IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 1, 2020, accepted April 30, 2020, date of publication May 14, 2020, date of current version June 3, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2994348

SDF-SLAM: Semantic Depth Filter SLAM for Dynamic Environments

LINYAN CUI AND CHAOWEI MA
Image Processing Center, School of Astronautics, Beihang University, Beijing 102206, China
Corresponding author: Linyan Cui (cuily@buaa.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61875003, and in part by the Fundamental Research Fund for the Central Universities of China under Grant YWF-20-BJ-J-425.

ABSTRACT Simultaneous Localization and Mapping (SLAM) has been widely applied in computer vision and robotics. For the dynamic environments which are very common in the real word, traditional visual SLAM system faces significant drop in localization and mapping accuracy due to the static world assumption. Recently, the semantic visual SLAM systems towards dynamic scenes have gradually attracted more and more attentions, which use the semantic information of images to help remove dynamic feature points. Existing semantic visual SLAM systems commonly detect the dynamic feature points by the semantic prior, geometry constraint or the combine of them, then map points corresponding to dynamic feature points are removed. In the visual SLAM framework, pose calculation is essentially around the 3D map points, so the essence of improving the accuracy of visual SLAM system is to build a more accurate and reliable map. These existing semantic visual SLAM systems are actually adopting an indirect way to acquire reliable map points, and several drawbacks exist. In this paper, we present SDF-SLAM: Semantic Depth Filter SLAM, a visual semantic SLAM system towards dynamic environments, which utilizes the technology of depth filter to directly judge whether a 3D map point is dynamic or not. First, the semantic information is integrated into the original pure geometry SLAM system by the semantic optical flow method to perform reliable map initialization. Second, design the semantic depth filter that satisfies the Gaussian Uniform mixture distribution to describe the inverse depth of each map point. Third, updating the inverse depth of 3D map point in a Bayesian estimation framework, and dividing the 3D map point into active one or inactive one. Last, only the active map points are utilized to achieve robust camera pose tracking. Experiments on TUM dataset demonstrate that our approach outperforms original ORB-SLAM2 and other state-of-the-art semantic SLAM systems.

INDEX TERMS Dynamic scenes, depth filter, semantic segmentation, simultaneous localization and mapping.

I. INTRODUCTION

Simultaneous Localization and Mapping (SLAM) constructs a map of the surrounding world using the data collected by the platform operating SLAM system, and simultaneously locates itself within the map. The SLAM technology using visual sensors is called visual SLAM. The research on visual SLAM technology has been more than 30 years. Many excellent visual SLAM systems have been developed, such as MonoSLAM [1], PTAM [2], ORB-SLAM [3], ORB-SLAM2 [4], LSD-SLAM [5], SVO [6], DSO [7]. The current SLAM system usually assumes that the environment is static. In order to make the visual SLAM system more practical, it is an urgent problem to improve the accuracy of the visual SLAM system in the dynamic scene.

In recent years, with the progress of deep learning algorithm and the improvement of computing performance, classic image processing tasks (such as image classification, target detection, semantic segmentation, etc.) can be well completed by computer. Among them, image semantic segmentation can get pixel level semantic classification results. According to these classification results, we can know the prior attributes of each pixel in the image. For example, the pixels with the semantic category of building hardly

The associate editor coordinating the review of this manuscript and approving it for publication was Yu-Huei Cheng.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

produce movement and change, while the pixels with the semantic category of human often produce movement and appearance change. These semantic information can provide the information of dynamic elements in the scene for SLAM.

The combination of traditional visual SLAM and semantic segmentation based on deep learning can greatly improve the robustness and accuracy of SLAM system in dynamic environment. Semantic visual SLAM is a new research field, but there is no mature and consistent scheme about how to use semantic information.

A. RELATED WORK

Recently, the semantic visual SLAM for dynamic scenes has gradually attracted more and more attentions, which uses the semantic information of images to help remove dynamic feature points. The semantic SLAM algorithms can be mainly classified into three classes: the pure semantic SLAM which solely adopts the semantic information, the semantic SLAM which couples loosely the semantic information and geometry calculation, and the semantic SLAM which couples tightly the sematic information and geometry calculation.

Approaches that solely depend on the semantic information are straightforward. By applying the classic semantic segmentation networks, such as YOLO [8], SSD [9], SegNet [10], Mask-RCNN [11], PSPNet [12], and Deeplab [13], the semantic labels of the extracted image features in visual SLAM framework can be obtained. When the objects in the image are recognized as movable objects, such as people, cat, and car, the features located on these objects are thought as dynamic features and will be directly removed [14]–[16] or further processed through a selective tracking method in the tracking thread of SLAM [17] to determine whether they are retained or removed. The idea of using semantic information to detect dynamic feature points is very simple and direct, but it also has some limitations, mainly including two aspects: first, the semantic dynamic feature points do not completely coincide with the actual dynamic ones; second, the semantic segmentation results have errors especially in the boundary region of objects.

In view of the limitation of pure semantic visual SLAM which solely adopts the semantic information, some recent semantic SLAM works towards the dynamic scene couple the semantic information and geometry calculation. The semantic visual SLAM systems which couple loosely the semantic information and geometry calculation are proposed. For DS-SLAM [18], it used SegNet [10] to obtain the semantic labels of feature points in a separate thread. For example, if a feature point is classified as potentially movable, such as 'human', then epipolar geometric constraint is used to further detect its dynamics by checking the moving consistency. If the detection result with geometric constraint is dynamic, then all feature points with semantic category of 'human' will be considered dynamic, and then be removed. The essence of this method is to take the intersect of the results of semantic prior and geometric constraint: only the feature points which are both dynamic in semantics and geometry are considered as dynamic feature points. For DynaSLAM proposed by Bescos *et al.* [19], it used Mask-RCNN [11] to obtain semantic segmentation results and then judged the dynamic characteristics of feature points. Meanwhile, it detected the dynamic characteristics of feature points according to multi-view geometry consistency. Then it took the union of the two detection results. For a feature point, as long as either of the two detection results is dynamic, the feature point is considered to be dynamic and removed. For the PSPNet-SLAM proposed by Han and Xi [20], it combined the PSPNet [12] and optical flow to detect and eliminate dynamic feature points. The optical flow is firstly used to judge and cull the dynamic point, and then dynamic characteristics of the remaining feature points are detected by judging whether they fall within the semantic dynamic object which is obtained by the PSPNet semantic segmentation. For the method proposed by Zhao *et al.* [21], it firstly used the Mask-RCNN and edge refinement to obtain the contour of potentially dynamic object, and then the optical flow is implemented to further detect the state of potentially dynamic object by checking the consistency of potentially dynamic object and background areas. In general, the 'loosely coupled' scheme takes either intersection or union of the two detection results from the semantic information and the geometry calculation. These two parts are implemented independently and perform independent functions respectively. There is no interaction between them, which may lead to the insufficient use of semantic information and geometry calculation.

