Degeneration-Aware Outlier Mitigation for Visual Inertial Integrated Navigation System in Urban Canyons

Xiwei Bai, Weisong Wen, and Li-Ta Hsu, *Member, IEEE*

***Abstract*—In this article, we proposed a graduated nonconvexity (GNC) aided outlier mitigation method for the improvement of the visual-inertial integrated navigation system (VINS) to face the challenge of dynamic environments with numerous unexpected outlier measurements. A GNC optical flow algorithm was proposed for the detection of the outliers of feature tracking in the front-end of VINS by iteratively estimating the optical flow and the optimal weightings of feature correspondences. Then the feature correspondences with small weightings were excluded. However, excessive outlier exclusion may cause insufficient constraints on the state, causing degeneration of VINS. To solve the problem, this article proposed to detect the potential degeneration based on the degree of constraint in different directions of the pose estimation. Then the number of features being considered was intelligently adapted based on the degeneration level to improve the geometry constraint in the coming epochs. We evaluated the effectiveness of the proposed method by using two challenging datasets (including challenging night scenarios) collected in urban canyons of Hong Kong. The results show that the proposed method can effectively reject the potential outlier visual measurements, and alleviate the degeneration, leading to improved positioning performance in both evaluated datasets.**

***Index Terms*—Graduated nonconvexity (GNC), navigation, optimization method, outlier measurements, urban canyons, visual-inertial integrated navigation system (VINS), visual odometry.**

I. Introduction

THE visual-inertial integrated navigation system (VINS) is widely studied in the past few years aiming to provide accurate state estimation of autonomous systems, e.g., autonomous driving vehicles (ADVs) [1] and unmanned aerial vehicles (UAVs) [2], [3]. Significant achievements have been obtained from the research on the VINS, such as the filtering-based methods, including multistate constraint Kalman filter (MSCKF) [4], robust visual-inertial odometry (ROVIO) [5], and open source for the visual-inertial navigation system (Openvins) [6]. The other research stream is the optimization-based VINS pipelines, including the oriented brief simultaneous localization and mapping (ORB-SLAM3) [7], open keyframe-based visual-inertial SLAM (OKVIS) [8], and monocular visual-inertial systems (VINS-Mono) [9]. The recent work in [10] extensively evaluates the performances of these existing VINS pipelines by using the popular European robotics challenge (EuRoC) datasets [11] with satisfactory illumination conditions and sufficient environment features. According to the conclusion provided in the work [10], if the resource budget of computation for the state estimation is sufficient, VINS-Mono can provide the best accuracy and robustness among all of the evaluated hardware platforms and datasets.

However, the realistic urbanized road scenarios face more challenges, such as unexpected dynamic objects (e.g., moving vehicles and pedestrians) [12]–[14] and motion blur caused by fast vehicle movement [15]. To further study the performance of the VINS in the challenging outdoor urban canyons, we evaluated and analyzed the VINS-Mono [9] based on the datasets collected in urban canyons of Hong Kong. According to the result [16], the accuracy of VINS was significantly decreased in the evaluated urban canyons with the accumulated error reaching 34.21 m in a driving distance of 2.1 km. The main reason accounting for the large errors is that the outliers caused by dynamic objects and motion blur are used for further positioning. Specifically, the existence of the dynamic objects can lead to incorrect feature tracking between consecutive images, thereby resulting in large errors in data association in the back-end optimization of VINS. On the other hand, the motion blur may increase the noise of visual measurements and even fail the feature tracking. Typically, in the front-end of VINS, the optical flow [17] is commonly used to track the feature correspondences between consecutive images. Compared with the descriptor-based feature tracking (e.g., ORB descriptor [7]), the optical flow-based tracking is characterized by lightweight and satisfactory accuracy [9] when the consecutive images are sufficient in texture. Therefore, one of the keys to the performance improvement of VINS in the urban canyon is to isolate the outlier measurements in the feature tracking of the front-end. In this article, we propose a graduated nonconvexity (GNC)-aided optical flow (GNC-OF) for the feature tracking in the front-end of VINS to detect the potential outlier measurements by using a coarse-to-fine process. The detected outlier measurements are then excluded from the back-end optimization of VINS. However, based on our previous work in [12], the excessive exclusion of visual measurements may lead to degeneration of the state

Manuscript received July 19, 2021; revised September 28, 2021; accepted October 22, 2021. Date of publication November 8, 2021; date of current version November 19, 2021. This work was supported by the Project under the Startup Project through The Hong Kong Polytechnic University under Grant BD63. The Associate Editor coordinating the review process for this article was Dr. Shovan Barma. *(Corresponding author: Li-Ta Hsu.)*

The authors are with the Department of Aeronautical and Aviation Engineering, The Hong Kong Polytechnic University, Hong Kong (e-mail: xiwei.bai@connect.polyu.hk; welson.wen@polyu.edu.hk; lt.hsu@polyu.edu.hk).

Digital Object Identifier 10.1109/TIM.2021.3126010

estimation. In view of this, this article proposes a method for the identification of the resulted degeneration by considering the degree of constraint in different directions of pose estimation. Then, the number of features being considered is intelligently adapted based on the degeneration level, thereby improving the geometry constraint in the coming epochs.

The main contributions of this article are listed as follows.

1) This article enables outlier visual measurement detection by using a proposed GNC-OF method without reliance on complicated semantic segmentation. Meanwhile, this article is a continuous work of [13] and enables outlier detection on an epoch-by-epoch basis.
2) This article proposes a novel method for the detection of the degeneration caused by outlier exclusion. Moreover, a solution to alleviate the caused degeneration is proposed.
3) This article validates the effectiveness of the proposed method based on two challenging datasets (including a night scene dataset) collected in urban canyons of Hong Kong.

The rest of this article is organized as follows. Related works are presented in Section II, which are followed by an overview of the proposed method in Section III. The derivation of the proposed GNC-OF is elaborated in Section IV. In Section V, the visual/inertial integration together with the degeneration detection and alleviation are presented. Besides, several real experiments were performed for the evaluation of the effectiveness of the proposed method in Section VI. Finally, the conclusions are drawn, and future work is suggested in Section VII.

II. Related Works

A. Existing Works on Visual Outlier Mitigation

To fill this gap, numerous works [18]–[20] have been done on improving the performance of the VINS in dynamic urban scenarios. It is a straightforward way to detect and remove the features arising from the dynamic objects by using the convolutional neural networks (CNNs), like semantic pixel-wise segmentation (SegNet) [21] and single-shot multibox detector (SSD) [22]. An object detection network SSD [18] was proposed for moving objects detection based on prior knowledge, and the detected dynamic features were removed to guarantee the accurate motion estimation. Additionally, a semantic optical flow simultaneous localization and mapping (SLAM) [20] was proposed to detect dynamic features by using the SegNet, thereby making full use of the feature's dynamic characteristic, and the dynamic features are removed in the optimization module.

Instead of the direct removal of the detected features from dynamic objects, we proposed to remodel the outlier features in [12], and the improved performance is obtained compared with the full removal. However, the studied methods in [12] rely on the accuracy of object detection, and the potential static vehicles detected by CNNs may also be removed. Therefore, a multilevel random sample consensus (ML-RANSAC) algorithm [23] was proposed to solve the problem of discriminating between static and dynamic objects. However, these methods heavily rely on the pretrained network model which could be time-consuming. Moreover, the outlier measurements arising from motion blur cannot be detected or mitigated by using the stream methods.

The other research stream lies in the utilization of the general time-correlated statistical model to detect the potential outlier measurements in the front-end or back-end of VINS. The previous work [13] proposed to adaptively tune the weightings of the visual measurements in the back-end optimization based on the quality of feature tracking in several consecutive epochs. The work argues that the uncertainty of the feature correspondence was highly correlated with the number of times for feature tracking. Moreover, an adaptive M-estimator [24] was proposed in [13] to mitigate the effects of the potential outlier measurements and obtain improved accuracy in the evaluated datasets. However, the improvement of the method relies on the percentage of the outlier measurements in the feature tracking of the front-end and parameter tuning of the adaptive M-estimator. The famous switchable constraint [25] was studied to probabilistically identify the potential outlier measurements inside a combined factor graph optimization (FGO) framework, and an improved result was achieved. However, the result relies heavily on the initial guess of switchable constraints. Recently, the research team from the Massachusetts Institute of Technology proposed a GNC aided robust and global outlier rejection method [26] to efficiently solve the problem of point cloud registration by formulating the robust least-square estimation as the combination of weighted least squares and the outlier process using the Black–Rangarajan duality [27]. The work solves the nonconvexity issue arising from the Geman McClure (GM) function via the GNC and enables the global and optimal estimation of the weightings of corresponding measurements simultaneously. However, a distinct boundary exists between the inlier and outlier measurements in the evaluated dataset, which limits the challenges for detecting the outlier measurements, while its potential in other fields is still needed to be explored. Inspired by the work [26], this article proposed to formulate a GNC-OF for visual outlier mitigation together with a degeneration detection and alleviation method.

B. Conventional Optical Flow for Feature Tracking

Feature tracking plays an important role in determining the performance of data association in the back-end of VINS. The objective of feature tracking is to find the correct feature correspondence between two consecutive frames of images. In general, the solutions to perform feature tracking mainly include two groups, i.e., the descriptor-based [7] and optical flow-based [28] methods. The former, such as the ORB-SLAM3 [7], represents the visual features using the ORB descriptors. Then, the features detected in two consecutive frames are matched based on corresponding descriptors in a one-to-many manner. However, brute descriptor-based matching may result in a high computational load. Different from the descriptor-based feature tracking, the optical flow-based method, such as the state-of-the-art Lucas–Kanade (LK) optical flow [17], track the features directly in a one-to-one manner, which is adopted in many VINS pipelines, such as MSCKF [4], ROVIO [5], Openvins [6], and VINS-Mono [9].

In theory, the traditional LK optical flow works under three key assumptions [17]: 1) Image brightness constancy: the same features within two consecutive images should have the same brightness; 2) Small motion: the features only involve short-term motion; and 3) Spatial smoothness: the pixels within a small window of the given features should have the same movement. Given a feature represented by $I(u, v, t)$, it is detected by using a typical corner-based descriptor [29] where $I(u, v, t)$ denotes the pixel intensity of the pixel (u, v) at time t. When the pixel moves between two consecutive frames over time dt, the corresponding displacement is denoted by (du, dv), which is a quite small movement [17]. Based on the first assumption of LK optical flow, the pixel intensity in two consecutive images satisfies the requirements of the following equation:

$$I(u, v, t) = I(u + du, v + dv, t + dt) \tag{1}$$

where $I(u, v, t)$ and $I(u + du, v + dv, t + dt)$ denote the intensity of the pixel (u, v) at time t and $(t + dt)$, respectively. By applying the first-order Taylor series expansion, the right side of (1) can be formulated as follows [17]:

$$\begin{aligned} &I(u + du, v + dv, t + dt) \\ &\qquad = I(u, v, t) + \frac{\partial I}{\partial u}du + \frac{\partial I}{\partial v}dv + \frac{\partial I}{\partial t}dt \end{aligned} \tag{2}$$

where $(\partial I/\partial u)$ and $(\partial I/\partial v)$ represent the gradient of the pixel intensity concerning u and v, respectively. $(\partial I/\partial t)$ denotes the gradient of the pixel intensity concerning time t. Again, based on the first assumption of LK optical flow, we can get

$$\frac{\partial I}{\partial u}\frac{du}{dt} + \frac{\partial I}{\partial v}\frac{dv}{dt} = -\frac{\partial I}{\partial t}. \tag{3}$$

