly perform temporal feature matching across frames and compute landmark positions via mid-point triangulation. Landmarks will be inserted to $\mathcal{M}$ if it could be observed from the SDF and $\mathcal{N}$ if not.

On the contrary to methods based on rigid registration, the local structure will be refined to better fit the epipolar constraints and the implicit global structure simultaneously, through minimizing the energy function in Eq. 9. To balance the relative weights of temporal multi-view stereo and the SDF, a coupling factor λ is introduced, thus the energy function is given as:

$$E = E_{\text{repro}} + \lambda \cdot E_{\text{sdf}}, \tag{9}$$

where

$$E_{\text{sdf}} = \sum_{\mathbf{p}_i \in \mathcal{M}} e_i^{\text{sdf}}(\mathbf{p}_i) = \sum_{\mathbf{p}_i \in \mathcal{M}} w_i^{\text{sdf}} \|\phi(\mathbf{p}_i)\|_\gamma \tag{10}$$

is the energy function over the distance penalty of all the landmarks observed in the map and

$$
\begin{aligned}
E_{\text{repro}} &= \sum_{\mathbf{p}_i \in \mathcal{P}} \sum_{j \in \mathcal{F}_i} e_{ij}^{\text{repro}}(\mathbf{p}_i, \mathbf{T}_j) \\
&= \sum_{\mathbf{p}_i \in \mathcal{P}} \sum_{j \in \mathcal{F}_i} w_{i,j}^{\text{repro}} \|\mathbf{u}_{i,j} - \pi(\mathbf{T}_j \mathbf{p}_i)\|_\gamma
\end{aligned} \tag{11}
$$

defines the visual constraints generated by temporal multi-view stereo. $\mathcal{F}_i$ is the set of keyframes in which $\mathbf{p}_i$ could be observed. $w_{i,j}^{\text{repro}}$ is dependent on the location variance of feature on the image plane, which is set to be one-pixel at each level of the image pyramid. w_i^{sdf} is associated with the uncertainty of the prior volumetric map and therefore we define it as $w_{\text{sdf}} = 1/\sigma_{\text{sdf}}^2$. This weighting in the meantime normalize the reprojection error and distance error, and thus λ becomes the only factor that makes a balance between both factors. We perform an optimization procedure similar to Section. IV-B to achieve the optimal local structure $\hat{\mathcal{P}}$.

D. Joint Optimization

For an accurate and drift-free localization, we use a stand-alone backend to jointly optimize the global poses and the local structure. The overall energy function is the same as Eq. 9 with a different set of optimization parameters, denoted as $\mathcal{X} = (\mathcal{P}, \mathcal{T})$, where $\mathcal{T}$ consists of the global poses for all the keyframes in the current optimization window. The incremental update is computed similar to Eq. 6 and Eq. 7. For simplification we assume the k-th residual term is relevant to i-th feature point $\mathbf{p}_i$ and j-th keyframe $\mathbf{T}_j$, thus the associated row $\mathbf{J}_k$ of Jacobian $\mathbf{J}$ defined as:

$$\mathbf{J}_k = [\lambda \mathbf{J}_{\text{sdf}} + \mathbf{J}_{\text{repro}}, \mathbf{J}_{\text{pose}}], \tag{12}$$

where $\mathbf{J}_{\text{sdf}}$ is the same in Eq. 8 and $\mathbf{J}_{\text{repro}} = \partial e_{ij}^{\text{repro}}/\partial \mathbf{p}$, $\mathbf{J}_{\text{pose}} = \partial e_{ij}^{\text{repro}}/\partial \mathbf{T}_j$ respectively. The optimal parameters are represented as:

$$\hat{\mathcal{X}} = \arg\min_{\mathcal{X}} E_{\text{repro}} + \lambda \cdot E_{\text{sdf}}. \tag{13}$$

A two-step optimization is performed for better filter out the outliers. In the first round, the system will converge regardless of outliers. Then outliers are detected and rejected according to Chi-Squared test. Ignoring the outliers, we perform another round of optimization. Finally, the outliers with large residuals will be determined and dismissed.

