

Point Cloud Change Detection With Stereo V-SLAM: Dataset, Metrics and Baseline

Zihan Lin¹, Jincheng Yu¹, Lipu Zhou², Xudong Zhang¹, Jian Wang¹, Yu Wang¹

Abstract—Localization and navigation are basic robotic tasks requiring an accurate and up-to-date map to finish these tasks, with crowdsourced data to detect map changes posing an appealing solution. Collecting and processing crowdsourced data requires low-cost sensors and algorithms, but existing methods rely on expensive sensors or computationally expensive algorithms. Additionally, there is no existing dataset to evaluate point cloud change detection. Thus, this paper proposes a novel framework using low-cost sensors like stereo cameras and IMU to detect changes in a point cloud map. Moreover, we create a dataset and the corresponding metrics to evaluate point cloud change detection with the help of the high-fidelity simulator Unreal Engine 4. Experiments show that our visual-based framework can effectively detect the changes in our dataset.

Index Terms—Visual-Inertial SLAM, Mapping, Data Sets for SLAM

I. INTRODUCTION

AUTONOMOUS ground and unnamed aerial vehicles have become popular over the last few years. These robotic vehicles are commonly used in various tasks, such as logistics delivery and remote sensing, with localization and navigation being mandatory sub-tasks requiring an accurate and up-to-date map. Thus, detecting the map changes is crucial.

This work is designed to exploit crowdsourced data to detect map changes in real-time. Crowdsourced data is collected from several vehicles equipped with sensors like a camera, and the location where vehicles pass also contributes to detecting map changes. When many vehicles collect data, the overall sensor cost and the collected data are enormous. Hence, these vehicles should be equipped with low-cost sensors to reduce hardware costs, and crowdsourced data should be processed using an algorithm with low computational complexity.

Current change detection methods utilize an expensive hardware setup. For example, remote sensing tasks utilize expensive sensors like vehicle-based laser scanner (VLS), airborne laser scanner (ALS), and mobile laser scanner (MLS)

to detect point cloud changes [1]–[3]. However, the price of a laser scanner ranges from thousands to millions of dollars. Alternative solutions exploit the structure from motion (SfM), and multi-view stereo (MVS) technology [4], but these methods impose a huge computational complexity, typically requiring several hours to days of calculations, depending on the point cloud size. For example, Yew and Lee [4] developed a SfM-based algorithm, where the reconstruction and registration steps need three days for a medium-size urban environment.

Considering the available datasets, EuRoC [5] and KITTI [6], and their corresponding metrics focus on evaluating the performance of localization and mapping. Moreover, the SHREC 2021 point cloud change detection dataset [7] focuses on change detection of point clouds collected by LiDAR. Hence, to the best of our knowledge, datasets and metrics to evaluate visual-based point cloud change detection do not exist, and thus evaluating the performance of change detection is a challenging task.

Hence, this paper proposes a novel visual-based framework to achieve low-cost change detection in point clouds and a dataset with the appropriate metrics to evaluate the performance of point cloud change detection. The main contributions of this paper are as follows:

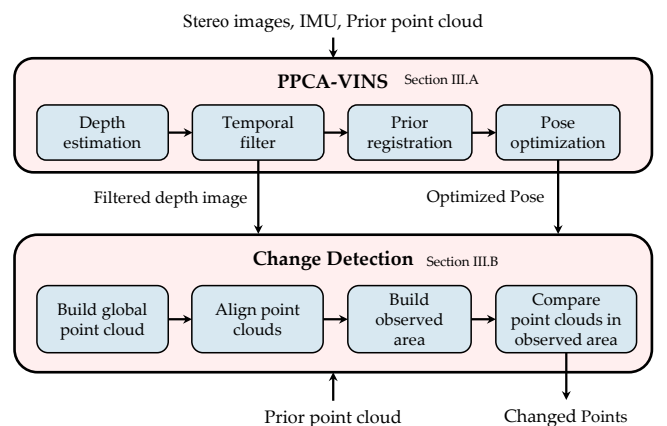


Fig. 1. Proposed change detection framework comprising the PPCA-VINS and Change Detection modules. PPCA-VINS provides 6D poses and depth images for change detection, which are then input to the change detection pipeline to build a global point cloud. Then the global point cloud and the prior point clouds of an observed area are compared, with the change detection pipeline providing the point cloud vertices that have changed.

- Proposing a prior point cloud-assisted SLAM framework named PPCA-VINS (prior point cloud-assisted VINS), based on VINS-Fusion [8] and assuming having an accurate initial pose. Compared to standard VINS-Fusion, our frame fuses prior information from prior point cloud

Manuscript received: June 28, 2022; Revised September, 26, 2022; Accepted October 20, 2022.

This paper was recommended for publication by Editor Sven Behnke. Name upon evaluation of the Associate Editor and Reviewers' comments.

This work was supported by National Natural Science Foundation of China (U20A20334, U19B2019 and M-0248), Tsinghua-Meituan Joint Institute for Digital Life, Tsinghua EE Independent Research Project, Beijing National Research Center for Information Science and Technology (BNRist), and Beijing Innovation Center for Future Chips.

¹Department of Electronic Engineering, Tsinghua University, Beijing, 100084, China. jian-wang@tsinghua.edu.cn

²Meituan, 7 Rongda Road, Chaoyang District, Beijing, 100012, China.

Corresponding author: Jian Wang

Our code is in the repository: <https://github.com/lnexen/PPCA-VINS.git>, and dataset is also released on this page.

Digital Object Identifier (DOI): see top of this page.

and affords an appealing performance over the standard VINS-Fusion on our dataset. (A prior point cloud is a point cloud of the scene that vehicle is in, but the prior point cloud is out-dated and needs change detection.)

