A Visual-Inertial Localization Method for Unmanned Aerial Vehicle in Underground Tunnel Dynamic Environments

DONGIANG LI\textsuperscript{1}, (Graduate Student Member, IEEE), WEI YANG\textsuperscript{1}, XUESONG SHI\textsuperscript{2}, (Member, IEEE), DAN GUO\textsuperscript{3,4}, QIWEI LONG\textsuperscript{3,5}, (Student Member, IEEE), FEI QIAO\textsuperscript{6}, (Member, IEEE), AND QI WEI\textsuperscript{6}, (Member, IEEE)

\textsuperscript{1}School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China
\textsuperscript{2}Intel Labs, Beijing 100190, China
\textsuperscript{3}Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
\textsuperscript{4}International School, Beijing University of Posts and Telecommunications, Beijing 100876, China
\textsuperscript{5}School of Software Engineering, Beijing Jiaotong University, Beijing 100044, China
\textsuperscript{6}Department of Precision Instrument, Tsinghua University, Beijing 100084, China

Corresponding authors: Wei Yang (weiyang@bjtu.edu.cn) and Fei Qiao (qiaofei@tsinghua.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 51874299, and in part by the Natural Science and Technology Pillar Program of China under Grant 2013BAK06B03.

ABSTRACT Unmanned Aerial Vehicles (UAVs) can significantly improve the autonomy of the mining industry, and self-localization is the key to autonomous flights of underground UAVs. A localization method of visual-inertial sensor data fusion is proposed in this paper. The method aims to improve the localization accuracy and robustness of underground UAVs in dynamic environments. First, an algorithm for dynamic point detection and rejection is presented, which combines a semantic segmentation neural network, an optical flow method, and an epipolar constraint method. Second, a visual-inertial sensor fusion algorithm is used to enhance performance in areas lacking static visual features. It can also provide absolute scales to the localization results, as opposed to monocular systems. Finally, a hand-held multi-sensor data collection system is developed with accurate calibration, for imitating flights of underground UAVs and easing data collection in real underground tunnels. We evaluate our proposed localization method and compare it with state-of-art method VINS-Mono on both the public EuRoC dataset and our own collected data in underground tunnels. Experimental results show that the proposed visual-inertial localization method can improve the accuracy by more than 67\% over VINS-Mono in high dynamic environments, and it can be applied to underground dynamic scenes with high robustness and accuracy.

INDEX TERMS Dynamic environment, sensor fusion, semantic segmentation, UAV pose estimation, underground tunnel.

I. INTRODUCTION

In recent years, Unmanned Aerial Vehicles (UAVs) technology has been introduced to the mining industry for topographical survey and safety investigation [1]. Underground monitoring and control system is usually fixed and limited ranges, so it is difficult to fully meet dynamic monitoring requirements of underground safety production. However, UAVs with vision and communication equipment can quickly fly to instructed locations and perform visual inspection missions, which have been an important component to complete the underground monitoring system [2]. Accurate pose estimation is a crucial part for flight control and path planning of underground UAVs.

Although the Global Positioning System (GPS) is frequently used in general localization system [3], it is unsuitable for mining scenarios because of the signal blockage. Another commonly used localization technique is Inertial Navigation System (INS). It uses Inertial Measurement Units (IMUs) composed of accelerometer and gyroscope for tracking pose by inertial navigation principle. But it is not accurate enough for UAV pose estimation due to accumulated sensor
errors. In real world deployment of UAVs, the integrated navigation system of GPS/INS is usually adopted [4], [5], but they cannot work underground. In order to develop robust and accurate localization systems for UAV applications, recent research has been focusing on localization systems with perception sensors such as RADAR, LiDAR, ultrasonic, camera, etc.

Visual Simultaneous Localization and Mapping (Visual-SLAM) uses monocular or stereo cameras to track the features of consecutive images and estimate the robotic relative orientation and translation, which has been broadly researched [6]–[9]. However, when the UAV armed with camera moves too fast, the robustness of pure Visual-SLAM system is low due to motion blur and image degradation. On the other hand, the dynamic objects like working people in underground scenes are also challenges for Visual-SLAM algorithms. In order to meet the reliability demand of system, the method based on camera and IMU sensor data fusion is adopted to estimate the pose of underground UAVs.

In this paper, a novel visual-inertial localization method for UAVs in underground tunnel dynamic environments is proposed. Fig. 1 provides an overview of our proposed method. Firstly, it eliminates the influence of dynamic environment to the pose estimation by combing geometric and deep learning method. And then a visual-inertial sensor fusion algorithm is used to improve the robustness of the proposed localization system. Experiments show that our proposed localization method is accurate and robust in underground dynamic environments.

The contributions of this paper include:

1) It proposes a dynamic point detection and rejection algorithm, which combines a semantic segmentation neural network ENet [10], a Lucas–Kanade (LK) optical flow method [11] and an epipolar constraint method [12], to eliminate the interference of dynamic objects during pose estimation.

2) A visual-inertial sensor fusion algorithm is used to enhance the system robustness in areas lacking static visual features and eliminate the impacts of the underground UAVs fast movement. It can also provide absolute scales to the localization results, as opposed to monocular systems. And the proposed localization system is integrated with Robot Operating System (ROS) [13], which is beneficial for the implementation in the real scene.

3) It develops a hand-held multi-sensor data collection system, which can imitate the flights of underground UAVs and collect data in real underground tunnels. The underground UAV coordinate systems are first built and then the data collection system is accurately calibrated according to the related transformation relationship of coordinate systems. The collected data is used to verify the effectiveness of the proposed method.

In the rest of this paper, the structure is as follows. Related works are given in Section II. Section III presents the underground UAV system model, namely transformation relations of underground UAV coordinate systems and the calibration of hand-held sensor data collection device, which is used for collecting data and verifying the proposed algorithm. Section IV describes the proposed dynamic point detection and rejection algorithm in detail. Then section V introduces underground UAV pose estimation method based on visual-inertial sensor fusion. Subsequently, Section VI presents experimental results. Finally, a brief conclusion and a discussion are given in section VII.

II. RELATED WORKS

UAVs have achieved great attentions, because they play a vital role in various sectors ranging from industrial and agricultural sectors to surveillance and public safety [14]. Chen et al. [15] gave an overall understanding about using Artificial Neural Networks (ANNs) for addressing key challenges in UAV-based applications. Nodland et al. [16] studied the use of a neural network (NN) for trajectory tracking of UAVs. In [17], the authors proposed an image matching system by a NN using aerial images and aerial geo-referenced images to estimate the positions of UAVs in a situation of GPS failure. The robustness and accuracy of mentioned NN-based localization methods are acceptable in large scale scenes. However, they can not meet the demand of high accuracy of localization in a limited space like underground scene.

For a potential localization in underground tunnel environment, SLAM technique can be a viable option, which can be used for mapping and tracking robotic position instantaneously [18]. The perception input of SLAM can be LI DAR and camera sensors. For example, Chow et al. [18] made a comparative analysis of the three selected laser-based SLAM approaches toward underground localization with a UAV. However, the rough surface within a deep tunnel might not be optimum for a laser sensor. In order to improve the robustness of localization system, the fusion of multiple sensors has been a researched topic [19]. In [20], a UAV system for environment exploration and mapping by fusing ultrasonic and camera sensors was developed. It uses a camera sensor to detect marker and measures $y$ positions of UAV. And $z$ position is obtained from a bottom ultrasonic sensor. But inferring the position from a marker in mining scenes does not work. Bavle et al. [21] proposed a modular EKF-based multisensor fusion approach for vertical localization of UAVs in unstructured indoor environments. However, for the sake of safety, 6DoF pose estimation of UAVs is necessary in underground environment.