Due to the artifacts in pre-built map, non-converged points might be an inlier but is not able to fit the global structure. Therefore, we keep a tolerant occlusion strategy for outliers. If $w_i^{\text{sdf}} \|e_i^{\text{sdf}}(\mathbf{p}_i)\|_\gamma > th_{\text{sdf}}$, instead of classifying $\mathbf{p}_i$ as an outlier, we remove the factor constrained by SDF in the optimization. If in the next optimization step, the related visual factor $w_{i,j}^{\text{repro}} \|e_{i,j}^{\text{repro}}(\mathbf{p}_i, \mathbf{T}_j)\|_\gamma > th_{\text{repro}}$ for any $j \in \mathcal{F}_i$, $\mathbf{p}_i$ is considered as an outlier and all the relevant observation and factors will be removed from the system. Similar to [23], we set the outlier rejection thresholds via $\mathcal{X}^2$-test, with the assumption of 95% inliers and 1-pixel or 1-volume variance for visual factor and SDF factor respectively. Accordingly, we set $th_{\text{sdf}} = 3.841$ and $th_{\text{repro}} = 5.991$.

V. EXPERIMENTAL RESULTS AND DISCUSSIONS

We evaluated our algorithm on public datasets EuRoC [17] and KITTI Odometry [25]. We used Voxblox [22] for the volumetric SDF reconstruction. For EuRoC dataset, the global maps were established from dense stereo depth map and ground truth poses, with the voxel size $\Delta_{\text{voxel}} = 0.1m$. For KITTI odometry, along with ground truth poses, Velodyne HDL-64E LiDAR scans within a radius of $25m$ were extracted for SDF reconstruction. As we did not leverage complex dense stereo reconstruction techniques for a consistent and high-accuracy map, some artifacts were manually corrected. For all the sequences, variance of SDF σ_{sdf} was set to be identical as Δ_{voxel}. The coupling factor λ was tuned manually dependent on the scene depth and camera fps. For indoor sequences of EuRoC, we set $\lambda = 1.0$. For a large-scale outdoor environment with relatively low camera frequency such as KITTI, λ was set to 10.0 to emphasize localization constraints.

We first compared the general localization performance against the state-of-the-art methods on both datasets to validate the effectiveness of the proposed method. To further demonstrate the capability of refining the local structure and initialization with a coarse initial guess, we then evaluated the accuracy of the sparse feature map and the consistency in localization performance against different initial guesses.

A. Localization Results

1) The EuRoc MAV Dataset: The EuRoC MAV Dataset provides sequences of stereo images and IMU data streams in three different indoor scenes, with extrinsic information, ground truth poses and structures. To compare our algorithm against the state-of-the-art methods, for each stage we built a localizable map on one sequence, and then evaluated the localization accuracy on the other. Tab. I shows the localization results on the three scenes.

We compare the Root Mean Square Error (RMSE) of Absolute Trajectory Error (ATE) for the translation against

TABLE I: Comparison of Translation ATE (m) on EuRoC. The proposed method is compared against state-of-the-art localizers (marked as *).

	MH01 (Map: MH02)	V201 (Map: V202)	V101 (Map: V102)
Ours*	0.032	0.041	0.034
ROVIOLI*	0.082	0.057	0.110
ORB Loc*	0.464	X	X
VINS-Mono	0.12	0.081	0.068
ORB SLAM2	0.037	0.035	0.035
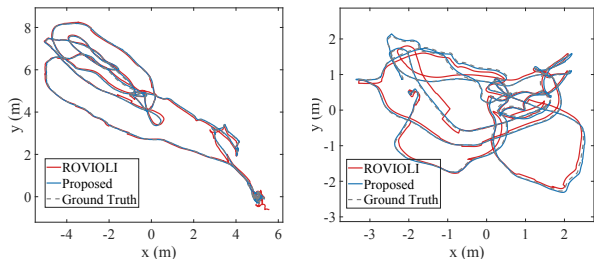
Fig. 5: Trajectory comparison on EuRoC. Left: MH01; Right: V101.

