Keyframe Selection with Information Occupancy Grid Model for Long-term Data Association

Weinan Chen[#], Hanjing Ye[#], Lei Zhu, Chao Tang, Changfei Fu, Yonggang Chen, Hong Zhang*

***Abstract*— As the basics of Visual Simultaneous Localization And Mapping (VSLAM), keyframes play an essential role. In previous works, keyframes are selected according to a series of view change-based strategies for short-term data association (STDA). However, the texture enrichment of frames is always ignored, resulting in the failure of long-term data association (LTDA). In this paper, we propose an information enrichment selection strategy with an information occupancy grid model and a deep descriptor. Frame is expressed by a deep global descriptor for a statistical explainable abstraction, in which the texture enrichment is indicated. Based on the abstraction, an information occupancy grid model is established to measure the information enrichment and the potential LTDA ability. Evaluations on variant datasets are conducted, showing the advantage of our proposed method in terms of keyframe selection and tracking precision. Also, the statistical explainability of the deep descriptor is provided. The proposed keyframe selection strategy can improve LTDA and tracking precision, especially in situations with repeated observations and loop-closures.**

VSLAM, Keyframe Selection, Occupancy Grid Model

I. Introduction

In the past ten years, VSLAM has been studied densely [1] [2], which plays an essential role in visual navigation [3] and self-driving [4]. Especially with the development of computer vision, VSLAM has become a popular research topic. Among the existing popular VSLAM studies, for example, the indirect method-based systems [5], the direct method-based systems [6], as well as the semi-direct method-based systems [7], most of them are built based on keyframes. Keyframes are the frames selected from a continuous frame sequence. As introduced in [8] [9], several standard steps are included in a keyframe-based VSLAM system, such as keyframe selection, pair-wise tracking (STDA), spatial projection, place recognition (LTDA), and graph optimization. In all the processes, the keyframe selection plays an essential role in providing reference markers. A visual map is updated after inserting a new keyframe, and the built map can satisfy STDA. Besides providing landmarks for STDA, as the popularity of submap-based VSLAM [10] [11] and lifelong VSLAM [12] in recent years, the role of keyframes in terms of LTDA is gaining focus. Regarding LTDA in VSLAM, it helps in loop-closure detection and submap alignment for the global consistency (Fig. 1).

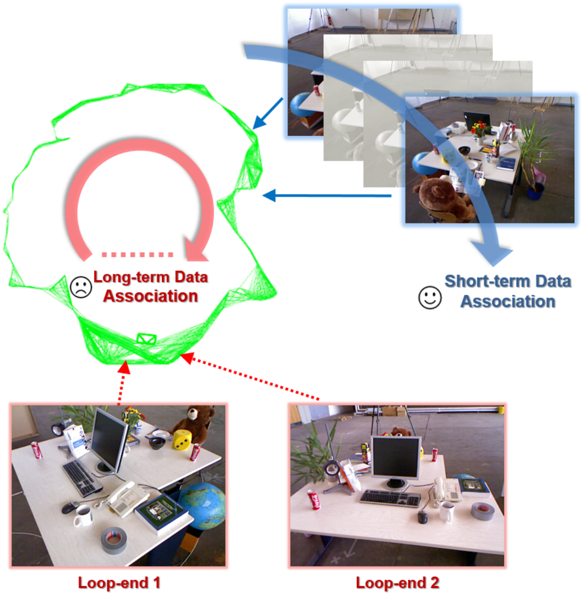


Fig. 1. In VSLAM, the existing keyframe selection strategies are designed to satisfy short-term data association only, shown as the blue arrows. However, long-term data association is ignored, which results in the missing of loop-end keyframes (red arrow) and the failure of loop-closure detection. In our work, the information enrichment of frames is measured to enhance long-term data association.

Due to the importance of keyframes, the strategy of keyframe selection is widely studied. In an indirect method-based VSLAM system [5], the strategy in terms of feature correspondence is proposed for local tracking. As for the direct method-based systems [6], the insertion of keyframe is decided by a weighted sum from pair-wise optical-flow tracking. In the study of keyframe selection comparison [13], lots of related methods have been analyzed, and different types of similarity metrics are introduced.

As is mentioned above, most of the metrics are designed to satisfy the requirement of STDA, while LTDA is always ignored. In an attempt to fill this gap, we propose to enhance LTDA by selecting texture-rich keyframes, which has better LTDA ability than the texture-less one. The texture enrichment is measured by information in this paper. To calculate the information enrichment, we introduce an information occupancy grid model based on a statistical explainable descriptor. The information-rich frame is defined as the one that increases the total occupancy probability much.

The work in this paper is supported by the National Natural Science Foundation of China (62103179), the China Postdoctoral Science Foundation (2021M701576), the Leading Talents Program of Guangdong Province (2016LJ06G498), and the Dongguan Science and Technology of Social Development Program (20211800904472)

Weinan Chen and Hanjing Ye contribute equally to this paper. Hanjing Ye, Chao Tang, Changefei Fu and Hong Zhang are with the Department of Electronic and Electrical Engineering of the Southern University of Science and Technology in Shenzhen, China. e-mail: zhangh33@sustech.edu.cn

Weinan Chen and Lei Zhu are with Guangdong University of Technology, Guangzhou, China.

Yonggang Chen is with DongGuan Polytechnic, Dongguan, China

This paper is organized as follows. Section II presents the related work and the motivation of this study. Section III details the proposed keyframe selection strategy. Section IV shows the experimental results and the discussion. Finally, section V concludes this paper.

The contribution of our study is threefold:
1. We propose to enhance long-term data association by selecting texture-rich keyframes;
2. To measure the texture enrichment, an information occupancy grid model and a statistical explainable deep descriptor are introduced;
3. In situations with repeated observation and loop-closure, the proposed strategy shows advanced precision.