Hence, the objective of the optical flow [17] is to solve $((du/dt), (dv/dt))$ to determine the pixel displacement over time dt. To simplify, we define $\Delta u = (du/dt)$, $\Delta v = (dv/dt)$, $I_u = (\partial I/\partial u)$, $I_v = (\partial I/\partial v)$, and $I_t = (\partial I/\partial t)$. Then (3) can be rewritten as follows [17]:

$$\begin{bmatrix} I_u & I_v \end{bmatrix}\begin{bmatrix} \Delta u \\ \Delta v \end{bmatrix} = -I_t. \tag{4}$$

There is only one equation but two unknown variables $(\Delta u, \Delta v)^T$, therefore, additional constraints are needed to solve the optical flow problems. To fill this gap, the third assumption of spatial smoothness is proposed [17], which means all neighboring pixels of the detected feature pixel have the same movement. Taking a small window of $n \times n$ around the detected feature (u, v) and referring to the spatial smoothness, all $n \times n$ pixels have the same movement $(\Delta u, \Delta v)^{\mathrm{T}}$. Therefore, there will be $n \times n$ equations similar to (4). The set of equations is represented as follows:

$$\begin{bmatrix} I_{u1} & I_{v1} \\ I_{u2} & I_{v2} \\ \vdots & \vdots \\ I_{ui} & I_{vi} \end{bmatrix}\begin{bmatrix} \Delta u \\ \Delta v \end{bmatrix} = -\begin{bmatrix} I_{t1} \\ I_{t2} \\ \vdots \\ I_{ti} \end{bmatrix}, \quad i \in (1, n \times n) \tag{5}$$

where I_{ui}, I_{vi}, andI_{ti} denote the image gradients (difference of pixel value) along the u, v-axis, and over time t of ith pixel in the small window of the image. $n \times n$ represents the size of the small window. According to (5), there are two unknowns with $n \times n$ equations, which are over-determined. To address the over-determination, the least-squares estimation is used to solve (6) as follows:

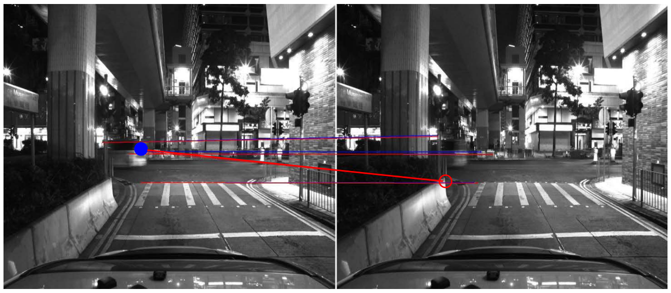


Fig. 1. Example of a failure of feature tracking of optical flow.

$$\begin{gathered} \begin{bmatrix} \Delta u \\ \Delta v \end{bmatrix} = \left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\right)^{-1}\mathbf{A}^{\mathrm{T}}(-b) \\ \text{with } \mathbf{A} = \begin{bmatrix} I_{u1} & I_{v1} \\ I_{u2} & I_{v2} \\ \vdots & \vdots \\ I_{ui} & I_{vi} \end{bmatrix} b = \begin{bmatrix} I_{t1} \\ I_{t2} \\ \vdots \\ I_{ti} \end{bmatrix}. \end{gathered} \tag{6}$$

Specifically, (6) can be further simplified into a compact form, expressed as follows:

$$\begin{bmatrix} \Delta u \\ \Delta v \end{bmatrix} = \begin{bmatrix} \sum_i I_{ui}^2 & \sum_i I_{ui}I_{vi} \\ \sum_i I_{vi}I_{ui} & \sum_i I_{vi}^2 \end{bmatrix}^{-1}\begin{bmatrix} -\sum_i I_{ui}I_{ti} \\ -\sum_i I_{vi}I_{ti} \end{bmatrix}. \tag{7}$$

Therefore, the $[\Delta u \Delta v]^{\mathrm{T}}$ can be estimated by solving (7). Satisfactory accuracy can be obtained by using the LK optical flow in the scenarios with sufficient textures and stable environmental conditions, and the three listed assumptions can be easily satisfied. Unfortunately, its performance significantly deteriorates in the highly dynamic urban canyons with an obvious change in illumination and multiple motions in a single localized region [30] which easily violates the assumptions of spatial smoothness. To increase the robustness of the LK optical flow against the unexpected large motion, the image pyramid aided LK method is proposed, which can separate large motion into small movements. However, the performance of LK optical is still not guaranteed in complex dynamic urban canyons [16].

Fig. 1 shows a scene where the LK optical flow is employed to track the features between two consecutive images collected in an urban canyon during the night. One of the features is located on the car (blue shaded circle) shown in the left figure. We can see that the strong motion blur exists on the car from the left (first) to the right (second) figure. Consequently, the feature is incorrectly tracked to the curb of the road on the right side (as shown by the red circle). To be specific, it is caused by the violation of the first assumption of LK optical flow because the pixel associated with the same pixel is not the same due to the motion blur. Therefore, the incorrect feature tracking may cause large errors in data association of the back-end of VINS. To detect such incorrect feature tracking, and further improve the performance of VINS, this article proposes an outlier-aware GNC optical flow presented in Section III.

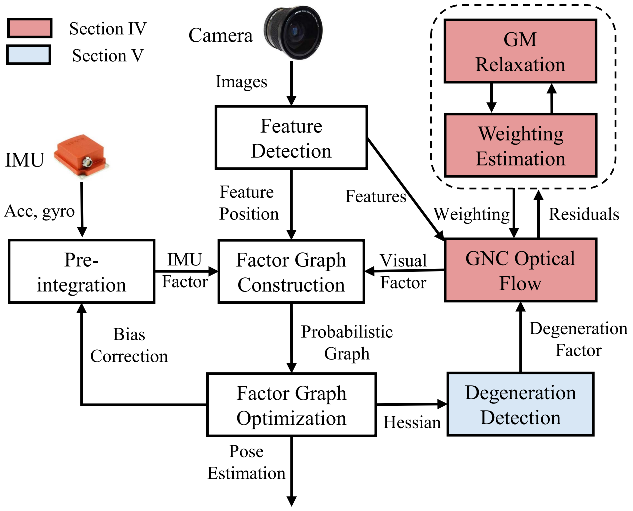


Fig. 2. Overview of the proposed method.

III. Overview of the Proposed Method

The overview of the proposed method is shown in Fig. 2 which is developed on top of the work in [9]. The inputs of the framework are raw images and acceleration as well as gyroscope measurements provided by the monocular camera and inertial measurement unit (IMU), respectively. While the output of the framework is the pose estimation. The framework starts with the measurement preprocessing, including IMU preintegration [31] and feature detection modules [29], presented in Sections V-B and V-C, respectively. These two modules follow the work in [9]. Subsequently, the factor graph construction is derived based on the IMU factor and visual factor, and then the formulation of FGO is presented in Section V-D. The proposed GNC-OF is shown in the red-shaded box (first contribution of this article) in Fig. 2, which enables the removal of the outlier features from the feature detection module. The blue-shaded box indicates the proposed degeneration detection and alleviation method (second contribution of this article). The degeneration factor derived from the degeneration detection module can be further utilized to benefit the alleviation of the degenerated cases in the coming epochs.

To make the presentation clear, in this article, matrices are denoted as uppercase with bold letters, while the vectors are denoted as lowercase with bold letters. Moreover, the variable scalars are denoted as italic letters, and the constant scalars are denoted as lowercase letters.

IV. GNC Optical Flow

A. Problem Formulation

Specifically, (7) can be expressed as an optimization oriented objective function as follows:

$$\min_{\Delta u^*,\Delta v^*} \sum_{i=1}^{n^2}\left(\left\| r\left(\mathbf{\Omega}_{t,i},\begin{bmatrix}\Delta u\\ \Delta v\end{bmatrix}\right)\right\|^2_{\sigma_t^i}\right)$$
$$\text{with } r\left(\mathbf{\Omega}_{t,i},\begin{bmatrix}\Delta u\\ \Delta v\end{bmatrix}\right)=\left(I_{ti}-h\left(\mathbf{\Omega}_{t,i},\begin{bmatrix}\Delta u\\ \Delta v\end{bmatrix}\right)\right) \quad (8)$$

where $\mathbf{\Omega}_{t,i}$ denotes a set of observation measurements associated with the ith pixel inside the window, including the position of the feature in the first image frame, the neighboring pixels, and the next image frame that is required to estimate the optical flow. $[\Delta u^* \ \Delta v^*]^T$ refers to the optimized state that we wish to estimate. σ_t^i stands for the uncertainty associated with the pixel inside the window. n^2 represents the number of observation measurements involved in the window. And the function $h(*)$ denotes the observation function connecting the state and the pixel observation, which can be written as follows:

$$h\left(\mathbf{\Omega}_{t,i},\begin{bmatrix}\Delta u\\ \Delta v\end{bmatrix}\right) = \frac{\partial(I(u+\Delta u, v+\Delta v, t+\Delta t)-I(u,v,t))}{\partial t}. \quad (9)$$

Therefore, the robustified objective function of (8) can be expressed as follows:

$$\min_{\Delta u^*,\Delta v^*} \sum_{i=1}^{n^2}\left(\rho\left(\left\| r\left(\mathbf{\Omega}_{t,i},\begin{bmatrix}\Delta u\\ \Delta v\end{bmatrix}\right)\right\|_{\sigma_t^i}\right)\right) \quad (10)$$

where $\rho(*)$ refers to the applied robust function, i.e., the GM function [32] in this article. According to Black–Rangarajan duality [27], a robust nonlinear least square problem (10) is equivalent to the following decoupled formulation:

$$\min_{\Delta u^*,\Delta v^*,\omega_{t,i}\in\mathcal{W}} \sum_{i=1}^{n^2}\left(\omega_{t,i}\left\| r\left(\mathbf{\Omega}_{t,i},\begin{bmatrix}\Delta u\\ \Delta v\end{bmatrix}\right)\right\|^2_{\sigma_t^i}+\emptyset_\rho(\omega_{t,i})\right) \quad (11)$$

where $\omega_{t,i}$ denotes the weighting for a given pixel measurement from the neighboring window at the epoch t, satisfying $\omega_{t,i} \in [0,1]$. The variable $\mathcal{W}$ is a set of weightings of $\omega_{t,i}$. The function $\emptyset_\rho(\omega_{t,i})$ represents the outlier process that encodes the penalty on the weighing $\omega_{t,i}$, determined by the chosen robust function. Therefore, the unknowns of the system involve $\Delta u^*, \Delta v^*$ and the optimal weighting $(\omega_{t,i})$ of the visual measurements. The solving of (11) is equivalent to the finding of the optimal state estimation of the optical flow and the optimal weightings of pixel measurements to minimize the summation of the residuals. To simplify the derivation in the rest of this article, we represent the weighted residual

$$\left\| r\left(\mathbf{\Omega}_{t,i},\begin{bmatrix}\Delta u\\ \Delta v\end{bmatrix}\right)\right\|_{\sigma_t^i}$$

using $\tilde{r}_{t,i}$.

Typically, the loss function using the GM function [32] for the given error function $\tilde{r}_{t,i}$ corresponding to the ith pixel measurement can be formulated as follows:

$$\Psi(\tilde{r}_{t,i}) = \frac{(c_{\mathrm{GM}})^2(\tilde{r}_{t,i})^2}{(c_{\mathrm{GM}})^2+(\tilde{r}_{t,i})^2} \quad (12)$$

where c_{GM} refers to the parameter that determines the shape of the GM function. Fig. 3 shows the GM loss corresponding to residual $(\tilde{r}_{t,i})$ ranging from $(-30, 30)$ with different c_{GM}. The smaller c_{GM} introduces stronger resistance against the outliers because the impacts of the enormous outliers are mitigated by the low curvature long tail. However, this may lead to a highly nonconvex surface. Consequently, it is hard to globally solve (11) by using typical nonlinear least square estimation. Thus, we formulate the GNC-OF to solve (11) in a coarse-to-fine manner in Section IV-B.

