Scalable Fiducial Tag Localization on a 3D Prior Map via Graph-Theoretic Global Tag-Map Registration

Kenji Koide[1], Shuji Oishi[1], Masashi Yokozuka[1], and Atsuhiko Banno[1]

***Abstract*— This paper presents an accurate and scalable method for fiducial tag localization on a 3D prior environmental map. The proposed method comprises three steps: 1) visual odometry-based landmark SLAM for estimating the relative poses between fiducial tags, 2) geometrical matching-based global tag-map registration via maximum clique finding, and 3) tag pose refinement based on direct camera-map alignment with normalized information distance. Through simulation-based evaluations, the proposed method achieved a 98 % global tag-map registration success rate and an average tag pose estimation accuracy of a few centimeters. Experimental results in a real environment demonstrated that it enables to localize over 110 fiducial tags placed in an environment in 25 minutes for data recording and post-processing.**

I. Introduction

In recent years, map-based visual localization methods have been actively studied and widely used for autonomous navigation systems [1], [2] and user interaction applications (e.g., augmented reality [3]). These methods employing precise 3D maps enable accurate localization and navigation in a large variety of environments with affordable equipment. These visual localization methods are, however, still error-prone in feature-less and dynamic environments, and it is sometimes necessary to rely on visual fiducial tags [4], [5] for initialization and fail-safe. Deploying a number of fiducial tags in the environment and combining them with visual localization methods allows us to build a robust and accurate localization and navigation system [6], [7], [8].

Deploying many fiducial tags on a 3D prior map is, however, sometimes difficult and tedious. Because the tag localization accuracy directly affects the navigation accuracy, we need to precisely determine the tag positions as accurate as possible. Furthermore, we need to place many fiducial tags (several hundreds, possibly) to cover the entire environment. However, fiducial tag positions on a prior map are often measured by hand in many works, which results in large human effort and inaccurate localization.

In this work, we propose an automatic method for fiducial tag localization on a 3D prior map. The proposed method enables to accurately determine the poses of many fiducial tags on a 3D prior map in a short time (e.g., more than 100 tags in less than 25 minutes) as a pre-installation process of automatic navigation and user interaction systems. We consider utilizing the precisely localized fiducial tags makes it easy to build a system based on vision-based localization like that shown in Fig. 1.

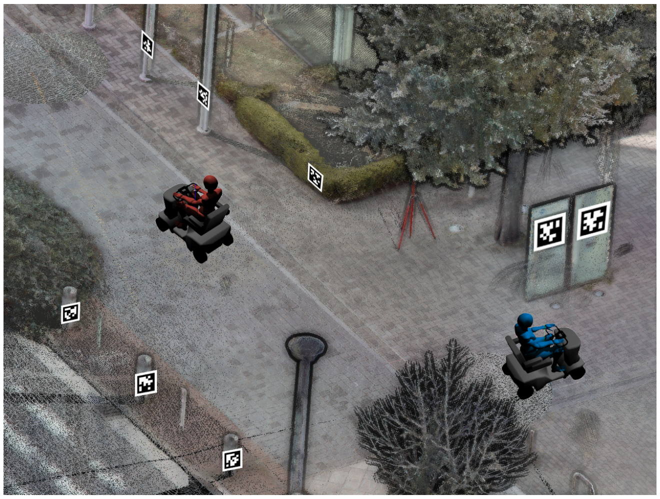


Fig. 1: With the proposed fiducial tag localization method, we aim to make it easy to develop a system based on robust vision-based localization using precisely localized tags on a precise 3D environmental map.

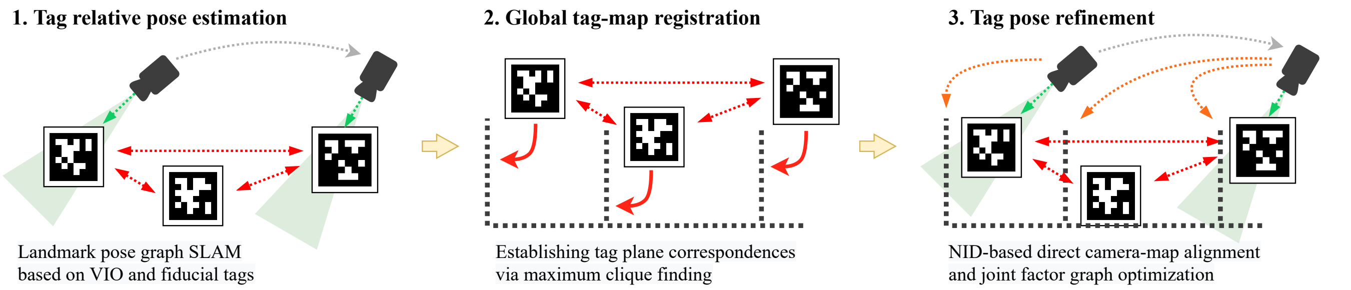
The proposed method comprises three steps: 1) We first estimate the relative poses between fiducial tags via visual inertial odometry (VIO)-based landmark pose graph SLAM. We observe each fiducial tag with an agile camera and construct a pose graph in which fiducial tag poses are bridged by VIO trajectory edges. 2) We then roughly align the fiducial tags with a 3D prior map (i.e., global tag-map registration). Inspired by the recent graph-theoretic global registration methods [9], [10], [11], we propose a tag-map matching method based on robust tag-plane correspondence estimation via maximum clique finding. 3) Finally, we refine the estimated tag poses by directly aligning agile camera images with the prior map using normalized information distance (NID), a mutual-information-based cross-modal distance metric.