II. Related Work

Our work intersects keyframe selection, deep global descriptor, and information consideration in visual navigation. To clarify the motivation of this study, we provide a survey of related studies in this section and discuss the shortcomings of existing works.

A. Keyframe Selection in VSLAM

As the basis of VSLAM systems, keyframes enable visual tracking and loop-closure detection. Also, with the development of submap-based and lifelong VSLAM systems, keyframe selection determines the global consistency between submaps. Therefore, many researchers have investigated this problem.

Based on the estimated camera motion with respect to the existing map, LSD-SLAM [14] and SVO [7] make keyframe decisions by transformation estimation. A similar metric is also leveraged in Code-SLAM [15] and RGBDSLAM [16]. The keyframe selection based on feature correspondence is also widely studied. In feature-based VSLAM systems [5] and submap-based VSLAM [11], the feature correspondence is established by local descriptor matching. To select good keyframes, [17] proposes to insert keyframes with a strong feature temporal disparity. To improve the quality and consistency of feature matching, in [18], a selection method maximizing the number of feature correspondence by storing observation data is proposed. Considering a combination of the mentioned strategy, DSO [6], LDSO [19] and DSM [20] use a weighted sum of camera motion, optical-flow matching, and illumination change as the selection judgment.

Regarding all the mentioned strategies above, they use one or more indicators to measure the view change of camera observation with respect to the existing map [9]. Because visual tracking is achieved by the overlap between frames, a new keyframe is needed to update the visual map when the view change is too significant to provide enough overlap. However, due to the non-sequential data association of loop-closure detection and submap alignment, the keyframe should also provide rich information for LTDA instead of the overlap for visual tracking only. In this paper, we propose a selection strategy considering information enrichment, and LTDA can be improved by the selected texture-rich keyframes.

B. Image Global Descriptor

In terms of image global descriptors, two types of global descriptors have been proposed. The first type is the hand-craft descriptor, which is extracted by a series of rules and designed filters. One of the classical hand-craft descriptors is the VLAD [21], which generates a low-dimensional global descriptor through a cluster-like aggregating method. Another type of hand-craft descriptor is the VBoW (Visual Bag of Word) class [22] [23]. These methods classify local feature descriptors into a pre-trained dictionary, and a frame is expressed by the counting scope of the dictionary words. GIST is also a popular global image descriptor [24], which uses a filter bank [25] to extract the image pixel response to different types of texture patterns. All the hand-craft descriptors are based on human experience and the raw image texture. Due to the lack of high-level knowledge, the hand-craft descriptors show the shortcoming on the variance of illumination and perspectives [26].

To solve the problem of hand-craft descriptors, learning-based deep descriptors are proposed. According to the backbone of network architecture, the exiting works are classified into two types: the pre-trained classification network type [27] and the end-to-end training model type [28] [29]. Both two types of descriptors are based on CNNs. As mentioned in [30], CNN-based deep descriptors can be regarded as the response to a series of texture patterns, which is similar to the GIST descriptors. However, due to the multiple layers perception in neural networks, both the low-level and high-level knowledge is represented in a learning-based deep descriptor. Therefore, the extracted deep descriptor indicates a more appropriate abstraction of a frame than the hand-craft descriptor.

Taking the superiority of deep descriptors, we proposed to calculate the information enrichment on top of it.

C. Information Consideration in Visual Navigation

As for the information consideration in visual navigation, there are two popularly studied applications: autonomous exploration and observation completion evaluation.

In the application of autonomous exploration, robots find the next best view to obtain as much information as possible. Therefore, the observation information with respect to an existing map is leveraged as the indicator in [31]. Also, to achieve autonomous visual mapping and ensure texture enrichment for visual tracking, some active VSLAM systems are proposed [32]. In these studies, entropy is always defined as the information introduced from the current observation. The next best view is found by minimizing the entropy.

As for the evaluation of visual observation completion, to our knowledge, the most related work is [33]. In this work, the information enrichment is leveraged to judge the completion of multiple camera observations and decide the update of a visual map. However, this study focuses on the multiple camera system for object reconstruction. Inspired by the mentioned work above, we propose to determine the keyframe selection using a novel information occupancy grid model and the statistical explainability of the deep descriptor.

III. Methodology

We propose to enhance LTDA by selecting keyframes with rich texture, which is measured by the information enrichment in this paper. For the efficient calculation, we introduce a low-dimension projection method using the deep descriptor, in which the texture enrichment is indicated. Based on the statistical explainability of the deep descriptor, a novel information occupancy grid model is proposed to calculate the information enrichment in a form of occupancy probability. Lastly, a keyframe selection strategy considering both the information enrichment and the view change is proposed. The framework is shown in Fig. 2.

A. Deep Descriptor Extraction

To efficiently calculate the information enrichment of a frame, a low-dimension projection is needed. Due to the superiority of deep descriptors over hand-craft descriptors [26], we utilize a deep global descriptor for low-dimension projection. The network architecture for the deep descriptor extraction is shown in Fig. 3. Firstly, a feature map is extracted from a high-level layer of a pre-trained CNN. The feature map is then normalized by a layer normalization [34] and fed into a convolutional autoencoder (CAE). CAE is a convolutional-based aggregating method. Our CAE consists of three encoder layers and three decoder layers. Each layer in the encoder/decoder is composed of a convolutional/deconvolutional filter, a batch normalization [35] and a parametric rectified linear unit [36]. With the sliding window procedure of convolutional filters, the spatial information of different level features can be kept and represented. In the training procedure, the CAE is trained by reconstructing the normalized feature map using a MSE (mean squared error) loss. However, the decoder is dropped in the inference step, and only the encoder is kept to produce the encoded descriptor. Lastly, the encoded descriptor is flattened and L2 normalized to get the deep descriptor.