- Designing a new *dataset and the corresponding metrics* for visual-based point cloud change detection tasks. The dataset fills the vacancy in the point cloud change detection task and provides raw data from multiple sensor types onboard a car or an unnamed aerial vehicle.
- Developing a visual-based change detection architecture that utilizes depth images and poses provided by PPCA-VINS to build a global point cloud. We detect changes by comparing the global point cloud to the prior point cloud. Our change detection pipeline is evaluated on our change detection dataset, verifying its capability for point cloud change detection.

The proposed framework is illustrated in Fig.1 and comprises two parts. The first part is PPCA-VINS (Section III-A), which estimates depth from stereo images and then builds filtered depth images. Then we convert the filtered depth image to a point cloud and register it to the previous point cloud. Finally, we optimize the vehicle's pose by fusing the pose obtained from VINS-Fusion and the prior registration.

The second part of our architecture is the change detection module (Section III-B). In this module, we first build a global point cloud using the filtered depth images and optimized poses from PPCA-VINS, and then we align the global point cloud to the prior point clouds. Then we build the point cloud of the observed area and compare the global point cloud to the prior point clouds using point cloud registration to obtain the changed vertices.

The remainder of this paper is as follows: Section II presents the related work, while Section III includes the proposed method. Section IV builds the suggested dataset using the high-fidelity simulator Unreal Engine 4 [9] due to ease of changing scene and data collection. Moreover, this section suggests some evaluation metrics. Section V conducts several experiments to demonstrate the framework's and dataset's effectiveness on map change detection tasks using low-cost sensors and algorithms. Finally, section VI concludes this work.

II. RELATED WORK

A. Visual Localization with Prior Map

Visual localization has been researched for several years. For example, Zuo *et al.* [10] proposed a tightly-coupled Multi-State Constraint Kalman Filter (MSCKF) [11] scheme using stereo vision, where MSCKF is a SLAM framework employing Kalman filter and sliding window. In this work, the local point clouds were generated from a stereo camera setup and utilized Normal Distribution Transform (NDT) [12] to align the local point clouds and the prior LiDAR map, where the location in the prior map was considered as a state in the MSCKF's extended Kalman filter. Lynen *et al.* [13] developed a pipeline based on SfM. This method was also MSCKF-based and tightly coupled. The difference between [13] and [10] is the prior localization algorithm, as [13]

employed SfM to generate the prior point cloud map and stored the visual landmark descriptors as a database in the meantime. The descriptors obtained from the images were calculated and matched against the database landmarks, and then the pose relationship between the 2D feature point and 3D landmarks was filtered using an extended Kalman filter to obtain the final estimated pose. Previous works localize using image sequences from a camera as temporal information, but the following work provides localization result with only single image. Li and Lee [14] suggested a deep learning method named DeepI2P to register images to point clouds. In this work, the authors converted the cross-modal registration problem into a classification problem, where a deep neural network is trained to classify whether a point in a point cloud is within the image frustum. In that case, a post-optimization method is applied to estimate the images' poses.

B. Change Detection

Yew and Lee [4] combined SfM and neural networks, where the SfM built a global point cloud and the neural network performed non-rigid registration between the point clouds. The author claimed that their algorithm takes several days to calculate, which highlights that SfM-based methods are extremely time-consuming. In addition, this method did not distinguish the unknown and empty areas. Xu *et al.* [15] utilized ALS and octrees. Specifically, this method collected point clouds utilizing ALS and the octree stored and indexed the irregularly-distributed points. Then, the points were clustered and classified into different kinds, followed by a registration process to reveal the changed areas. Although this work did not distinguish between unknown and empty areas, it is reasonable to assume that collecting LiDAR data from airborne sensors do not produce many unknown areas.

Besides from change detection with point cloud, there are some other image-based techniques to detect changes from images. Sakurada *et al.* [16] propose a probabilistic model to detect changes. Different from [4], they build a probabilistic model to represent the structure of an outdoor scene and evaluate the probability of changes. Sakurada and Okatani [17] develop a neural network based method to develop changes between images taken at same place but at different time. They exploit neural network to output image features and segmentation results, and they detect changes in images using these outputs. Taneja *et al.* [18] builds a 3D model with multi-view stereo first. Then they collect images from low resolution cameras and detect changes against the 3D model by detecting inconsistencies between images and the model. Taneja *et al.* [19] then develops a methods detecting changes in city's cadastral model using images. Based on previous work [18], they propose a new method to detect changes in large scale with sparse imagery.

Our work employs VINS-Fusion, a unified framework for pose estimation, to provide the VIO pose for our pipeline. VINS-Fusion is a general optimization-based SLAM framework which can fuse multiple sensors (stereo cameras, IMU, Magnetometer etc.). First, VINS-Fusion calculates 6D VIO

poses with inputs from cameras and IMU. Then a global pose optimization is made to fuse 6D VIO poses and inputs from other sensors, and finally VINS-Fusion outputs global 6D poses after optimization. Besides, VINS-Fusion uses keyframe technique, a keyframe is a frame containing timestamp, images, IMU readings and a 6D pose, and there should be enough parallax between two keyframes' images.

In addition, VINS-Fusion use a sliding window to reduce processing time. Thus a tightly-coupled VIO requires a quick registration between the local and the prior point clouds—otherwise, the keyframe is discarded from the sliding window. Opposing current works, the proposed method relies exclusively on cheap sensors like stereo camera, GPS and IMU. Moreover, our pipeline is computationally efficient; thus, our entire framework is low-cost.

III. FRAMEWORK

A. PPCA-VINS

1) *Depth Estimation*: VINS-Fusion can track feature points and generate a sparse point cloud from these points. However, using a sparse point cloud does not provide an accurate localization result through registration. Thus, additional work producing a semi-dense point cloud is expected.

One standard method of cross-modal data processing is converting different modalities into one. In our pipeline, we convert stereo images into point clouds, as stereo depth estimation has been a popular research topic for several years. Semi-global matching (SGM) [20] is a classic method for disparity estimation, based on minimizing the mutual information (MI) between two images. Recently, many neural networks (NNs) based on stereo depth estimation methods have emerged, such as PSMNet [21] and AnyNet [22]. However, these methods have a significant time complexity, and the depths produced are not better than the traditional SGM method. Thus, we leverage SGM to produce depth estimation for stereo images.