The main contribution of this work is three-fold:

1) We propose an accurate and scalable fiducial tag localization method that enables deploying a massive amount of tags on a 3D prior map in a short time.
2) To robustly perform global tag-map registration, a graph-theoretic tag-plane correspondence estimation method is proposed.
3) We show that the combination of NID-based direct camera-map alignment and maximum clique finding-

*This work was supported in part by a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

[1] All the authors are with the Department of Information Technology and Human Factors, the National Institute of Advanced Industrial Science and Technology, Umezono 1-1-1, Tsukuba, 3050061, Ibaraki, Japan, k.koide@aist.go.jp


1. Tag relative pose estimation

Landmark pose graph SLAM based on VIO and fiducial tags

2. Global tag-map registration

Establishing tag plane correspondences via maximum clique finding

3. Tag pose refinement

NID-based direct camera-map alignment and joint factor graph optimization

Fig. 2: Processing flow of the proposed system.

based outlier filtering enables to further improve the fiducial tag localization accuracy.

II. Related work

There have been proposed many monocular camera localization methods in a 3D prior map for vision-based navigation. Caselitz et al. reconstructed the surrounding environment from camera images using a visual SLAM technique and estimated the camera pose in a given map by matching reconstructed points with map points [1]. Pascoe et al., used the NID metric, a mutual-information-based cross-modal distance metric, to directly align camera images with the 3D prior map [12]. Ye et al., combined surfel-based map rendering and direct photometric comparison to keep tracking the camera trajectory on a prior map [13]. While these methods enable accurate map-referenced camera localization and affordable vision-based navigation with a single camera, they can sometimes be unreliable in featureless and dynamic environments. Several practical systems thus often combine vision-based camera localization with fiducial tag detection for reliability and for fail-safe [6], [14], [15].

As we can robustly detect fiducial tags on an image, by using a visual odometry technique, their poses with respect to the visual odometry reference frame can easily be estimated in the form of the landmark pose graph SLAM [16], [17]. However, aligning the estimated tag poses with a 3D prior map (i.e., tag-map global registration) is not straightforward because of the difference of modalities between visually detected fiducial tags and a 3D point cloud map. The modality difference makes it difficult to apply image-to-image matching methods [18], [19] nor geometry-based global registration methods [9], [20] to estimate the transformation between tag and map reference frames.

If a 3D prior map is recorded in an ordered point cloud format (e.g., PTX format), it would be possible to generate images from points and perform visual image matching (e.g., [21]). However, many 3D map datasets provide only unordered point clouds that make it difficult to generate images with good quality resulting in deteriorated accuracy of visual image matching.

The proposed method robustly determines the tag-map transformation across different modalities by combining geometry-based tag-plane correspondence hypothesis making and graph-theoretic outlier hypothesis rejection. While the geometrical hypothesis making yields many false correspondences, the graph-theoretic algorithm robustly filters out wrong hypotheses and finds the best subset of correspondence hypotheses that gives the best explanation for the tag placement in the map.

III. Methodology

Fig. 2 shows an overview of the proposed method. In the tag relative pose estimation step, we observe each fiducial tag using an agile camera and estimate the relative poses between fiducial tags in the form of the landmark visual SLAM. In the following step, we roughly align fiducial tags with a 3D prior map by establishing tag-plane correspondences via maximum clique finding. We then refine tag and camera poses by directly aligning camera images with the map.

A. Tag Relative Pose Estimation based on Landmark Pose Graph SLAM

In this step, we use an agile camera to observe fiducial tags placed in an environment, and reconstruct the relative poses between tags with a standard landmark pose graph SLAM approach [22]. We estimate the camera ego-motion using a VIO algorithm (e.g., VINS-Mono [23]) while detecting fiducial tags on images [24]. Let $T_{WC}(t)$ be the camera pose estimated by VIO at time t, and $T^i_{CT}(t)$ be the pose of a detected fiducial tag with tag ID $= i$. We estimate the camera pose at every time step $\widetilde{T}_{WC}(t)$ and fiducial tag poses $\widetilde{T}^i_{WT}$ by minimizing the following objective function that combines odometry factors e_{odom} and tag observation factors e_{tag}:

$$e_{\text{slam}} = e_{\text{odom}} + e_{\text{tag}}, \tag{1}$$

$$e_{\text{odom}} = \sum_{t=1}^{N-1} \| \log \left(\widetilde{T}_{WC}(t,t+1)^{-1} T_{WC}(t,t+1) \right) \|^2, \tag{2}$$

$$e_{\text{tag}} = \sum_{t=1}^{N} \sum_{i}^{M} \| \log \left(\widetilde{T}_{WC}(t)^{-1} \widetilde{T}^i_{WM} T^i_{CM}(t)^{-1} \right) \|^2, \tag{3}$$

where $T_{WC}(t,t+1) = T_{WC}(t)^{-1} T_{WC}(t)$ is the relative camera pose between t and $t+1$, and log is the SE3 logarithmic map. Here, we intentionally decouple VIO estimation and tag pose estimation and fuse them on a pose graph so that we can easily change the VIO algorithm depending