We choose an illumination-invariant feature maps as the output, which are found in previous works. Similar to [29], we choose the pre-trained VGG16 [37] as our backbone. Given a frame $\boldsymbol{f_t}$ at time t, the network output descriptor $\boldsymbol{F_t}$ is computed as:

$$\boldsymbol{F_t} = N_\theta(\boldsymbol{f_t}) \tag{1}$$

where N_θ is the network. Specifically, $\boldsymbol{F_t}$ is from the last layer of CNNs. Then, $\boldsymbol{F_t}$ is normalized and fed into a CAE for the final deep descriptor $\boldsymbol{M_t} = [d_1^t \dots d_i^t \dots d_{dim(\boldsymbol{M_t})}^t]$.

B. Information Occupancy Grid Model

Information enrichment of a frame is calculated on top of the deep descriptor. To analyze the frame abstraction $\boldsymbol{M_t}$, we draw the learned filters in Fig. 4 using the method in [38]. As is shown, the learned filters represent a series of basic texture patterns, such as line, circle, rectangle, corner. Therefore, the output descriptor is the count of responses to the patterns. When a large number of the corresponding pattern is detected in a frame, the response value is significant. Therefore, the value in each descriptor dimension indicates the enrichment of certain types of texture patterns; the number of non-zero descriptor dimensions indicates the enrichment of texture pattern diversity. The data distribution of the deep descriptor indicates the texture enrichment. Based on the statistical explainability of $\boldsymbol{M_t}$, an information occupancy grid model is proposed to calculate the information enrichment.

The occupancy grid model is a common method for the enrichment calculation. In the proposed information occupancy grid model, each grid is defined by a dimension of the deep descriptor. The occupancy probability P_i of grid i is given by d_i^t, which indicates the enrichment of variant patterns. Obviously, an information-rich frame is the one that increases the total occupancy probability much with respect to the existing model, which is indicated by the change of entropy in our method.

The occupancy probability $P_{i,\boldsymbol{S_t}}$ is calculated with $\boldsymbol{S_t}$, where $\boldsymbol{S_t}$ is a subset of all the existing keyframes $\boldsymbol{S_{all}}$. Firstly, we calculate the total dimension value $D_{i,\boldsymbol{S_t}}$ for each dimension of the deep descriptor. Since the deep descriptor is normalized before, $D_{i,\boldsymbol{S_t}}$ is the sum of the corresponding d_i^k of each frame $\boldsymbol{f_k} \in \boldsymbol{S_t}$, as shown in Formula 2. In addition, we limit $D_{i,\boldsymbol{S_t}}$ to a given threshold T^D, making $D_{i,\boldsymbol{S_t}}$ satisfy the occupancy grid model and balances the weight between each dimensions. Then, $P_{i,\boldsymbol{S_t}}$ is calculated in a form of logarithmical probability (Formula 3).

$$D_{i,\boldsymbol{S_t}} = min\{T^D, \sum_{\boldsymbol{f_k} \in \boldsymbol{S_t}} d_i^k\} \tag{2}$$

$$P_{i,\boldsymbol{S_t}} = log\frac{D_{i,\boldsymbol{S_t}}}{1 - D_{i,\boldsymbol{S_t}}} \tag{3}$$

Lastly, the Shannon Entropy $H_{i,\boldsymbol{S_t}}$ of grid i is calculated as shown in Formula 4:

$$H_{i,\boldsymbol{S_t}} = -P_{i,\boldsymbol{S_t}} \cdot logP_{i,\boldsymbol{S_t}} \tag{4}$$

After obtaining $H_{i,\boldsymbol{S_t}}$, the total entropy of $\boldsymbol{S_t}$, written as $E_{info}(\boldsymbol{S_t})$, is calculated as shown in Formula 5:

$$E_{info}(\boldsymbol{S_t}) = \sum_{i < dim(\boldsymbol{M_t})} H_{i,\boldsymbol{S_t}} \tag{5}$$

The information enrichment G_t of $\boldsymbol{f_t}$ at time t is the change of $E_{info}(*)$ introduced by $\boldsymbol{f_t}$. Because the input of VSLAM is sequential, we pick the existing keyframes within the last n seconds as $\boldsymbol{S_t}$. Therefore, G_t is defined with respect to the neighbor of $\boldsymbol{f_t}$. The calculation of G_t is shown in Formula 6, where $e \oplus B$ represents inserting an element e into a set B.

$$G_t = E_{info}(\boldsymbol{f_t} \oplus \boldsymbol{S_t}) - E_{info}(\boldsymbol{S_t}) \tag{6}$$

C. Keyframe Selection Strategy

We introduce the keyframe selection strategy based on the information occupancy grid model in this section. Due to the variance of perception, each frame can contribute to the information enrichment. However, the keyframe redundancy is not considered strictly, which influences the real-time performance. To balance the requirement of information enrichment and redundancy, we consider both the information