ROVIOLI [26] and ORB SLAM2 in localization mode (denoted as ORB Loc). Due to illumination and viewpoint variances, ORB Loc fails on two of the sequences (marked as X). Along with the comparison of global localization, we list results of state-of-the-art visula-inertial, VINS-Mono [27], and stereo SLAM , ORB_SLAM2 [23], to prove that our method improves the accuracy of local monocular odometry.

For all the three sequences, the localization error of the proposed method is better than ROVIOLI, indicating a more accurate global pose estimation, with an average RMSE lower than 0.04m. In addition, our method achieves comparable pose estimation results against VINS-Mono and ORB SLAM2. This comparison indicates that our approach can automatically recover the scale powered by combining the SDF cost with the local reprojection error.

2) The KITTI Odometry Dataset: We validated our algorithm on the KITTI odometry dataset. For sequences containing loopy trajectories, the groundtruth poses lack such loop-closure consistency. Therefore, we only estimated the results on the loop-less sequences. We compared the average translation and rotation error along with the variances against a recently published stereo localizer [12] and stereo SLAM [23].

The results are shown in Tab. II. The translational error comparison demonstrates a comparable localization accuracy of the proposed method against the stereo localizer, while only using temporal monocular information to localize camera in the map. The success tracking on KITTI 01, a highway sequence, indicates the robustness of our system. On the other hand, the rotational estimations from our method outperform the stereo localizer. This is because our
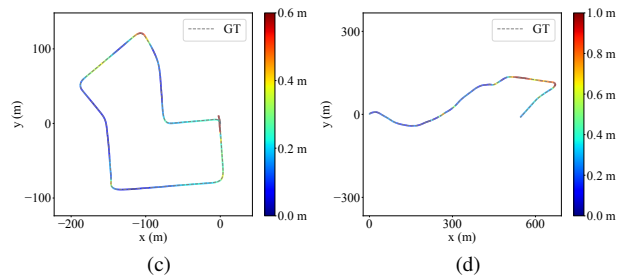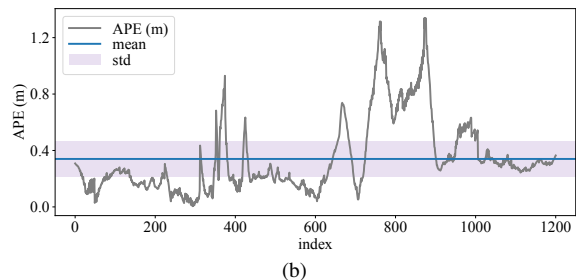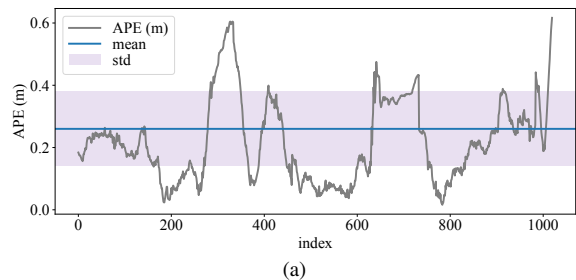

Fig. 6: Localization error on KITTI. (a) and (b) are Absolute Pose Errors (APEs) for KITTI07 and KITTI10 respectively. (c) and (d) are estimated trajectories on KITTI07 and KITTI10, coded by ATEs.

system better binds the structure from the pre-built map. As we take advantage of temporal multi-view constraints, the variances of localization error are generally much lower than the stereo localizer.

Additionally, with the comparison of ORB SLAM2 (stereo), improvement of the proposed system to local odometry is noticed. For sequence 04, both localization methods fail to outperform ORB SLAM2 under a structure-less rural environment. This indicates that approaches emphasizing geometry-based matching require diversity of the structure.