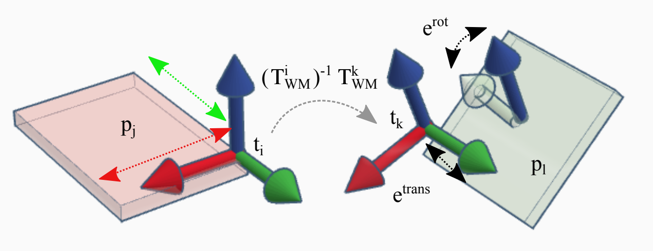


Fig. 3: Tag-plane correspondence consistency.

on the use scenario (e.g., using another VIO running on a smartphone [25] instead of VINS-Mono).

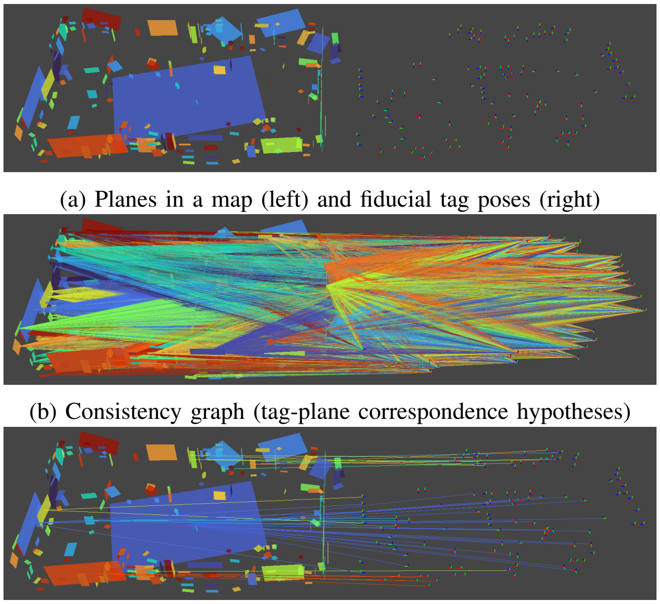
B. Tag-Map Global Registration via Graph-Theoretic Tag-Plane Correspondence Establishment

Given the estimated relative poses between fiducial tags, we roughly align the reference frame of the tags (i.e., VIO origin) with the map reference frame (i.e., tag-map global registration). The challenge here is that we need to deal with the difference of modalities between the sparse point cloud map and visually detected fiducial tags. Due to the difference of modalities, traditional vision-based image matching methods [18], [19] are not applicable.

In this work, we assume that most of the fiducial tags are placed on a plane in the environment, and solve the global registration problem by establishing tag-plane correspondences. This assumption can naturally be held in many practical use scenarios because most fiducial tags are required to be placed on a flat surface for accurate detection and localization [24].

Inspired by the recent success of graph-theoretic approaches for global registration [9], [10], [11], we robustly estimate tag-plane correspondences via maximum clique finding. We first construct a consistency graph, in which vertices represent a hypothesis of tag-plane correspondence, and edges represent the consistency between two tag-plane correspondence hypotheses. By finding the largest subset of hypotheses that are all mutually consistent (i.e., the maximum clique in the consistency graph), we can robustly filter out outlier correspondences and determine a set of tag-plane correspondences that gives the best explanation for the tag placement in the prior map.

To construct a consistency graph, we first extract planes from the map point cloud using region growing segmentation [26] and then list all possible tag-plane correspondences. We evaluate the geometrical consistency of every combination of correspondence hypotheses as illustrated in Fig. 3. Let $h_{ij} = (t_i, p_j)$ be a hypothesis that a fiducial tag t_i corresponds to a plane p_j in the map and consider its consistency with another hypothesis $h_{kl} = (t_k, p_l)$. We transform the pose of t_i such that its normal becomes aligned with the normal of p_j. Given the relative pose between t_i and t_k, we shift and rotate t_i on the plane p_j such that the distance between t_k and p_l is minimized. If the distance between t_k and p_l is smaller than th^{trans} and the angle error of their normals is smaller than th^{rot} (e.g., 0.4 m and 10°), we consider h_{ij} and h_{kl} are mutually consistent. For the algorithmic detail of the consistency check procedure, see the appendix.


(a) Planes in a map (left) and fiducial tag poses (right)

(b) Consistency graph (tag-plane correspondence hypotheses)

(c) Maximum clique in the consistency graph

Fig. 4: Tag-plane correspondence estimation via maximum clique finding.

With all combinations of tag-plane correspondence hypotheses that pass the mutual geometrical consistency check, we construct a consistency graph and find the maximum clique using the parallel heuristic maximum clique finding algorithm [27].

Fig. 4 shows an example of tag-plane correspondence estimation results. For all possible combinations of planes and tags, we evaluated the correspondence consistency and constructed a consistency graph (Fig. 4 (b)) and then found the maximum clique (Fig. 4 (c)). While the consistency graph contained a massive amount of tag-plane correspondence hypotheses (429,735 hypothesis pairs), the maximum clique (56 tag-plane correspondences) was efficiently found in 92 msec.