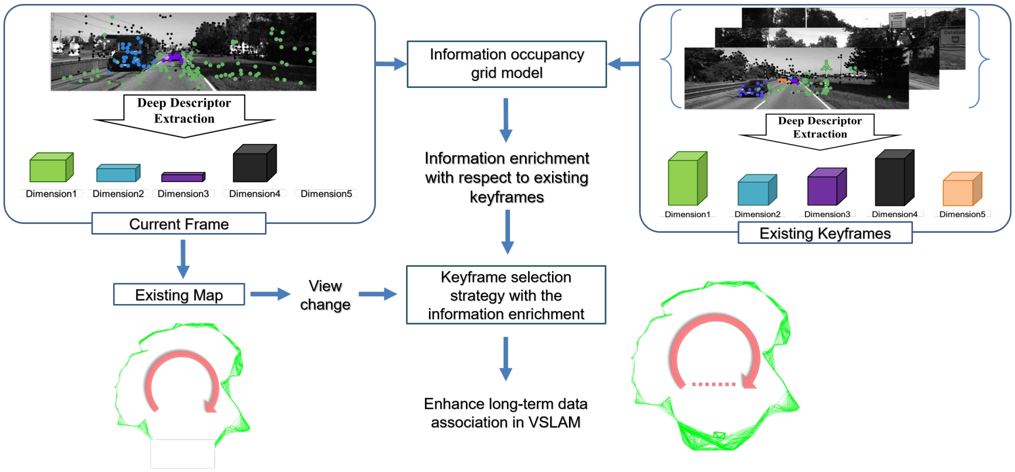

Deep Descriptor Extraction
Dimension1
Dimension2
Dimension3
Dimension4
Dimension5
Current Frame
Information occupancy grid model
Information enrichment with respect to existing keyframes
Deep Descriptor Extraction
Dimension1
Dimension2
Dimension3
Dimension4
Dimension5
Existing Keyframes
Existing Map
View change
Keyframe selection strategy with the information enrichment
Enhance long-term data association in VSLAM

Fig. 2. The framework of our proposed method. Both the current frame and the existing keyframes are expressed by a deep descriptor abstraction (shown as the histograms). The information enrichment of the current frame with respect to the existing keyframes is calculated in the information occupancy grid model. After that, a keyframe selection strategy considering both the information enrichment and the view change is designed for enhancing LTDA.

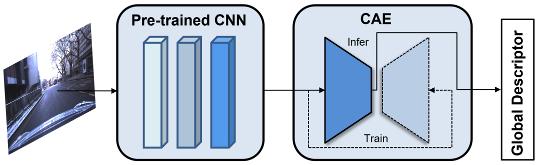


Fig. 3. The architecture of our deep descriptor network.

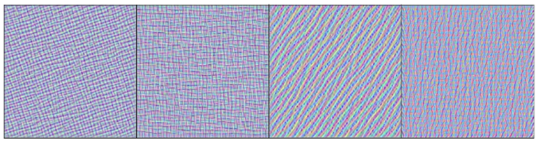


Fig. 4. Learned filters of the deep descriptor network, which represent a series of texture patterns.

enrichment and the view change in our strategy. Referred to the existing work [9], the keyframe redundancy is always indicated by the view change V_t. V_t represents the diversity between the current frame and the existing map. With the calculated V_t in the VSLAM, the proposed selection strategy is shown in Algorithm 1.

Firstly, before the extraction of $\boldsymbol{M_t}$, we remove the frames with motion blurry [39] due to the fake information introduced by blurry pixels. Then, the information enrichment G_t is calculated. To combine the view change and the information enrichment, we tune the threshold on view change ϕ^V by G_t. Since ϕ^V has been given in most of the existing VSLAM systems, the strategy on top of ϕ^V makes our algorithm general. Then, the average information enrichment $\overline{G}$ learned from an offline test is involved to judge the information-poor frames. Finally, the decision is made by comparing V_t with the tuned threshold $(\phi^V + (G_t - \overline{G}) \cdot w)$. The requirement of diversity is released when the frame information is rich; and the frame redundancy is also considered to limit the number of selected keyframes.

Algorithm 1 Keyframe Selection Strategy

```
Require: Newly captured frame, $\boldsymbol{f_t}$;
    All the existing keyframes, $\boldsymbol{S_{all}}$;
    Given threshold on view change, $\phi^V$;
    Given weight for judgment combination, $w$;
    Average information enrichment, $\overline{G}$;
Check the Motion blurry of $\boldsymbol{f_t}$;
Extract the deep descriptor $\boldsymbol{M_t}$ from $\boldsymbol{f_t}$;
Get neighbor keyframe subset $\boldsymbol{S_t}$ from $\boldsymbol{S_{all}}$;
Extract the deep descriptor of each frame in $\boldsymbol{S_t}$;
Calculate $G_t$ between $\boldsymbol{f_t}$ and $\boldsymbol{S_t}$;
Calculate view change $V_t$ of $\boldsymbol{f_t}$ with respect to $\boldsymbol{S_{all}}$;
return $V_t > \phi^V + (G_t - \overline{G}) \cdot w$;
```

IV. Experiments

To verify the performance of our proposed keyframe selection strategy, based on a state-of-the-art VSLAM system, we conduct experiments on different datasets. Firstly, we introduce the setup of experiments. Secondly, the effectiveness of the deep descriptor and its statistical explainability is demonstrated. Then, we show the feasibility of the proposed information occupancy grid model. Lastly, a comparison with other existing keyframe selection strategies is provided to show the superiority of our proposed strategy.

A. Implementation

In the training procedure of the deep descriptor network, we use part of Oxford RobotCar [40], TUM dataset for the CAE training. Within these datasets, challenging appearance-change views are captured.

As for the keyframe selection strategy, we implement the proposed method based on ORB-SLAM3. ORB-SLAM3 is an indirect method-based system in which the camera view change is measured by the number of feature matching. Also, ORB-SLAM3 is a submap-based system whose tracking precision can be highly affected by LTDA. In this modification, the information enrichment term is added in the “c4” condition threshold (a given threshold in terms of matched features) in the code. As for the given parameters, the probability threshold $T^D = 0.8$, the combination weight $w = 45.0$, the average information enrichment $\overline{G} = 6.0$, and the size of slide-window $n = 30$.

B. Experiments Setup

In evaluations of the deep descriptor, we draw the learned filters for visualizing the frame abstraction procedure. As for the verification metric of frame abstraction, we follow the common evaluation metric introduced in [29]. The long-term dataset: Oxford RobotCar, excluding the training images, is used for evaluation. Then, the data distribution of the deep descriptor extracted from frame examples is provided, showing the statistical explainability of the deep descriptor.