Another challenge of the application of SLAM algorithms in underground scenes is that underground environment is dynamic. However, most of the current SLAM algorithms assume that the environment is static [22]. Therefore, most state-of-the-art approaches, which are initially designed for static environments, are not capable of handling underground dynamic scenarios containing working people. Identifying the dynamic objects from the static parts, and then discarding them before pose estimation are necessary [23]. In our previous work [23], a semantic segmentation thread by SegNet [24] is added to ORB-SLAM2 framework, which can remove the dynamic features extracted from images captured...
by Kinect2 and improve the accuracy of pose estimation in dynamic environments. However, this method can not be directly transplanted to a UAV system because of limited computational resources and weight limitations.

III. UNDERGROUND UAV SYSTEM MODEL

A. TRANSFORMATION OF UNDERGROUND UAV COORDINATE SYSTEMS

The underground UAV with sensors including monocular camera, IMU and markers can collect the tunnel environment data. Among sensors, markers with reflective material are used for obtaining the ground truth of UAV pose. The transformation relationships of mentioned sensors are important to estimate the UAV pose and verify the effectiveness of the proposed localization method. On the other hand, the sensors are all fixed on the body of underground UAV. The relative pose of sensors and the UAV body remains unchanged. Theoretically, we can establish the UAV body coordinate system with any sensors. In this paper, it is assumed that the UAV body coordinate system is consistent with the camera’s and the pose estimation of the underground UAV is the same as the camera pose estimation. In the following part, the coordinate systems are established for each sensor, and the transformation relations among the established coordinate systems are analyzed.

1) ESTABLISHMENT OF UNDERGROUND UAV COORDINATE SYSTEMS

Fig. 2 is the diagram of the UAV flight in coal mine tunnel. It shows the coordinate transformation model of visual-inertial navigation system equipped in the underground UAV. $o_c$ is the optical centre of the monocular camera, which is taken as the origin of the camera coordinate system $o_c = x_cy_cz_c$, where $z_c$ axis points towards the front of the camera, $x_c$ axis points towards the right direction, and $y_c$ points towards the ground. $o' - x'y' , o - uv$ are physical imaging plane and pixel coordinate system of the camera. $o_i$ is the origin of the IMU coordinate system $o_i = x_idx_iz_i$, in which $x_i$, $y_i$ and $z_i$ axes are parallel to $x_c$, $y_c$ and $z_c$ axes. $o_w$ is the origin of the world coordinate system $o_w = x_wy_wz_w$ set at the entrance of tunnel, where $x_w$ axis is vertical to the tunnel wall and $z_w$ axis is vertical to the ground. $o_m$ is the origin of the Marker coordinate system $o_m = x_my_mz_m$, in which $z_m$ axis points upward and is vertical to the surface of the UAV body, $y_m$ axis points toward the front of UAV, and $x_m$ axis is vertical to the surface consisting of $y_m$ and $z_m$ axes.

2) TRANSFORMATION RELATIONS OF COORDINATE SYSTEMS

In Fig. 2, $P$ observed by the camera of underground UAV is a point on underground tunnel wall $l$. $P'$ point is the projection of $P$ through the optical centre of the camera onto the physical imaging plane. Let the coordinates of point $P$ be $P_w[x_{wp}, y_{wp}, z_{wp}]$ in the world coordinate system, be $P_c[x_{cp}, y_{cp}, z_{cp}]$ in the camera coordinate system, be $P'[x'_p, y'_p]$ in the physical coordinate system, be $P_m[x_{mp}, y_{mp}, z_{mp}]$ in the pixel coordinate system, be $P_i[x_{ip}, y_{ip}, z_{ip}]$ in the IMU coordinate system, and be $P_m[x_{mp}, y_{mp}, z_{mp}]$ in the Marker coordinate system. The transformation relations of the coordinates of point $P$ in different coordinate systems are as follows.

(1) Camera Coordinate System – Physical Imaging Plane Coordinate System

The camera transforms 3D spatial points observed in the camera coordinate system into 2D points in the physical imaging plane coordinate system. According to the pin-hole camera model [25] and the similar triangle principle, the transformation relation can be expressed as,

$$
\begin{align*}
    & x' = \frac{x_{cp}}{z_{cp}}, \quad y' = \frac{y_{cp}}{z_{cp}} \\
    \Rightarrow & \begin{cases} 
    x_p = f \cdot \frac{x_{cp}}{z_{cp}} \\
    y_p = f \cdot \frac{y_{cp}}{z_{cp}}
    \end{cases}
\end{align*}
$$
where \( f \) is the focal length of the camera.

(2) Physical Imaging Plane Coordinate System – Pixel Coordinate System

The pixel coordinate system and the physical imaging plane coordinate system are in the same plane. \( u \) and \( v \) axes of pixel coordinate system are parallel to the \( x' \) and \( y' \) axes of physical imaging plane coordinate system. It is assumed that pixel coordinates on \( u \) and \( v \) axes are scaled by \( \alpha \) and \( \beta \) times, respectively, the origin point is translated to \([cx', cy']^T\). The relation of the physical imaging plane coordinates and pixel coordinates of \( P' \) can be expressed as,

\[
\begin{align*}
    up &= \alpha x' + cx \\
    vp &= \beta y' + cy
\end{align*}
\]

(3) Camera Coordinate System – Pixel Coordinate System

Combine (1) with (2) and the transformation relation can be expressed as,

\[
\begin{align*}
    up &= af \frac{x_p}{z_p} + cx \\
    vp &= \frac{y_p}{z_p} + cy
\end{align*}
\]

\[
P_{wp} = \begin{bmatrix} up \\ vp \\ 1 \end{bmatrix} = \frac{1}{z_p} \begin{bmatrix} af & 0 & c_y \\ 0 & \beta f & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_p \\ y_p \\ z_p \end{bmatrix} \triangleq KP_c
\]  

(4) Transformation Relations among World Coordinate System, Camera Coordinate System, IMU Coordinate System and Marker Coordinate System

The world coordinate system, the camera coordinate system, the IMU coordinate system and the Marker coordinate system can be regarded as rigid bodies. And coordinates of a point in any of these coordinate systems can be transformed to another coordinate system through rotation and translation operations. Take the transformation from world coordinate system to camera coordinate system as an example, let the rotation and translation matrices be \( R \) and \( t \), then the coordinates of point \( P \) can be expressed as,

\[
P_c = \begin{bmatrix} x_p \\ y_p \\ z_p \end{bmatrix} = R \begin{bmatrix} x_w \\ y_w \\ z_w \end{bmatrix} + t \triangleq T_{CW} \begin{bmatrix} P_w \\ 1 \end{bmatrix} \triangleq T_{CW}P_w
\]  

where \( T_{CW} \) is the transformation matrix from the world coordinate system to the camera coordinate system. \( T_{CW} \) is the inverse matrix of \( T_{WC} \) shown in Fig. 2. The coordinates transformation of point \( P \) from the camera coordinate system to the world coordinate system can be expressed as,

\[
P_w = T_{WC}P_c
\]  