2) *Temporal Filter*: The depth images produced by SGM incorporate significant noise. Thus, we adopt the temporal filter proposed in [23] and a similar method named *depth correspondence matching* [10]. We follow these methods and implement a simplified version of the temporal filter.

Let the current keyframe be i . We re-project the estimated depth of every pixel (u, v) in keyframe j ($i - 2 \leq j \leq i + 2, j \neq i$) to the current keyframe i and obtain the projected depth $d_j^i(u, v)$ and pixel coordinate (u', v') . If the distance between the projected depth $d_j^i(u, v)$ and the current frame depth $d_i(u', v')$ is smaller than a threshold δ_d , we call this a successful projection. When a pixel is successfully projected over α times, we average the projected depth and the current frame depth of this pixel to obtain a filtered depth, which is added to the filtered depth image.

3) *Prior Registration*: In this subsection, we register local point clouds to prior point cloud to get prior localization results. We increase the point cloud's density by aggregating the point cloud of every five keyframes to a local cloud, producing a denser point cloud.

For our pipeline, we need to register two point clouds, each including tens of thousands of points. Thus, we need

a fast and accurate registration algorithm. Normal distribution transform (NDT) [12] is a registration algorithm based on normal distribution. It divides points into small cells, and uses normal distribution to represent points' distribution in cells, thus accelerating the computation. But NDT is unsuitable for our pipeline, as it often produces results suffering from a considerable drift. Alternatively, the iterative closest point (ICP) [24] is an accurate registration algorithm but it requires much time even to align two small point clouds. Besides, the Generalized ICP (GICP) [25] is an improved ICP-based algorithm that balances speed and accuracy. Koide *et al.* [26] boosted GICP by implementing the algorithm on multiple threads. Indeed, we conducted several tests comparing a 4-threaded GICP and other single threaded algorithms involving a point cloud and its variant translated by 0.2m. And we found that 4-threaded GICP can run fast and achieve a good result.

Registration is not always accurate, and therefore we evaluate the registration results using the fitness metric, which is the mean distance from each point in the local cloud to its closest point in the prior cloud. We evaluate every registration based on the following:

- 1) Registration should be converged.
- 2) Fitness of registration should be less than a threshold.
- 3) Registration's translation length should be less than a threshold.

Note that GICP is formulated as a Maximum-Likelihood Estimation (MLE) problem, so the Hessian matrix in the GICP registration is the negative of the Fisher information matrix. The registration's covariance matrix is the inverse of the Fisher information matrix, written as follows:

$$\text{cov}(\theta) = \mathcal{I}^{-1}(\theta) = (-\mathbf{H}_\theta)^{-1} \quad (1)$$

The covariance matrix is used in the following global pose estimation.

4) *Pose Optimization*: Based on VINS-Fusion, we consider our registration results as outputs from a pseudo sensor. Then we fuse registration results and 6D VIO poses to get global 6D poses.

The prior localization states considering the prior position and orientation of the k -th keyframe are:

$$\mathbf{p}_k^p, \mathbf{q}_k^p \quad (2)$$

In VINS-Fusion, the global pose graph optimization problem is modelled as a maximum likelihood estimation (MLE) problem. The variables to be optimized are global poses:

$$\mathcal{X} = [\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_n] \quad (3)$$

$$\mathbf{x}_k = [\mathbf{p}_k, \mathbf{q}_k], k \in [0, n] \quad (4)$$

Assuming that the measurement uncertainties are Gaussian distributions, the problem is formulated as :

$$\mathcal{X}^* = \arg \max_{\mathcal{X}} \prod_{t=0}^n \prod_{k \in \mathbf{S}} p(\mathbf{z}_t^k | \mathcal{X}) \quad (5)$$

$$= \arg \min_{\mathcal{X}} \sum_{t=0}^n \sum_{k \in \mathbf{S}} \|\mathbf{z}_t^k - h_t^k(\mathcal{X})\|_{\Omega_t^k}^2 \quad (6)$$

Since we use the same VIO factor as in the VINS-Fusion’s original paper, we do not explain it here to enhance readability. Considering the prior localization factor, we divide it into two parts. One is the global position factor:

$$\mathbf{z}_{t,p}^p - h_{t,p}^p(\mathcal{X}) = \mathbf{z}_{t,p}^p - h_{t,p}^p(\mathbf{x}_t) = \mathbf{p}_t^p - \mathbf{p}_t \quad (7)$$

and the other one is the relative rotation factor from the initial frame to current frame t:

$$\mathbf{z}_{t,q}^p - h_{t,q}^p(\mathcal{X}) = \mathbf{z}_{t,q}^p - h_{t,q}^p(\mathbf{x}_1, \mathbf{x}_t) = (\mathbf{q}_1^p)^{-1} \mathbf{q}_t^p \ominus (\mathbf{q}_1)^{-1} \mathbf{q}_t \quad (8)$$

Here \mathbf{z} denotes observations from sensors, and h denotes observation function transforming poses to observed values.

Considering the prior localization factor, we build a pose graph and optimize it to obtain the global pose. The covariance of the prior localization is given in eq.1, which is used to calculate the confidence of the prior localization.

B. Change Detection

This subsection builds a simple change detection algorithm that comprises the following steps:

- 1) build the global point cloud \mathcal{P}_g
- 2) align the global point cloud to the prior point cloud \mathcal{P}_p using GICP
- 3) build the observed area \mathcal{A}_{obs}
- 4) build the observed prior point cloud (\mathcal{P}_p^{obs}) and compare \mathcal{P}_p^{obs} and \mathcal{P}_g

Initially, we use the keyframes’ 6D pose and filtered depth image to build a global point cloud. Then we register the global point cloud to the prior point cloud to align the two clouds.