To further utilize the semantic information and geometric constraint, the semantic visual SLAM which tightly couples the semantic information and geometry calculation is proposed to detect the dynamic feature points. For SOF-SLAM proposed by Cui and Ma [22], it coupled the semantic information and geometric information in a unified framework. The SegNet was firstly used to get pixel-wise semantic segmentation of each image and used to get a relatively reliable fundamental matrix. Then the fundamental matrix is used to further detect dynamic features through geometry constraint. In this approach, fundamental matrix serves as the bridge that links these two sources of information in a unified framework and only one decision is made whether a feature is dynamic or not. The hidden dynamic characteristic in semantic and geometry information is further utilized to remove dynamic feature more effectively.

B. MOTIVATION

The semantic visual SLAM algorithms mentioned above have common problems: they detect the dynamic scenes by judging whether the feature points are dynamic or not from the 2D image level and commonly use the information of adjacent frames to eliminate the dynamic points.

There are two problems in doing so:

1) If the motion of dynamic elements in the scene is not fast enough, or the frame rate of the image is very high, the image feature points on the dynamic elements will not show particularly obvious motion in the adjacent

two frames. In this case, these dynamic feature points are easily confused with the static ones.

2) The dynamic information in adjacent frames is limited. It is easy to be interfered by noise which may arises from semantic segmentation error, fundamental matrix calculation, slow motion of objects, and so on.

In order to solve the above problems, we use a probabilistic framework to continuously accumulate image data inputs. In this probabilistic framework, we maintain the probability distribution of each map point's inverse depth.

C. CONTRIBUTION AND OUTLINE

In this paper we propose a visual semantic SLAM system toward dynamic environment, i.e. Semantic Depth Filter SLAM (SDF-SLAM), which is built on ORB-SLAM2. We use the RGB-D version of ORB-SLAM2 which takes both RGB image and depth map as input. This framework aims at making the system more accurate in dynamic environments. This work is the continuation of our previous work [22], and it is superior than [22] by solving the problems mentioned in section I.B. The proposed SDF-SLAM system can highly reduce the influence of dynamic objects in the environment through two modules we added to the original ORB-SLAM2 framework, including the map initialization module based on tightly coupling semantic information and geometry information, and the camera pose tracking module based on depth filter.

Our contribution can be summarized as follows:

1) Utilize the technology of depth filter to directly judge whether a 3D map point is dynamic or not, while the existing semantic visual SLAM systems commonly detect the dynamic features points from the 2D image level. More reliable dynamic information is obtained.
2) Build a reliable initial map based on the semantic optical flow. The semantic information is integrated into the original pure geometry SLAM system in order to make the depth filter work more normally.
3) Adopt the probability framework to maintain the map. For a map point, all the image data related to it contribute its dynamic judgment. More abundant information about the dynamic characteristics of map points can be obtained, which lead to better robustness to noise and single outlier observation.

The rest of the paper is structured as follows: the proposed SDF-SLAM is described in Section 2. First, the system overview is presented. Second, the necessity of a semantic information aided map initialization module is discussed and how to initialize map is also presented. Third, the procedure of how to use depth filter in the framework of SLAM is demonstrated. Section 3 evaluates the accuracy of our system on TUM RGB-D dataset and compares our system with the state-of-the-art semantic visual SLAM systems toward dynamic environments. Finally, a summary is provided in Section 4.

II. SEMANTIC DEPTH FILTER SLAM

In this section, the proposed SDF-SLAM system will be clarified in detail. In the proposed system, there are two important modules that work together to make the whole system operate robustly in dynamic scenes, including the map initialization module based on semantic optical flow and the robust tracking module based on semantic depth filter. In the following part of this chapter, we will first give an overview about the whole system. Then the two modules mentioned above will be demonstrated separately.

A. SYSTEM OVERVIEW

The overall architecture of the proposed SDF-SLAM is shown in Fig.1. The whole SDF-SLAM system is built upon ORB-SLAM2. The local mapping and loop closing threads are the same as ORB-SLAM2.

In the visual SLAM framework, pose calculation is essentially around the 3D map points in this map, so the essence of improving the accuracy of visual SLAM system is to build a more accurate and reliable map. Therefore, the proposed SDF-SLAM firstly initializes a more reliable map with the semantic optical flow. A separate semantic segmentation thread is added to provide semantic label for map point and helps to build a reliable initial map. Then, updating the map with the depth filter which is designed and integrated into the tracking thread to generate reliable map points, and the dynamic 3D map points are removed more reasonably and effectively. The updated 3D map where the dynamic map points are removed is used by the tracking thread, and more robust camera pose calculation is performed towards dynamic environments.

In our approach, the map initialization based on the semantic optical flow and the map update based on the depth filter are the two most important modules, so they will be stated in detail.

B. SEMANTIC OPTICAL FLOW BASED MAP INITIALIZATION IN DYNAMIC SCENE

In SLAM system, the map is the reconstruction of surrounding environment based on the acquired sensor data. When the platform that operates the SLAM system moves around and sensor data about new surroundings are acquired, the current camera pose within the map and the map itself will expand with the newly acquired sensor data. Therefore, when the system just starts to run, a map initialization module is needed to build an initial map, then subsequent calculation will be around this map. As the initial map is the origin of all subsequent calculation, the quality of initial map is of vital importance to the system accuracy. In dynamic scene the initial map is easy to be disturbed by dynamic objects, so we propose to initialize the map with the semantic optical flow method proposed in our previous work [22].

The procedure of map initialization is shown in Fig.2. First, once current input image frame is chosen to initialize the map, SegNet [10] is used to get semantic segmentation