Given the tag-plane correspondences, we estimate the transformation between the tag and map reference frames by minimizing the symmetric point-to-normal ICP distance [28] between corresponding tags and planes.

C. Estimation Refinement via Information-theoretic Direct Camera-Map Alignment

We then refine the tag and camera pose estimates by directly aligning each camera image with the global map using the NID [12], a mutual-information-based cross-modal distance measure. We insert the camera poses aligned with the map into the landmark pose graph as prior factors, and re-optimize the graph to improve the camera and tag pose estimates.

Let $\mathcal{P}$ be a map point cloud, I_r be a camera image, and $\widetilde{T}_{WC}(t)$ be the camera pose with respect to the map reference frame. Given an initial estimate of $\widetilde{T}_{WC}(t)$, we first remove points from $\mathcal{P}$ that should not be visible from the current viewpoint using direct visibility assessment [29], and then estimate $\widetilde{T}_{WC}(t)$ by minimizing the NID metric between $\mathcal{P}$ and I_r using the BFGS algorithm. The NID is defined as follows:

$$\mathrm{NID}(I_r, I_s) = \frac{\mathrm{H}(I_r, I_s) - \mathrm{MI}(I_r; I_s)}{\mathrm{H}(I_r, I_s)}, \tag{4}$$

$$\mathrm{MI}(I_r; I_s) = \mathrm{H}(I_r) + \mathrm{H}(I_s) - \mathrm{H}(I_r, I_s), \tag{5}$$

where I_s is a map image created by projecting $\mathcal{P}$ on the image space of I_r, $\mathrm{H}(I_r, I_s), \mathrm{H}(I_r), \mathrm{H}(I_s)$ are the joint and marginal entropies of I_r and I_s, and $\mathrm{MI}(I_r, I_s)$ is the mutual information between I_r and I_s. Because this metric does not directly compare pixel and point colors but measure co-occurrence of them, it enables to measure the distance between data across different modalities. Following [12], we use B-spline based weighted histogram voting to make Eq. 4 differentiable.

The NID enables to accurately determine the camera pose with respect to a map point cloud. It is, however, very sensitive to the initial guess and often gets corrupted. Because the NID is a dimensionless quantity, it is not easy to remove corrupted results with simple thresholding of the NID value.

To robustly filter out corrupted camera-map alignment results, we again use a graph-theoretic approach to find the maximum mutually consistent subset of them. Let $h_t = (\widetilde{T}_{WC}(t), \widehat{T}_{WC}(t))$ be a pair of initial and refined camera poses. To determine the consistency between h_t and h_k, we calculate the camera pose displacements they are declaring:

$$\Delta\widehat{T}_{WC}(t) = \widetilde{T}_{WC}^{-1}(t)\widehat{T}_{WC}(t), \tag{6}$$

$$\Delta\widehat{T}_{WC}(t, k) = \Delta\widehat{T}_{WC}(t)^{-1}\Delta\widehat{T}_{WC}(k). \tag{7}$$

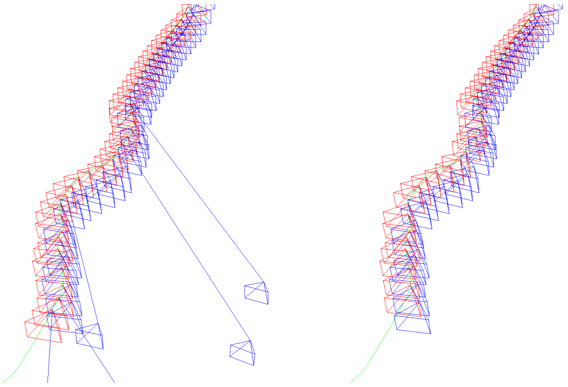
If the translational and rotational errors of $\Delta\widehat{T}_{WC}(t, k)$ are smaller than threshold values (e.g., 0.5 m and 5°), we consider h_t and h_k are mutually consistent. We construct a consistency graph for all combinations of camera-map alignment results and find the maximum clique to filter out corrupted results (i.e., outliers).

Fig. 5 (a) shows an example of NID-based camera-map alignment results. Red frustums show the initial camera poses of frames where the BFGS optimization converges while blue ones show estimated camera poses. We can see that the NID-based optimization sometimes gets corrupted. Fig. 5 (b) shows a result of outlier filtering, in which corrupted camera-map alignment results are filtered out and only mutually consistent results remain.

We insert the refined camera poses in the factor graph created in the tag relative pose estimation as pose prior factors, and re-optimize tag and camera poses with all the constraints:

$$e_{\text{refine}} = e_{\text{odom}} + e_{\text{tag}} + e_{\text{NID}}, \tag{8}$$

$$e_{\text{NID}} = \sum_{i}^{N} \| \log\left(\widetilde{T}_{WC}(t)^{-1}\widehat{T}_{WC}(t)\right) \|^2. \tag{9}$$



(a) Camera alignment results (b) Outlier filtering result

Fig. 5: Camera-map alignment and outlier filtering results. Red: initial camera poses, Blue: refined camera poses.