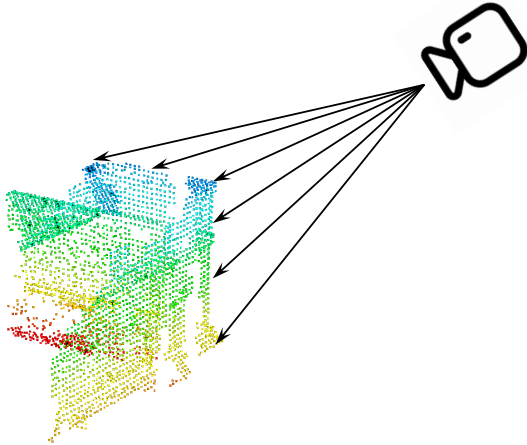


Fig. 2. Illustration of how \mathcal{A}_{obs} is built. For every keyframe, we have a corresponding point cloud and 6D camera pose. We cast rays from camera to point cloud, and the voxels which rays passed are considered a part of \mathcal{A}_{obs}

In step 3, we must know what part of the point cloud has been observed to determine the part of the point cloud that has changed. So we leverage the octomap [27] of resolution ρ_o to build an observed area \mathcal{A}_{obs} . For the pixels in the filtered depth image with a depth smaller than th_d , we cast a ray from the camera to the point. (Fig.2) For those pixels without a depth,

we also cast a ray from the camera with a certain depth th_f to add some free space to octomap. The optimal value of th_f is experimentally defined.

In step 4, we build the observed prior point cloud \mathcal{P}_p^{obs} . Given the depth estimation inaccuracy, we use distance to \mathcal{P}_g to judge whether a point of \mathcal{P}_p is observed. So we consider a point of \mathcal{P}_p is observed when it is in \mathcal{A}_{obs} or is within a distance th_{ch} of \mathcal{P}_g , as shown in Algorithm 1.

Finally, we compare the point clouds by calculating the point-to-cloud distance between \mathcal{P}_g and \mathcal{P}_p^{obs} . If a point in \mathcal{P}_g is not within a distance th_{ch} of \mathcal{P}_p^{obs} , we classify this point as a new point. If a point in \mathcal{P}_p^{obs} is not within a distance th_{ch} of \mathcal{P}_g , we classify this point as a removed point. Finally, a set of new points \mathcal{P}_{new} and a set of removed points \mathcal{P}_{rm} are built, as shown in Algorithm 2.

Algorithm 1: Building the observed prior point cloud

Data: $\mathcal{P}_p, \mathcal{P}_g, \mathcal{A}_{obs}, th_{ch}$
Result: Observed prior point cloud \mathcal{P}_p^{obs}
for $p \in \mathcal{P}_p$ **do**
 if $p \in \mathcal{A}_{obs}$ **or** $calcNearestDist(pt, \mathcal{P}_g) \leq th_{ch}$ **then**
 $\mathcal{P}_p^{obs}.append(p)$

Algorithm 2: Classification of points

Data: $\mathcal{P}_g, \mathcal{P}_p^{obs}$
Result: $\mathcal{P}_{new}, \mathcal{P}_{rm}$
for $p \in \mathcal{P}_g$ **do**
 if $calcNearestDist(p, \mathcal{P}_p^{obs}) \geq th_{ch}$ **then**
 $\mathcal{P}_{new}.append(p)$
for $p \in \mathcal{P}_p^{obs}$ **do**
 if $calcNearestDist(p, \mathcal{P}_g) \geq th_{ch}$ **then**
 $\mathcal{P}_{rm}.append(p)$

IV. DATASET AND METRICS

A. Dataset

To evaluate our change detection pipeline, we build a dataset using the simulated scene in Unreal Engine 4 with the AirSim [28] plugin. The simulated dataset affords quickly changing scenes and collecting data in a simulated world. We have built three scenes in good light condition, namely the original scene and two changed scenes (S1, S2, and in each one we remove a building and add a new building (Fig.3)). Each scene has a size of $250 \times 250 \times 40$ meters (length \times width \times height), and for each scene, we also build its ground truth point cloud using LiDAR. The groundtruth point cloud of original scene is used as prior point cloud.

In each scene, we record 2 MAV’s trajectory (Fig.3) at different height (4m and 15m) with 30 Hz 800×450 stereo pairs, 200 Hz IMU and ground truth 6D pose data. The stereo pairs and IMU are noiseless. All these trajectories have an approximately 170s duration and a length about 700 meters. And we name these trajectories using scene name and height. (e.g. ”S1_01”, ”S2_02”, ”S1” means this trajectory is collected in S1 scene, and ”01” means this trajectory is recorded at a

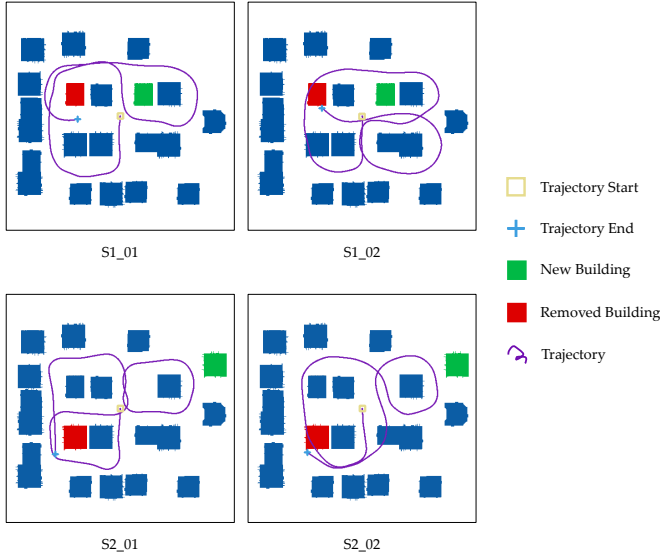


Fig. 3. Top view of scenes and trajectories.

height about 4m, "02" means this trajectory is recorded at a height about 15m)

Except for these 4 trajectories, we also collected 2 trajectories with more difficulties: S1_mirror with mirrors in the scene and S1_dark with bad lighting condition. For S1_mirror, we place one mirror on the outer-wall of one building and the other mirror on the ground of removed building in S1 scene. For S1_dark, we decrease sun light and sun height, so it's much darker than original S1 scene. These two trajectories are with the same trajectory as S1_01.