UAV pose estimation, which means estimating the pose of the camera optical centre in the world coordinate system, is equivalent to solving the transformation matrix \( T_{WC} \). Similarly, the transformation relations of the point \( P \) from the IMU coordinate system to the camera coordinate system, to the Marker coordinate system and to the world coordinate system can be expressed as,

\[
\begin{align*}
    P_c &= T_{CI}P_i \\
    P_m &= T_{MC}P_c \\
    P_w &= T_{WM}P_m
\end{align*}
\]

where \( T_{CI} \) is the transformation matrix from the IMU coordinate system to the camera coordinate system, \( T_{MC} \) is the transformation matrix from the camera coordinate system to the Marker coordinate system and they should be accurately calibrated based on the experimental device. \( T_{WM} \) is the transformation matrix from the Marker coordinate system to the world coordinate system and \( T_{WC} \) can be obtained through the MCS located in underground tunnel. Let \( T'_{WC} \) be the ground truth of the underground UAV pose in the camera coordinate system and it can be expressed as,

\[
T'_{WC} = T_{WM}T_{MC}
\]
The average localization accuracy of MCS used in this paper can reach to 0.167mm and data transfer delay is lower than 5ms. In the above equation, $T_{WM}$ is captured by the MCS, $T_{MC}$ is needed to be calibrated based on the experimental device. $T_{WC}$ can be used for evaluating the effectiveness of the proposed underground UAV localization system. Next, the transformation matrices $T_{CI}$ and $T_{MC}$ will be calibrated.

### B. CALIBRATION OF HAND-HELD SENSOR DATA COLLECTION DEVICE

In this paper, a hand-held sensor data collection device shown in Fig. 3 is developed to simulate flights of the underground UAV. It provides a verification platform for the proposed visual-inertial localization method. The device is equipped with battery, monitor, an Intel RealSense D435i camera, markers and an Intel NUC. The D435i has a built-in IMU sensor, which can combine the linear accelerometer with the gyroscope to detect the rotation and translation of the x, y and z axes. During the data collection, the time stamp of the D435i and the built-in IMU sensor are aligned by hardware.

The transfer matrix $T_{CI}$ has been calibrated by Intel manufacturer. Therefore, the transfer matrix $T_{MC}$ is the only one needed to be calibrated.

Fig. 4 shows the calibration of the transfer matrix $T_{MC}$. Before the calibration, the MCS system is initialized including the configuration of the world coordinate system (World), and then a calibration plate is placed before the D435i and kept still. While the hand-held sensor device is moving in the area covered by the MCS, the equation can be obtained below,

$$T_{WM(i)}T_{MC}T_{CP(i)} = T_{WM(i+1)}T_{MC}T_{CP(i+1)}$$

where $T_{WM(i)}$ and $T_{WM(i+1)}$ are the translation matrices from the Marker coordinate system to world coordinate system captured by the MCS at time $i$ and time $i + 1$. $T_{CP(i)}$ and $T_{CP(i+1)}$ are the translation matrices from the origin $o_1$ of the calibration plate to the camera coordinate system (Cam), which are captured by the OpenCV ArUco module at time $i$ and time $i + 1$. According to matrix transformation,

$$T_{MC}T_{CP(i)}^{-1}T_{CP(i+1)} = T_{WM(i)}^{-1}T_{WM(i+1)}T_{MC}$$

(10)

where $T_{CP(i)}^{-1}T_{CP(i+1)}$ and $T_{WM(i)}^{-1}T_{WM(i+1)}$ are respectively corresponding to the translation matrices of $Bi$ and $Ai$ as shown in Fig. 4. Equation (10) can be expressed as,

$$T_{MC}Bi = AiT_{MC}$$

(11)

However, when the data collection device moves in the area covered by the MCS, there are multiple groups of translation matrices $Ai$ and $Bi$ that satisfy (11). Therefore, the least squares method is adopted to calculate the optimized translation matrix $T_{MC}$. Table 1 shows the calibration result of $T_{MC}$.

| Pinhole camera parameters | Camera distortion parameters |
|---------------------------|-----------------------------|
| $a_f$ | $596.950638$ | $k_1$ | $0.15852$ |
| $b_f$ | $597.579445$ | $k_2$ | $-0.33979$ |
| $c_*$ | $322.253667$ | $p_1$ | $0.003706$ |

### IV. PROCESSING METHOD OF UNDERGROUND DYNAMIC ENVIRONMENT

#### A. PROBLEM DESCRIPTION OF POSE ESTIMATION TOWARDS UNDERGROUND DYNAMIC ENVIRONMENT

The pose estimation of underground UAV utilizes feature points extracted from images collected by the D435i and measured data of the built-in IMU sensor. Currently, the methods...
of extracting image feature points mainly include SIFT [26], SURF [27] and ORB [28]. Among them, ORB integrates the improved Fast [29] corner point and the BRIER [30] descriptors and has much better performance than SIFT and SURF in terms of real time. Therefore, in this paper, ORB is adopted to extract feature points from the D435i images. ORB has considered the attributes of scale invariance and rotation consistency of feature points, but not the dynamic and static properties. Therefore, the dynamic objects in underground environments will cause the low accuracy of pose estimation.

In order to solve this problem, a dynamic point detection and rejection algorithm is proposed to eliminate the dynamic abnormal points for underground vision tracking. This method combines the semantic segmentation neural network ENet, the LK optical flow method and epipolar constraint method, which can improve the localization accuracy and the stability of underground UAV localization system in the underground dynamic environment.

### B. DYNAMIC POINT DETECTION AND REJECTION ALGORITHM

The idea of underground dynamic environment processing method is similar to the corresponding part in our previous work DS-SLAM [23]. But according to the characteristics of underground environment, some improvements are performed compared to original method. The improvements are mainly reflected in two aspects. First, considering data collection in the underground environment, Intel RealSense D435i camera is adopted to against the limitations of underground UAV to the size and quality of sensors. What’s more, compared with Kinect V2 camera used in previous work, D435i consumes less power, which is a more reasonable configuration for UAV. Second, because the underground localization system must have a high real-time performance, a lightweight semantic segmentation network ENet is adopted.

Dynamic point detection and rejection algorithm includes four parts: ENet semantic segmentation, dynamic point detection, ORB feature points extraction and ORB feature points refinement. When the hand-held device shown in Fig. 3 collects the tunnel data, ORB features are extracted from images collected by D435i, which are stored in ORB feature points set A. Meanwhile, ENet semantic segmentation and dynamic point detection are performed to acquire semantic segmentation image and dynamic point set S respectively. Both semantic segmentation image and dynamic point set S are used to refine ORB feature points set A, then the static ORB feature points can be obtained. In the following, the important parts of the process, including ENet semantic segmentation, dynamic point detection and ORB feature points refinement, are introduced.

1) **ENet SEMANTIC SEGMENTATION**

The first step of eliminating the dynamic interference of the underground workers is to detect whether there are operating workers in underground tunnel. Semantic segmentation network is adopted to process the collected underground images and then the corresponding semantic segmentation images with semantic labels can be obtained. If the segmentation images contain human semantic labels, it can be determined that there are operators in the underground environment. However, applying semantic segmentation network to the underground environment must consider both the real-time need of the system and the scene features. It is quite different from our previous work which applied semantic segmentation network to the robot localization in indoor dynamic scenes and the mapping of semantic environments.