To verify the proposed information occupancy grid model, we provide the curve of information enrichment with a sequential frame input. Both the change of information enrichment and its relationship to the frame are shown. Finally, we list the results of the proposed strategy and the existing strategy on different datasets to demonstrate our advantage. The precision of visual tracking is indicated by the root mean square of absolute trajectory error (RMSE), and the size of selected keyframes is also recorded. In terms of the conducted datasets for VSLAM validations, two datasets are conducted: TUM [41] (exclude the training sequences) and EuRoC [42]. The datasets are recorded by a handheld sensor and a camera mounted on a UAV in indoor environments. The ground truth is provided by a motion capture system. To show the influence of LTDA, we select the sequences with repeated observation and loop-closure for the evaluation of long-term data association.

As for the computer configuration, we run all the evaluations on a desktop computer with an Intel Core i9 (2.80 GHz) CPU, Nvidia 3090 GPU and 128GB memory.

C. Experimental Results

1) Evaluations of Deep Descriptor and Statistical Explainability: In the comparison of frame abstraction (Tab. I), our output dimension is set to 4096. We can observe that compared to the SoTA: NetVLAD and VGG16, our deep descriptor achieves the best results with higher AP. Even in an appearance-change dataset, such as RobotCar (dbNight vs. qSnow), our recall@1 is 0.758 versus NetVLAD’s 0.642, VGG16’s 0.523 and AlexNet’s 0.660. As for the extraction of deep descriptors, we provide examples of the frames with or without repeated texture and their corresponding descriptors. As shown in Fig. 5, compared with the original frame, the deep descriptor extracted from repeated texture frame has a smaller number of peaks and the distribution of peaks are more centralized. In addition, as the patch size decreases (the size order is: b-c-d-e-f), the number of peaks is less and the peaks are more centralized. Such results support the discussion mentioned in Sec. III-B, that the data distribution of the deep descriptor indicates the texture enrichment.

Method	AP	R@1	R@5	R@10	R@20
Ours	**0.948**	**0.758**	**0.832**	**0.862**	**0.895**
NetVLAD	0.853	0.642	0.765	0.814	0.864
VGG	0.745	0.530	0.652	0.706	0.765
AlexNet	0.902	0.660	0.761	0.800	0.840

TABLE I
THE COMPARISONS OF PLACE RECOGNITION PERFORMANCE BETWEEN OURS AND OTHER EXISTING WORK ON ROBOTCAR. OUR DEEP DESCRIPTOR ACHIEVES THE BEST RESULTS.

2) Evaluations of Information Occupancy Grid Model: We show the entropy calculated from the information occupancy grid model. With the repeated observation of a particular room, the entropy is shown in Fig. 6. The entropy increases at the beginning and converges to 0 after a certain duration of frame capturing. Meanwhile, as the converging of the entropy, the environment modeling is also completed. Such a result shows the astringency of the information occupancy grid model.

Also, to show the proposed strategy intuitively, the influence of frame texture is provided. The experiment is conducted on frei1_room. The relationship between the information enrichment and the number of non-zero dimensions (written as n^d), and the newly introduced n^d from a frame, is shown in Fig. 7. As is shown (texture-less frame: 2, 3, 5, and texture-rich frame: 1, 4), with the increment of n^d and Δn^d, more information is obtained. It is because the diverse texture pattern response indicates information enrichment. Also, little information enrichment is calculated when the corresponding frame is texture-less. The information-rich frame is supposed to enhance LTDA.

3) Evaluations of Keyframe Selection Strategy in VSLAM: To evaluate the proposed keyframe selection strategy, we implement our strategy based on a popular VSLAM system and show the tracking performance in Tab. II, including the tracking precision, the tracking time consumption and the size of the keyframe set. Compared with the existing keyframe selection strategy used in ORB-SLAM3 and LDSO, the advantage of our strategy over the second best one is shown. The time consumption of deep descriptor extraction is 0.012s per frame, which is not included in the table. The degree of repeated observation (repeated in the table) is the ratio between the trajectory coverage area and the trajectory length. The results in the table indicate the superiority of the proposed strategy with lower tracking error, where the improved precision by ours is shown in the “Advance” column.

As shown in Tab. II, our advantage in terms of RMSE is obvious when the repeated observation is significant. It is because more LTDA can be established when the number of repeated observations is larger. This result verifies the enhancement of LTDA by our method. To show the relationship between our advance and the degree of repeated observation, we plot the curve in Fig. 8. Because of the different environments among sequences, we plot the results

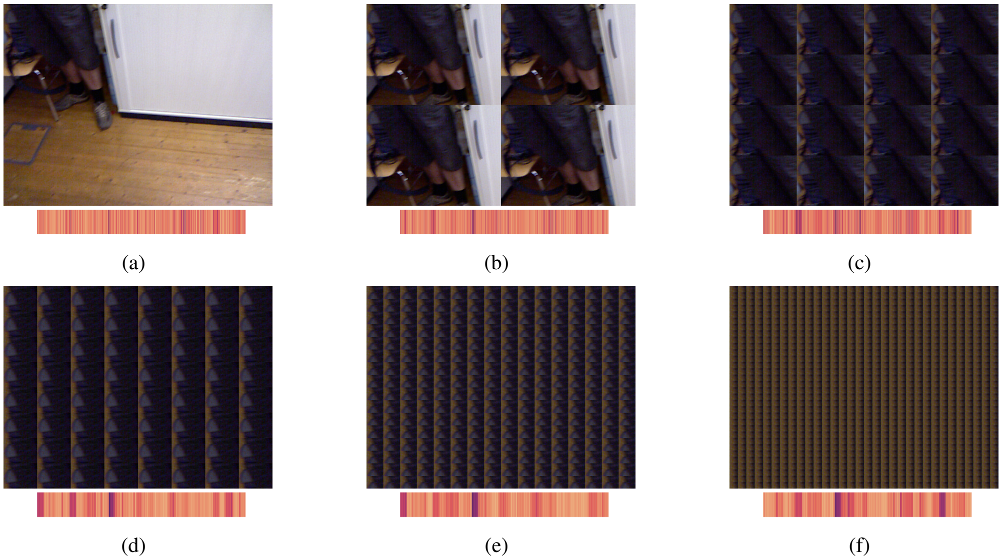

(a)
(b)
(c)
(d)
(e)
(f)