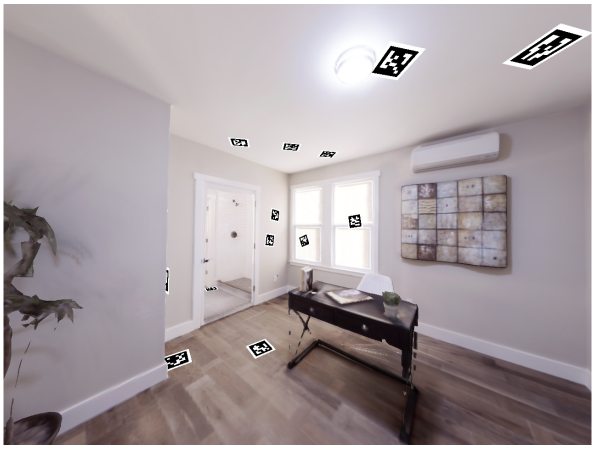


Fig. 6: Fiducial tags randomly placed on the Replica dataset.

IV. Experiments

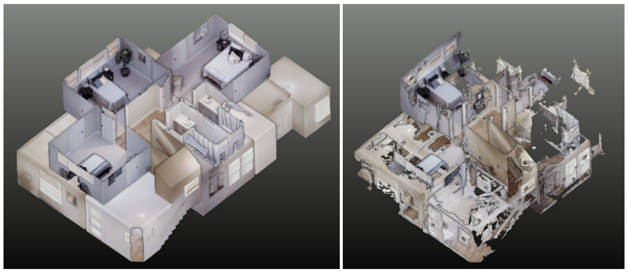
A. Evaluation in a Simulated Environment

1) Global registration evaluation: To evaluate the proposed method, we used *apartment_0* model in the *Replica* dataset [30], [31]. We generated 50 camera trajectories by randomly sampling waypoints in the map and interpolating them using SE3 B-spline interpolation. Along with camera images, we synthesized IMU data using [32]. For each trajectory, we randomly placed 200 fiducial tags on planes in the map and estimated the poses of fiducial tags, which were observed by the camera more than once, using the proposed method.

As a baseline, we compared the proposed method with feature-based global registration methods. We first ran colmap [21] on the camera image stream to obtain a dense 3D point cloud of the environment. We then extracted FPFH features [33] respectively from the reconstructed point cloud

TABLE I: Global registration success rate

Method	RANSAC [20]	Teaser [9]	Proposed
Success rate	26% (13 / 50)	70% (35 / 50)	98% (49 / 50)



(a) Replica model (b) Reconstructed model

Fig. 7: Prior environmental map and reconstructed point cloud for *apartment_0* model.

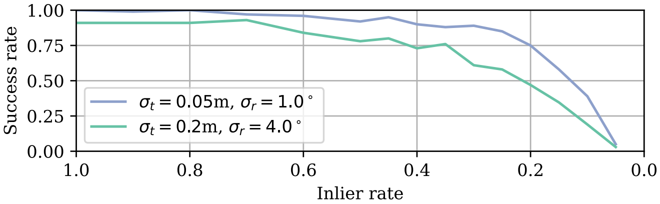


Fig. 8: Fiducial tag inlier rate vs global registration success rate. The proposed method is robust to outlier tags that do not lie on a plane.

and the 3D prior model and estimated the transformation between them using RANSAC [20] and Teaser [9].

We first evaluated the global registration success rate of each method. If the translational and rotational errors of a global registration result are smaller than threshold values (1.0 m and 15°), we consider the registration is succeeded. Table I shows the success rate of each global registration method. We can see that, even with Teaser [9], a state-of-the-art transformation estimation algorithm, the global registration failed for 30 % of the sequences. Fig. 7 shows the 3D prior map model and the reconstructed model. While the reconstructed model well captures the overall shape of the environment, the detailed shapes and densities of points are largely different from those of the prior map model. We consider that these differences make it difficult to obtain consistent features between the reconstructed and the prior models, which results in the global registration failures.

The proposed tag-plane matching method successfully estimated the tag-map transformation except for a sequence where fiducial tags were placed very symmetrically and wrong tag-plane correspondences were given via maximum clique finding.

2) Fiducial tag localization accuracy: We calculated tag localization errors for sequences where global registration succeeded. For RANSAC and Teaser, assuming the perfect tag localization accuracy is given on the reconstructed model, we calculated transformation errors of point cloud registration results as fiducial tag localization errors.

Table II summarizes the fiducial tag localization accuracy of each method. We can see that Teaser exhibited better localization errors (0.180 m and 2.807°) than that of RANSAC (0.416 m and 7.847°) thanks to its robust feature matching mechanism. The proposed method showed the best localization accuracy among compared methods (0.110 m and 1.870°) owing to the robust tag-plane matching. With the refinement step, the localization accuracy was further improved, and we achieved an average translation error of a few centimeters (0.039 m and 1.021°).

3) Robustness to outliers: To evaluate the robustness of the proposed method to outlier fiducial tags that are not lying on a plane, we evaluated the global registration success rate while changing the number of inlier tags. For each inlier tag rate setting R^{in}, we generated $100R^{\text{in}}$ fiducial tags on randomly selected planes in the environment and $100(1 - R^{\text{in}})$ tags with random poses, and repeated random tag placement and global registration for 100 times. To see how the global registration success rate changes depending on tag relative pose estimation errors, we added two levels of random pose noise ($\sigma_t = 0.05$m / $\sigma_r = 1.0°$, and $\sigma_t = 0.2$m / $\sigma_r = 4.0°$) to the tag poses.