Firstly, considering the real-time application, a lightweight semantic segmentation network ENet is adopted to segment the underground images. ENet is a new neural network architecture optimized for fast inference and high accuracy. ENet is trained on Pascal VOC dataset [31] in this paper. The Pascal VOC dataset is used for image recognition challenge. It contains 21 categories and each category is marked with different label, for example, the airplane label is 1 and the human label is 15. Take the human label as an example, if there are operating workers in the underground tunnel scene, the pixel values of semantic segmentation image corresponding to the operators pixels in the image are all 15.

Secondly, considering that there will be no animal categories such as cow, horse and indoor scene object categories such as bottle, chair and sofa appearing in the actual underground tunnel environment scenario, therefore, in this paper, only human labels in Pascal VOC dataset are loaded for training ENet. The trained underground segmentation network ENet only processes the operating workers in underground scenes. ENet and SegNet semantic segmentation networks are all trained on Pascal VOC dataset. The comparison diagram performing on the underground collected images is showed in Fig. 5.

In the underground scene, more operating workers mean the greater dynamic environment influence on UAV pose estimation. Compared with SegNet, it can be seen that the underground semantic segmentation network ENet trained by our method has a more accurate and complete segmentation effect on underground operators for both low and high dynamic scenes. In addition, the average running time of ENet and SegNet on 1000 collected images with the same hardware equipment is tested. The average time of processing one image size 480 × 640 by ENet is 26.96 ms, the frame rate is up to 33.4 fps. However, the average time by SegNet is 150.78ms, the frame rate is about 6.6 fps. Therefore,

---

**TABLE 3.** Translation matrix $T_{CI}$ from IMU coordinate system to camera coordinate system.

| Rotation Matrix | Translation Matrix |
|-----------------|--------------------|
| 0.9999600       | 0.0041929          |
| -0.0041856      | 0.9999000          |
| -0.0082871      | 0.0008636          |
|                 | 0.9999700          |
|                 | 0.01209129         |
TABLE 4. Dynamic point detection Algorithm [23].

| Input: | Previous frame, $F_1$; Previous frame’s feature points, $P_1$; Current frame, $F_3$; |
| --- | --- |
| Output: | The set of dynamic points, $S$; |
| 1: | Current frame’s feature points $P_2 = \text{CalcOpticalFlowPyrLK}(F_1, F_2, P_1)$ |
| 2: | Remove outliers in $P_2$ |
| 3: | $FM = \text{FindFundamentalMatrix}(P_1, P_2)$ |
| 4: | for each matched pairs $p_1, p_2$ in $P_1, P_2$ do |
| 5: | $I_1 = \text{FindEpipolarLine}(p_1, FM)$ |
| 6: | $D = \text{CalcDistanceFromEpipolarLine}(p_2, I_1)$ |
| 7: | if $D > r$ then |
| 8: | Append $p_2$ to $S$ |
| 9: | end if |
| 10: | end for |

ENet is adopted to detect whether the underground environment contains operating workers. It is more beneficial to the real-time realization of underground UAV localization system.

2) DYNAMIC POINT DETECTION AND ORB FEATURE POINTS REFINEMENT

After determining the presence of operators in the underground tunnel, it is necessary to determine whether the operators are in motion because only dynamic workers impact UAV pose estimation. Therefore, the dynamic point detection algorithm is adopted in Table 4, which has been well explained in DS-SLAM. It first obtains and traverses the dynamic point set $S$. If there exists a point in set $S$ corresponding to the label 15 in the ENet segmentation image, the points with the pixel label of 15 in set $A$ are all removed. And then the remaining feature point set $A$ is the refined static ORB points. In this way, the dynamic interference of operating workers is eliminated.

V. UNDERGROUND UAV POSE ESTIMATION METHOD BASED ON VISUAL-INERTIAL SENSOR FUSION

In most cases, the process of the refinement of feature points can lead to tracking lost of UAV pose and low system reliability because of the decrease in the amount of feature points [32]. To address this problem, an underground UAV pose estimation method based on visual-inertial sensor fusion is proposed. When the hand-held sensor device shown in Fig. 3 moves in the underground tunnel with sensors, assuming that at time $t$, camera pose is $x_t$, measurement of built-in IMU sensor is $u_t$, system noise is $w_t$ and then $x_t$ can be expressed as,

$$x_t = f(x_{t-1}, u_t) + w_t$$  \hspace{1cm} (12)

where the function $f$ represents the transformation from IMU measurement and previous camera pose to current camera pose. Noise $w_t$ obeys the Gaussian distribution $w_t \sim N(0, R_t)$. It calculates the underground UAV pose according to IMU measurement and is called motion equation. Assuming at time $t$, underground tunnel landmark observed by D435i is $y_t$, observation value is $z_t,j$ and then $z_t,j$ can be expressed as,

$$z_{t,j} = h(y_t, x_t) + v_{t,j}$$ \hspace{1cm} (13)

where the function $h$ describes the coordinates transformation from landmark and camera pose to camera observation. The observation noise $v_{t,j}$ obeys the Gaussian distribution $v_{t,j} \sim N(0, Q_{t,j})$. It transforms 3D landmark coordinates in underground tunnel scene into 2D pixel values and is called observation equation.

Camera pose with scale uncertainty can be estimated by observation equation. However, the pose without absolute scales is meaningless to autonomous flights of underground UAV. In order to solve this problem, method of monocular visual observation and IMU measurement fusion is adopted. Observation equation is first used for estimating the underground UAV initial pose and then IMU measured data between two key frames is fused to recover absolute scale of the estimated UAV pose.

The specific form of observation equation is needed to be determined during the UAV initial pose estimation. The transformation relationship of 3D landmarks and 2D pixel values can be acquired by combining (5) with (4), and it can be expressed as,

$$P_{uw} = KT_{CW}P_w$$ \hspace{1cm} (14)

Assuming that $P^1_w$ is one of the 3D landmarks, $P_{1uw}$ is the corresponding pixel coordinates in the first images. When the first image is collected, the camera is assumed as static and $T_{CW}$ is a unit matrix. According to (14), the transformation relationship is

$$P_{1uw} = KP^1_w$$ \hspace{1cm} (15)

Assuming that $P_{2uw}$ is the corresponding pixel coordinates of $P^1_w$ in the second image, transformation matrix is $t$ and rotation matrix is $R$ of $T_{CW}$. According to (14), the transformation
relationship is,
\[ P_{2uv} = K (R P_{uv} + t) \]  
(16)

Combine (15) and (16),
\[ K^{-1} P_{2uv} = R K^{-1} P_{1uv} + t \]  
(17)

Outer Product of \( t \), denoted as \( t \wedge \), is added to each side of the equation and (17) should be written,
\[ t \wedge K^{-1} P_{2uv} = t \wedge R K^{-1} P_{1uv} \]  
(18)

Left multiply each side of the equation by \( (K^{-1} P_{uv})^T \), and (18) should be written as,
\[ P_{2uv}^T K^{-T} t \wedge K^{-1} P_{2uv} = P_{2uv}^T K^{-T} t \wedge R K^{-1} P_{1uv} \]  
(19)

The left side of the equation equals to zero, and then,
\[ P_{2uv}^T K^{-T} t \wedge R K^{-1} P_{1uv} = 0 \]  
(20)

Therefore, we can use matched feature points to calculate the initial camera pose, namely \( R \) and \( t \). Additionally, the essential matrix is defined as,
\[ E = t \wedge R \]  
(21)

The essential matrix \( E \) has 5 degrees of freedom and can be solved using at least 5 pairs of detected points. However, the form of calculating essential matrix \( E \) by 5 pairs of points is complicated, because of the nonlinear property of the \( E \). Therefore, the classic Eight-point method is adopted [33].