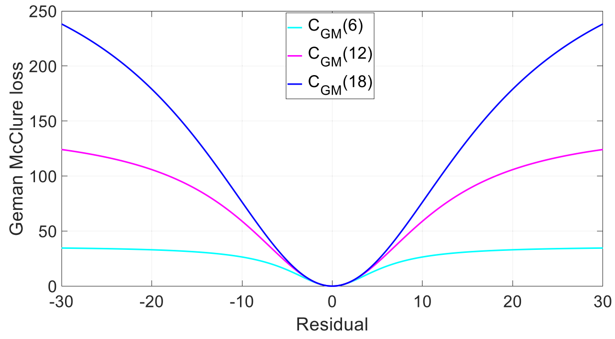


Fig. 3. Illustration of the GM function with different parameters c_{GM} annotated with different colors (cyan: $c_{\text{GM}} = 6$, magenta: $c_{\text{GM}} = 12$, blue: $c_{\text{GM}} = 18$).

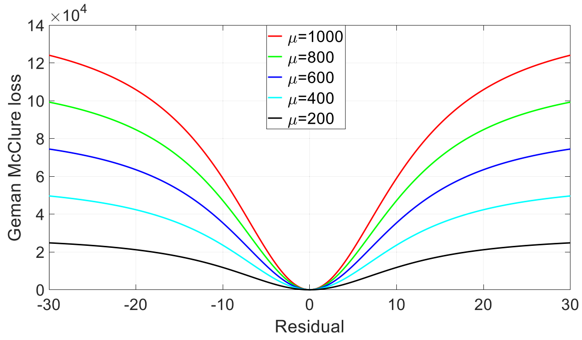


Fig. 4. Illustration of the surrogate function for GM with different control parameters μ annotated with different colors (red: $\mu = 1000$, green: $\mu = 800$, blue: $\mu = 600$, cyan: $\mu = 400$, black: $\mu = 200$).

B. Solution to GNC-OF

The GNC is a popular method for the optimization of a universal nonconvex cost function [26], and the main idea is that a surrogate cost function $\rho_\mu(\cdot)$ is introduced to replace the general nonconvex cost function $\rho(\cdot)$. The new cost function $\rho_\mu(\cdot)$ is convex for a certain μ which changes gradually till the original nonconvex cost function $\rho(\cdot)$ is recovered. During the process, GNC can provide a solution to the nonconvex problem.

According to the selected GM estimator, $\emptyset_{\rho_\mu}(\omega_{t,i})$ is derived as follows:

$$\emptyset_{\rho_\mu}(\omega_{t,i}) = \mu c_{\text{GM}}^2 \left(\sqrt{\omega}_{t,i} - 1\right)^2. \tag{13}$$

As μ tends to $+\infty$, $\rho_\mu(\cdot)$ is convex, and $\rho_\mu(\cdot)$ recovers to be nonconvex as μ decreases and get close to 1, as shown in Fig. 4.

Optimize the GNC-OF problem by alternating the following four steps.

Step 1: Initialization: The variable is initialized by least squares, and the weightings $(\omega_{t,1}, \omega_{t,2}, \ldots, \omega_{t,i})$ are initialized by setting all of them to 1.

Step 2: Variable update: Let weighting $\omega_{t,i}$ be fixed, and optimize $[\begin{smallmatrix}\Delta u\\ \Delta v\end{smallmatrix}]$. Minimize (14) concerning $[\begin{smallmatrix}\Delta u\\ \Delta v\end{smallmatrix}]$

$$\min_{\Delta u^*, \Delta v^*, \omega_{t,i\in\mathcal{W}}} \sum_{i=1}^{n^2} \left(\omega_{t,i}\tilde{r}_{t,i}^2 + \emptyset_{\rho_\mu}(\omega_{t,i})\right). \tag{14}$$

Step 3: Weight update: Let $[\begin{smallmatrix}\Delta u\\ \Delta v\end{smallmatrix}]$ be fixed, and optimize $\omega_{t,i}$ which can then be solved in a closed-form as

$$\omega_{t,i} = \left(\frac{\mu c_{\text{GM}}^2}{\tilde{r}_{t,i}^2 + \mu c_{\text{GM}}^2}\right)^2 \tag{15}$$

where $\tilde{r}_{t,i}$ denotes the residual of pixel value corresponding to the ith pixel.

Step 4: $\mu = (\mu/1.4)$, repeat Steps 2 to 4, until $\mu < 1$.

Therefore, the state $[\Delta u^* \;\; \Delta v^*]^T$ together with the associated weightings set $\mathcal{W}$ are obtained for a certain feature located at $I(u, v, t)$ and $I(u + \Delta u^*, v + \Delta v^*, t + dt)$, respectively. Ideally, the weightings of all the pixels located inside the window reach or get close to 1 if the feature is correctly tracked with all the listed three assumptions satisfied. On the contrary, in the case that most of the weightings are close to 0, the detected feature tends to be the outlier. The recent work in [33] extends their previous work in [26] by using the Chi-square test to find the boundary between the inlier and outlier. On this basis, we set a threshold of weighting to distinguish those outlier pixels as follows:

$$\omega_{t,i} < \omega_{\text{thresh}}, \quad \omega_{t,i} \in \mathcal{W} \tag{16}$$

where ω_{thresh} denotes the threshold of weighting. If $\omega_{t,i}$ is smaller than the threshold, the corresponding pixel is determined to be the outlier pixel. The percentage of such an outlier pixel is accumulated to more than half of all pixels in a small window, and the corresponding detected feature is determined to be the outlier. All the existing features are evaluated by using GNC-OF following the same way, and the detected outliers are excluded from the front-end of VINS.

V. Degeneration-Aware Visual/Inertial Integration

A. System States

In this study, the proposed method is based on VINS [9], and the considered state vector is defined as follows:

$$\begin{aligned}\chi &= \left[\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{x}_c^b, \lambda_1, \lambda_2, \ldots \lambda_M\right]\\ \mathbf{x}_k &= \left[\mathbf{p}_{b_k}^w, \mathbf{v}_{b_k}^w, \mathbf{q}_{b_k}^w, \mathbf{b}_a, \mathbf{b}_g\right], \quad k\epsilon[0, n]\\ \mathbf{x}_c^b &= \left[\mathbf{p}_c^b, \mathbf{q}_c^b\right]\end{aligned} \tag{17}$$

where w denotes the world frame and b_k represents the body (IMU) frame. And $\mathbf{x}_k$ refers to the state of IMU when the kth image is captured. IMU state involves the position, velocity, and orientation, denoted by $\mathbf{p}_{b_k}^w$, $\mathbf{v}_{b_k}^w$, and $\mathbf{q}_{b_k}^w$, respectively, as well as the acceleration bias $(\mathbf{b}_a)$ and the gyroscope bias $(\mathbf{b}_g)$ denoted in the body frame. It should be noted that the orientation is represented by a quaternion, and the coordinate transformation is transformed from the subscript to the superscript frame. n refers to the used keyframes for optimization, and M stands for the sum of features considered for optimization. λ_l refers to the inverse depth of the lth feature observed for the first time, $l \in (1, M)$. $\mathbf{x}_c^b$ represents the transformation matrix that transforms the camera frame to the body frame. In this study, we directly use the extrinsic parameter calibrated previously.

B. IMU Modeling With Preintegration

The IMU measurements involve the acceleration bias ($\mathbf{b}_{a_t}$), the gyroscope bias ($\mathbf{b}_{\omega_t}$) and the additive noise ($\mathbf{n}_a, \mathbf{n}_\omega$). It is worth noting that the noise is assumed to be Gaussian white noise. The raw gyroscope ($\hat{\omega}_t$) and accelerometer ($\hat{a}_t$) measurements modeling is expressed at Epoch t as follows:

$$\hat{\mathbf{a}}_t = \mathbf{a}_t + \mathbf{R}_w^t \mathbf{g}^w + \mathbf{b}_{a_t} + \mathbf{n}_a \tag{18}$$

$$\hat{\boldsymbol{\omega}}_t = \boldsymbol{\omega}_t + \mathbf{b}_{\omega_t} + \mathbf{n}_\omega \tag{19}$$

where $\mathbf{a}_t$ and $\boldsymbol{\omega}_t$ denote the expected measurements of the accelerometer and gyroscope, and the gravity is represented by $\mathbf{g}^w$ in the world frame. $\mathbf{R}_w^t$ stands for the rotation matrix that transforms the world frame into the body frame at Epoch t.

The IMU measurements are utilized to constrain the relative motion between two consecutive epochs. Thanks to the high frequency of the IMU, there are plenty of inertial measurements between the time interval (t_k, t_{k+1}). Therefore, the IMU preintegration technique [31] is employed to integrate the several measurements into a single factor between two consecutive frames of b_k and b_{k+1}. Through the given bias estimation, the IMU preintegration is integrated into the b_k frame as follows:

$$\boldsymbol{\alpha}_{b_{k+1}}^{b_k} = \iint_{t\in[t_k, t_{k+1}]} \mathbf{R}_t^{b_k}\left(\hat{\mathbf{a}}_t - \mathbf{b}_{a_t}\right)dt^2 \tag{20}$$

$$\boldsymbol{\beta}_{b_{k+1}}^{b_k} = \int_{t\in[t_k, t_{k+1}]} \mathbf{R}_t^{b_k}(\hat{\mathbf{a}}_t - \mathbf{b}_{a_t})dt \tag{21}$$

$$\boldsymbol{\gamma}_{b_{k+1}}^{b_k} = \int_{t\in[t_k, t_{k+1}]} \frac{1}{2}\boldsymbol{\Omega}(\hat{\boldsymbol{\omega}}_t - \mathbf{b}_{\omega_t})\boldsymbol{\gamma}_t^{b_k} dt \tag{22}$$

$$\boldsymbol{\Omega}(\omega) = \begin{bmatrix} 0 & -\omega_z & \omega_y & \omega_x \\ \omega_z & 0 & -\omega_x & \omega_y \\ -\omega_y & \omega_x & 0 & \omega_z \\ \omega_x & \omega_y & \omega_z & 0 \end{bmatrix} \tag{23}$$

where ($\boldsymbol{\alpha}_{b_{k+1}}^{b_k}, \boldsymbol{\beta}_{b_{k+1}}^{b_k}, \boldsymbol{\gamma}_{b_{k+1}}^{b_k}$) refer to the preintegration items that denote the change of position, velocity, and orientation, respectively. $\mathbf{R}_t^{b_k}$ and $\boldsymbol{\gamma}_t^{b_k}$ represent the rotation matrix and quaternion, respectively, which transform the body frame at Time t into the reference frame b_k. (ω_x, ω_y, ω_z) stand for the angular velocity in the IMU frame.