Fig. 5. Six groups of frames with/without repeated texture and the corresponding deep descriptor. The repeated texture frames (figure b, c, d, e and f) are generated from a part of the original frame (figure a). The results with variant patch size are provided to show the multi-scale perception in the deep descriptor. Compared with the original frame's descriptor, the deep descriptors extracted from the repeated texture frames have fewer peaks, and the peaks are also more centralized. Such a result shows the statistical explainabily for occupancy grid model establishment.

Sequence			Proposed Selection Strategy			Selection Strategy in ORB-SLAM3			Selection Strategy in LDSO		
Name	Repeated	Advance	Time	Keyframe	RMSE	Time	Keyframe	RMSE	Time	Keyframe	RMSE
V1_01	++	0.007	0.028	228	**0.086**	0.030	182	0.093	0.021	704	0.098
V1_02	+++	0.011	0.029	233	**0.051**	0.029	190	0.062	0.020	741	1.527
V1_03	+++	0.043	0.028	307	**0.063**	0.027	286	0.106	0.020	1103	0.367
V2_01	+	0.000	0.029	254	0.060	0.029	183	0.060	0.019	432	0.073
V2_02	++	0.016	0.029	291	**0.056**	0.028	212	0.072	0.022	889	0.107
V2_03	+++	0.307	0.028	357	**0.160**	0.028	268	0.467	-	-	-
MH_01	++	0.004	0.031	381	**0.046**	0.031	256	0.050	0.021	708	0.052
MH_02	++	0.007	0.030	286	**0.053**	0.029	245	0.060	0.019	840	0.064
MH_03	+++	0.008	0.029	255	**0.044**	0.027	200	0.052	0.020	758	0.070
MH_04	+	0.001	0.030	290	**0.084**	0.030	212	0.085	0.020	468	0.086
MH_05	+	0.001	0.029	308	**0.057**	0.029	224	0.058	0.021	565	0.098
frei2_p360	+	-0.003	0.019	147	0.133	0.019	136	**0.130**	0.016	335	1.694
frei2_slam	++	0.007	0.019	101	**0.117**	0.018	76	0.124	0.015	1244	1.973
frei2_slam2	+	0.000	0.018	76	0.030	0.018	63	0.030	-	-	-
frei2_slam3	+	0.005	0.018	109	**0.098**	0.018	125	0.103	-	-	-

TABLE II

COMPARISONS BETWEEN THE PROPOSED STRATEGY AND THE STRATEGY IN ORB-SLAM3 AND LDSO. THE TIME CONSUMPTION OF PER FRAME (S), THE NUMBER OF KEYFRAME AND THE PRECISION INDICATED BY RMSE (M) ARE SHOWN. ALL THE RESULTS ARE THE AVERAGE OF 5 TESTS.

(five tests) on EuRoC_V1 and EuRoC_V2 separately. Since the difficulty of visual tracking on V1_03 and V2_03 is serious, the corresponding results have a large variance.

D. Discussion

As shown in Tab. I, the proposed deep descriptor achieves advanced performance in the aspect of place recognition. Compared to the baseline, the AP of our deep global descriptor is improved by 10.5% on average. Such a result indicates an appropriate frame abstraction. Also, the strong place recognition ability enhance LTDA in VSLAM. Besides, the statistical explainability of the deep descriptor is demonstrated in Fig. 5. The deep descriptor from a repeated feature frame has a smaller number of peaks than the original frame, and the peaks are also more concentrated. Such a result can be explained by the deep descriptor extraction, which is the response to different texture patterns. Because the repeated texture frame responds heavily to one of the patterns and has little response to the non-existed patterns, the distribution of non-zero values is within a small range of dimensions. Also, the deep descriptor from the small patch repeated frame has fewer peaks than the one with a big patch because the former has more repeated patterns, and the pattern diversity is less. Such results also verify that the network has a multi-scale perception ability. According to the experimental results, we can conclude that the value in each dimension indicates the enrichment of certain texture patterns, and the number of non-zero dimensions indicates the texture pattern diversity. Therefore, the deep descriptor is able to indicate the texture enrichment of a frame.

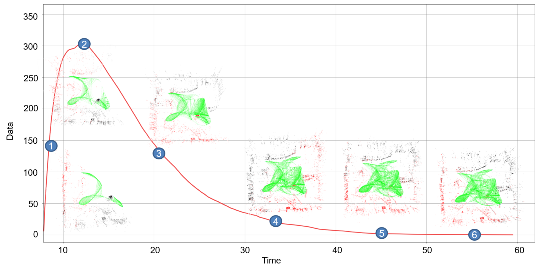


Fig. 6. Curve of entropy, as well as the environment modeling results at the corresponding moments. The experiment is conducted on V1_02.