The ORB mentioned in Section IV is first adopted to extract 1000 features for each underground pixel-wise image. The dynamic points detection and rejection algorithm proposed in Section 3.2 is adopted to refine ORB feature points, which eliminates interference of underground dynamic scenes. Then, the Fast Library for Approximate Nearest Neighbour (FLANN) is used for matching features. For a pair of matched ORB feature points \( P_1 \) and \( P_2 \), their normalized coordinates are assumed as \( x_1 = [u_1, v_1, 1]^T \), \( x_2 = [u_2, v_2, 1]^T \), the relation of the \( E \) and matched points can be expressed as,
\[ \begin{pmatrix} (u_2, v_2, 1) \end{pmatrix} \begin{pmatrix} e_1 & e_2 & e_3 \\ e_4 & e_5 & e_6 \\ e_7 & e_8 & e_9 \end{pmatrix} \begin{pmatrix} u_1 \\ v_1 \\ 1 \end{pmatrix} = 0 \]  
(22)

For 8 pairs of matched features, the vector form of the essential matrix \( E \) is,
\[ \begin{pmatrix} u_1^1 u_2^1 & u_1^1 v_2^1 & v_1^1 v_2^1 & u_1^1 v_1^1 & v_1^1 v_1^1 & v_1^1 v_1^1 & v_1^1 v_1^1 & v_1^1 v_1^1 & v_1^1 v_1^1 \\ u_2^1 u_2^1 & u_2^1 v_2^1 & v_2^1 v_2^1 & u_2^1 v_1^1 & v_2^1 v_1^1 & v_2^1 v_1^1 & v_2^1 v_1^1 & v_2^1 v_1^1 & v_2^1 v_1^1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ u_1^8 u_2^8 & u_1^8 v_2^8 & v_2^8 v_2^8 & u_2^8 v_1^8 & v_2^8 v_1^8 & v_2^8 v_1^8 & v_2^8 v_1^8 & v_2^8 v_1^8 & v_2^8 v_1^8 \end{pmatrix} \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \\ e_8 \\ e_9 \end{pmatrix} = 0 \]  
(23)

The above equation is linear and each element of \( E \) can be calculated. Then, \( R \) and \( t \) of \( E \) are calculated with Singular Value Decomposition (SVD) method and optimized by performing re-projection error optimization with the remaining matched feature points. The relation of Lie Group and Lie Algebra is adopted to express \( R \) and \( t \) with \( \exp(t \wedge) \). Therefore, the re-projection error between measured data and estimated value for a single landmark point \( y_j \) can be defined as,
\[ e_{y,j,t} = z_{t,j} - \frac{1}{s} \cdot K \exp(t \wedge) y_j \]  
(24)

The sum-squared error can be expressed as,
\[ J_{cam}(x) = \sum_i \sum_j e^T_{y,j,t} \Omega^{-1}_{t,j} e_{y,j,t} \]  
(25)

The Levenberg-Marquardt method [34] is adopted to optimize above equation and then UAV initial pose can be obtained. After UAV initial pose optimization, the current frame is defined as the key frame. Next, the absolute scale information will be recovered. Considering noise and system error, the built-in IMU sensor model can be expressed as,
\[ \begin{pmatrix} \dot{a} \\ \dot{\omega} \end{pmatrix} = \begin{pmatrix} a \\ \omega \end{pmatrix} + \begin{pmatrix} b_a \\ b_\omega \end{pmatrix} + \begin{pmatrix} \eta_a \\ \eta_\omega \end{pmatrix} \]  
(26)

where subscript \( W \) represents the variable’s transformation from IMU to world coordinate system, \( W^a \) and \( W^\omega \) represent the accelerated velocity and the acceleration of gravity of the UAV in world coordinate system, respectively. \( \omega \) with head \( \sim \) represents the angle velocity of IMU measurement and \( \omega \) represents the actual value without noise. Similarly, \( a \) with head \( \sim \) represents the measure value of the accelerated velocity, \( b \) and \( \eta \) represent the noise and the zero-drift in the IMU measurement, where zero-drift \( b \) follows the random walk model and noise \( \eta \) is the white Gaussian noise. By measuring the angular and accelerated velocities in IMU coordinate system, equation (12) can be expressed by dynamic differential model with rotation matrix \( R \), velocity \( v \) and position \( P \) as follows,
\[ \begin{pmatrix} \dot{R}_{WI} \\ \dot{w} \\ \dot{P} \end{pmatrix} = \begin{pmatrix} R_{WI} \omega \wedge \\ w \dot{v} = a \\ w \dot{P} = w v \end{pmatrix} \]  
(27)

The pre-integration [35] is used to calculate the IMU measurement. Then, \( R, v \) and \( P \) can be expressed as,
\[ \begin{align*}
R_{WI}^{k+1} &= R_{WI}^k \exp \left( (\dot{\omega}_{W}^k - b_\omega - \eta_\omega) \Delta t \right) \\
w v_{W}^{k+1} &= w v_{W}^k + w g \Delta t + R_{WI}^k (\dot{\omega}_{W}^k - b_\omega - \eta_\omega) \Delta t \\
w P_{W}^{k+1} &= w P_{W}^k + w v_{W}^k \Delta t + \frac{1}{2} w g \Delta t^2 + \frac{1}{2} R_{WI}^k (\dot{\omega}_{W}^k - b_\omega - \eta_\omega) \Delta t \tag{28}
\end{align*} \]

where \( R_{WI}^k \) represents the rotation matrix of the \( k \)th key frame from IMU to world coordinate system, \( \exp(t \wedge) \) represents exponent mapping relation from Lie Algebra to Lie Group, \( \Delta t \) is
the interval between two adjacent key frames, $v_I$ and $wP_I^k$, represent the velocity and position of IMU sensor in world coordinate system.

The sampling frequency of the built-in IMU is 263Hz, D435I camera frequency is 30Hz. The frequency of camera is greater than IMU’s. The IMU measurement between the consecutive key frames $k$ and $k+1$ can be calculated as,

$$
\begin{align*}
R_{W_k}^{k+1} &= R_{W_k}^k R_{k,k+1} \exp \left( J_g^k b^k \right) \\
 v_{I_k}^{k+1} &= v_{I_k}^k + w \gamma \Delta t + R_{W_k}^{k+1} \left( v_{k,k+1} + J_g^k b^k + J_a^k b^k \right) \\
 wP_I^{k+1} &= wP_I^k + wv_I^k \Delta t + \frac{1}{2} w \gamma \Delta t^2 \\
 &+ R_{W_k}^{k+1} \left( P_{k,k+1} + J_g^k b^k + J_a^k b^k \right)
\end{align*}
$$

(29)

where $J_g$ and $J_a$ represent the first order approximations of gyroscope and acceleration deviation. Let the scale factor of the IMU measured data, the target loss function of tightly-visual observation data [36]. Based on (24), after fusing successfully recovered, the triangulation method is adopted for defining the transformation from camera to IMU coordinate system be $s$, the space relation equation between the camera and IMU coordinate system can be expressed,

$$
wP_I = swP_C + R_{WCC} P_I
$$

(30)

Substitute (30) into (29),

$$
swP_C^{k+1} = swP_C^k + wv_I^k \Delta t + \frac{1}{2} w \gamma \Delta t^2 \\
&+ R_{W_k}^{k+1} \Delta P_{k,k+1} + \left( R_{W_k}^{k+1} - R_{W_k}^{k+1} \right) C P_I
$$