Fig. 8 shows a plot of global registration success rate vs inlier rate. We can see that the proposed method is robust to outlier tags and achieved a success rate of over 90% with 60% outlier tags under a low-level noise (0.05 m and 1.0°). Even under larger tag pose noise (0.2 m and 4.0°), the proposed method achieved a success rate about 75% with 60% outlier tags.

B. Evaluation in a real environment

To demonstrate that the proposed method enables robust fiducial tag localization on a 3D prior map with a small effort, we placed 117 fiducial tags in the environment shown in Fig. 9. The red circles in the figure show the positions of placed fiducial tags. Note that only tags that are visible in the figure are drawn just for visualization. We recorded two environmental map point clouds with and without fiducial tags using a 3D LiDAR (FARO Focus). We manually annotated fiducial tag positions on the environmental map to obtain the ground-truth tag poses. We then estimated the tag poses with the proposed method using the environmental map without tags.

For VIO, we recorded a stream of monocular images and IMU measurements using a MYNTEYE camera. Each fiducial tag was observed by the agile camera at least once during the recording. The duration of the image stream was about 970 s.

Table III summarizes the processing time of each step in the proposed method. The tag relative pose estimation step was performed on the fly while recording the image stream. The global tag-map registration step took only about 1.5 s thanks to the efficient maximum clique finding algorithm. The refinement step took about 381.3 s in total, and the NID-based camera pose estimation was the most computationally demanding process in this step (379.6 s). Note that the current implementation uses only a CPU, and the processing time of the NID optimization can be improved by 10 to 50 times faster by using a GPU implementation [2], [12], resulting in reducing the total processing time to about 10 s. Furthermore, because we can independently perform the

TABLE II: Fiducial tag localization errors

Method	RANSAC [20]	Teaser [9]	Proposed	
			w/o refinement	w/ refinement
Translational error [m]	0.416 ± 0.214	0.180 ± 0.190	0.110 ± 0.078	**0.039** ± **0.060**
Rotational error [°]	7.847 ± 4.081	2.807 ± 3.042	1.870 ± 1.629	**1.021** ± **1.629**

TABLE III: Processing time

Step	Process	Time [s]
Tag pose estimation	Visual inertial odometry Fiducial tag detection Pose graph optimization	on-the-fly
Global registration	Consistency graph creation Maximum clique finding Transformation optimization	1.395 0.092 0.004
	Total	1.491
Estimation refinement	NID camera alignment Outlier filtering Pose graph optimization	379.6 0.072 1.614
	Total	381.3

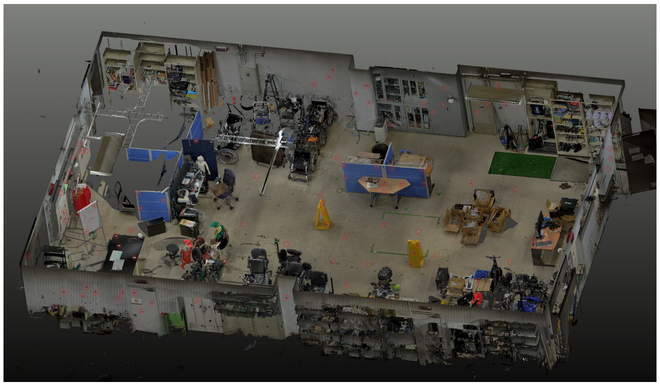


Fig. 9: Experimental environment. The red circles indicate the positions of fiducial tags. 117 fiducial tags were placed in the environment.

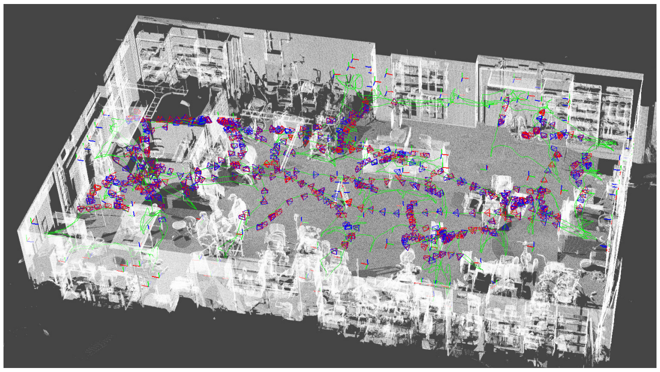


Fig. 10: Estimation result. RGB thick lines: estimated tag poses, Green thin line: camera trajectory, Red frustums: camera poses where the NID optimization converged, Blue frustums: NID-based camera-map alignment results.

image alignment frame-by-frame, it can easily be accelerated by using, e.g., cloud computing.

Fig. 10 shows the estimation result. The thick RGB lines indicate the estimated fiducial tag poses, and the thin green line shows the estimated camera trajectory. The average translational and rotational errors of the estimated fiducial tag poses were respectively 0.019 ± 0.014 m and 2.382 ± 4.093 °. This result would be sufficiently accurate for the requirement for vision-based navigation systems.