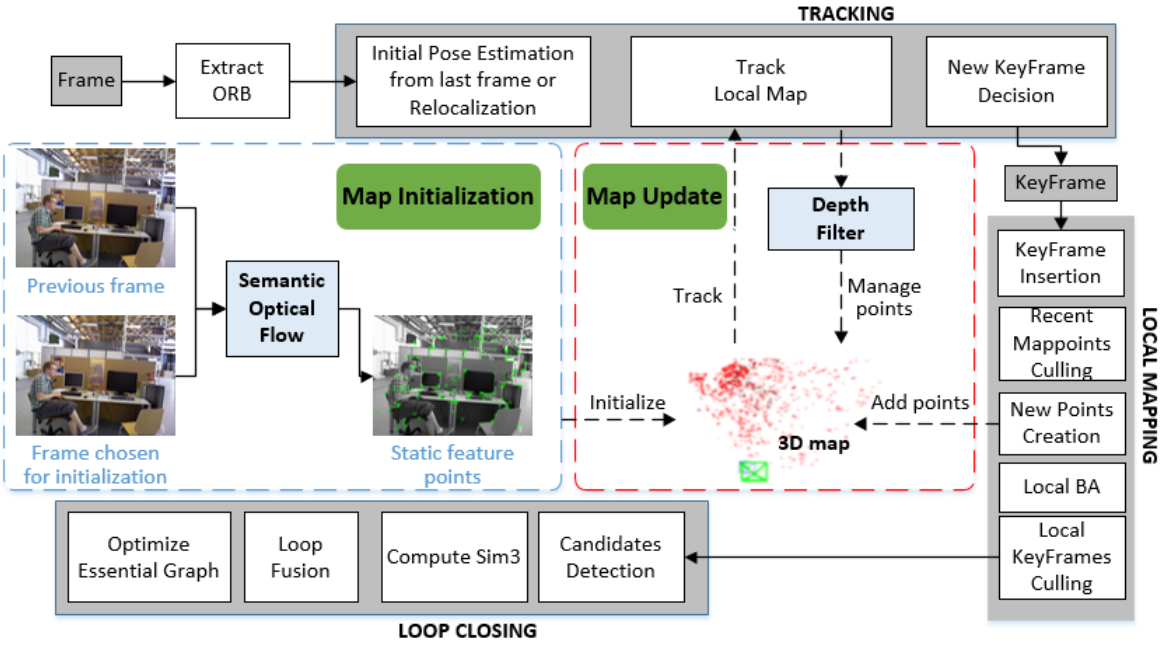


FIGURE 1. Overall-architecture for SDF-SLAM. The local mapping and loop closing threads are the same as ORB-SLAM2. A separate semantic segmentation thread is added to provide semantic label for map point and helps to build a reliable initial map. Depth filter is designed and integrated into the tracking thread to generate reliable map points where the dynamic map points are removed. Then, estimate camera pose solely depending on the reliable map points.

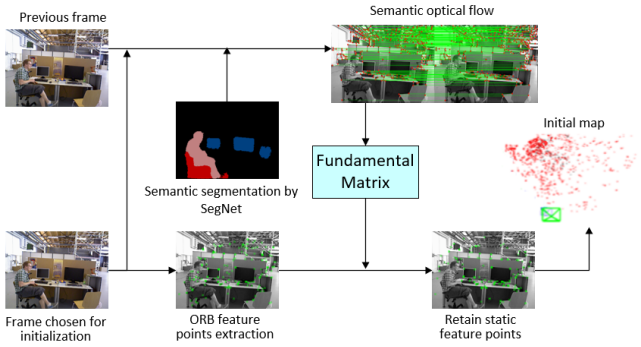


FIGURE 2. Map Initialization based on tightly coupling semantic and geometry information.

result, which is implemented in caffe framework and trained on PASCAL VOC dataset [23]. Twenty classes are obtained, including airplane, bicycle, bird, boat, bottle, bus, car, cat, chair, cow, dining table, dog, horse, motor bike, person, potted plant, sheep, sofa, train, and monitor. Based on the semantic results, the prior knowledge of each pixel can be roughly classified as static, dynamic, and potentially dynamic according to the human common sense.

Second, calculate the optical flow of current image frame and previous adjacent image frame to acquire feature point correspondences. Then the correspondences of both dynamic and potentially dynamic feature points are removed with the semantic segmentation prior. Only the semantic static feature points are used to calculate the relatively reliable fundamental matrix $\boldsymbol{F}$ with the 8-points algorithm [24].

Third, with the calculated $\boldsymbol{F}$, the epipolar line constraint is used to further judge the motion characteristics of all feature points in current image frame. If the feature point in current frame is static, its corresponding feature in last frame should resides closely to the epipolar line. We choose 1 pixel as the distance threshold, if the feature point in current frame whose corresponding feature in last frame is more than one pixel away from the epipolar line, it is considered as dynamic.

Fourth, construct a reliable initial map. Through the above three steps, we make good use of semantic information and geometry information to retain static feature points in current image frame. As we take RGB-D image as input data, when depth data of these feature points is added, they turn into 3D points and form the initial map. These 3D map points are almost all on the static background, which makes the static world assumption satisfied in dynamic scenes.

C. INVERSE DEPTH FILTER BASED MAP UPDATE IN DYNAMIC SCENE

After a reliable initial map is acquired, the SLAM system is able to start locating the camera within the map and expanding the map simultaneously. As for the image coming immediately after the map initialization procedure, the corresponding camera pose calculation is reliable as the initial map has been specially processed toward dynamic scene. However, these subsequent map expansion procedure will still be disturbed by the dynamic objects in the environment. If dynamic map points are added, the map will be polluted and not reliable anymore for further camera pose calculation.

A very straightforward idea to solve this problem is to adopt a more cautious strategy to expand the map.

In order to keep the updated map reliable for camera pose calculation, we firstly define two kinds of map points: activated map points and inactive map points. Map points which are confirmed static and reliable are determined as active, and the other map points are determined as inactive. Only active map points are involved in the calculation of camera pose. Therefore the key is to adopt a proper method to determine the active or inactive state of newly added map points.

In the following part, we will firstly introduce the technology of inverse depth filter. Then how to utilize inverse depth filter to detect dynamic map points is clarified. At last, we will demonstrate how to integrate inverse depth filter into the framework of ORB-SLAM2 to achieve proper map points management and achieve robust camera pose tracking in dynamic environment.

1) INVERSE DEPTH FILTER INTRODUCTION

Depth filter is proposed by Vogiatzis *et al.* in their video based real-time multi-view stereo system [25]. Depth filter adopts a probabilistic depth estimation scheme that updates posterior depth distributions with every new image frame in the video sequence. In order to deal with large scene depth, Forster *et al.* [6] replace the depth in depth filter with the inverse depth and implement inverse depth filter. Inverse depth filter is based on the assumption that the inverse depth of a map point in the corresponding image follows Gaussian Uniform mixture distribution:

$$p\left(x|Z,\pi\right)=\pi N\left(x|Z,\tau^2\right)+(1-\pi)\mathrm{U}\left(x|Z_{\min},Z_{\max}\right) \quad (1)$$

x is the measurement of the inverse depth of a map point, which is modeled as a random variable. Z is the true inverse depth of the map point, and is the value to be estimated. π is the probability that x is a good measurement that follows Gaussian distribution $N\left(x|Z,\tau^2\right)$ with the real inverse depth Z as the mean and τ^2 as the variance. $1-\pi$ is the probability that x is an outlier measurement that follows Uniform distribution $\mathrm{U}\left(x|Z_{\min},Z_{\max}\right)$ in the interval $[Z_{\min},Z_{\max}]$. According to (1), the distribution of measurement x is determined by two parameters: the true inverse depth Z and good measurement probability π, which is obvious in the expression $p\left(x|Z,\pi\right)$. Based on the model described by (1), if a series of independent inverse depth measurements of a map point are given, they can be regarded as samples of random variable x and be used to estimate parameters Z and π.