(31)

The scale factor $s$ can be solved by solving this linear equation. Meanwhile, the pre-integration is also used for defining the measured error of IMU. The IMU error vector is defined as,

$$
e_I^k = \left[ e_{R}^k, e_v^k, e_p^k \right]^T
$$

(32)

Each element of the error vector is defined as,

$$
\begin{align*}
e_R &= \log \left( R_{k-1,k} \exp \left( J_g^k b^k \right) \right) \\
e_v &= R_{IW}^k wv_I^k - wv_I^{k-1} - w \gamma \Delta t \\
&- \left( \Delta v_{k-1,k} + J_v^k b^k + J_a^k b^k \right) \\
e_p &= R_{IW}^k wP_I^k - wP_I^{k-1} - wv_I^{k-1} \Delta t - \frac{1}{2} w \gamma \Delta t^2 \\
&- \left( \Delta P_{k-1,k} + J_p^k b^k + J_p^k b^k \right)
\end{align*}
$$

(33)

After the scale information of initial pose estimation is successfully recovered, the triangulation method is adopted to estimate space positions of feature points. And then the tightly-coupled optimization method is adopted to perform the UAV pose estimation utilizing IMU measured data and visual observation data [36]. Based on (24), after fusing the IMU measured data, the target loss function of tightly-coupled optimization can be expressed as a combination of two measurement error terms,

$$
J(x) = \sum_k \sum_j e_{y,j,k}^T Q_{i,j}^{-1} e_{y,j,k}
$$

(34)

where $Q_{i,j}$ is covariance matrix of D435I visual measurement, $\sum_l$ is information matrix of the IMU pre-integration and $\sum_R$ is the random walk of zero-shift error. By minimizing the target loss function $J(x)$, the optimal solution of pose estimation is obtained as follows,

$$
x = \arg\min(J(x))
$$

(35)

The problem defined in the above equation is a nonlinear least square problem. The optimization methods including the steepest descent method [37], the Gauss-Newton method [38] and the Levenberg-Marquardt method can be used. The steepest descent method is so greedy that falling into saw-tooth route, lead to the increase of the iteration numbers. Thus, the Levenberg-Marquardt method is eventually adopted in this paper to estimate the UAV pose after the fusion of IMU measured data and visual observation data.

VI. EXPERIMENTAL RESULTS

In this section, experimental results are presented to demonstrate the effectiveness of our method. The performance of our method is tested on the public EuRoC dataset collected by a UAV. The sequences in EuRoC dataset containing motion blur and poor illumination are challenging, which could be used for evaluating the adaptiveness of our method in real UAV data. We also integrated the proposed localization system with ROS and qualitatively demonstrated it using dynamic tunnel data collected by hand-held sensor device. All the experiments are performed on a computer with Intel i7 CPU, 32GB Memory and GeForce GTX 1080. The average frame rate of processing images of our method is measured to be 25fps, which can satisfy the real-time requirement. Besides, the comparison of our method with the state-of-art system VINS-Mono [19] for visual-inertial localization is performed.

A. EVALUATION USING EUROC DATASET

The EuRoC dataset [39] has a total of eleven sequences collected on-board a micro aerial vehicle. Sequences are measured in a machine hall and two different rooms, and are classified into easy, medium, and difficult levels according to the illumination, texture, and motion speed. The EuRoC dataset provides synchronized stereo images, IMU measurement, and ground truth. The imaging system uses a global shutter camera, supports the size of $752 \times 480$, and the image data frequency of 20 fps, and IMU data supports the data frequency of 200 Hz. The ground-truth data is used for comparison when analyzing the accuracy of the estimated trajectory.

In this paper, the Absolute Trajectory Error (ATE) and the Relative Pose Error (RPE) [40] are selected to evaluate the stability and the robustness of each system. Metric ATE
TABLE 5. Results of absolute trajectory error (ATE).

| Data       | VINS-Mono | Our work | Improvement |
|------------|-----------|----------|-------------|
| MH_01_easy | 0.08993   | 0.11859  | -31.87%     |
| MH_02_easy | 0.10235   | 0.05112  | 50.05%      |
| MH_03_medium | 0.06759  | 0.05354  | 20.79%      |
| MH_04_difficult | 0.20507 | 0.14307  | 30.23%      |
| V2_02_medium | 0.12999  | 0.06437  | 50.48%      |
| V1_03_difficult | 0.20089 | 0.09844  | 50.10%      |

TABLE 6. Results of metric translational drift (RPE).

| Data       | VINS-Mono | Our work | Improvement |
|------------|-----------|----------|-------------|
| MH_01_easy | 0.13221   | 0.03066  | 76.81%      |
| MH_02_easy | 0.13392   | 0.02822  | 78.95%      |
| MH_03_medium | 0.24869  | 0.07192  | 71.08%      |
| MH_04_difficult | 0.25189 | 0.08050  | 68.04%      |
| V2_02_medium | 0.12532  | 0.04319  | 65.54%      |
| V1_03_difficult | 0.13444 | 0.07955  | 40.83%      |

TABLE 7. Results of metric rotational drift (RPE).

| Data       | VINS-Mono | Our work | Improvement |
|------------|-----------|----------|-------------|
| MI_01_easy | 0.08789   | 0.02021  | 77.01%      |
| MH_02_easy | 0.08447   | 0.02028  | 75.99%      |
| MH_03_medium | 0.10432  | 0.02967  | 71.56%      |
| MH_04_difficult | 0.08692 | 0.02302  | 73.52%      |
| V2_02_medium | 0.16973  | 0.06107  | 64.02%      |
| V1_03_difficult | 0.18752 | 0.09093  | 51.51%      |

stands for global consistency of trajectory, whereas RPE is applicable for estimating the drift of translation and rotation. The value of root-mean-square (RMSE) is presented. The value of improvement of our method compared to the VINS-Mono is also shown in the tables. The improvement value is calculated as follows,

\[
\eta_V = \frac{o_V - u}{o_V} \times 100\% \tag{36}
\]

where \(o_V\) and \(u\) are the errors of VINS-Mono and our work, respectively. \(\eta_V\) represent the improvement percentages of the localization accuracy of our work compared to VINS-Mono.

We evaluate the mentioned localization systems in two scenes of the EuRoC dataset including sequences MH_01_easy, MH_02_easy, MH_03_medium, MH_04_difficult, V2_02_medium and V1_03_difficult. Tables from 5 to 7 indicate the testing results on six sequences. In terms of ATE and RPE, improvement can reach up to 50.48% and 78.93%. On sequences MH_04_difficult and V1_03_difficult representing dynamic flights with motion blur and poor illumination, our method outperforms VINS-Mono. The results indicate that our method can improve the robustness and stability in real UAV scenarios. On sequence MH_01_easy, the result of ATE of VINS-Mono is better than ours, but fortunately, our approach also achieve a high accuracy with a small system drift of RPE about 3cm.