V. Conclusions

We have proposed an accurate and scalable method for fiducial tag localization on a 3D prior environmental map. We first estimate the relative poses between fiducial tags using VIO-based landmark graph SLAM, and then roughly align the fiducial tags with a 3D prior map using a graph-theoretic tag-plane correspondence estimation. We refine the estimated tag and camera poses by directly aligning camera images with the environmental map using an information-theoretic metric. Through simulation-based experiments, the proposed method achieved a global registration success rate of 98% and tag estimation accuracy of a few centimeters. The real experiment demonstrated that the proposed method can accurately localize over 100 fiducial tags on a prior map in 16 minutes for data recording and 6 minutes for post-processing.

Appendix

A. Tag-Plane Correspondence Consistency Check

Alg. 1 describes the algorithm to determine the consistency of two tag-plane correspondence hypotheses. $\mathbf{n}_*, R_*, \mathbf{p}_*$ are the normal, rotation, and translation of a tag or plane, respectively. $\mathbf{l}_*$ is the length of a plane along XYZ axes (Z = normal). Lines 2 and 3 swap h_{ij} and h_{kl} if the normal of p_j is almost vertical to avoid indeterminacy of the tag-plane rotation (Because we can assume that both t_i and p_j are gravity-aligned, the tag-plane rotation can be determined from only their normals as long as the plane normal is not vertical). Lines 4-7 calculate the transformation that transforms t_i such that its normal is aligned with the normal of p_j. Note that we consider only rotation along the gravity direction assuming tags and planes are gravity-aligned. Line 8 checks if the angle error between the normals of p_l and rotated t_k. If it is larger than a threshold, we consider h_{ij} is inconsistent with h_{kl}. Lines 10-13 calculate the translation to minimize the distance between t_k and p_l such that t_i remains on p_j. Then, if the distance between transformed t_k and p_l is smaller than a threshold, we consider h_{ij} and h_{kl} are mutually consistent.

Algorithm 1 Tag-plane correspondence consistency check

```
function CONSISTENCY_CHECK($h_{ij}, h_{kl}$)
    if $\mathbf{n}_j \cdot [0,0,1]^T > 1 - \epsilon$ then
        SWAP($h_{ij}, h_{kl}$)
    $\mathbf{n}_i^{\mathrm{XY}} = \mathbf{n}_{t_i} \circ [1,1,0]^T$
    $\mathbf{n}_j^{\mathrm{XY}} = \mathbf{n}_{p_j} \circ [1,1,0]^T$
    $R_{ji} = \mathrm{ALIGN\_VECTORS}(\mathbf{n}_i^{\mathrm{XY}}, \mathbf{n}_j^{\mathrm{XY}})$
    $\mathbf{p}_{ji} = \mathbf{p}_{p_j} - R_{ji}\mathbf{p}_{t_i}$
    if $\mathrm{ANGLE}(R_{ji}\mathbf{n}_{t_k}, \mathbf{n}_{p_l}) > \mathrm{th}^{\mathrm{rot}}$ then
        return False
    $\mathbf{p}_{lk} = \mathbf{p}_{p_l} - (R_{ji}\mathbf{p}_{t_k} + \mathbf{p}_{ji})$
    $\mathbf{p}_{jk} = R_{p_j}^{-1}\mathbf{p}_{lk}$
    $\mathbf{p}'_{jk} = \mathrm{CLAMP}(\mathbf{p}_{jk}, -\mathbf{l}_{p_j}/2, \mathbf{l}_{p_j}/2)$
    $\mathbf{p}'_{t_k} = R_{ji}\mathbf{p}_{t_k} + \mathbf{p}_{ji} + R_{p_j}\mathbf{p}'_{jk}$
    if $\mathrm{DISTANCE}(\mathbf{p}'_{t_k}, p_l) > \mathrm{th}^{\mathrm{trans}}$ then
        return False
    else
        return True
function ALIGN_VECTORS($\mathbf{a}, \mathbf{b}$)
    $\mathbf{v} = \mathbf{a} \times \mathbf{b}$
    $s = \|\mathbf{v}\|$
    $c = \mathbf{a} \cdot \mathbf{b}$
    if $s < \epsilon$ then
        return $I$
    $S = \mathrm{skew}(\mathbf{v})$
    return $I + S + \frac{1-c}{s^2} S^2$
```

References

[1] T. Caselitz, B. Steder, M. Ruhnke, and W. Burgard, "Monocular camera localization in 3d LiDAR maps," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, Oct. 2016.

[2] S. Oishi, Y. Kawamata, M. Yokozuka, K. Koide, A. Banno, and J. Miura, "C*: Cross-modal simultaneous tracking and rendering for 6-DoF monocular camera localization beyond modalities," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 5229–5236, Oct. 2020.

[3] J. Park, D. Lee, and J. Park, "Digital map based pose improvement for outdoor augmented reality," in *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, nov 2012.

[4] M. Krogius, A. Haggenmiller, and E. Olson, "Flexible layouts for fiducial tags," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, Nov. 2019.

[5] J.-K. Huang, S. Wang, M. Ghaffari, and J. W. Grizzle, "LiDARTag: A real-time fiducial tag system for point clouds," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4875–4882, July 2021.