Employing the preintegration items, the position, velocity, and orientation of the b_{k+1} in the world frame can be formulated as follows:

$$\mathbf{p}_{b_{k+1}}^w = \left(\mathbf{p}_{b_k}^w + \mathbf{v}_{b_k}^w \Delta t_k - \frac{1}{2}\mathbf{g}^w \Delta t_k^2\right) + \mathbf{R}_{b_k}^w \boldsymbol{\alpha}_{b_{k+1}}^{b_k} \tag{24}$$

$$\mathbf{v}_{b_{k+1}}^w = \left(\mathbf{v}_{b_k}^w - \mathbf{g}^w \Delta t_k\right) + \mathbf{R}_{b_k}^w \boldsymbol{\beta}_{b_{k+1}}^{b_k} \tag{25}$$

$$\boldsymbol{\gamma}_{b_{k+1}}^{b_k} = \mathbf{q}_w^{b_k} \otimes \mathbf{q}_{b_{k+1}}^w \tag{26}$$

where the symbol $\otimes$ refers to the multiplication between two quaternions. Finally, the residual $\mathrm{r}_{\mathcal{B}}(\cdot)$ for IMU preintegration and system states can be formulated as follows:

$$\begin{aligned} &\mathrm{r}_{\mathcal{B}}\left(\hat{Z}_{b_{k+1}}^{b_k}, \chi\right) \\ &= \begin{bmatrix} \delta\boldsymbol{\alpha}_{b_{k+1}}^{b_k} \\ \delta\boldsymbol{\beta}_{b_{k+1}}^{b_k} \\ \delta\boldsymbol{\theta}_{b_{k+1}}^{b_k} \\ \delta\mathbf{b}_a \\ \delta\mathbf{b}_\omega \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{R}_w^{b_k}\left(\mathbf{p}_{b_{k+1}}^w - \mathbf{p}_{b_k}^w + \frac{1}{2}\mathbf{g}^w \Delta t_k^2 - \mathbf{v}_{b_k}^w \Delta t_k\right) - \boldsymbol{\alpha}_{b_{k+1}}^{b_k} \\ \mathbf{R}_w^{b_k}\left(\mathbf{v}_{b_{k+1}}^w + \mathbf{g}^w \Delta t_k - \mathbf{v}_{b_k}^w\right) - \boldsymbol{\beta}_{b_{k+1}}^{b_k} \\ 2\left[\mathbf{q}_{b_k}^{w^{-1}} \otimes \mathbf{q}_{b_{k+1}}^w \otimes \left(\boldsymbol{\gamma}_{b_{k+1}}^{b_k}\right)^{-1}\right]_{xyz} \\ \mathbf{b}_{a,b_{k+1}} - \mathbf{b}_{a,b_k} \\ \mathbf{b}_{\omega,b_{k+1}} - \mathbf{b}_{\omega,b_k} \end{bmatrix} \end{aligned} \tag{27}$$

where $\mathcal{B}$ denotes the set of IMU measurements. $\hat{Z}_{b_{k+1}}^{b_k}$ represents the observation measurements of the IMU between (b_k, b_{k+1}). $\delta\boldsymbol{\alpha}_{b_{k+1}}^{b_k}$, $\delta\boldsymbol{\beta}_{b_{k+1}}^{b_k}$, and $\delta\boldsymbol{\theta}_{b_{k+1}}^{b_k}$ stand for the position, velocity, and orientation residual constraints, respectively. The operator $[\cdot]_{xyz}$ extracts the imaginary part of a quaternion. $\delta\mathbf{b}_a$ and $\delta\mathbf{b}_\omega$ represent the accelerometer and gyroscope biases constraints, respectively.

C. Visual Measurements Modeling

The visual measurement used in our study is a set of features detected by the Shi–Tomasi corner algorithm [29]. In this article, the proposed robust GNC-OF is employed to track the existing features. The number of features and spatial distribution is based on the work of [9] where the maximum number of features is set to 150 to guarantee real-time performance, and the distance between two features is 30 pixels to keep features uniformly distributed. Considering that lth feature is first observed in the eth image, and it is observed again in jth image. Let $(\hat{u}_l^{c_e}, \hat{v}_l^{c_e})$ denote the pixel position of the lth feature in the eth image of camera frame c, and let $(\hat{u}_l^{c_j}, \hat{v}_l^{c_j})$ denotes the pixel position of the lth feature in the jth image of camera frame c. Then the expected observation of the lth feature in the jth image is derived as follows:

$$\begin{bmatrix} x^{c_j} \\ y^{c_j} \\ z^{c_j} \\ 1 \end{bmatrix} = \left(\mathbf{T}_c^b\right)^{-1}\left(\mathbf{T}_{b_j}^w\right)^{-1}\mathbf{T}_{b_e}^w \mathbf{T}_c^b \pi_c^{-1}\frac{1}{\lambda_l}\begin{bmatrix} \hat{u}_l^{c_e} \\ \hat{v}_l^{c_e} \\ 1 \\ \lambda_l \end{bmatrix}. \tag{28}$$

Equation (28) follows the pinhole camera projection model [34]. $(x^{c_j}, y^{c_j}, z^{c_j})^{\mathrm{T}}$ is the 3-D coordinates of the lth feature in the jth camera frame c. b denotes the body frame. b_e and b_j denote the eth and the jth body frame, respectively. $\mathbf{T}_c^b$ is the transformation matrix that transforms the camera frame into the body frame. Similarly, $\mathbf{T}_{b_e}^w$, $\mathbf{T}_{b_j}^w$, and $\mathbf{T}_c^b$ transform the coordinates of the subscript to the superscript one. π_c is the camera projection function, which is related to camera intrinsics, and λ_l denotes the inverse depth of the lth feature in the eth image.

The $\mathbf{T}$ is the transformation matrix including translation matrix $\mathbf{p}$ and rotation matrix $\mathbf{R}$. Therefore, (28) can further be formulated as

$$\begin{bmatrix} x^{c_j} \\ y^{c_j} \\ z^{c_j} \end{bmatrix} = \mathbf{R}_b^c\left(\mathbf{R}_w^{b_j}\left(\mathbf{R}_{b_e}^w\left(\mathbf{R}_c^b\frac{1}{\lambda_l}\pi_c^{-1}\left(\begin{bmatrix}\hat{u}_l^{c_e}\\ \hat{v}_l^{c_e}\end{bmatrix}\right)+\mathbf{p}_c^b\right)+\mathbf{p}_{b_e}^w-\mathbf{p}_{b_j}^w\right)-\mathbf{p}_c^b\right). \tag{29}$$

Let $p_l^{c_j}$ denote the 3-D coordinates $(x^{c_j}, y^{c_j}, z^{c_j})^{\mathrm{T}}$

$$\bar{p}_l^{c_j} = \frac{p_l^{c_j}}{\|p_l^{c_j}\|} \tag{30}$$

where $\bar{p}_l^{c_j}$ is the expected observation in the normalized plane. Let the observation measurement of the lth feature in the jth image be $\hat{\bar{p}}_l^{c_j}$

$$\hat{\bar{p}}_l^{c_j} = \pi_c^{-1}\left(\begin{bmatrix}\hat{u}_l^{c_j}\\ \hat{v}_l^{c_j}\end{bmatrix}\right). \tag{31}$$

Hence, the residual model of the reprojection can be derived as follows:

$$\mathrm{r}_{\mathcal{C}}\left(\hat{Z}_l^{c_j}, \chi\right) = \left(\hat{\bar{p}}_l^{c_j} - \bar{p}_l^{c_j}\right) \tag{32}$$

where $\mathcal{C}$ denotes the set of features that have been observed at least twice, $\mathrm{r}_{\mathcal{C}}(\cdot)$ represents the residual of the lth feature measurement between the two images, and $\hat{Z}_l^{c_j}$ denotes the measurement of the observation in the jth image.

D. Factor Graph Optimization

The goal of FGO [35] is to minimize the sum of all sensor measurement residuals to achieve a maximum posterior estimation. The residuals in this article contain three parts: 1) the residual from marginalization; 2) the residual from IMU preintegration; and 3) the residual from the visual reprojection, consequently the objective function of the system can be formulated as follows:

$$\min_{\chi}\left\{\left\|\mathrm{r}_p - \mathrm{H}_p\chi\right\|^2 + \sum_{k\epsilon\mathcal{B}}\left\|\mathrm{r}_{\mathcal{B}}\left(\hat{Z}_{b_{k+1}}^{b_k}, \chi\right)\right\|^2_{\mathrm{P}_{b_{k+1}}^{b_k}} + \sum_{(l,j)\epsilon\mathcal{C}}\rho\left(\left\|\mathrm{r}_{\mathcal{C}}\left(\hat{Z}_l^{c_j}, \chi\right)\right\|^2_{\mathrm{P}_l^{c_j}}\right)\right\} \tag{33}$$

where $\{\mathrm{r}_p, \mathrm{H}_p\}$ is the prior information from the marginalization operation [36]. Since the sliding window optimization technique is adopted in the system, the marginalization operation is introduced to convert the marginalized states into a prior. $\mathrm{r}_{\mathcal{B}}(\cdot)$ and $\mathrm{r}_{\mathcal{C}}(\cdot)$ are residuals for IMU and visual measurements, respectively. The detailed information on the residuals is presented in Sections V-B and V-C. $\mathrm{P}_{b_{k+1}}^{b_k}$ and $\mathrm{P}_l^{c_j}$ are the information matrix of IMU measurement and visual reprojection residuals. $\rho(\cdot)$ denotes the robust M-estimator [37], and Huber is adopted here. l denotes the lth feature, and c_j denotes the jth camera frame.

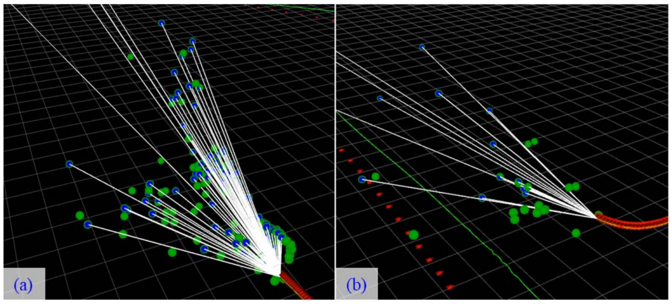


Fig. 5. Illustration of the visual landmarks distribution. The green circles denote the position of the landmark. The white lines denote the connection between the camera and the landmarks. (a) State estimation is constrained by more and decentralized visual landmarks. (b) State estimation is constrained by fewer and centralized visual landmarks.

E. Degeneration Detection and Alleviation

While the rejection of the outlier can help to improve the overall system performance by mitigating the impacts of incorrect features correspondence association, however, this can result in a new degeneration problem. Theoretically, the pose of the system is mainly constrained by the visual landmarks. More features normally lead to stronger constraints on the state estimation. Moreover, more decentralized visual landmark distribution also leads to better constraints [13]. Fig. 5(a) shows the scene with constraints from decentralized visual landmarks. Conversely, Fig. 5(b) shows the case in which very limited visual landmarks are available as constraints to the system after the outlier rejection.