Inverse depth filter takes depth measurement uncertainty into account and adopts a probabilistic to fuse multiple measurements, which leads to a more noise-robust and reliable inverse depth estimation result.

2) DYNAMIC MAP POINTS DETECTION BASED ON DEPTH FILTER

Inspired by the idea of inverse depth filter, we extend the assumption of inverse depth's Gaussian Uniform mixture distribution to the SLAM application in dynamic environment. Although the model described by (1) is originally proposed to merge inverse depth measurements acquired in different baseline, it is also very suitable to detect dynamic map points in feature based SLAM system.

In ORB-SLAM2, ORB features are used to perform stereo matching procedure, which is very accurate compared with stereo matching algorithm based on photometric consistency. In dynamic environment, if the matching results are accurate enough between multiple views, which means there are almost no outlier measurements, the estimated π will be close to 1 when using (1) to merge multiple measurements of inverse depth. However, situation is different in dynamic environment. Even though matching results are accurate, multiple measurements aren't consistent if the map point is moving, which will lead to a low estimation value of π. Low value of π means the corresponding map point tends to produce bad measurement, but the measurement is actually right. Therefore, it is more reasonable to define π as the probability of map point being dynamic when we use the idea of inverse depth filter to detect dynamic map points. Then we can use the newly defined π to reinterpret (1): π is the probability that x is a measurement produced by a static map point that follows Gaussian distribution $N\left(x|Z,\tau^2\right)$ with the real inverse depth Z as the mean and τ^2 as the variance. $1-\pi$ is the probability that x is a measurement from a dynamic map point that follows Uniform distribution $\mathrm{U}\left(x|Z_{\min},Z_{\max}\right)$ in the interval $(Z_{\min},Z_{\max})$. As for a map point, a series of measurements of its inverse depth can be merged to get a more reliable estimation of Z, as well as the estimation of π which indicates its probability of being dynamic. The details are as follows:

$x_1,x_2,\cdots,x_n$ are a series of measurements of the inverse depth of a map point, they are independent of each other and all follow the distribution described by (1), now we want to estimate the parameters Z and π. This problem can be solved through maximum posterior approach:

$$\arg\max_{Z,\pi} p\left(Z,\pi|x_1,\ldots,x_n\right) \quad (2)$$

According to Bayes formula, $p\left(Z,\pi|x_1,\ldots,x_n\right)$ can be expanded as follows:

$$\begin{aligned} p\left(Z,\pi|x_1,\ldots,x_n\right) &= \frac{p\left(Z,\pi,x_1,\ldots,x_n\right)}{p\left(x_1,\ldots,x_n\right)} \\ &= \frac{p(Z,\pi)p\left(x_1,\ldots,x_n|Z,\pi\right)}{p\left(x_1,\ldots,x_n\right)} \\ &\propto p(Z,\pi)p\left(x_1,\ldots,x_n|Z,\pi\right) \end{aligned} \quad (3)$$

As the measurements are independent of each other, (3) can be further written as:

$$\begin{aligned} &p\left(Z,\pi|x_1,\ldots,x_n\right) \\ &\propto p(Z,\pi)p\left(x_1,\ldots,x_n|Z,\pi\right) \\ &= p(Z,\pi)p\left(x_1|Z,\pi\right)p\left(x_2|Z,\pi\right)\ldots p\left(x_n|Z,\pi\right) \end{aligned} \quad (4)$$

The range of Z is set to $(Z_{\min},Z_{\max})$ and the range of π is set to $(0,1)$. In the absence of a more reliable prior knowledge, it can be considered that $p(Z,\pi)$ follows a

two-dimensional uniform distribution [25]. Suppose that we sample m points in $(Z_{\min}, Z_{\max})$ and sample n points in (0, 1), then the distribution of $p(Z, \pi)$ can be approximated as:

$$p(Z, \pi) = p(Z)p(\pi) = \frac{1}{m} \cdot \frac{1}{n} \tag{5}$$

Substitute (5) into (4):

$$p(Z, \pi | x_1, \ldots, x_n) \propto \frac{1}{mn} p(x_1|Z, \pi) p(x_2|Z, \pi) \ldots p(x_n|Z, \pi) \tag{6}$$

With (6), the maximum posterior estimation problem described by (2) can be transformed into a maximum likelihood estimation problem:

$$\arg\max_{Z,\pi} p(x_1|Z, \pi) p(x_2|Z, \pi) \ldots p(x_n|Z, \pi) \tag{7}$$

However in SLAM system, the measurements are obtained in the form of video. If every time a new measurement is obtained, a new maximum likelihood estimation problem is calculated according to (7), many items are actually calculated repeatedly. Therefore, it is more reasonable to change formula (4) to recursive form:

$$p(Z, \pi | x_1, \ldots, x_n) \propto p(Z, \pi | x_1, \ldots, x_{n-1}) p(x_n|Z, \pi) \tag{8}$$

$p(Z, \pi | x_1, \ldots, x_{n-1})$ is the posterior in last moment, $p(x_n|Z, \pi)$ is the likelihood of current inverse depth measurement. If the distribution form is unknown, (8) is still hard to be solved. The authors in [25] approximate the true posterior with a Beta Gaussian distribution:

$$p(Z, \pi | x_1, \ldots, x_n) \approx q\left(Z, \pi | a_n, b_n, \mu_n, \sigma_n^2\right) \tag{9}$$

where $\left(a_n, b_n, \mu_n, \sigma_n^2\right)$ are the parameters of approximated Beta Gaussian distribution at current moment. Substitute (9) into (8):

$$q\left(Z, \pi | a_n, b_n, \mu_n, \sigma_n^2\right) \times \propto q\left(Z, \pi | a_{n-1}, b_{n-1}, \mu_{n-1}, \sigma_{n-1}^2\right) p(x_n|Z, \pi) \tag{10}$$

Then by matching the first and second order moments for Z and π between the left and right of (10), the recursive update formula of $\left(a, b, \mu, \sigma^2\right)$ can be acquired. The details on the derivation can be found in the original work [25].

Then $q\left(Z, \pi | a_n, b_n, \mu_n, \sigma_n^2\right)$ can be used to calculate the first moment for π, and this first moment can be approximated as the estimate of π:

$$\pi = \frac{a_n}{a_n + b_n} \tag{11}$$

$q\left(Z, \pi | a_n, b_n, \mu_n, \sigma_n^2\right)$ can also be used to calculate the first moment for Z, and this first moment can be approximated as the estimate of Z:

$$Z = \mu_n \tag{12}$$

Each time we get a new inverse depth measurement x_n, the inverse depth estimation value can be updated using (12). In the meanwhile, the probability of the map point being dynamic can be updated according to (11). If the change value of the estimation of Z is smaller than a predefined threshold, it is considered to be convergent. When the estimation of Z has converged: if π is higher than a predefined threshold, we consider the corresponding map point being static, and it will not be updated anymore. We admit there is small chance that this map point may turn to be dynamic, but it can be handled by the map point maintaince module of ORB-SLAM2, so it is resource consuming and worthless to keep updating it; Otherwise, it is considered as dynamic. If the estimation of Z hasn't converged, it means that more measurements are need to determine the motion characteristics of the corresponding map point.