Algorithm 3: Build clouds from the true positive points

```

Data:  $\mathcal{P}_{p,rm}^{obs}, \mathcal{P}_{p,new}^{obs}, \mathcal{P}_{rm}, \mathcal{P}_{new}$ 
Result:  $\mathcal{P}_{p,rm}^{obs,TP}, \mathcal{P}_{p,new}^{obs,TP}, \mathcal{P}_{rm}^{TP}, \mathcal{P}_{new}^{TP}$ 
for  $p \in \mathcal{P}_{p,rm}^{obs}$  do
    if  $calcNearestDist(p, \mathcal{P}_{rm}) \leq th_{ch}$  then
         $\mathcal{P}_{p,rm}^{obs,TP}.append(p)$ 
for  $p \in \mathcal{P}_{p,new}^{obs}$  do
    if  $calcNearestDist(p, \mathcal{P}_{new}) \leq th_{ch}$  then
         $\mathcal{P}_{p,new}^{obs,TP}.append(p)$ 
for  $p \in \mathcal{P}_{rm}$  do
    if  $calcNearestDist(p, \mathcal{P}_{p,rm}^{obs}) \leq th_{ch}$  then
         $\mathcal{P}_{rm}^{TP}.append(p)$ 
for  $p \in \mathcal{P}_{new}$  do
    if  $calcNearestDist(p, \mathcal{P}_{p,new}^{obs}) \leq th_{ch}$  then
         $\mathcal{P}_{new}^{TP}.append(p)$ 
    
```

B. Metrics

In our pipeline, the changed area is presented as a point cloud. Hence, we consider the change detection problem as a

classification problem of points, and therefore, we introduce some standard classification metrics to the change detection problem.

We employ the following metrics for change detection evaluation:

- 1) Recall of class new: R_{new}
- 2) Precision of class new: P_{new}
- 3) Recall of class removed: R_{rm}
- 4) Precision of class removed: P_{rm}

Besides, we denote $N(\cdot)$ as the number of points in a point cloud.

To evaluate the algorithms on the same standards, all point clouds are downsampled to a certain resolution ρ_p . In our paper, $\rho_p = 0.4m$.

We must know the prior (i.e. groundtruth) new and the removed points to calculate these metrics. Hence, in our change detection algorithm, we use the same algorithm as in Algorithm 2. Specifically, we replace $\mathcal{P}_g, \mathcal{P}_p^{obs}$ with the prior point cloud of the changed and the original scene, and the algorithm outputs prior new points $\mathcal{P}_{p,new}$ and the prior removed points $\mathcal{P}_{p,rm}$ of the changed scene.

Similarly, for the change detection algorithms, we must know which points are observed in the prior new points and prior removed points. To obtain these points, we replace \mathcal{P}_p with $\mathcal{P}_{p,new}$ and $\mathcal{P}_{p,rm}$ in Algorithm 1 and obtain the observed prior new and removed points $\mathcal{P}_{p,new}^{obs}, \mathcal{P}_{p,rm}^{obs}$.

Superscript "TP" marks the point cloud with correctly classified points. For example, $\mathcal{P}_{p,rm}^{TP}$ means points in $\mathcal{P}_{p,rm}$ that are within a distance th_{ch} of $\mathcal{P}_{p,rm}^{obs}$, as presented in Algorithm 3.

Thus the metrics can be written as:

$$R_{new} = \frac{N(\mathcal{P}_{p,new}^{obs,TP})}{N(\mathcal{P}_{p,new}^{obs})}, \quad P_{new} = \frac{N(\mathcal{P}_{new}^{TP})}{N(\mathcal{P}_{new})} \quad (9)$$

$$R_{rm} = \frac{N(\mathcal{P}_{p,rm}^{obs,TP})}{N(\mathcal{P}_{p,rm}^{obs})}, \quad P_{rm} = \frac{N(\mathcal{P}_{rm}^{TP})}{N(\mathcal{P}_{rm})} \quad (10)$$

Besides, we challenge our dataset against other point cloud change detection datasets like the SHREC 2021 [7] and the dataset built by Gélis *et al.* [29]. The corresponding results are reported in Table I.

TABLE I
COMPARISON OF POINT CLOUD CHANGE DETECTION DATASETS

Dataset	Scene type	Sensor
Ours	simulated urban	Stereo camera, IMU
SHREC 2021 [7]	real urban	LIDAR
Gélis <i>et al.</i> [29]	simulated urban	LIDAR

Dataset	Vehicle type	Metrics
Ours	aerial	recall, precision of points' classes
SHREC 2021	ground	accuracy, IoU of objects' classes
Gélis <i>et al.</i>	aerial	IoU of points' classes

V. EXPERIMENTAL RESULTS

Our experiments can be divided into two parts: First, the prior localization test shows our PPCA-VINS's advantage over no prior information-assisted methods. Second, the change detection test demonstrates that our change detection dataset and metrics can effectively evaluate a change detection algorithm and that our visual-based baseline pipeline works successfully on the dataset. In addition, in our experiment we assume that we have an accurate initial guess of the vehicle's pose. All experiments are conducted on a desktop PC with an Intel i9-11900K, an Nvidia RTX 3090, and 32 GB RAM.

A. Prior Localization Test

1) *EuRoC MAV Dataset*: EuRoC MAV Dataset [5] is an indoor dataset containing 20Hz stereo camera images, 200Hz IMU, ground truth 6D pose, and ground truth laser scan data (prior LiDAR map). Hence, this dataset is suitable for the prior localization test. The prior LiDAR map is downsampled to a resolution of 5 cm. We assume that we have an accurate initial pose of MAV, and we simultaneously perform VIO localization and prior map-assisted localization.