1) Jacobian Formulation: Theoretically, the constraint between the visual landmarks and the state of the system is connected by the Jacobian matrix of the visual reprojection residual concerning the $\mathrm{r}_{\mathcal{C}}(\hat{Z}_l^{c_j}, \chi)$. Therefore, the work in [38] proposed the detection of the potential degeneration via the Jacobian matrix. Given the reobserved lth feature in b_j, the reprojection error is associated with two frames b_e and b_j, then the Jacobian of the lth feature can be derived as follows:

$$\mathbf{H}_{j,l}^e = \begin{bmatrix} \dfrac{\partial \mathrm{r}_{\mathcal{C}}^l}{\partial\delta\mathbf{p}_{b_e}^w} & \dfrac{\partial \mathrm{r}_{\mathcal{C}}^l}{\partial\delta\mathbf{q}_{b_e}^w} \\ \dfrac{\partial \mathrm{r}_{\mathcal{C}}^l}{\partial\delta\mathbf{p}_{b_j}^w} & \dfrac{\partial \mathrm{r}_{\mathcal{C}}^l}{\partial\delta\mathbf{q}_{b_j}^w} \end{bmatrix} \tag{34}$$

where $\mathrm{r}_{\mathcal{C}}^l$ denotes the reprojection residual of the lth feature between frames b_e and b_j. Specifically, the Jacobian component for the position and orientation of the frame b_e can be expressed as follows [9]:

$$\frac{\partial \mathrm{r}_{\mathcal{C}}^l}{\partial\delta\mathbf{p}_{b_e}^w} = \mathbf{R}_b^c\mathbf{R}_w^{b_j} \tag{35}$$

$$\frac{\partial \mathrm{r}_{\mathcal{C}}^l}{\partial\delta\mathbf{q}_{b_e}^w} = -\mathbf{R}_b^c\mathbf{R}_w^{b_j}\mathbf{R}_{b_e}^w\left(\mathbf{R}_c^b\frac{1}{\lambda_l}\hat{\bar{p}}_l^{c_e} + \mathbf{p}_c^b\right)^{\wedge}$$

$$\text{with } \hat{\bar{p}}_l^{c_e} = \pi_c^{-1}\left(\begin{bmatrix}\hat{u}_l^{c_e}\\ \hat{v}_l^{c_e}\end{bmatrix}\right). \tag{36}$$

Similarly, the Jacobian component for the position and orientation of the frame b_j is as follows:

$$\frac{\partial \mathrm{r}_{\mathcal{C}}^{l}}{\partial \delta \mathbf{p}_{b_j}^{w}} = -\mathbf{R}_b^c \mathbf{R}_w^{b_j} \tag{37}$$

$$\frac{\partial \mathrm{r}_{\mathcal{C}}^{l}}{\partial \delta \mathbf{q}_{b_j}^{w}} = \mathbf{R}_b^c. \tag{38}$$

Therefore, the combined Jacobian matrix considering all the visual constraints associated with the current epoch $\mathbf{x}_n = [\mathbf{p}_{b_n}^w, \mathbf{q}_{b_n}^w]$ can be formulated as follows:

$$\mathbf{H}_{\mathcal{C}} = \begin{bmatrix} \mathbf{H}_{j,0}^e \\ \vdots \\ \mathbf{H}_{j,E}^e \end{bmatrix} \tag{39}$$

where $\mathbf{H}_{\mathcal{C}}$ denotes the Jacobian of all the reobserved features at the current epoch. The E denotes the number of constraints associated with the current (latest) epoch $\mathbf{x}_n$. The size of the $\mathbf{H}_{\mathcal{C}}$ is $2E \times 6$. Note that we only considered the degeneration in the position and the orientation estimation, since the other states are also associated with the position or orientation.

2) Degeneration Detection and Alleviation: To further identify the level of constraints in the given measurements, the eigenvalue of the associated Jacobian matrix is employed as an indicator in both the global navigation satellite systems (GNSS) [39] field, and the Robotic field [38]. Recently, the research team from Carnegie Mellon University robotics institute proposed to use the associated eigenvalues in the evaluation of the degeneracy of the system built by visual and light detection and ranging (LiDAR), and the experimental results showed an improvement in the robustness [38]. The work in [38] argued that degeneration occurs when the minimum eigenvalues of the $\mathbf{H}_{\mathcal{C}}$ is smaller than a given threshold λ_{thresh}. However, there is difficulty in adapting a certain value of the λ_{thresh} to different scenarios. For example, a given λ_{thresh} can be suitable for an indoor scenario, while its usability in outdoor scenarios is limited. To fill this gap, we proposed the evaluation of both the minimum eigenvalue and the ratio between the maximum and the minimum eigenvalues.

Given a matrix $\mathbf{H}_{\mathcal{C}}$, the singular value decomposition (SVD) [40] can be expressed as follows:

$$\mathbf{H}_{\mathcal{C}}^{\mathbf{T}} \mathbf{H}_{\mathcal{C}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{40}$$

where the matrix $\mathbf{U}$ is a real 6×6 orthogonal matrix. Meanwhile, the $\mathbf{V}$ is a real 6×6 orthogonal matrix. The matrix $\mathbf{\Sigma}$ is a real 6×6 diagonal matrix with nonnegative real numbers on the diagonal. The diagonal entries $\lambda_s = \mathbf{\Sigma}_{ss}$ are considered to be the eigenvalues. The s denotes the index of the six eigenvalues associated with the position and orientation, as follows:

$$\boldsymbol{\lambda} = \left[\, \lambda_1 \ \lambda_2 \ \lambda_3 \ \lambda_4 \ \lambda_5 \ \lambda_6 \,\right]^T. \tag{41}$$

Therefore, degeneration is detected if $\lambda_{\min}$ is smaller than an experimentally determined threshold λ_{thresh} or the ratio $(\lambda_{\max}/\lambda_{\min})$ is larger than a given threshold λ_{ratio}. The $\lambda_{\min}$ and $\lambda_{\max}$ denote the minimum and maximum eigenvalues within the $\boldsymbol{\lambda}$, respectively. Therefore, the degeneration detection above considers both the absolute and relative values involved in the eigenvalues. Compared with the single $\lambda_{\min}$ considered, the benefits of the introduced ratio is to avoid $\lambda_{\max}$ even smaller than λ_{thresh} in some extreme conditions.

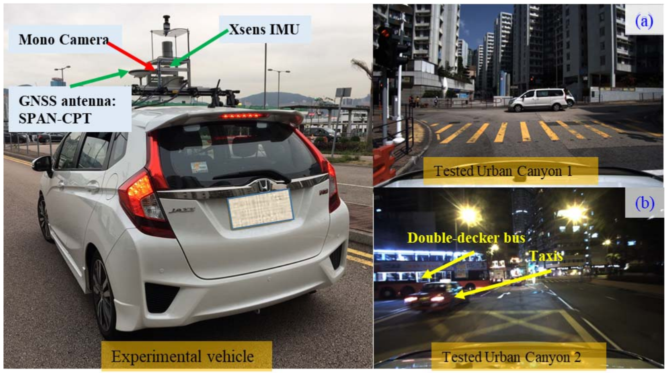


Fig. 6. Experimental setup and the evaluated scenes (a) and (b).

To alleviate the degeneration of the system arising from the removal of the outlier, we propose to adaptively increase the number of features based on the degeneration levels associated with related eigenvalues. Considering that the minimum eigenvalue is a powerful indicator of degeneration, we propose to define the level of degeneration as follows:

$$D_\lambda = \|\lambda_{\min} - \lambda_{\text{thresh}}\|, \quad \text{with } \lambda_{\min} < \lambda_{\text{thresh}} \tag{42}$$

where the D_λ denotes the degeneration factor encoding the level of degeneration. Larger D_λ means that stronger degeneration occurs and vice versa. Then the total number of features to be detected and tracked will increase in the next epoch as follows:

$$N_f^* = N_f + \frac{D_\lambda}{10}, \quad \text{with } \lambda_{\min} < \lambda_{\text{thresh}} \tag{43}$$

where the N_f^* denotes the total number of features after adaptively increasing, and N_f denotes the number of features remaining after the removal of outliers. Therefore, the degeneration will be alleviated in the subsequent epochs after the addition of more features. Fortunately, the additional features can easily be detected in an outdoor environment, and these features are also extracted using the Shi–Tomasi corner algorithm, and the distance from the existing features is set to 30 pixels to keep the features uniformly distributed.

VI. Experiment Results and Discussion

A. Experiment Setup

1) Experimental Scenes: Two real datasets were collected in typical urban canyons of Hong Kong to verify the feasibility of the proposed method in this article. All the data are postprocessed and the experimental sensor setup is presented on the left side of Fig. 6. Fig. 6(a) and (b) illustrates the scenes of the tested urban canyons. A commercial level Xsens MTi 10 IMU sensor was utilized in the collection of raw IMU data at a frequency of 200 Hz. The monocular camera was used to collect raw images at a frequency of 10 Hz. The ground truth of the pose estimation was provided by the NovAtel SPAN-CPT, which is a GNSS (GPS, GLONASS, and BeiDou) real-time kinematic (RTK)/inertial navigation system (INS) with fiber-optic gyroscopes integrated navigation system. In addition, the well-known *Inertial Explorer* software [41] was used to postprocess the data from NovAtel SPAN-CPT to maximize the accuracy of the ground truth of positioning. All the collected measurements were recorded and synchronized based on the timestamp provided by the robot operation system (ROS) [42] platform. The baseline distance between the rover and the GNSS base station is about 7 km. The intrinsic parameters of the camera and the extrinsic parameters between the applied camera and the IMU sensor are calibrated based on the recommendation of [43]. Different from the extensively evaluated EuRoC dataset [11] which was mainly collected in indoor scenarios, the applied datasets (even includes a night scene) collected from urban canyons in this article comprises numerous dynamic objects and unstable illumination conditions, which can cause numerous unexpected outlier visual measurements. To benefit the research community, we open-sourced the evaluated dataset [44] in this article.

2) Experimental Parameters: We set the threshold λ_{thresh} to an experimentally determined value of 200 based on our recently published urbanNav dataset [45]. The ω_{thresh} is set to 0.5.

To stepwise verify the contributions of the proposed method, several methods were compared as follows.

1) **VINS-Mono [9]:** The original VINS solution from [9].
2) **ORB-SLAM3 [7]:** The VINS solution from [7] where the ORB features are employed for visual feature detection and association.
3) **VINS-ac-ME [13]:** VINS aided by adaptive covariance estimation and adaptive M-estimator proposed in our previous work [13].
4) **VINS-GNC-OF:** The original VINS solution from [9] is aided by the visual outlier rejection in the front end using the proposed GNC in this article. This is to verify the first contribution of this article.
5) **VINS-DAOM:** The proposed degeneration-awareness outlier mitigation for VINS in this study. Note that the proposed optical flow, **GNC-OF**, is included in the front-end of this method.

The improvement from the VINS-ac-ME compared with the original VINS-Mono for the positioning estimation has been extensively studied in our previous work [13], thus we present the results of the VINS-Mono, and VINS-ac-ME directly. In this article, we analyzed the proposed method from two parts: the outlier mitigation in the front-end and the degeneration-awareness in the back-end. Interestingly, we combined the geometry of the visual feature distribution and the quality of the visual feature tracking to estimate the uncertainty of visual measurements to further mitigate the effects of outlier measurements in the previous work [13], while we aim to dive into the fundamental problem of optical flow for feature tracking in this study by proposing the GNC-OF detection of outliers and the mitigation their effects for positioning estimation.

To evaluate the experimental results, we used the evaluation of odometry and SLAM (EVO) [46] tool, which is extensively used for the SLAM algorithms. The mean error is defined by the relative pose error (RPE) in the EVO. Besides, the final total positioning error is provided, which is calculated by the final epoch of the positioning error, denoted by FPE. The experimental results are evaluated in the local frame, and the first frame is regarded as a reference frame.

TABLE I
POSITIONING PERFORMANCE OF THE LISTED METHODS IN URBAN CANYON 1

Items	VINS-Mono	ORB-SLAM3	VINS-AC-ME	VINS-GNC-OF	VINS-DAOM
MEAN (m)	0.71	0.86	0.71	0.45	0.40
FPE (m)	86.09	71.52	65.38	51.63	51.63
STD (m)	0.98	2.26	0.86	0.54	0.46
Max (m)	4.03	23.82	3.88	3.02	3.02
Improvement%			0%	36.6%	43.6%

B. Experimental Evaluation in Urban Canyon 1

1) Positioning Performance Analysis: The first experiment is conducted in a typical urban canyon (Whampoa in Hong Kong) to verify the performance of the proposed method. The positioning results are listed in Table I. With the help of the proposed degeneration-awareness and outlier mitigation method, the mean error decreased from 0.71 to 0.40 m, and the standard deviation (STD) also dropped to 0.46 m. Interestingly, we found that the proposed optical flow method can significantly improve the performance when compared to the previous method and VINS-Mono results and ORB-SLAM3 results, also there was a slight improvement in performance due to further degeneration awareness and mitigation. To further validate our proposed method, another experiment is conducted in a more challenging environment.