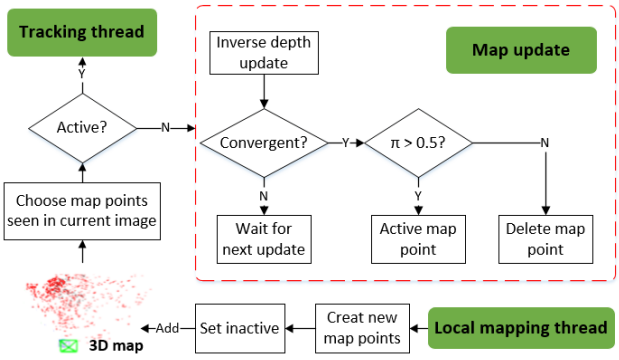


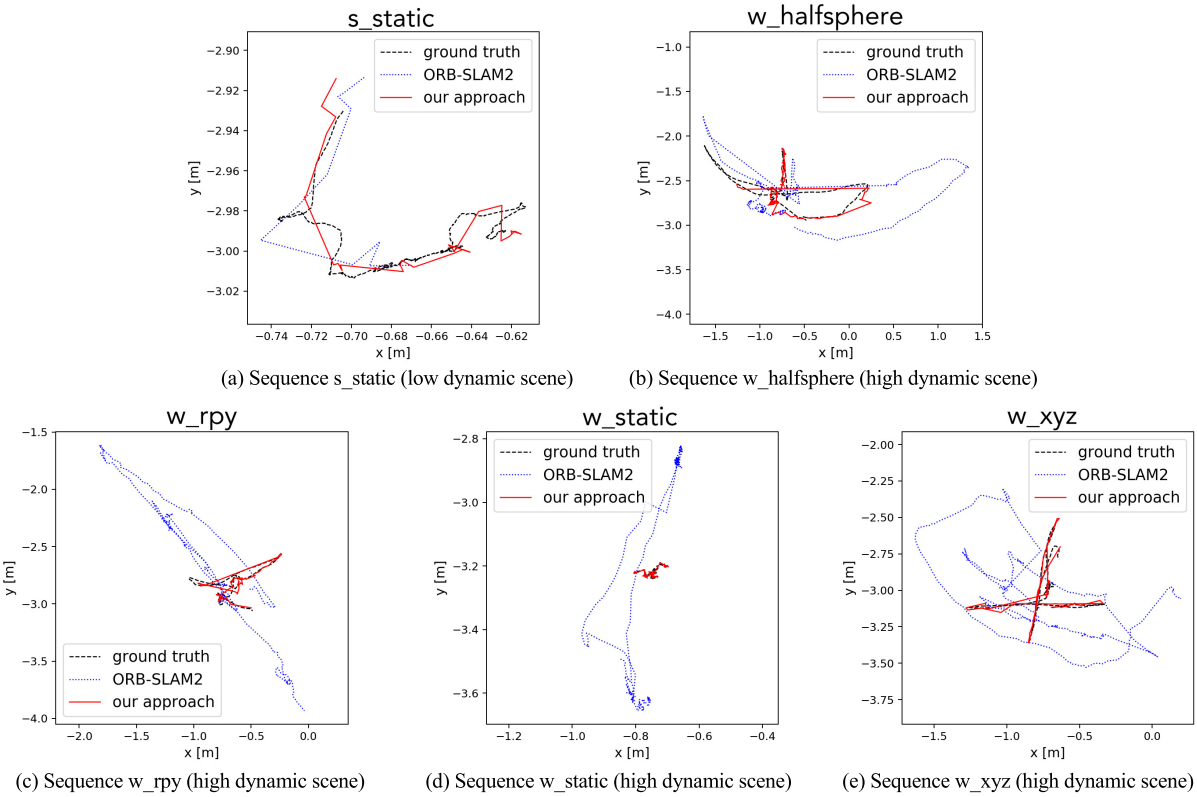
FIGURE 3. Map update and robust camera pose tracking.

3) MAP UPDATE AND rOBUST CAMERA POSE TRACKING

In this part, we will demonstrate how to integrate inverse depth filter into the framework of ORB-SLAM2 to perform map update and achieve robust camera pose tracking in dynamic environment.

On the basis of the reliable initial map, as for the image coming immediately after the map initialization procedure, the corresponding camera pose calculation is accurate. However, these subsequent map expansion procedure in dynamic environment can be properly handled by the inverse depth filter described in previous part.

As is shown in Fig.3, active map points that can be seen in current image frame are used by the tracking thread to calculate camera pose. At the same time, there are new map points created by local mapping thread. These newly created points are all set to be inactive, which means that they will not be immediately used to calculate camera pose, until they are confirmed static by the inverse depth filter. Among these inactive map points, there are some map points that can be seen in current image frame, therefore new measurement are avaliable to perform inverse depth filter. If the inverse depth of a map point is not convergent, more measurements are needed to confirm its motion characteristics, so it has to wait for next update. When the inverse depth of a map point converges and π is low (for example, lower than 0.5), the map point is still considered as a dynamic point and will



(a) Sequence s_static (low dynamic scene)
(b) Sequence w_halfsphere (high dynamic scene)
(c) Sequence w_rpy (high dynamic scene)
(d) Sequence w_static (high dynamic scene)
(e) Sequence w_xyz (high dynamic scene)

FIGURE 4. **Comparison of output camera trajectories of our SDF-SLAM, ORB-SLAM2 and ground-truth for the five dynamic sequences.**

be deleted. Only when the inverse depth of a map point converges and π is high (for example, higher than 0.5), the map point will be considered as a reliable static map point and is activated. Now the active map points in the map is updated, when next image comes, active map points which are reliable are used to calculate corresponding camera pose and the map update procedure is conducted in the same way.

In view of the above research principles, our approach is suitable to the problem that the motion of the dynamic feature points between two adjacent frames is not obvious: 1) the new method considers all the image frames that can observe the map points. Although the relative motion between two adjacent frames is not obvious, the relative motion of the dynamic elements shown by the two images with a long time span is relatively obvious; 2) the new method adopts probability framework. This framework constantly accumulates the new observation data, so even if the single movement between adjacent frames is relatively small, whose impact on the map point's probability of being static is limited, but the accumulation of these small effects will significantly reduce its static probability, so as to detect dynamic map points.