B. EVALUATION USING COLLECTED DYNAMIC UNDERGROUND DATA

In order to verify the performance of the proposed UAV localization method with the real dynamic underground tunnel scene data, the underground tunnel of Orbital Vibration Reduction and Control Laboratory of Beijing Jiaotong University is chosen. The tunnel is a horseshoe-shaped structure. Eight cameras connected to a PC through a switch are used for obtaining the pose of markers on the sensor device. Fig. 6 shows the topology of tunnel cameras. In order to maximize the MCS covering space, four cameras \(C_1, C_3, C_5\) and \(C_7\) with interval of 4m are set at the tunnel entrance and close to the left side of tunnel wall. \(C_2\) is set at the central point between \(C_1\) and \(C_3\) on the right side of tunnel wall. \(C_4, C_6\) and \(C_8\) follow \(C_2\) with the same interval of 4m. When the hand-held sensor device moves into the circular area in Fig. 7, markers’ pose is captured by \(C_1, C_2\) and \(C_3\). Once the device leaves the detection range of \(C_1\), the pose can be captured by \(C_2, C_3\) and \(C_4\). This topology can ensure maximum coverage of MCS and obtain markers’ pose in real-time.

In the test process, testers wear the safety helmets and walk in the tunnel to simulate the operating workers. The more testers are, the more challenging data is. So our method is tested with different number of testers in front of the devices. Table 8 shows types of data collection. Fig. 7 shows the real scene recorded by hand-held device. During the data collection process, the MCS captures the ground truth of markers’ pose which is sent to Intel NUC carried by hand-held sensor device. Meanwhile, Intel UNC runs ubuntu16.04 and ROS Kinetic, records the built-in IMU data with the rate of 263Hz and the D435i camera data with the image size 480 × 640 and the frame rate of 30 fps.
Tables from 9 to 11 indicate the testing results on collected dynamic underground data. For VINS-Mono, the RMSE value of ATE on the collected data can reach up to 38cm, and is almost twice as large as that tested on EuRoC dataset, which demonstrated that the collected underground dynamic data impose challenge to localization algorithms. In terms of average ATE, for high-dynamic data P3_D0 and P3_D1, the RMSE improvement values can reach up to 46.33% and 67.84% respectively. Meanwhile, compared to VINS-Mono, our method has a smaller translational and rotational drift of RPE. The results demonstrate that the proposed method has a greater ability to process the dynamic environment and has a greater system stability in real tunnel scene. Next, VINS-Mono and our work are evaluated utilizing the representative dynamic data P2_D0 from these aspects: tracking global 2D trajectory and tracking trajectory along x, y and z axes.

Fig. 8 shows the results of tracking global 2D trajectory heat map of each system. The grey line represents ground truth of trajectory and the colored line represents the trajectory of corresponding algorithm. Compared to VINS-Mono, our method has a smaller drift about 6mm almost during the whole trajectory, and has a relatively large drift about 59cm only in a small period of trajectory, which is caused by a rapid movement of sensor device. It can be concluded that our method is more robust than VINS-Mono during the whole trajectory.

Fig. 9 shows the results of tracking trajectory along x, y and z axes. The grey line represents the ground truth of the trajectory, and the blue and green lines represent system trajectories of VINS-Mono and our work respectively. For one thing, the maximum of drift distance of VINS-Mono reaches

**TABLE 9. Results of absolute trajectory error (ATE).**

| Data  | VINS-Mono | Our work | Improvement |
|-------|-----------|----------|-------------|
| P0_D0 | 0.16911   | 0.14933  | 11.69%      |
| P0_D1 | 0.15394   | 0.14549  | 5.49%       |
| P1_D0 | 0.21915   | 0.13729  | 37.35%      |
| P1_D1 | 0.24420   | 0.13538  | 44.56%      |
| P2_D0 | 0.32631   | 0.12154  | 62.75%      |
| P2_D1 | 0.30251   | 0.15517  | 48.71%      |
| P3_D0 | 0.38359   | 0.20588  | 46.33%      |
| P3_D1 | 0.37679   | 0.12116  | 67.84%      |
up to 0.7m in the direction of z axes. For another, VINS-Mono has a poor performance during the whole trajectory in the direction of y axes. It can be seen that the accuracy of the proposed localization method outperforms VINS-Mono in either direction of x, y and z axes.

VII. CONCLUSION

In this paper, a visual-inertial localization method for underground UAVs in dynamic environments is proposed. In order to improve the performance and increase the robustness of the localization system in underground high-dynamic scenes, a dynamic points detection and rejection algorithm is presented, which combines a semantic segmentation neural network ENet, an LK optical flow method and an epipolar constraint method to refine ORB feature points set. After eliminating the interference of dynamic features on e.g. walking workers, the underground UAV pose estimation method based on visual-inertial sensor fusion is adopted to overcome the drawbacks of monocular visual system lacking scale information and low system reliability caused by the decrease of the feature points. Furthermore, a hand-held mobile sensor data collection device is developed and calibrated, which can imitate flights of underground UAVs in underground tunnels. The proposed localization system is integrated with ROS and extensive experiments are performed to verify the robustness and stability of the localization system. The results indicate that the proposed localization method can be applied to underground dynamic scenes with high robustness and accuracy requirements of localization system. It will be integrated into real UAV platforms in our future works.

REFERENCES

[1] S. Lee and Y. Choi, “Reviews of unmanned aerial vehicle (drone) technology trends and its applications in the mining industry,” Geosyst. Eng., vol. 19, no. 4, pp. 197–204, Jul. 2016.
[2] Y. Wei, S. Chunyan, and L. Junbo, “Underground unmanned aerial vehicles cluster system for monitoring and its key technologies,” Ind. Mine Autom., vol. 45, no. 4, pp. 9–15, Apr. 2019.
[3] S. Kuutti, S. Fallah, K. Katsaros, M. Dianati, F. Mccullough, and A. Mouzakitis, “A survey of the state-of-the-art localization techniques and their potentials for autonomous vehicle applications,” IEEE Internet Things J., vol. 5, no. 2, pp. 829–846, Apr. 2018.
[4] K. Feng, J. Li, X. Zhang, X. Zhang, C. Shen, H. Cao, Y. Yang, and J. Liu, “An improved strong tracking cubature Kalman filter for GPS/INS integrated navigation systems,” Sensors, vol. 18, no. 6, p. 1919, Jun. 2018.
[5] Y. Guo, M. Wu, K. Tang, J. Tie, and X. Li, “Covert spoofing algorithm of UAV based on GPS/INS-integrated navigation,” IEEE Trans. Veh. Technol., vol. 68, no. 7, pp. 6557–6564, Jul. 2019.
[6] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, “MonoSLAM: Real-time single camera SLAM,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 29, no. 6, pp. 1052–1067, Jun. 2007.
[7] R. Mur-Artal and J. D. Tardos, “ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras,” IEEE Trans. Robot.,
vol. 33, no. 5, pp. 1255–1262, Oct. 2017.

[8] J. Engel, T. Schöps, and D. Cremers, “LSD-SLAM: Large-scale direct monocular SLAM,” in Proc. ECCV, Zurich, Switzerland, 2014, pp. 834–849.

[9] J. Engel, V. Koltun, and D. Cremers, “Direct sparse odometry,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 3, pp. 611–625, Mar. 2018.

[10] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, “ENet: A deep neural network architecture for real-time semantic segmentation,” 2016, arXiv:1606.02147. [Online]. Available: http://arxiv.org/abs/1606.02147

[11] B. D. Lucas and T. Kanade, “An iterative image registration technique with an application to stereo vision,” in Proc. IJCAI, Vancouver, BC, Canada, 1997, pp. 674–679.