The trajectories of the listed methods and the ground truth trajectory are shown in Fig. 7. The length of the trajectory is 546.131 m. The trajectory of the proposed method (blue curve) is the closest to the reference trajectory (black curve). In contrast, the trajectory of the ORB-SLAM3 (cyan curve) has the highest deviation from the reference point. The positioning error of the listed methods is shown in Fig. 8. There is a significant improvement in the accuracy of the proposed between epoch 0 and epoch 50.

2) Rotation Performance Analysis: Table I shows that there is a significant improvement in positioning accuracy using the proposed method. To further validate the effectiveness of the proposed method in improving the rotational accuracy, the performance comparisons are shown in Table II. Interestingly, the mean errors of rotation from the listed methods are almost the same except for ORB-SLAM3. We found that the initialization of ORB-SLAM3 is not stable, and its drift is heavy in urban canyon 1. Therefore, we take the VINS-Mono methods as the baseline, which is more robust. The maximum

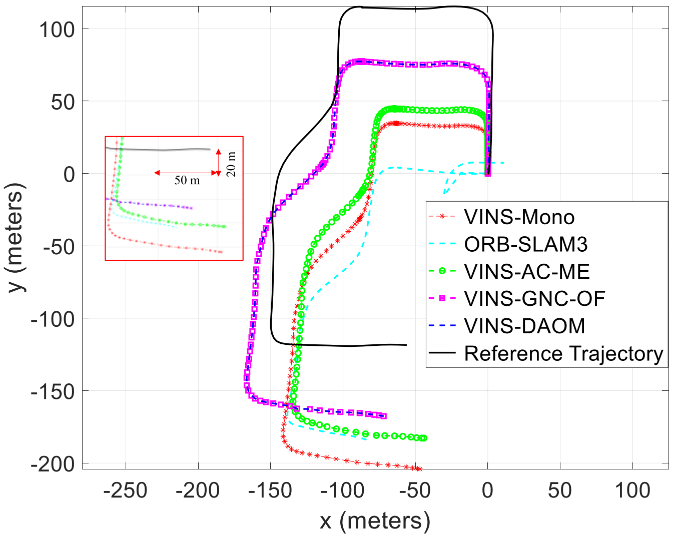


Fig. 7. Estimated trajectories of the VINS-Mono and the listed methods and reference trajectory in urban canyon 1.

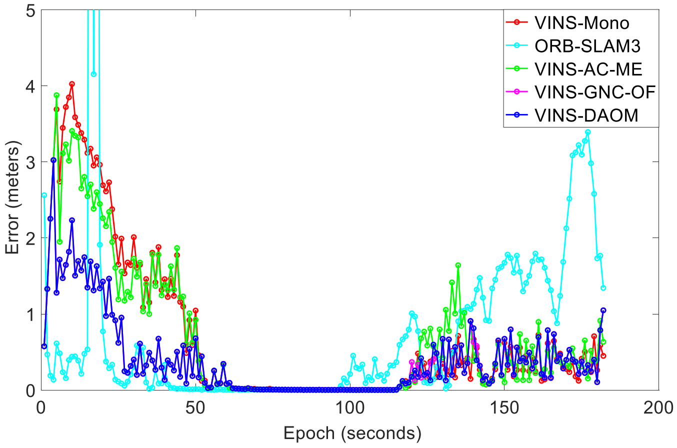


Fig. 8. Positioning errors of the listed methods in urban canyon 1.

TABLE II

ROTATION PERFORMANCE OF THE LISTED METHODS IN URBAN CANYON 1

Items	VINS-Mono	ORB-SLAM3	VINS-AC-ME	VINS-GNC-OF	VINS-DAOM
MEAN (°)	0.89	2.04	0.84	0.89	0.87
FPE (°)	8.42	255.98	7.59	7.46	7.46
STD (°)	0.94	11.09	0.85	0.98	0.90
Max (°)	4.81	119.86	4.77	6.79	4.80
Improvement%			4.82%	0.22%	2.13%

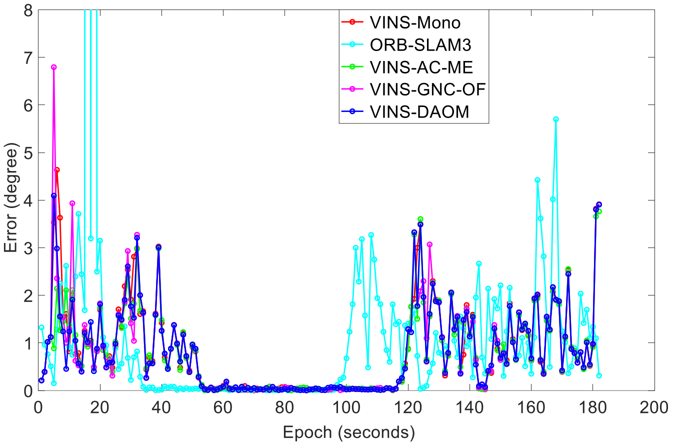


Fig. 9. Rotation errors of the listed methods in urban canyon 1.

TABLE III

POSITIONING PERFORMANCE OF THE LISTED METHODS IN URBAN CANYON 2

Items	VINS-Mono	ORB-SLAM3	VINS-AC-ME	VINS-GNC-OF	VINS-DAOM
MEAN (m)	0.79	Fail	0.59	0.54	0.52
FPE (m)	38.81	Fail	81.79	36.91	37.20
STD (m)	0.96	Fail	0.75	0.60	0.58
Max (m)	5.58	Fail	7.26	3.51	3.94
Improvement%			25.3%	31.6%	34.2%

value increases from 4.81° to 6.79° after the detected outliers are removed based on the proposed GNC-OF in the front-end, and this change means that the removal of excessive outliers can also lead to degeneration in rotation. The maximum error drops to 4.80° from 6.79° based on the proposed degeneration alleviation method, and the improvement can also be seen in Fig. 9. The rotation error of VINS-DAOM denoted by the blue curve declined compared to the VINS-GNC-OF curve denoted by magenta during the first 20 epochs. Therefore, the supplemented features based on (43) can effectively provide more constraints in the alleviation of the degenerated epoch.

Generally, the improvement in the rotation estimation is limited after using the proposed method. On the one hand, the rotation usually offers better constraints with the help of the gyroscope sensor, which is significantly higher in accuracy than the accelerometer inside the employed IMU sensor. Moreover, the pitch and the roll angle are globally observable [9] which further enhances the accuracy of the rotation estimation. Thus, the partial outlier visual measurement removal does not necessarily lead to the degeneration of the rotation estimation [38].

C. Experimental Evaluation in Urban Canyon 2

1) Positioning Performance Analysis: To validate the reliability of the proposed method, another experiment is conducted in urban canyon 2 (Tsim Sha Tsui in Hong Kong) during the night, the scene incorporated numerous dynamic objects

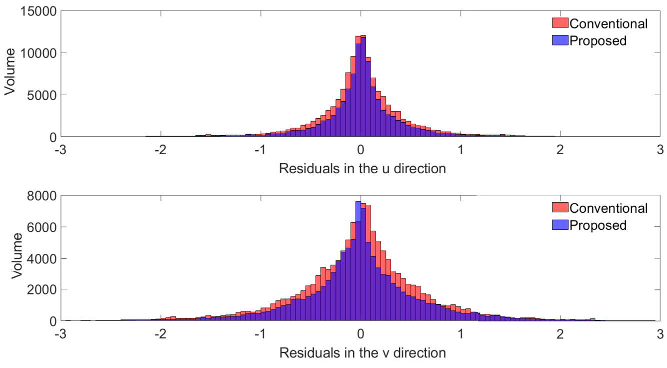


Fig. 10. Residuals of visual reprojection in the u- and v-directions of conventional (VINS-Mono) and the proposed method (VINS-GNC-OF).

and unstable illumination conditions. The positioning results for the listed methods are shown in Table III. The mean error of VINS-Mono is 0.79 m, with the maximum error reaching 5.58 m. Based on the previous work (VINS-ac-ME), the mean error decreases to 0.59 m. The improvement can reach 25.32%. In the previous work, we focused on the visual measurement model based on the quality of the feature tracking to improve the performance of VINS in urban canyons, and thus in this study, we continue to explore the quality of feature tracking. The mean error of the proposed optical flow VINS-GNC-OF decreased to 0.54 m and the maximum error dropped to 3.51 m. Furthermore, by increasing the features in the back-end of the VINS, the mean error further decreased to 0.52 m compared to the 0.79 m of the VINS-Mono, with an improvement of 34.2%, and the maximum error decreased to 3.94 m. The STD was also reduced to 0.58 m.

Typically, the outlier visual measurements usually involve larger residuals. To further elaborate on the reason behind the improvement of the proposed GNC-OF in improving the VINS through the rejection of the visual measurement outlier, we present the residuals of the visual reprojection in the back-end of the VINS corresponding to the conventional VINS and the GNC-OF aided VINS as shown in Fig. 10. The top and bottom figures show the residuals in u- and v-directions, respectively. The top of Fig. 10 shows that the majority of the residuals lie within -3 to 3. With the help of the GNC-OF, the histogram tends to be thinner with a smaller STD which shows the effectiveness of the proposed method in rejecting the visual measurements outliers with larger residuals. A similar phenomenon can be found in the v-direction as shown at the bottom of Fig. 10.

The trajectories of the listed methods and reference trajectory are shown in Fig. 11. The total length of the trajectory in urban canyon 2 is about 1984.448 m. The trajectory of the proposed method VINS-DAOM (blue curve) is the closest to the reference trajectory (black curve). The positioning error of the listed methods is shown in Fig. 12. Thus, improved performance in positioning is obtained by the proposed method (blue line) compared to the original VINS-Mono (red line). Since the VINS can only provide the relative pose estimation continuously, the smaller attitude estimation can lead to significant drift in the long term, as shown by the green curve in Fig. 11. To mitigate the overall drift in VINS, one promising solution is to integrate the globally referenced GNSS positioning and the locally smooth estimation from VINS, and this will be the focus of one of our future works.

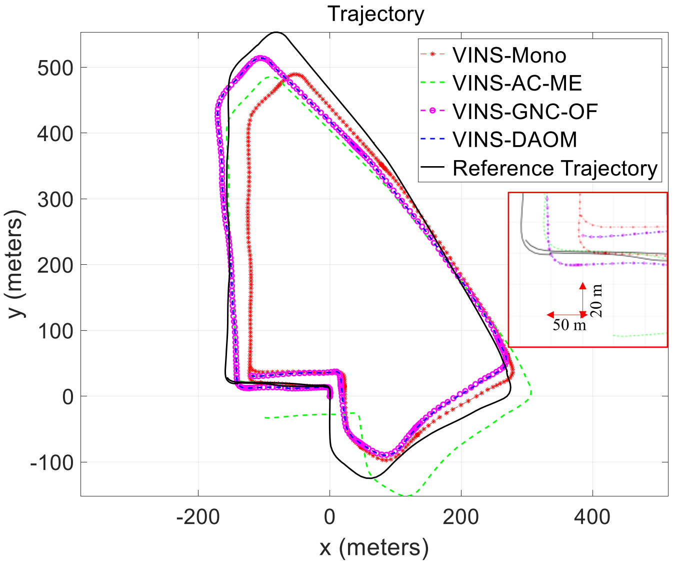


Fig. 11. Estimated trajectories of the VINS-Mono and the listed methods and reference trajectory in urban canyon 2.