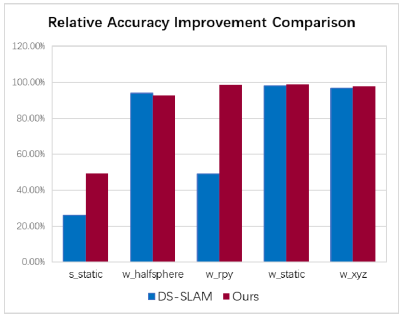
III. EVALUATION

In this section, experiments are performed to verify the effectiveness of SDF-SLAM towards the dynamic environments. All the experiments were performed on a computer with Intel i9 9940X CPU, TITAN RTX GPU, and 48GB memory. First, we compare our SDF-SLAM with the baseline framework, i.e. ORB-SLAM2, to verify the improvement of our system. Second, we compare our system with other state-of-the-art visual SLAM systems towards dynamic environment. The possible results published in the original papers are adopted directly.

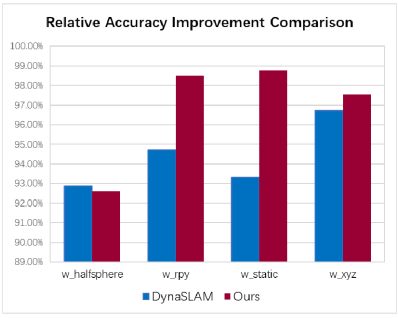
A. DATASET

We evaluate the accuracy of SDF-SLAM on the public TUM RGB-D dataset [26] which is a novel benchmark to evaluate the visual SLAM system. Five dynamic scene video sequences, i.e., freiburg3_sitting_static (s_static), freiburg3_walking_halfsphere (w_ halfsphere), freiburg3 _walking_rpy (w_rpy), freiburg3_walking_static (w_static), and freiburg3_walking_xyz (w_xyz), are chosen. These sequences are captured at 30Hz, and they contain 640 × 480.8-bit RGB images and 640 × 480.16-bit depth images. Also, the ground-truth camera trajectory is obtained by a high-accuracy motion-capture system with eight high-speed tracking camera (100Hz).

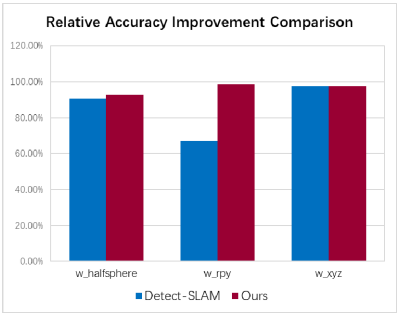
These five chosen dynamic sequences were taken in the 'desk' scene, and two persons are walking or sitting. They are applicable to verify the efficiency of our approach towards dynamic environment. 's_static' sequence represents the scene of two persons are sitting at the desk and the camera is kept in place manually. These two persons sit at the desk, talk, and gesticulate a little bit. This sequence is intended to



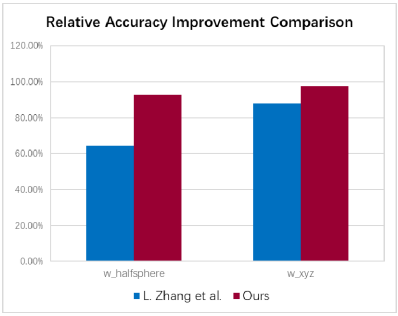
(a) comparison with DS-SLAM [18]



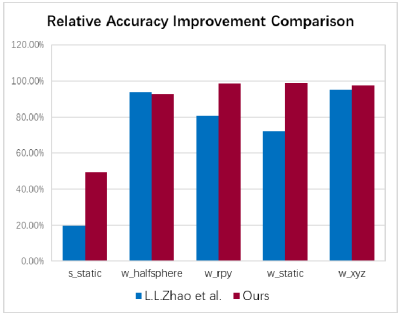
(b) comparison with DynaSLAM [19]



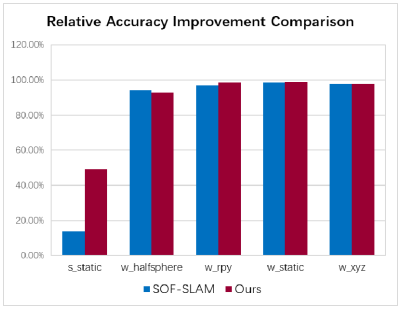
(c) comparison with Detect-SLAM [14]



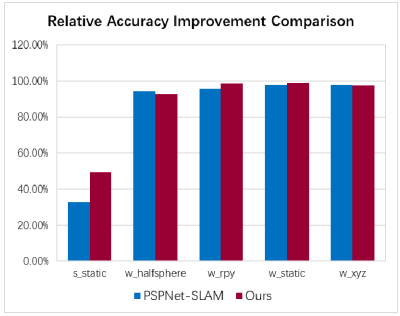
(d) comparison with L.Zhang et al. [15]



(e) comparison with L.L.Zhao et al. [21]



(f) comparison with SOF-SLAM [22]



(g) comparison with PSPNet-SLAM [20]

FIGURE 5. Comparison of relative accuracy improvement in dynamic environments between our system and other state-of-the-art semantic SLAM systems.

evaluate the robustness of visual SLAM to slowly moving dynamic objects. Therefore, 's_static' can be seen as low dynamic scene. The other four sequences describe the scenes of two persons are walking through the office and the camera moves along different directions. These four sequences can be seen as high dynamic scenes. In detail, 'w_halfsphere' sequence means two persons are walking through an office and the camera moves on a small half sphere of approximately one meter diamter. In 'w_rpy' sequence, the camera rotates along the principal axes (roll-pitch-yaw) at the same position. And in 'w_xyz' sequence, the camera moves along three directions (xyz) while keeping the same orientation.

B. EXPERIMENTS AND ANALYSIS

First, we compare our SDF-SLAM with the baseline system, i.e. ORB-SLAM2. The output camera trajectories of our approach, ORB-SLAM2 and ground truth for the five dynamic sequences are plotted and shown in Figure 4. In this figure, we project the 3D trajectories into 2D plane to exhibit the results more intuitively. As shown, our result exhibits much higher similarity to ground truth than ORB-SLAM2. To further evaluate qualitatively our approach, the qualitative calculation results are listed in Table 1. The evaluation metric is RMSE (Root Mean Squared Error) of ATE (Absolute Trajectory Error). All the video sequences run for five times to obtain the median, mean, minimum, and maximum of RMSE results, which can reduce the impact of system's non-deterministic nature. As seen, for the five dynamic video sequences, our approach owns lower RMSE values regardless in the low dynamic environment or in the high dynamic environment. The median, mean, minimum and maximum RMSE are reduced obviously. In addition, the relative accuracy improvement of our system against the original ORB-SLAM2 is also calculated and shown in Table 1. Compared with original ORB-SLAM2 system, our approach can improve the accuracy greatly in all the five

TABLE 1. Comparisons of RMSE [m] in dynamic sequences of TUM RGB-D dataset for ORB-SLAM2 and our approach.