Fig. 6 illustrates the translation error and trajectory comparison of our approach on KITTI07 and KITTI10. For both sequences, the results exhibit the capability of our method to correct the odometry drift and recover the real scale from the map. For KITTI10 covering a path over 1 km, the localization error is always lower than 1.0 m while the average error is 0.34 m. Large localization error mainly occurrs in structure-less regions, or when the camera is rotated with large angular velocity, for that our algorithms require a structural scene to correct the drift and recover scale for monocular odometry.

TABLE II: Comparison of ATE on KITTI. The ATE of each sequence is shown as $mean \pm std$.

Seq #	Environments	Ours		Stereo Localizer [12]		ORB SLAM2 (Stereo)	
		Translation (m)	*Rotation* (deg)	*Translation* (m)	*Rotation* (deg)	*Translation* (m)	*Rotation* (deg)
01	Highway	**5.66 ± 3.24**	1.55 ± **0.49**	X		10.54 ± 3.37	**1.44** ± 0.59
03	Country	1.59 ± 0.92	0.97 ± **0.16**	**0.24 ± 0.23**	**0.41** ± 0.33	0.59 ± 0.35	0.60 ± 0.17
04	Country	0.42 ± 0.27	**0.81 ± 0.07**	0.45 ± 0.35	0.88 ± 0.39	**0.17 ± 0.07**	0.99 ± **0.07**
06	Urban	**0.22 ± 0.12**	**0.32** ± 0.39	0.38 ± 0.67	0.85 ± 1.80	0.55 ± 0.19	0.40 ± **0.09**
07	Urban	0.28 ± 0.26	**0.18 ± 0.08**	**0.13 ± 0.12**	0.49 ± 0.47	0.40 ± 0.15	0.39 ± 0.18
10	Urban+Country	0.34 ± **0.26**	**0.36 ± 0.25**	**0.24** ± 0.68	0.49 ± 0.77	0.90 ± 0.40	0.90 ± 0.36

TABLE III: Comparison of sparse reconstruction RMSE (m) for the proposed method, ORB SLAM (mono) and ORB SLAM (stereo).

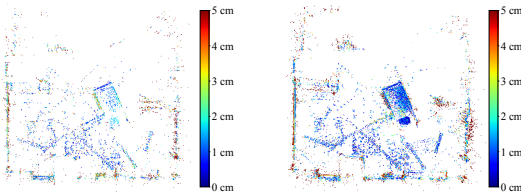|	Ours	ORB (mono)	ORB (stereo)
V101	0.0609	0.2177	**0.0561**
V201	**0.0836**	0.1176	0.1008

Fig. 7: The sparse reconstruction on sequence V201 colored by metric error. The left is from ORB SLAM2 (stereo) and the right is from the proposed method. A comparable local structure to stereo SLAM boosts both tracking and localization, explaining the localization performance of our system.

B. Local Structure Refinement

To demonstrate the capability of refining local structure, we evaluated the accuracy of the sparse reconstruction on EuRoC Vicon Room sequences. These sequences provide a highly-accurate model from Leica Scan that could serve as ground truth. The sparse feature map reconstructed were saved and then aligned with the global model. For monocular SLAM, the structure was aligned with scale correction. We searched the nearest neighbourhood for each point in the sparse reconstruction and used Euclidean distance between them as the definition of metric error. By calculating RMSE between sparse features and the ground truth, we compare the local reconstruction accuracy against both monocular SLAM and stereo SLAM. The quantitative results are reported in Tab. III. For all the sequences, monocular SLAM system estimates a less accurate structure which in turn leads to less accurate local localization performance. On the contrary, our system achieves the lowest average RMSE below 0.1m, explaining the comparable localization accuracy against the stereo SLAM.