[12] J. Domke and Y. Aloimonos, “A probabilistic notion of correspondence and the epipolar constraint,” in Proc. 3rd Int. Symp. 3D Data Process., Visualization, Transmission. (3DPVT), Chapel Hill, NC, USA, Jun. 2006, pp. 41–48.

[13] M. Quigley, “ROS: An open-source robot operating system,” in Proc. ICRA, Kobe, Japan, 2009, p. 5.

[14] U. Challita, F. Ferdowski, M. Chen, and W. Saad, “Machine learning for wireless connectivity and security of cellular-connected UAVs,” IEEE Wireless Commun., vol. 26, no. 1, pp. 28–35, Feb. 2019.

[15] M. Chen, U. Challita, W. Saad, C. Yin, and M. Debbah, “Artificial neural networks-based machine learning for wireless networks: A tutorial,” IEEE Commun. Surveys Tuts., vol. 21, no. 4, pp. 3039–3071, 4th Quart., 2019.

[16] D. Nodland, H. Zargarzadeh, and S. Jagannathan, “Neural network-based optimal adaptive output feedback control of a helicopter UAV,” IEEE Trans. Neural Netw. Learn. Syst., vol. 24, no. 7, pp. 1061–1073, Jul. 2013.

[17] J. R. G. Braga, H. F. C. Velho, G. Conte, P. Doherty, and E. H. Shiguemori, “An image matching system for autonomous UAV navigation based on neural network,” in Proc. 14th Int. Conf. Control, Automat., Robot. Vis. (ICARCV), Phuket, Thailand, Nov. 2016, pp. 1–6.

[18] J. Fatt Chow, B. Bahadir Kocer, J. Henawy, G. Seet, Z. Li, W. Y. Yau, and M. Pratama, “Toward underground localization: Lidar inertial odometry enabled aerial robot navigation,” 2019, arXiv:1910.13085. [Online]. Available: http://arxiv.org/abs/1910.13085

[19] 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.

[20] S.-J. Lee and J.-H. Kim, Advances in Intelligent Systems and Computing, New York, NY, USA: Springer, 2013, pp. 477–484.

[21] H. Bavle, J. L. Sanchez-Lopez, A. Rodriguez-Ramos, C. Sampredo, and P. Campoy, “A flight altitude estimator for multicopter UAVs in dynamic and unstructured indoor environments,” in Proc. Int. Conf. Unmannned Aircr. Syst. (ICUAS), Miami, FL, USA, Jun. 2017, pp. 1044–1051.

[22] D. Zou and P. Tan, “CoSLAM: Collaborative visual SLAM in dynamic environments,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 2, pp. 354–366, Feb. 2013.

[23] C. Yu, Z. Liu, X.-J. Liu, F. Xie, Y. Yang, Q. Wei, and Q. Fei, “DS-SLAM: A semantic visual SLAM towards dynamic environments,” in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), Madrid, Spain, Oct. 2018, pp. 1168–1174.

[24] 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.

[25] P. Sturm, Pinhole Camera Model. New York, NY, USA: Springer, 2014, pp. 610–613.

[26] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” Int. J. Comput. Vis., vol. 60, no. 2, pp. 91–110, Nov. 2004.

[27] H. Bay, T. Tuytelaars, and L. Van Gool, “Surf: Speeded up robust features,” in Proc. ECCV, Graz, Austria, 2006, pp. 404–417.

[28] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. “ORB: An efficient alternative to SIFT or SURF,” in Proc. Int. Conf. Int. Vis. Comput. Conf. Vis., Barcelona, Spain, Nov. 2011, pp. 2564–2571.

[29] E. Rosten and T. Drummond, “Fusing points and lines for high performance tracking,” in Proc. 10th IEEE Int. Conf. Comput. Vis. (ICCV), Beijing, China, 2005, pp. 1508–1515.

[30] M. Calonder, “Brief: Binary robust independent elementary features,” in Proc. ECCV, Berlin, Germany, 2010, pp. 778–792.

[31] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The Pascal visual object classes challenge: A retrospective,” Int. J. Comput. Vis., vol. 111, no. 1, pp. 98–136, Jan. 2015.

[32] X. Shi, D. Li, P. Zhao, Q. Tian, Y. Tian, Q. Long, C. Zhu, J. Song, F. Qiao, L. Song, Y. Guo, Z. Wang, Y. Zhang, B. Qin, W. Yang, F. Wang, R. H. M. Chan, and Q. She, “Are we ready for service robots? The OpenLORIS-scene datasets for lifelong SLAM,” 2019, arXiv:1911.05603. [Online]. Available: http://arxiv.org/abs/1911.05603

[33] R. I. Hartley, “In defense of the eight-point algorithm,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 19, no. 6, pp. 580–593, Jun. 1997.

[34] A. Ranganath, “The Levenberg-Marquardt algorithm,” Tutorial LM Algo- rithm, vol. 11, no. 1, pp. 101–110, Jun. 2004.

[35] 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.

[36] A. Martinelli, “Closed-form solution of visual-inertial structure from motion,” Int. J. Comput. Vis., vol. 106, no. 2, pp. 138–152, Jan. 2014.

[37] J. Fliege and B. F. Svaiter, “Steepest descent methods for multicriteria optimization,” Math. Methods Oper. Res., vol. 51, no. 3, pp. 479–494, Aug. 2000.

[38] D. Salas-Gonzalez, J. M. Górriz, J. Ramírez, A. Lass, and C. G. Puntonet, “Improved gauss-Newton optimisation methods in affine registration of SPECT brain images,” Electron. Lett., vol. 44, no. 22, pp. 1291–1292, Oct. 2008.

[39] M. Burni, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omani, M. W. Achtelik, and R. Siegwart, “The EuRoC micro aerial vehicle datasets,” Int. J. Robot. Res., vol. 35, no. 10, pp. 1157–1163, Sep. 2016.

[40] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, “A benchmark for the evaluation of RGB-D SLAM systems,” in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst., Vilamoura-Algarve, Portugal, Oct. 2012, pp. 573–580.
DAN GUO is currently pursuing the bachelor’s degree in telecommunication engineering and management with the Beijing University of Posts and Telecommunications, Beijing, China. Her research interests include semantic segmentation of computer vision, digital signal and image processing, and wireless communication networks.

QI WEI received the Ph.D. degree from Tsinghua University, Beijing, China, in 2010. He is currently an Assistant Professor with the Department of Electronic Engineering, Tsinghua University. His research interests are MEMS inertial sensors ASIC design and high-performance data converters.

FEI QIAO (Member, IEEE) received the bachelor’s degree from Lanzhou University, China, in 2000, and the Ph.D. degree from Tsinghua University, China, in 2006. He is currently an Associate Professor with the Department of Electronic Engineering, Tsinghua University. He has authored around 90 articles and holds over 30 invented patents. His research interests include low power circuits design, and energy-efficient integrated perception circuits and systems for intelligent robots, and wearables and the Internet of Things (IoT) devices.

QI WEI (Member, IEEE) received the bachelor’s degree from Tsinghua University, Beijing, China, in 2006. He is currently an Associate Professor in the Department of Electronic Engineering, Tsinghua University. His research interests include low power circuits design, and energy-efficient integrated perception circuits and systems for intelligent robots, and wearables and the Internet of Things (IoT) devices.

QI WEI (Member, IEEE) received the Ph.D. degree from Tsinghua University, Beijing, China, in 2010. He is currently an Assistant Professor with the Department of Electronic Engineering, Tsinghua University. His research interests are MEMS inertial sensors ASIC design and high-performance data converters.