Sequence	ORB-SLAM2				Ours				Improvement of our approach against ORB-SLAM2			
	Median	Mean	Min	Max	Median	Mean	Min	Max	Median	Mean	Min	Max
s_static	0.012	0.012	0.010	0.012	**0.0061**	**0.0062**	**0.0053**	**0.0071**	**49.16%**	**48.33%**	**47.00%**	**40.83%**
w_halfsphere	0.497	0.576	0.375	0.826	**0.0367**	**0.0358**	**0.0260**	**0.0441**	**92.62%**	**93.78%**	**93.07%**	**94.66%**
w_rpy	0.916	0.976	0.828	1.210	**0.0137**	**0.0149**	**0.0113**	**0.0202**	**98.50%**	**98.46%**	**98.64%**	**98.33%**
w_static	0.437	0.429	0.394	0.445	**0.0053**	**0.0053**	**0.0047**	**0.0059**	**98.79%**	**98.76%**	**98.81%**	**98.67%**
w_xyz	0.771	0.726	0.590	0.800	**0.0190**	**0.0174**	**0.0131**	**0.0196**	**97.54%**	**97.60%**	**97.78%**	**97.55%**

TABLE 2. Comparisons of relative RMSE [m] reduction for our system against the state-of-the-art in dynamic sequences of TUM dataset.

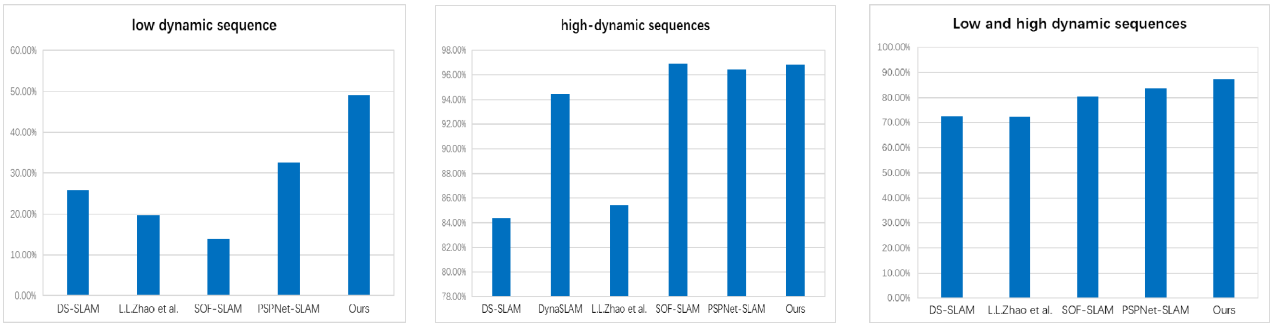
Sequence	DS-SLAM [18]	DynaSLAM [19]	Detect-SLAM [14]	L. Zhang et al. [15]	L.L.Zhao et al. [21]	SOF-SLAM [22]	PSPNet-SLAM [20]	**Ours**
s_static	25.94%	-	-	-	19.70%	13.87%	32.58%	**49.16%**
w_halfsphere	93.76%	92.88%	90.72%	64.31%	93.90%	94.25%	94.33%	92.62%
w_rpy	48.97%	94.71%	66.94%	-	80.80%	97.03%	95.58%	**98.50%**
w_static	97.91%	93.33%	-	-	72.00%	98.49%	97.87%	**98.79%**
w_xyz	96.71%	96.73%	-	87.92%	95.10%	97.71%	98.05%	97.54%

dynamic sequences. Specifically, more than 40% improvement is achieved for the low dynamic sequence. For the high dynamic scenes, the improvements are more obvious and more than 90% improvement is obtained. The results indicate that our approach can further remove the disturb of dynamic objects and thus degrades the pose error during optimization.

Second, we compare our SDF-SLAM with other state-of-the-art semantic SLAM systems which were proposed in recent two years and towards the dynamic environment. In specific, the DS-SLAM [18], DynaSLAM [19], Detect-SLAM [14], the system proposed by Zhang *et al.* [15], the system proposed by Zhao *et al.* [21],SOF-SLAM [22], and PSPNet-SLAM [20] are adopted for comparisons. All the above systems are built on ORB-SLAM2, and are tested on the dynamic sequences from TUM dataset. The relative RMSE reduction (i.e. relative accuracy improvement) of each system with respect to ORB-SLAM2 is calculated as the evaluation metric just like former researchers did in their works [20], [22]. The comparison results are shown in Table 2. As seen, all these semantic visual SLAM systems can further discard the moving objects and reach higher precisions compared with the original ORB-SLAM2. For the 's_static' sequence which owns limited dynamic objects and the movement between frames is slight, the relative accuracy improvement of our approach (49.16% RMSE reduction against the ORB-SLAM2) is more obvious than the other methods. That is because our method adopts the 3D map to detect the dynamic object and considers all the image frames that can observe the map points. Although the relative motion between two adjacent frames is not obvious, the relative motion of the dynamic elements shown by the two images with a long time span is relatively obvious. In this case, the low dynamic objects in 's_static' can be further removed by our approach. Among the five dynamic sequences, our approach has the highest accuracy improvements in three sequences (s_static, w_rpy, and w_static), and in the other two sequences (w_halfsphere, w_xyz) the differences between our method and PSPNet-SLAM that achieves the best results for these two sequences are very small. To further compare qualitatively our system with the state-of-the-art semantic SLAM systems, the average accuracy improvements of these systems in low dynamic sequence, four high dynamic sequences, and all the five dynamic sequences are also calculated and shown in Table 3. As shown, only in the high dynamic sequences, our approach is 0.01% relative accuracy reduction compared with the SOF-SLAM which achieves the highest average accuracy improvement in high dynamic sequences. In the five dynamic

TABLE 3. Comparisons of average relative RMSE [m] reduction for our system against the state-of-the-art in dynamic sequences of TUM dataset.

Sequence	DS-SLAM [18]	DynaSLAM [19]	Detect-SLAM [14]	L. Zhang et al. [15]	L.L.Zhao et al. [21]	SOF-SLAM [22]	PSPNet-SLAM [20]	**Ours**
low dynamic sequence	25.94%	-	-	-	19.70%	13.87%	32.58%	**49.16%**
High dynamic sequences	84.34%	94.41%	-	-	85.45%	**96.87%**	96.46%	96.86%
Low and high dynamic sequences	72.66%	-	-	-	72.30%	80.27%	83.68%	**87.32%**



(a) comparison in low dynamic sequence (b) comparison in high dynamic sequences (c) comparison in all the dynamic sequences

FIGURE 6. Comparison of average relative accuracy improvement in dynamic environments between our system and other state-of-the-art semantic SLAM systems.

sequences which contain both low and high dynamic scenes, our approach gets the highest average accuracy improvement. The reason why our algorithm achieves better performance is that we detect the dynamic objects directly from the more reliable 3D map which is updated with the depth filter. For a map point, all the image data related to it contribute its dynamic judgment. More abundant information about the dynamic characteristics of map points can be obtained, which lead to better robustness to noise and single outlier observation.