Since our work is based on the state-of-the-art SLAM frame VINS-Fusion, we compare our work against the VINS-Fusion. Considering the evaluation metrics, we leverage Evo [30] to calculate the root mean squared error (RMSE) and standard deviation (STD) of the absolute trajectory error (ATE). To avoid randomness during feature tracking and keyframe selection, we perform each method five times on one trajectory and calculate the average RMSE and STD of ATE. Depending on the trajectory, our configuration slightly differs, with the related code available in our Github repository. The corresponding results are reported in Table II.

TABLE II
AVERAGE RMSE (M) AND STD (M) OF ATE ON EUROC FOR
PPCA-VINS AND VINS-FUSION

Trajectory	PPCA-VINS		VINS-Fusion	
	RMSE	STD	RMSE	STD
V1_01_easy	0.1464	0.0197	0.1470	0.0231
V1_02_medium	0.0904	0.0232	0.1170	0.0448
V1_01_difficult	0.1372	0.0393	0.1859	0.0576
V2_01_easy	0.1785	0.1001	0.1790	0.1013
V2_02_medium	0.1603	0.0678	0.1690	0.0760
V2_03_difficult	0.2387	0.1080	0.2445	0.1095

Our PPCA-VINS performs better than VINS-Fusion on every trajectory ranging from easy to difficult. Hence, our method is more appealing than the standard VINS-Fusion affording a higher localization accuracy and a more stable trajectory error. However, our method's performance on trajectory V2_03_difficult is inferior to V1_03_difficult because the former trajectory is more dynamic and recorded in darker light conditions, limiting our point cloud generation scheme and creating more noise in our prior locate result.

2) *Change Detection Dataset*: We also compare our work to the standard VINS-Fusion on our dataset. Unlike the EuRoC dataset, our change detection dataset is collected in outdoor scenes, and thus we use the original scene's point cloud as the prior point cloud in PPCA-VINS. This means that we use an outdated point cloud as prior information.

TABLE III
SMALL AND BIG NOISE PARAMETERS IN EXPERIMENTS

noise	small	big
$\sigma_g (rad \cdot s^{-1/2})$	4.0e-4	4.0e-3
$\sigma_a (m \cdot s^{-3/2})$	3.0e-3	3.0e-2
$\sigma_{bg} (rad \cdot s^{-3/2})$	4.0e-5	4.0e-4
$\sigma_{ba} (m \cdot s^{-5/2})$	3.0e-4	3.0e-3
σ_{img}	0.02	0.04

TABLE IV
AVERAGE RMSE (M), STANDARD DEVIATION (M), AND MAXIMUM (M)
OF ATE ON OUR DATASET FOR PPCA-VINS AND VINS-FUSION

Trajectory/noise	PPCA-VINS			VINS-Fusion		
	RMSE	STD	MAX	RMSE	STD	MAX
S1_01/small	1.0969	0.5316	2.0758	1.7680	0.5029	2.8961
S1_01/big	1.8654	0.8419	4.5178	2.7155	1.0426	5.2554
S1_mirror/small	1.1999	0.6040	2.9738	2.1328	0.8652	4.3606
S1_mirror/big	2.3591	1.1215	4.7216	3.5658	1.3468	6.2925
S1_02/small	1.9786	1.0629	4.0199	3.0335	1.4665	5.8233
S1_02/big	2.6830	1.3906	6.2569	4.2080	1.9974	8.6237
S2_01/small	0.8584	0.4018	1.8421	1.0998	0.5024	2.5601
S2_01/big	1.0623	0.5183	4.1442	2.3007	1.2102	5.4237
S2_02/small	2.0957	1.0231	4.6308	2.8582	1.3670	5.3185
S2_02/big	2.6681	1.4537	6.7778	3.3960	1.6929	7.4251

As our original dataset is noiseless, we add small noise and big noise to images (gaussian noise) and IMU readings (gaussian noise and random walk noise), parameters are shown in Table.III. Each trajectory is tested five times with both small and big noise, and the metrics are averaged over five runs. The corresponding results are presented in Table IV, highlighting that even with an outdated prior point cloud, PPCA-VINS performs much better on each trajectory than the standard VINS-Fusion. Moreover, the results reveal that our method runs successfully in both small noise and big noise conditions, and shows greater robustness than original VINS-Fusion. And for trajectory S1_mirror, our PPCA-VINS's localization accuracy doesn't decrease much, but VINS-Fusion's accuracy has a big fall. Notice that for trajectory "S1_dark", VINS-Fusion failed to run on it, so we will not analyze this trajectory in following sections.

B. Change Detection Test

In this section, we employ our suggested change detection dataset and set th_{ch} to 3.2 m, ρ_o to 0.8 m, and all point clouds are downsampled to 0.8m.

1) *Free Space Ray Casting Test*: As mentioned in III-B, for pixels in filtered depth images that have an empty depth value, we cast a ray from the camera to depth th_f . This section evaluates various th_f depths to determine the optimal value.

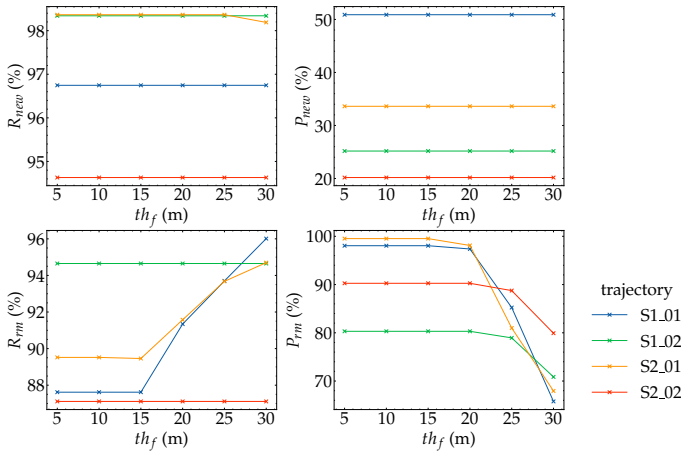


Fig. 4. Recall and precision results of different classes when free space ray casting depth th_f has various values.