Visualization of structure error is shown in Fig. 7. Since our strategy includes features with large distance and narrow
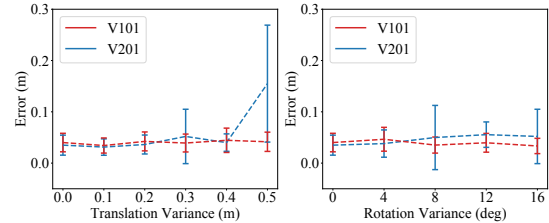

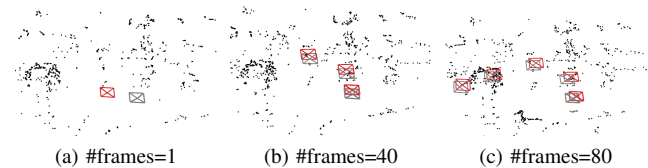Fig. 8: Localization Error with respect to Initialization Variances.

Fig. 9: A sample initialization process with translational error of 0.25m and rotational error of 10 degree.

baseline across frames, the reconstruction from our system is denser and more outliers are noticed. Despite this fact, our system still achieves a comparable structure accuracy.

C. Evaluation of Initialization Robustness

We evaluated the localization error under different initial poses. For EuRoC V101 and V201, we added variance to the initial guess along a specific axis for a consistent evaluation.

The quantitative results is shown in Fig. 8, although the initialization variance increases, our system maintains a consistent accuracy under 0.1m with acceptable localization variances, except when translation error is set to 0.5m. As noticed in Fig. 8, localization variance on V201 is larger than on V101. The explanation is that the initial scene of V101 is a degeneration case, a combination of planes, thus optimization direction is typically fixed regardless of initial guess. Fig. 9 illustrates how the proposed system recovers from a initialization variance. With one single frame, the structure and absolute scale is ambiguous for the monocular image. However, when the scene structure is gradually observed, the local structure is scaled and aligned with the global model, with the depth of features in the first frame being corrected seamlessly. Finally, with enough observations, the system converges to a relatively accurate global pose with fewer outliers in the local structure.

VI. Conclusions

This work presents a hierarchical metric visual localization method based on the SDF-based map representation. With a monocular camera, the proposed system exhibits a comparable performance against the state-of-the-art approaches in terms of both accuracy and robustness. Further study on the local reconstruction accuracy and robustness on initialization proves the capability of our method to refine local structure for a better global pose estimation.

To achieve a more reliable localization system, we will investigate a more robust initialization strategy such as integrating visual SFM and the SDF observation. Another issue of the current SDF-based localization is the large memory usage. We believe a hierarchical map structure is significant for large-scale localization. Additionally, further study on handling strutural changes might improve the long-term robustness.

Acknowledgment

References

[1] Y. Chen, C. Liu, L. Tai, M. Liu, and B. E. Shi, "Utilizing eye gaze to enhance the generalization of imitation network to unseen environments," in *ICML 2019 Workshop on Understanding and Improving Generalization in Deep Learning*, June 2019.

[2] W. Hengli, S. Yuxiang, and L. Ming, "Self-supervised drivable area and road anomaly segmentation using rgb-d data for robotic wheelchairs," in *IEEE Robotics and Automation Letters (RA-L)*. IEEE, 2019.

[3] J. Zhang and S. Singh, "Loam: Lidar odometry and mapping in real-time." in *Robotics: Science and Systems*, vol. 2, 2014, p. 9.

[4] H. Ye, Y. Chen, and M. Liu, "Tightly coupled 3d lidar inertial odometry and mapping," in *2019 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2019.

[5] Y. Yu, W. Gao, S. Shen, and M. Liu, "A gps-aided omnidirectional visual-inertial state estimator in ubiquitous environments," in *2019 IEEE International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019.

[6] Y. Sun, M. Liu, and M. Q.-H. Meng, "Improving rgb-d slam in dynamic environments: A motion removal approach," *Robotics and Autonomous Systems*, vol. 89, pp. 110–122, 2017.

[7] G. Pascoe, W. Maddern, and P. Newman, "Direct visual localisation and calibration for road vehicles in changing city environments," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015, pp. 9–16.