Last, the accuracy superiorities of our approach against the state-of-the-art semantic SLAM systems towards dynamic scenes are shown in Fig.5 and Fig.6 more intuitively in the form of bar chart.

IV. CONCLUSIONS AND DISCUSSIONS

We have presented a new semantic visual SLAM, i.e. SDF-SLAM, towards the dynamic environment. It is built on ORB-SLAM2, and detects the dynamic scene directly from the 3D map points with the Bayesian filtering framework. Two modules, namely the semantic map initialization module with the semantic optical flow and the dynamic map points detection module with inverse depth filter, are introduced to the ORB-SLAM2 framework. Our system can overcome the drawbacks of dynamic feature points detection from the 2D image level, and more reliable dynamic characteristics of objects are detected. Experiments in public TUM dataset demonstrate that our approach outperforms the ORB-SLAM2 and other state-of-the-art semantic SLAM systems towards dynamic scenes. In low dynamic sequence, our system can achieve 49.16% accuracy improvement against the ORB-SLAM2, and this improvement is very obvious compared with other state-of-the-art SLAM systems. In high dynamic sequences, our approach obtains averagely 96.86% accuracy improvement and is only 0.01% accuracy reduction compared with the SOF-SLAM which achieves the highest average accuracy improvements in the high dynamic scene. When considering the low dynamic scene and high dynamic scene synthetically, our algorithm obtains the highest average accuracy improvement and more suitable to deal with the SLAM problem in dynamic environment.

Our work can be further improved in the following aspects: 1) Degenerate situations, such as geometrical degenerate correspondences (e.g. all the observed features lie on a plane or lie on a ruled quadric) and degenerate camera motion (e.g. pure rotation) haven't been carefully handled; 2) More in-depth and detailed research on the convergent properties of the inverse depth of the map point is needed; 3) Our approach adopts the classic ORB features of image in the whole SLAM framework. With the development of deep learning for feature extraction, such as the MagicPoint [27] and GCN features [28], in future we will try to replace the ORB features with the deep-learning features to further improve the robust of our system.

REFERENCES

[1] 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.

[2] G. Klein and D. Murray, "Parallel tracking and mapping for small AR workspaces," in *Proc. 6th IEEE ACM Int. Symp. Mixed Augmented Reality*, Nov. 2007, pp. 225–234.

[3] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE Trans. Robot.*, vol. 31, no. 5, pp. 1147–1163, Oct. 2015.

[4] 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.

[5] J. Engel, T. Schöps and D. Cremers, "LSD-SLAM: Large-scale direct monocular SLAM," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 834–849.

[6] C. Forster, M. Pizzoli, and D. Scaramuzza, "SVO: Fast semi-direct monocular visual odometry," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2014, pp. 15–22.

[7] 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.

[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[9] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multiBox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[10] 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.

[11] K. He, G. Gkioxari, P. Dollar, and R. B. Girshick, "Mask R-CNN," in *Proc. Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2980–2988.

[12] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.

[13] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[14] F. Zhong, S. Wang, Z. Zhang, C. Chen and Y. Wang, "Detect-SLAM: Making object detection and SLAM mutually beneficial," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1001–1010.

[15] L. Zhang, L. Wei, P. Shen, W. Wei, G. Zhu, and J. Song, "Semantic SLAM based on object detection and improved octomap," *IEEE Access*, vol. 6, pp. 75545–75559, 2018.

[16] Z. Wang, Q. Zhang, J. Li, S. Zhang, and J. Liu, "A computationally efficient semantic SLAM solution for dynamic scenes," *Remote Sens.*, vol. 11, no. 11, p. 1363, 2019.

[17] 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. Auto. Syst.*, vol. 117, pp. 1–16, Jul. 2019.

[18] 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)*, Oct. 2018, pp. 1168–1174.

[19] B. Bescos, J. M. Facil, J. Civera, and J. Neira, "DynaSLAM: Tracking, mapping, and inpainting in dynamic scenes," *IEEE Robot. Autom. Lett.*, vol. 3, no. 4, pp. 4076–4083, Oct. 2018.

[20] S. Han and Z. Xi, "Dynamic scene semantics SLAM based on semantic segmentation," *IEEE Access*, vol. 8, pp. 43563–43570, 2020.

[21] L. Zhao, Z. Liu, J. Chen, W. Cai, W. Wang, and L. Zeng, "A compatible framework for RGB-D SLAM in dynamic scenes," *IEEE Access*, vol. 7, pp. 75604–75614, 2019.

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

[23] M. Everingham and J. Winn, "The PASCAL visual object classes challenge 2007 (VOC 2007) development kit," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, 2006.

[24] 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.

[25] G. Vogiatzis and C. Hernández, "Video-based, real-time multi-view stereo," *Image Vis. Comput.*, vol. 29, no. 7, pp. 434–441, Jun. 2011.

[26] 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.*, Oct. 2012, pp. 573–580.

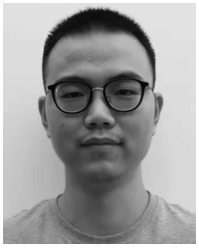
[27] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Toward geometric deep SLAM," 2017, *arXiv:1707.07410*. [Online]. Available: http://arxiv.org/abs/1707.07410

[28] J. Tang, L. Ericson, J. Folkesson, and P. Jensfelt, "GCNv2: Efficient correspondence prediction for real-time SLAM," *IEEE Robot. Autom. Lett.*, vol. 4, no. 4, pp. 3505–3512, Jul. 2019.

LINYAN CUI was born in Hengshui, Hebei, China. She received the B.S., M.S., and Ph.D. degrees from the School of Astronautics, Beihang University, Beijing, China, in 2006, 2008, and 2013, respectively.

She joined the Image Processing Center, Beihang University, as an Assistant Professor, in 2013, where she is currently an Associate Professor with the Image Processing Center. She published more than 36 SCI articles in optics express and other international journals. Her research interests include SLAM, computer vision, turbulence-degraded image restoration, and theoretical modeling of optical waves in atmospheric turbulence.

Dr. Cui was awarded the Excellent National Doctoral Dissertations in China in the field of Aeronautical and Astronautical Science and Technology, in 2016.



CHAOWEI MA received the bachelor's degree from Beihang University, Beijing, China, where he is currently pursuing the master's degree with the School of Astronautics. His research interests include computer vision, 3-D reconstruction, and SLAM.