We set th_f to 5, 10, 15, 20, 25, and 30 m and calculate the recall and precision metrics in IV-B to determine the best th_f value. This experiment utilizes the 6D pose and the corresponding filtered depth image from PPCA-VINS with small sensor noise. Each metric is averaged over five runs. Fig.4 illustrates the results revealing that th_f has a great influence on P_{rm} .

As presented in the figures above, setting th_f to 15 or 20 meters attains the best performance. Thus, in the subsequent section we set th_f to 20 meters.

2) Comparison of Metrics using Different Pose Estimations:

This section calculates the metrics mentioned above with the 6D poses calculated from our method and the standard VINS-Fusion. The metrics are calculated for every trajectory and are averaged over five runs.

In Fig.5, we illustrate five different conditions when compare global point cloud to prior point cloud, thus making it easier to understand how these metrics are calculated.

Fig.6 highlights that PPCA-VINS affords a better performance on nearly every trajectory and metric, especially on \mathcal{P}_{new} and \mathcal{P}_{rm} , indicating that our dataset and metrics are successful on point cloud change detection tasks. Besides, our visual-based baseline pipeline is verified effective on point cloud change detection tasks, affording a high recall on both new and removed points. Besides, we also found that when big noise is introduced to our trajectories, PPCA-VINS performs more robust than VINS-Fusion. Also, in scene S1_mirror, \mathcal{P}_{new} and \mathcal{P}_{rm} is lower than that of scene without mirrors. This is because images with mirrors will produce wrongly estimated depth, thus making global point cloud has a poor quality.

Moreover, we find the \mathcal{P}_{new} is significantly lower than \mathcal{P}_{rm} , for \mathcal{P}_{new} is heavily influenced by the pose estimation accuracy. Indeed, when the estimated error increases, the re-projected point cloud from the camera to the real world drifts more,

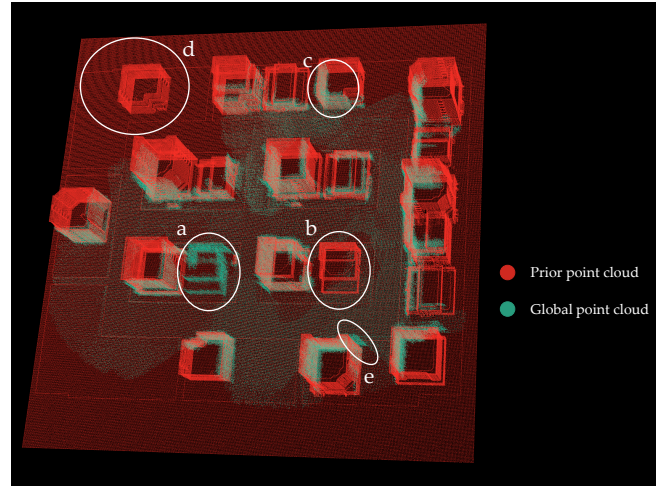


Fig. 5. An illustration of different conditions when compare global point cloud to prior point cloud. a) new building; b) removed building; c) place where global point cloud and prior point cloud are aligned well; d) place where our MAV hasn't observed; e) place where global point cloud and prior point cloud are poorly aligned (Zoom in to have a clear view)

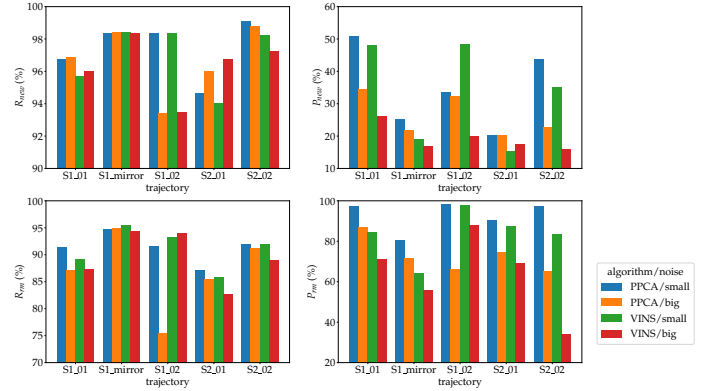


Fig. 6. Comparison of the change detection results using 6D poses from PPCA-VINS and VINS-Fusion.

leading to a wrongly estimated observed area and imposing more re-projected points to be falsely classified as new points. \mathcal{P}_{rm} is related to the prior point cloud and observed area. If the pose error is less than th_{ch} , the prior points in the wrongly observed area will also be within the threshold th_{ch} of our global point clouds. Besides, in our scene, free space is much more than space occupied by buildings. So a minor drift very likely leads vehicles from observing a free area to observing another free space, which has a minor impact on the final metrics.

C. Processing Time

Our baseline framework is processing efficient, as on our PC, PPCA-VINS completes every trajectory of our dataset in 240 seconds (each scene in our dataset has a size of $250m \times 250m \times 40m$). For the change detection pipeline, building point cloud having new and removed points requires less than five minutes per trajectory. Compared to Lidar-based and SfM-based methods, the proposed method has the advantages of

relying on low-cost hardware and affords a low computational complexity.

VI. CONCLUSIONS AND FUTURE WORK

This paper proposes a novel visual-based point change detection pipeline and a simple simulated point change detection dataset and corresponding metrics. Compared to not employing a prior map, our PPCA-VINS achieves a higher pose estimation accuracy and better point cloud change detection results. Our visual-based point cloud change detection pipeline achieves an appealing result on our dataset, and the point cloud change detection metrics represent the corresponding performance efficiently.

We found that change detection results are significantly influenced by localization accuracy. Future work will improve localization accuracy using other kind of map (e.g. feature map). Moreover, we will conduct some real-world indoor tests.