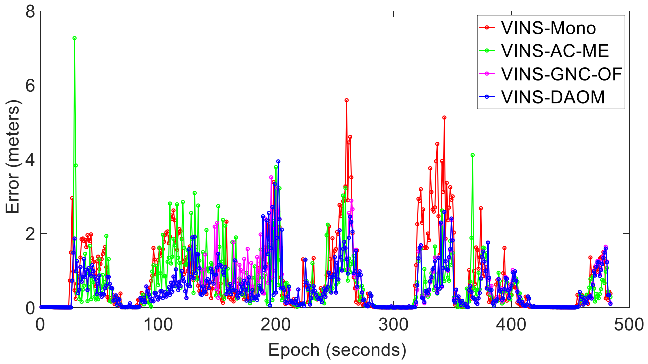
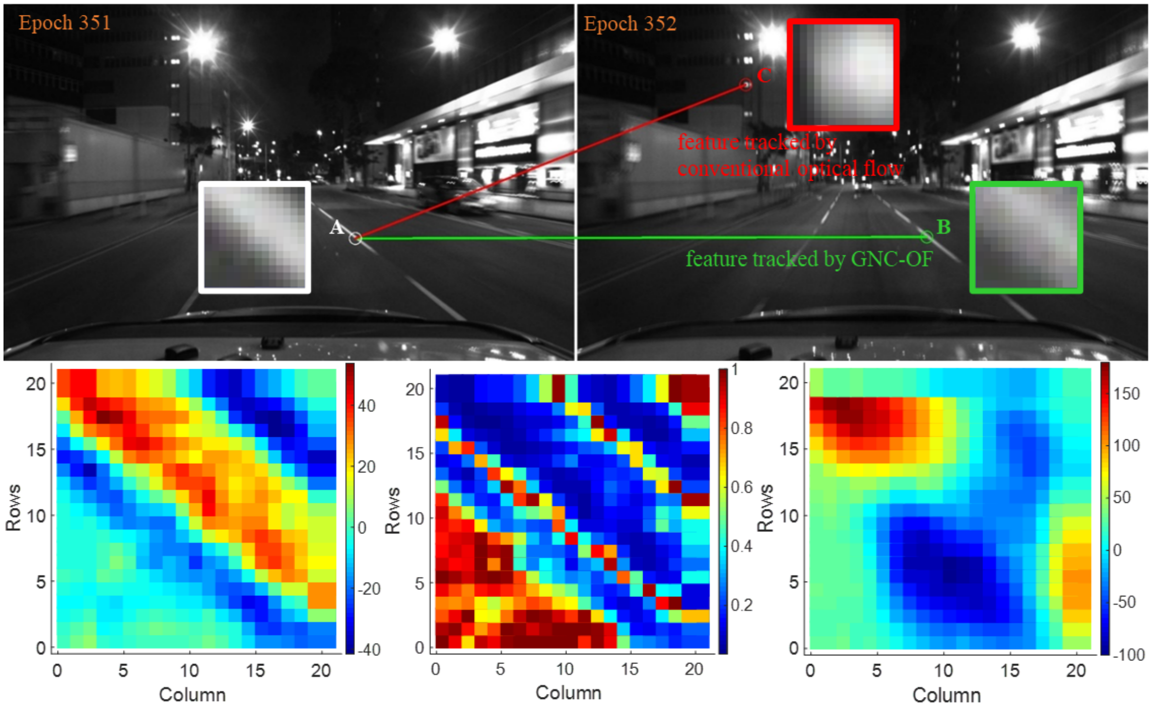


Fig. 12. Positioning errors of the listed methods in urban canyon 2.

2) Discussion: Analysis of Residuals and Weightings for GNC-OF in Front End of VINS: To show further details of the tracking feature using the conventional optical flow and the proposed GNC-OF, we selected a challenging case of urban canyon 2 as shown in Fig. 13. The left image and right images are two consecutive frames from epochs 351 and 352, respectively. Intuitively, the conventional optical flow-based feature tracking method finds an incorrect feature correspondence with a matching pair of feature A (in epoch 351) and feature C (in epoch 352) as feature A should be located on a road lane line. We can see that the incorrectly tracked feature C is located under a light condition with a very similar pixel value to the road lane line. As a result, the conventional optical flow-based tracking feature method gets into the local minimum leading to an incorrect tracking feature. Fig. 13(c) shows the residuals of the pixel values



(a) (b) (c)

Fig. 13. Analysis of the residuals and weightings of the feature tracking of conventional optical flow from OpenCV and the feature tracking from GNC-OF at epochs 351 and 352. (a) Residuals between features A and B (tracked by GNC-OF). (b) Weightings of each pixel of feature B (tracked by GNC-OF). (c) Residuals between feature A and C (tracked by conventional optical flow).

associated with the matching pairs of feature A (in epoch 351) and feature C (in epoch 352). We can see that the maximum residual reached 150 due to the incorrect tracking feature. Moreover, the incorrectly tracked feature C introduces a large error compared with the correctly tracked feature B which can significantly degrade the performance of the data association in the back-end of the VINS.

The proposed GNC-OF correctly tracked the feature with a matching pair of feature A (in epoch 351) and feature B (in epoch 352). Fig. 13(a) shows the detail of the residual associated with the tracking feature. Interestingly, we can see that the shape of the road lane line can also be seen in the residual heat map. The deeper color indicates larger residuals. Furthermore, the larger residuals mainly occurred on the boundary of the road lane line. Fig. 13(b) shows the estimated weightings of the pixel positions surrounding the feature pair A and B. The bluer color indicates the smaller weightings. As expected, the pixel positions with larger residuals are associated with smaller weightings, which subsequently leads to the rejection of the outlier measurements. As a result, feature A is correctly tracked as feature B in epoch 352.

3) Discussion: Degeneration Detection and Analysis in Back End of VINS: As mentioned in the experimental setup, we experimentally set the parameter of λ_{thresh} to 200 to detect the potential degeneration. Subsequently, we presented some of the detected degeneration scenes as shown in Fig. 14. We found that all the minimum eigenvalues in Fig. 14(a)–(c) are smaller than 200 and the related RPE is larger than the mean error of 0.54 m (Table III). This phenomenon shows that the positioning error tends to increase due to insufficient feature constraints (degeneration). However, many factors can cause large errors such as poor illumination, dynamic objects, and feature distribution. Fig. 14(d) shows that although the minimum eigenvalue is 192.34, with an RPE of 0.243 m. This is because the limited high-quality features are used as the constraints of the state. In addition, the vehicles in Fig. 14(d) have no movement, and thus there are no dynamic features. Fig. 14(g)–(i) are detected as healthy cases because the minimum eigenvalues are more than 200 with relatively small RPE values. Specifically, the detected feature in Fig. 14(g)–(i) are more uniformly distributed compared to the degeneration case in Fig. 14(a)–(c). Compared to the degeneration case defined using minimum eigenvalue, the maximum eigenvalues in Fig. 14(e) and (f) are even smaller than 200, and thus

TABLE IV
COMPUTATION COST STUDY OF THE VINS-MONO AND THE PROPOSED METHOD IN URBAN CANYON 1

Items	Conventional		Proposed	
	Front End	Back End	Front End	Back End
MEAN (s)	0.09	0.05	0.13	0.02
STD (s)	0.02	0.01	0.03	0.01
Max (s)	0.18	0.10	0.22	0.08

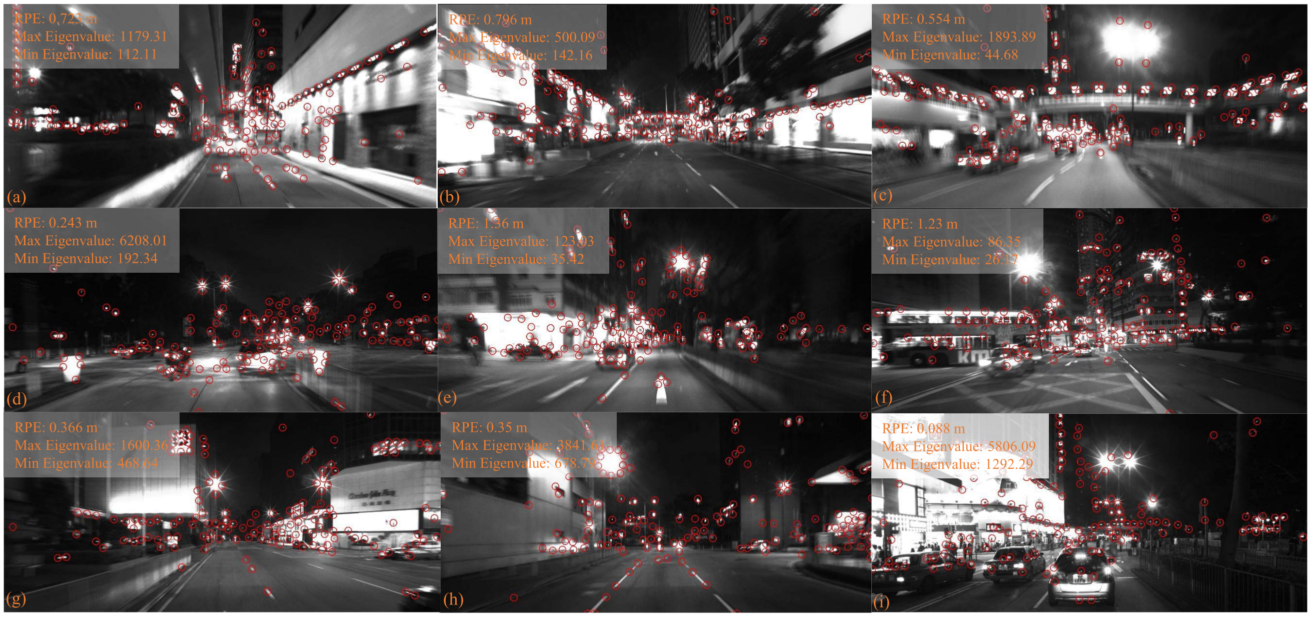


Fig. 14. Illustration of the degeneration and healthy case with associated maximum and minimum eigenvalues, and relative positioning errors. The red and blue circles are the detected features, and the red circle denotes that the feature is tracked more times than the blue one. (a)–(c) Minimum eigenvalue is smaller than 200, thereby the case is degenerated. (d) There are no dynamic features, the case is healthy. (e)–(f) Maximum eigenvalue is even smaller than 200, and thus the ratio between $\lambda_{\max}$ and $\lambda_{\min}$ are used to identify the degeneration case. (g)–(i) Minimum eigenvalue is larger than 200, thereby the case is healthy.

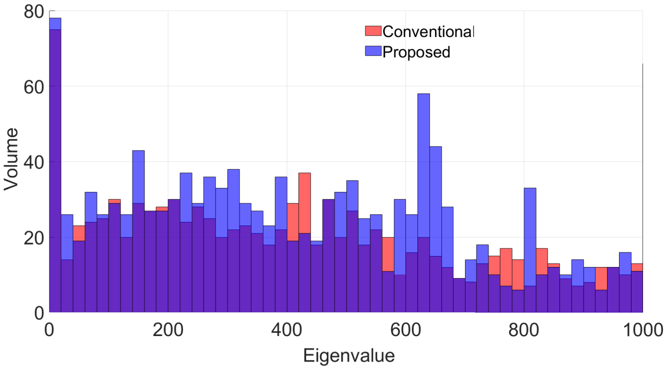


Fig. 15. Histogram of the minimum eigenvalues concerning the translation estimation before outlier removal (conventional VINS-Mono) and after outlier removal (proposed VINS-GNC-OF).

the ratio between $\lambda_{\max}$ and $\lambda_{\min}$ are used to identify the degeneration. The ratio is also obtained in the same way as Fig. 14.

To examine the degeneration case after the removal of outliers by the proposed GNC-OF, and we analyzed the histogram of the minimum eigenvalues concerning the translation estimation before outlier removal (conventional VINS-Mono) and after outlier removal (proposed VINS-GNC-OF), as shown in Fig. 15. The x-axis denotes the minimum eigenvalues for translation estimation. The y-axis represents the volume associated with each bin of the histogram. Statistically, we found that the number of minimum eigenvalues (near 0 to 200) increases after the rejection of the outlier feature using the proposed method. This is due to the enhanced degeneration caused by the rejection of the visual measurements, where the smaller eigenvalue means that the corresponding direction has fewer constraints than the larger one.