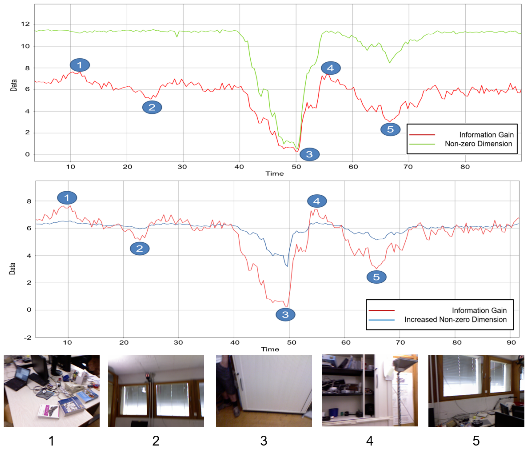
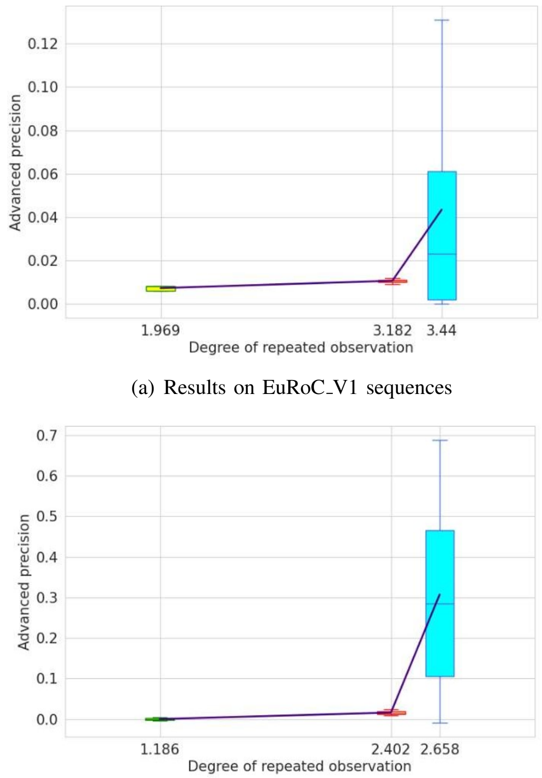


Fig. 7. Curve of information enrichment and the non-zero dimension in the deep descriptor. The top row is the information enrichment (red) and the number of non-zero dimension (green); the mid row is the information enrichment (red) and the increased number of non-zero dimension (blue) from a frame; and the down row is the corresponding frames during the experiment (texture-less frame: 2, 3, 5, and texture-rich frame: 1, 4).



(a) Results on EuRoC_V1 sequences

(b) Results on EuRoC_V2 sequences

Fig. 8. Plots of our advanced precision and the degree of repeated observation. The improved precision by our strategy is positive relative to the degree of repeated observation. Due to the fact that LTDA is significant in the situations with repeated observation, such a result verifies that the proposed strategy enhances LTDA for precision.

In the aspect of the proposed information occupancy grid model, we verify its feasibility by analyzing the calculated information enrichment. To show the information enrichment indication, we draw the curve of information enrichment with repeated observation. As is shown in Fig. 6, the calculated entropy convergences after a period of observation of a room, which means that the information enrichment from the information occupancy grid model can judge the completion of observation. In addition, we analyze the relationship between the information enrichment and the frame texture, which is shown in Fig. 7. With the increment of n^d, more information is obtained because of the diverse texture pattern in a frame. Also, as the increment of Δn^d, the higher information enrichment is achieved due to the new texture pattern response. Such a result shows that our proposed information occupancy grid model is able to measure the information enrichment of frames.

With the information enrichment calculated from the information occupancy grid model, we compare different types of keyframe selection strategies. As shown in Tab. II, our strategy achieves higher precision over the existing keyframe selection strategy, which indicates a better global consistency in a submap-based VSLAM system. Due to the consideration of information enrichment in our keyframe selection, the selected keyframes facilitate loop-closure detection and submap alignment. Therefore, the higher precision is achieved in the VSLAM performance. Moreover, our advance in terms of RMSE is larger when the repeated observation is more significant, as is shown in Fig. 8. Since more LTDA can be established when the degree of repeated observations is larger, this result verifies the enhancement of LTDA by the proposed strategy.

V. Conclusion

In this paper, we propose to enhance long-term data association by selecting texture-rich keyframes. To measure the texture enrichment, a selection strategy with a novel information occupancy grid model and a statistical explainable deep descriptor is introduced. The experiment results demonstrate the statistical explainability of the deep descriptor and the superiority of the proposed keyframe selection strategy. Compared to the existing work in terms of tracking precision, our superiority is significant in situations with repeated observation, showing the enhancement of LTDA introduced by our keyframe selection strategy. In the feature, we plan to build a cloud-edge VSLAM system offloading the feature extraction in the cloud for robot navigation.

References

[1] X. Yang, Z. Yuan, D. Zhu, C. Chi, K. Li, and C. Liao, "Robust and efficient rgb-d slam in dynamic environments," *IEEE Transactions on Multimedia*, pp. 1–1, 2020.

[2] B. Yang, W. Ran, L. Wang, H. Lu, and Y.-P. P. Chen, "Multi-classes and motion properties for concurrent visual slam in dynamic environments," *IEEE Transactions on Multimedia*, pp. 1–1, 2021.

[3] W. Chen, L. Zhu, C. Wang, L. He, and M. Q.-H. Meng, "Ceb-map: Visual localization error prediction for safe navigation," *IEEE Sensors Journal*, vol. 21, no. 10, pp. 11 769–11 780, 2021.

[4] W. Chen, L. Zhu, S. Y. Loo, J. Wang, C. Wang, M. Q.-H. Meng, and H. Zhang, "Robustness improvement of using pre-trained network in visual odometry for on-road driving," *IEEE Transactions on Vehicular Technology*, pp. 1–1, 2021.

[5] R. Mur-Artal and J. D. Tardós, "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[6] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE transactions on pattern analysis and machine intelligence*, vol. 4, 2017.

[7] C. Forster, Z. Zhang, M. Gassner, M. Werlberger, and D. Scaramuzza, "Svo: Semidirect visual odometry for monocular and multicamera systems," *IEEE Transactions on Robotics*, vol. 33, no. 2, pp. 249–265, 2017.