REFERENCES

- [1] J. Hyypä, A. Jaakkola, H. Hyypä, H. Kaartinen, A. Kukko, M. Holopainen, L. Zhu, M. Vastaranta, S. Kaasalainen, A. Krooks, P. Litkey, P. Lyytikäinen-Saarenmaa, L. Matikainen, P. Ronnholm, R. Chen, Y. Chen, A. Kivilahti, and I. Kosonen, "Map updating and change detection using vehicle-based laser scanning," in *2009 Joint Urban Remote Sensing Event*, 2009, pp. 1–6.
- [2] H. Murakami, K. Nakagawa, H. Hasegawa, T. Shibata, and E. Iwanami, "Change detection of buildings using an airborne laser scanner," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 54, no. 2-3, pp. 148–152, 1999.
- [3] R. Qin and A. Gruen, "3d change detection at street level using mobile laser scanning point clouds and terrestrial images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 90, pp. 23–35, 2014.
- [4] Z. J. Yew and G. Hee Lee, "City-scale scene change detection using point clouds," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 13 362–13 369.
- [5] 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>
- [6] 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.
- [7] T. Ku, S. Galanakis, B. Boom, R. C. Veltkamp, D. Banger, S. Gangisetty, N. Stagakis, G. Arvanitis, and K. Moustakas, "Shrec 2021: 3d point cloud change detection for street scenes," *Comput. Graph.*, vol. 99, no. C, p. 192–200, oct 2021. [Online]. Available: <https://doi.org/10.1016/j.cag.2021.07.004>
- [8] T. Qin, S. Cao, J. Pan, and S. Shen, "A general optimization-based framework for global pose estimation with multiple sensors," 2019. [Online]. Available: <http://arxiv.org/abs/1901.03642>
- [9] Epic Games, "Unreal engine." [Online]. Available: <https://www.unrealengine.com>
- [10] X. Zuo, P. Geneva, Y. Yang, W. Ye, Y. Liu, and G. Huang, "Visual-inertial localization with prior lidar map constraints," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 3394–3401, 2019.
- [11] A. I. Mourikis and S. I. Roumeliotis, "A multi-state constraint kalman filter for vision-aided inertial navigation," in *Proceedings 2007 IEEE International Conference on Robotics and Automation*, 2007, pp. 3565–3572.
- [12] M. Magnusson, "The three-dimensional normal-distributions transform: an efficient representation for registration, surface analysis, and loop detection," Ph.D. dissertation, Örebro universitet, 2009.
- [13] S. Lynen, T. Sattler, M. Bosse, J. Hesch, M. Pollefeys, and R. Siegwart, "Get out of my lab: Large-scale, real-time visual-inertial localization," in *Robotics: Science and Systems*, 2015.
- [14] J. Li and G. H. Lee, "Deepi2p: Image-to-point cloud registration via deep classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 15 960–15 969.
- [15] H. Xu, L. Cheng, M. Li, Y. Chen, and L. Zhong, "Using octrees to detect changes to buildings and trees in the urban environment from airborne lidar data," *Remote Sensing*, vol. 7, no. 8, pp. 9682–9704, 2015.
- [16] K. Sakurada, T. Okatani, and K. Deguchi, "Detecting changes in 3d structure of a scene from multi-view images captured by a vehicle-mounted camera," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 137–144.
- [17] C. JST, "Change detection from a street image pair using cnn features and superpixel segmentation," 2015.
- [18] A. Taneja, L. Ballan, and M. Pollefeys, "Image based detection of geometric changes in urban environments," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 2336–2343.
- [19] —, "City-scale change detection in cadastral 3d models using images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 113–120.
- [20] H. Hirschmüller, "Accurate and efficient stereo processing by semi-global matching and mutual information," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2, 2005, pp. 807–814 vol. 2.
- [21] J. Chang and Y. Chen, "Pyramid stereo matching network," *CoRR*, vol. abs/1803.08669, 2018. [Online]. Available: <http://arxiv.org/abs/1803.08669>
- [22] Y. Wang, Z. Lai, G. Huang, B. H. Wang, L. Van Der Maaten, M. Campbell, and K. Q. Weinberger, "Anytime stereo image depth estimation on mobile devices," *arXiv preprint arXiv:1810.11408*, 2018.
- [23] S. Matyunin, D. Vatolin, Y. Berdnikov, and M. Smirnov, "Temporal filtering for depth maps generated by kinect depth camera," in *2011 3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)*, 2011, pp. 1–4.
- [24] P. Besl and N. D. McKay, "A method for registration of 3-d shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, 1992.
- [25] A. Segal, D. Haehnel, and S. Thrun, "Generalized-icp." in *Robotics: science and systems*, vol. 2, no. 4. Seattle, WA, 2009, p. 435.
- [26] K. Koide, M. Yokozuka, S. Oishi, and A. Banno, "Voxelized gicp for fast and accurate 3d point cloud registration," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 11 054–11 059.
- [27] A. Hornung, K. M. Wurm, M. Bennewitz, C. Stachniss, and W. Burgard, "OctoMap: An efficient probabilistic 3D mapping framework based on octrees," *Autonomous Robots*, 2013, software available at <https://octomap.github.io>. [Online]. Available: <https://octomap.github.io>
- [28] S. Shah, D. Dey, C. Lovett, and A. Kapoor, "Airsim: High-fidelity visual and physical simulation for autonomous vehicles," in *Field and Service Robotics*, 2017. [Online]. Available: <https://arxiv.org/abs/1705.05065>
- [29] I. de Gélis, S. Lefèvre, and T. Corpetti, "Change detection in urban point clouds: An experimental comparison with simulated 3d datasets," *Remote Sensing*, vol. 13, no. 13, 2021. [Online]. Available: <https://www.mdpi.com/2072-4292/13/13/2629>
- [30] M. Grupp, "evo: Python package for the evaluation of odometry and slam." <https://github.com/MichaelGrupp/evo>, 2017.