D. Discussion: Computational Time Cost Analysis

To analyze the real-time performance of the proposed method, a computational cost study is provided in Table IV. Especially, our processor is based on Intel Core i7-9750H CPU at 2.60 GHz. Table IV compares the processing time in the front-end and back-end of the conventional method and proposed method, respectively. The feature tracking is time-consuming in the front-end, thereby our proposed method needs 0.04 s more than the traditional method. Overall, the performance of our proposed method can be real-time.

VII. Conclusion

Achieving satisfactory positioning of VINS in urban canyons is challenging due to the influence of numerous factors, such as dynamic objects and illumination conditions. Different from the previous work [13], this study excludes the outliers detected from the front-end of VINS, while also detecting and removing the resulting degeneration. Given the degeneration level, the actual number of features is considered to be significant in the mitigation of degenerated performance. The improved performance is demonstrated in both experiments in urban canyons 1 and 2.

Future studies will focus on investigating the integration of VINS positioning with a global navigation satellite system to provide more robust and accurate positioning for vehicular navigation.

References

[1] W. Wen *et al.*, "UrbanLoco: A full sensor suite dataset for mapping and localization in urban scenes," in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, May 2020, pp. 2310–2316.

[2] J. Surber, L. Teixeira, and M. Chli, "Robust visual-inertial localization with weak GPS priors for repetitive UAV flights," in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, May 2017, pp. 6300–6306.

[3] C. Zhang, L. Chen, and S. Yuan, "ST-VIO: Visual-inertial odometry combined with image segmentation and tracking," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 10, pp. 8562–8570, Oct. 2020.

[4] A. I. Mourikis and S. I. Roumeliotis, "A multi-state constraint Kalman filter for vision-aided inertial navigation," in *Proc. IEEE Int. Conf. Robot. Automat.*, Apr. 2007, pp. 3565–3572.

[5] M. Bloesch, S. Omari, M. Hutter, and R. Siegwart, "Robust visual inertial odometry using a direct EKF-based approach," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2015, pp. 298–304.

[6] P. Geneva, K. Eckenhoff, W. Lee, Y. Yang, and G. Huang, "OpenVINS: A research platform for visual-inertial estimation," in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, May 2020, pp. 4666–4672.

[7] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós, "ORB-SLAM3: An accurate open-source library for visual, visual–inertial, and multimap SLAM," *IEEE Trans. Robot.*, early access, May 25, 2021, doi: 10.1109/TRO.2021.3075644.

[8] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, "Keyframe-based visual-inertial odometry using nonlinear optimization," *Int. J. Robot. Res.*, vol. 34, no. 3, pp. 314–334, 2015.

[9] T. Qin, P. Li, and S. Shen, "VINS-mono: A robust and versatile monocular visual-inertial state estimator," *IEEE Trans. Robot.*, vol. 34, no. 4, pp. 1004–1020, Aug. 2018.

[10] J. Delmerico and D. Scaramuzza, "A benchmark comparison of monocular visual-inertial odometry algorithms for flying robots," in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, May 2018, pp. 2502–2509.

[11] M. Burri *et al.*, "The EuRoC micro aerial vehicle datasets," *Int. J. Robot. Res.*, vol. 35, no. 10, pp. 1157–1163, 2016.

[12] X. Bai, B. Zhang, W. Wen, L.-T. Hsu, and H. Li, "Perception-aided visual-inertial integrated positioning in dynamic urban areas," in *Proc. IEEE/ION Position, Location Navigat. Symp. (PLANS)*, Apr. 2020, pp. 1563–1571.

[13] X. Bai, W. Wen, and L.-T. Hsu, "Robust visual-inertial integrated navigation system aided by online sensor model adaption for autonomous ground vehicles in urban areas," *Remote Sens.*, vol. 12, no. 10, p. 1686, May 2020.

[14] W. Xie, P. X. Liu, and M. Zheng, "Moving object segmentation and detection for robust RGBD-SLAM in dynamic environments," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–8, 2021.

[15] Z. Zhu, F. Xu, M. Li, Z. Wang, and C. Yan, "Challenges from fast camera motion and image blur: Dataset and evaluation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 211–227.

[16] X. Bai, W. Wen, and L.-T. Hsu, "Performance analysis of visual/inertial integrated positioning in typical urban scenarios of Hong Kong," in *Proc. APCATS*, Taipei City, Taiwan, 2019, pp. 1–9.

[17] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. Imag. Understand. Workshop*, Vancouver, BC, Canada, 1981, pp. 121–130.

[18] L. Xiao, J. Wang, X. Qiu, Z. Rong, and X. Zou, "Dynamic-SLAM: Semantic monocular visual localization and mapping based on deep learning in dynamic environment," *Robot. Auton. Syst.*, vol. 117, pp. 1–16, Jul. 2019.

[19] D. V. Nam and K. Gon-Woo, "Robust stereo visual inertial navigation system based on multi-stage outlier removal in dynamic environments," *Sensors*, vol. 20, no. 10, p. 2922, May 2020.

[20] L. Cui and C. Ma, "SOF-SLAM: A semantic visual SLAM for dynamic environments," *IEEE Access*, vol. 7, pp. 166528–166539, 2019.

[21] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[22] W. Liu *et al.,* "SSD: Single shot MultiBox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37.

[23] M. S. Bahraini, A. B. Rad, and M. Bozorg, "SLAM in dynamic environments: A deep learning approach for moving object tracking using ML-RANSAC algorithm," *Sensors*, vol. 19, no. 17, p. 3699, Aug. 2019.

[24] W. Li and J. J. Swetits, "The linear l_1 estimator and the Huber M-estimator," *SIAM J. Optim.*, vol. 8, no. 2, pp. 457–475, May 1998.

[25] N. Sunderhauf and P. Protzel, "Switchable constraints for robust pose graph SLAM," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2012, pp. 1879–1884.

[26] H. Yang, P. Antonante, V. Tzoumas, and L. Carlone, "Graduated non-convexity for robust spatial perception: From non-minimal solvers to global outlier rejection," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 1127–1134, Apr. 2020.

[27] M. J. Black and A. Rangarajan, "On the unification of line processes, outlier rejection, and robust statistics with applications in early vision," *Int. J. Comput. Vis.*, vol. 19, no. 1, pp. 57–91, 1996.

[28] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artif. Intell.*, vol. 17, nos. 1–3, pp. 185–203, Aug. 1981.

[29] J. Shi and Tomasi, "Good features to track," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 1994, pp. 593–600.

[30] E. P. Simoncelli, E. H. Adelson, and D. J. Heeger, "Probability distributions of optical flow," in *Proc. CVPR*, vol. 91, 1991, pp. 310–315.

[31] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza, "On-manifold preintegration for real-time visual–inertial odometry," *IEEE Trans. Robot.*, vol. 33, no. 1, pp. 1–21, Feb. 2017.

[32] J. T. Barron, "A general and adaptive robust loss function," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4331–4339.

[33] P. Antonante, V. Tzoumas, H. Yang, and L. Carlone, "Outlier-robust estimation: Hardness, minimally tuned algorithms, and applications," 2020, *arXiv:2007.15109*.

[34] P. Sturm, "Pinhole camera model," Nat. Inst. Res. Digit. Sci. Technol., Paris, France, Tech. Rep., 2014.

[35] W. Wen, T. Pfeifer, X. Bai, and L.-T. Hsu, "It is time for factor graph optimization for GNSS/INS integration: Comparison between FGO and EKF," 2020, *arXiv:2004.10572*.

[36] G. Sibley, L. Matthies, and G. Sukhatme, "Sliding window filter with application to planetary landing," *J. Field Robot.*, vol. 27, no. 5, pp. 587–608, 2010.

[37] P. J. Huber, "Robust estimation of a location parameter," in *Breakthroughs in Statistics*. New York, NY, USA: Springer, 1992, pp. 492–518.

[38] J. Zhang, M. Kaess, and S. Singh, "On degeneracy of optimization-based state estimation problems," in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, May 2016, pp. 809–816.

[39] M. Tahsin, S. Sultana, T. Reza, and M. Hossam-E-Haider, "Analysis of DOP and its preciseness in GNSS position estimation," in *Proc. Int. Conf. Electr. Eng. Inf. Commun. Technol. (ICEEICT)*, May 2015, pp. 1–6.

[40] G. H. Golub and C. Reinsch, "Singular value decomposition and least squares solutions," in *Linear Algebra*. Berlin, Germany: Springer, 1971, pp. 134–151.

[41] S. Kennedy, D. Cosandier, and J. Hamilton, "GPS/INS integration in real-time and post-processing with NovAtel's SPAN system," in *Proc. Int. Global Navigat. Satell. Syst. Soc. Symp.*, 2007, pp. 4–6.

[42] M. Quigley *et al.*, "ROS: An open-source robot operating system," in *Proc. ICRA Workshop Open Source Softw.*, Kobe, Japan, 2009, vol. 3, no. 3.2, p. 5.

[43] J. Rehder, J. Nikolic, T. Schneider, T. Hinzmann, and R. Siegwart, "Extending kalibr: Calibrating the extrinsics of multiple IMUs and of individual axes," in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, May 2016, pp. 4304–4311.

[44] W. Wen and L.-T. Hsu, "3D LiDAR aided GNSS real-time kinematic positioning," presented at the ION GNSS, St. Louis, MO, USA, 2021.

[45] L.-T. Hsu, N. Kubo, W. Chen, Z. Liu, T. Suzuki, and J. Meguro, "UrbanNav: An open-sourced multisensory dataset for benchmarking positioning algorithms designed for urban areas," presented at the ION GNSS, Tallahassee, FL, USA, 2021.

[46] M. Grupp, "Python package for the evaluation of odometry and SLAM," Tech. Univ. Munich, Munich, Germany, Tech. Rep., 2017.

Xiwei Bai received the M.Sc. degree in engineering from China Agricultural University, Beijing, China, in 2018. After that, she worked as a Research Assistant with The Hong Kong Polytechnic University, Hong Kong, from 2018 to 2019.

She is currently pursuing the Ph.D. degree with The Hong Kong Polytechnic University. Her research interests include visual simultaneous localization and mapping (SLAM) and vision-aided global navigation satellite systems (GNSS) positioning in urban canyons for the intelligent transportation system, autonomous driving.

Weisong Wen was born in Ganzhou, Jiangxi, China. He received the Ph.D. degree in mechanical engineering from The Hong Kong Polytechnic University, Hong Kong, in 2020.

He was a Visiting Student Researcher with the University of California, Berkeley (UCB), Berkeley, CA, USA, in 2018. He is currently a Research Assistant Professor with the Department of Aeronautical and Aviation Engineering, The Hong Kong Polytechnic University. His research interests include multisensor integrated localization for autonomous vehicles, simultaneous localization and mapping (SLAM), and global navigation satellite systems (GNSS) positioning in urban canyons.

Li-Ta Hsu (Member, IEEE) received the B.S. and Ph.D. degrees in aeronautics and astronautics from the National Cheng Kung University, Tainan, Taiwan, in 2007 and 2013, respectively.

In 2012, he was a Visiting Scholar with the University College London, London, U.K. He served as a Post-Doctoral Researcher with the Institute of Industrial Science, The University of Tokyo, Tokyo, Japan. He is currently an Assistant Professor with the Department of Aeronautical and Aviation Engineering, The Hong Kong Polytechnic University, Hong Kong. His research interests include global navigation satellite systems (GNSS) positioning in challenging environments and localization for pedestrian, autonomous driving vehicle, and unmanned aerial vehicle.