[6] Y. Huang, J. Zhao, X. He, S. Zhang, and T. Feng, "Vision-based semantic mapping and localization for autonomous indoor parking," in *IEEE Intelligent Vehicles Symposium (IV)*. IEEE, June 2018.

[7] Z. Fang, Y. Chen, M. Zhou, C. Lu, N. Rottmann, R. Bruder, H. Xue, A. Schweikard, E. Rueckert, R. Nabati, *et al.*, "Marker-based mapping and localization for autonomous valet parking," in *IEEE/RSJ International Conference on Intelligent Robots and Systems WS (IROSWS)*, 2020.

[8] N. Kayhani, W. Zhao, B. McCabe, and A. P. Schoellig, "Tag-based visual-inertial localization of unmanned aerial vehicles in indoor construction environments using an on-manifold extended kalman filter," *Automation in Construction*, vol. 135, p. 104112, Mar. 2022.

[9] H. Yang, J. Shi, and L. Carlone, "TEASER: Fast and certifiable point cloud registration," *IEEE Transactions on Robotics*, vol. 37, no. 2, pp. 314–333, Apr. 2021.

[10] J. Shi, H. Yang, and L. Carlone, "ROBIN: a graph-theoretic approach to reject outliers in robust estimation using invariants," in *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, May 2021.

[11] P. C. Lusk, K. Fathian, and J. P. How, "CLIPPER: A graph-theoretic framework for robust data association," in *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, May 2021.

[12] G. Pascoe, W. Maddern, and P. Newman, "Robust direct visual localisation using normalised information distance," in *British Machine Vision Conference (BMVC)*. British Machine Vision Association, 2015.

[13] H. Ye, H. Huang, and M. Liu, "Monocular direct sparse localization in a prior 3d surfel map," in *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, May 2020.

[14] K. Eckenhoff, Y. Yang, P. Geneva, and G. Huang, "Tightly-coupled visual-inertial localization and 3-d rigid-body target tracking," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1541–1548, Apr. 2019.

[15] M. Beul, D. Droeschel, M. Nieuwenhuisen, J. Quenzel, S. Houben, and S. Behnke, "Fast autonomous flight in warehouses for inventory applications," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3121–3128, 2018.

[16] B. Pfrommer and K. Daniilidis, "Tagslam: Robust slam with fiducial markers," *arXiv preprint arXiv:1910.00679*, 2019.

[17] K. Koide and E. Menegatti, "Non-overlapping RGB-d camera network calibration with monocular visual odometry," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, Oct. 2020.

[18] D. Galvez-López and J. D. Tardos, "Bags of binary words for fast place recognition in image sequences," *IEEE Transactions on Robotics*, vol. 28, no. 5, pp. 1188–1197, Oct. 2012.

[19] D. Cattaneo, M. Vaghi, S. Fontana, A. L. Ballardini, and D. G. Sorrenti, "Global visual localization in LiDAR-maps through shared 2d-3d embedding space," in *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, May 2020.

[20] A. G. Buch, D. Kraft, J.-K. Kamarainen, H. G. Petersen, and N. Kruger, "Pose estimation using local structure-specific shape and appearance context," in *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, May 2013.

[21] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2016.

[22] G. Grisetti, R. Kummerle, C. Stachniss, and W. Burgard, "A tutorial on graph-based SLAM," *IEEE Intelligent Transportation Systems Magazine*, vol. 2, no. 4, pp. 31–43, Dec. 2010.

[23] 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, Aug. 2018.

[24] J. Wang and E. Olson, "AprilTag 2: Efficient and robust fiducial detection," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, Oct. 2016.

[25] E. Marder-Eppstein, "Project tango," in *ACM SIGGRAPH 2016 Real-Time Live!* ACM, July 2016.

[26] T. Rabbani, F. Van Den Heuvel, and G. Vosselmann, "Segmentation of point clouds using smoothness constraint," *International archives of photogrammetry, remote sensing and spatial information sciences*, vol. 36, no. 5, pp. 248–253, 2006.

[27] R. A. Rossi, D. F. Gleich, and A. H. Gebremedhin, "Parallel maximum clique algorithms with applications to network analysis," *Journal on Scientific Computing*, vol. 37, no. 5, pp. 589–616, Jan. 2015.

[28] S. Rusinkiewicz, "A symmetric objective function for ICP," *ACM Transactions on Graphics*, vol. 38, no. 4, pp. 1–7, Aug. 2019.

[29] S. Katz, A. Tal, and R. Basri, "Direct visibility of point sets," in *ACM SIGGRAPH 2007*. ACM, 2007.

[30] J. Straub, T. Whelan, L. Ma, Y. Chen, E. Wijmans, S. Green, J. J. Engel, R. Mur-Artal, C. Ren, S. Verma, *et al.*, "The replica dataset: A digital replica of indoor spaces," *arXiv preprint arXiv:1906.05797*, 2019.

[31] M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik, D. Parikh, and D. Batra, "Habitat: A platform for embodied AI research," in *IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2019.

[32] P. Geneva, K. Eckenhoff, W. Lee, Y. Yang, and G. Huang, "OpenVINS: A research platform for visual-inertial estimation," in *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, May 2020.

[33] R. B. Rusu, N. Blodow, and M. Beetz, "Fast point feature histograms (FPFH) for 3d registration," in *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, May 2009.