[8] G. Younes, D. Asmar, E. Shammas, and J. Zelek, "Keyframe-based monocular slam: Design, survey, and future directions," *Robotics & Autonomous Systems*, vol. 98, 2017.

[9] W. Chen, L. Zhu, X. Lin, Y. Guan, L. He, and H. Zhang, "Dynamic strategy of keyframe selection with pd controller for vslam systems," *IEEE/ASME Transactions on Mechatronics*, pp. 1–1, 2021.

[10] W. Chen, L. Zhu, Y. Guan, C. R. Kube, and H. Zhang, "Submap-based pose-graph visual slam: A robust visual exploration and localization system," in *arXiv:1807.01012, to appear in Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, 2018.

[11] C. Campos, R. Elvira, J. J. G. Rodriguez, J. M. M. Montiel, and J. D. Tardos, "Orb-slam3: An accurate open-source library for visual, visual¨cinertial, and multimap slam," *IEEE Transactions on Robotics*, p. 1¨C17, 2021. [Online]. Available: http://dx.doi.org/10.1109/TRO.2021.3075644

[12] N. Banerjee, R. C. Connolly, D. Lisin, J. Briggs, and M. E. Munich, "View management for lifelong visual maps," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 7871–7878.

[13] H. Zhang, B. Li, and D. Yang, "Keyframe detection for appearance-based visual slam," in *Ieee/rsj International Conference on Intelligent Robots and Systems*, 2010, pp. 2071–2076.

[14] J. Engel, T. Schöps, and D. Cremers, "Lsd-slam: Large-scale direct monocular slam," in *European Conference on Computer Vision*. Springer, 2014, pp. 834–849.

[15] M. Bloesch, J. Czarnowski, R. Clark, S. Leutenegger, and A. J. Davison, "Codeslam - learning a compact, optimisable representation for dense visual slam," 2019.

[16] F. Endres, J. Hess, J. Sturm, D. Cremers, and W. Burgard, "3-d mapping with an rgb-d camera," *IEEE Transactions on Robotics*, vol. 30, no. 1, pp. 177–187, 2014.

[17] M. Fanfani, F. Bellavia, and C. Colombo, "Accurate keyframe selection and keypoint tracking for robust visual odometry," *Machine Vision and Applications*, vol. 27, no. 6, pp. 833–844, 2016.

[18] J. Stalbaum and J. B. Song, "Keyframe and inlier selection for visual slam," in *International Conference on Ubiquitous Robots and Ambient Intelligence*, 2013, pp. 391–396.

[19] X. Gao, R. Wang, N. Demmel, and D. Cremers, "Ldso: Direct sparse odometry with loop closure," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 2198–2204.

[20] I. J. M. M. Zubizarreta, Jon£¬Aguinaga, "Direct sparse mapping," in *arXiv:1904.06577*, 2019.

[21] R. Arandjelovic and A. Zisserman, "All about vlad," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2013, pp. 1578–1585.

[22] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Computer Vision, IEEE International Conference on*, vol. 3. IEEE Computer Society, 2003, pp. 1470–1470.

[23] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *2007 IEEE conference on computer vision and pattern recognition*. IEEE, 2007, pp. 1–8.

[24] Y. Chen, W. Chen, L. Zhu, Z. Su, X. Zhou, Y. Guan, and G. Liu, "A study of sensor-fusion mechanism for mobile robot global localization," *Robotica*, pp. 1–15, 2019.

[25] J. K. Kamarainen, V. Kyrki, and H. Klviinen, "Invariance properties of gabor filter-based features - overview and applications," *IEEE Transactions on Image Processing*, vol. 15, no. 5, pp. 1088–1099, 2006.

[26] Z. Dai, X. Huang, W. Chen, L. He, and H. Zhang, "A comparison of cnn-based and hand-crafted keypoint descriptors," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019.

[27] N. Sünderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford, "On the performance of convnet features for place recognition," in *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2015, pp. 4297–4304.

[28] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[29] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5297–5307.

[30] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, "Visualizing higher-layer features of a deep network," University of Montreal, Tech. Rep. 1341, Jun. 2009, also presented at the ICML 2009 Workshop on Learning Feature Hierarchies, Montréal, Canada.

[31] B. Charrow, G. Kahn, S. Patil, S. Liu, K. Goldberg, P. Abbeel, N. Michael, and V. Kumar, "Information-theoretic planning with trajectory optimization for dense 3d mapping." in *Robotics: Science and Systems*, vol. 11. Rome, 2015, pp. 3–12.

[32] H. Carrillo, P. Dames, V. Kumar, and J. A. Castellanos, "Autonomous robotic exploration using occupancy grid maps and graph slam based on shannon and r¨¦nyi entropy," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, 2015, pp. 487–494.

[33] A. Das and S. L. Waslander, "Entropy based keyframe selection for multi-camera visual slam," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015, pp. 3676–3681.

[34] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[35] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[38] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, "Visualizing higher-layer features of a deep network," *University of Montreal*, vol. 1341, no. 3, p. 1, 2009.

[39] R. Bansal, G. Raj, and T. Choudhury, "Blur image detection using laplacian operator and open-cv," in *2016 International Conference System Modeling Advancement in Research Trends (SMART)*, 2016, pp. 63–67.

[40] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000 km: The oxford robotcar dataset," *The International Journal of Robotics Research*, vol. 36, no. 1, pp. 3–15, 2017.

[41] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of rgb-d slam systems," in *Proc. of the International Conference on Intelligent Robot Systems (IROS)*, Oct. 2012.

[42] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. W. Achtelik, and R. Siegwart, "The euroc micro aerial vehicle datasets," *The International Journal of Robotics Research*, vol. 35, no. 10, pp. 1157–1163, 2016.