[8] S. Lynen, T. Sattler, M. Bosse, J. A. Hesch, M. Pollefeys, and R. Siegwart, "Get out of my lab: Large-scale, real-time visual-inertial localization." in *Robotics: Science and Systems*, 2015.

[9] A. D. Stewart and P. Newman, "Laps-localisation using appearance of prior structure: 6-dof monocular camera localisation using prior pointclouds," in *2012 IEEE International Conference on Robotics and Automation*. IEEE, 2012, pp. 2625–2632.

[10] T. Caselitz, B. Steder, M. Ruhnke, and W. Burgard, "Monocular camera localization in 3d lidar maps," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 1926–1931.

[11] C. Forster, M. Pizzoli, and D. Scaramuzza, "Air-ground localization and map augmentation using monocular dense reconstruction," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2013, pp. 3971–3978.

[12] J. J. Youngji Kim and A. Kim, "Stereo camera localization in 3D LiDAR maps," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Madrid, Oct. 2018, pp. 1–9.

[13] Y. Sun, M. Liu, and M. Q.-H. Meng, "Motion removal for reliable rgb-d slam in dynamic environments," *Robotics and Autonomous Systems*, vol. 108, pp. 115–128, 2018.

[14] M. Klingensmith, I. Dryanovski, S. Srinivasa, and J. Xiao, "Chisel: Real time large scale 3d reconstruction onboard a mobile device using spatially hashed signed distance fields." in *Robotics: science and systems*, vol. 4, 2015, p. 1.

[15] M. Zucker, N. Ratliff, A. D. Dragan, M. Pivtoraiko, M. Klingensmith, C. M. Dellin, J. A. Bagnell, and S. S. Srinivasa, "Chomp: Covariant hamiltonian optimization for motion planning," *The International Journal of Robotics Research*, vol. 32, no. 9-10, pp. 1164–1193, 2013.

[16] E. Bylow, J. Sturm, C. Kerl, F. Kahl, and D. Cremers, "Real-time camera tracking and 3d reconstruction using signed distance functions." in *Robotics: Science and Systems*, vol. 2, 2013.

[17] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. W. Achtelik, and R. Siegwart, "The euroc micro aerial vehicle datasets," *The International Journal of Robotics Research*, 2016. [Online]. Available: http://ijr.sagepub.com/content/early/2016/01/21/0278364915620033.abstract

[18] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[19] R. W. Wolcott and R. M. Eustice, "Visual localization within lidar maps for automated urban driving," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2014, pp. 176–183.

[20] S. F. F. Gibson, "Using distance maps for accurate surface representation in sampled volumes," in *IEEE Symposium on Volume Visualization (Cat. No. 989EX300)*. IEEE, 1998, pp. 23–30.

[21] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon, "Kinectfusion: Real-time dense surface mapping and tracking," in *2011 IEEE International Symposium on Mixed and Augmented Reality*. IEEE, 2011, pp. 127–136.

[22] H. Oleynikova, Z. Taylor, M. Fehr, R. Siegwart, and J. Nieto, "Voxblox: Incremental 3d euclidean signed distance fields for on-board mav planning," in *2017 Ieee/rsj International Conference on Intelligent Robots and Systems (iros)*. IEEE, 2017, pp. 1366–1373.

[23] 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*, vol. 33, no. 5, pp. 1255–1262, 2017.

[24] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *Computer Vision (ICCV), 2011 IEEE international conference on*. IEEE, 2011, pp. 2564–2571.

[25] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[26] T. Schneider, M. Dymczyk, M. Fehr, K. Egger, S. Lynen, I. Gilitschenski, and R. Siegwart, "maplab: An open framework for research in visual-inertial mapping and localization," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1418–1425, 2018.

[27] T. Qin, P. Li, and S. Shen, "Vins-mono: A robust and versatile monocular visual-inertial state estimator," *IEEE Transactions on Robotics*, vol. 34, no. 4, pp. 1004–1020, 2018.