Robotica (2023), **41**, pp. 3724–3745
doi:10.1017/S026357472300125X

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

SPSVO: a self-supervised surgical perception stereo visual odometer for endoscopy

Junjie Zhao[1], Yang Luo[1], Qimin Li[1], Natalie Baddour[2] and Md Sulayman Hossen[3]

[1]College of Mechanical and Vehicle Engineering, Chongqing University, Chongqing, China, [2]Department of Mechanical Engineering, University of Ottawa, Ottawa, ON, Canada, and [3]College of civil engineering Chongqing university, Chongqing University, Chongqing, China
Corresponding author: Yang Luo; Email: yluo688@cqu.edu.cn.

Received: 4 May 2023; **Revised:** 28 July 2023; **Accepted:** 30 August 2023; **First published online:** 29 September 2023

Keywords: Visual odometer (VO); virtual endoscopy; self-supervision; pose estimation

Abstract

Accurate tracking and reconstruction of surgical scenes is a critical enabling technology toward autonomous robotic surgery. In endoscopic examinations, computer vision has provided assistance in many aspects, such as aiding in diagnosis or scene reconstruction. Estimation of camera motion and scene reconstruction from intra-abdominal images are challenging due to irregular illumination and weak texture of endoscopic images. Current surgical 3D perception algorithms for camera and object pose estimation rely on geometric information (e.g., points, lines, and surfaces) obtained from optical images. Unfortunately, standard hand-crafted local features for pose estimation usually do not perform well in laparoscopic environments. In this paper, a novel self-supervised Surgical Perception Stereo Visual Odometer (SPSVO) framework is proposed to accurately estimate endoscopic pose and better assist surgeons in locating and diagnosing lesions. The proposed SPSVO system combines a self-learning feature extraction method and a self-supervised matching procedure to overcome the adverse effects of irregular illumination in endoscopic images. The framework of the proposed SPSVO includes image pre-processing, feature extraction, stereo matching, feature tracking, keyframe selection, and pose graph optimization. The SPSVO can simultaneously associate the appearance of extracted feature points and textural information for fast and accurate feature tracking. A nonlinear pose graph optimization method is adopted to facilitate the backend process. The effectiveness of the proposed SPSVO framework is demonstrated on a public endoscopic dataset, with the obtained root mean square error of trajectory tracking reaching 0.278 to 0.690 mm. The computation speed of the proposed SPSVO system can reach 71ms per frame.

1. Introduction

Gastrointestinal cancer is the second leading cause of cancer death in the world and accounts for about 35% of all cancer-related deaths [1, 2]. Some hospitals are now equipped with two-dimensional endoscopic instruments for doctors, such as the da Vinci® surgical system (Intuitive Surgical, Inc., Sunnyvale, CA), to assist in performing minimally invasive surgery (MIS) of the gastrointestinal tract, abdominal cavity, chest cavity, and throat. The most direct and effective screening for gastrointestinal cancers is two-dimensional endoscopy, such as capsule endoscopy, upper gastrointestinal endoscopy, and colonoscopy [3–6].

In traditional endovascular MIS processes, the position of diseased tissue is generally estimated by visually examining 2D endoscope images. However, the endoscope images usually lack sufficient texture. When combined with irregular illumination, extensive, similar areas, and low contrast, it becomes difficult for surgeons to quickly and accurately locate lesions. Other problems due to hand-eye coordination and visual misdirection may also occur during operation [7]. Recently, computer vision-based algorithms have attracted much attention for success in stereoscopic endoscope position tracking and

providing intraoperative reconstruction of surgical scenes. Tatar et al. [8] attempted to use a depth camera combined with a time-of-flight method to locate positions of surgical instruments. Lamata et al. [9] investigated the features (mutual reflection, diffuse reflection, highlight parts, and colors) of human liver photos based on the Lambert-body method and tried to reconstruct a 3D model of the liver by adjusting the albedo and light intensity of the endoscopic images. Wu et al. [10] aimed to track geometric constraints of surgical instruments and reconstruct 3D structures from 2D endoscopic images with a constrained decomposition method. Seshamani et al. [11] combined a video mosaic method and an online processing technique to expand the field of view to better assist surgeons in performing surgeries and lesion diagnosis. Due to the complex features of an enterocele, endoscopic images often have strong illumination variation and feature sparsity, resulting in difficulties for the aforementioned methods to realize precise organ 3D reconstruction and lesion localization.

Recently, the Structure from Motion (SfM) approach was proposed to construct high-quality 3D models of human organs based on endoscopic images. The SfM approach mainly consists of feature extraction, keypoint matching, attitude estimation, and beam adjustment. Based on the SfM technique, Thormaehlen et al. [12] generated a 3D model of the human colon with surface texture features. Koppel et al. [13] developed an automated SfM approach to reconstruct a 3D model of the colon from endoscopic images to assist surgeons in surgical planning. Mirota et al. [14] proposed a direct SfM approach to track endoscope position using video data to improve the accuracy of Endonasal Skull Base Surgery navigation. Kaufman et al. [15] applied a direct Shape from Shading (SfS) algorithm to better extract detailed information of surface textures from endoscopic images and combined the SfM method to reconstruct a refined 3D model of human organs. Assisted by manual drawing of the outline of the major colonic folds, Hong et al. [16] reconstructed a virtual colon segment based on an individual colonoscopy image to aid surgeons in detecting colorectal cancer lesions. However, accurate reconstruction of human organs based on SfM methods requires stable camera motion since it needs to match feature points between multiple images and calculate the camera pose. Furthermore, data obtained from sensors such as monocular cameras, Inertial Measurement Units, ultrasonic lidar, etc., are usually large, thus requiring computing resources to perform batch data processing. Hence, SfM techniques are usually applied offline. For actual surgical operation, real-time feedback plays an important role in providing surgeons with timely and accurate information to allow them to make optimal decisions and adapt their approach as necessary during the procedure. A real-time online computer vision-based algorithm is hence highly desirable to improve accuracy and precision of surgical interventions and reduce the risk of complications or adverse outcomes.

The Visual Simultaneous Localization and Mapping (VSLAM) method is a real-time online data processing technique which requires less computing resources compared to the SfM approach. VSLAM utilizes endoscopic video or image sequences to estimate the pose and location of the endoscope and to reconstruct the abdominal cavity and other scenes of the MIS [17–19]. The goal of VSLAM is to improve the visual perception of surgeons, and it plays an important role in developing online surgical navigation systems and medical augmented reality technology. Much research in recent years has focused on improving the accuracy and efficiency of VSLAM methods for medical applications, particularly in the context of MIS systems. Mountney et al. [20] first explored the application of VSLAM in MIS by extending the Extended Kalman Filter SLAM (EKF-SLAM) framework to handle complex light reflection and low-texture environments. However, the obtained point clouds were too sparse and could not represent 3D shapes and detailed surface textures of human organs. Mountney and Yang [21] proposed a novel VSLAM method to online estimate tissue deformation and motion of the laparoscopic camera by establishing a periodic tissue deformation parameter model and generating a joint registered 3D map with preoperative data. However, the slow speed of the system's map-building algorithm can lead to poor real-time tracking and loss of feature points. In [22], Klein and Murray proposed a Parallel Tracking and Mapping (PTAM) algorithm, a monocular VSLAM approach based on keyframes. The PTAM can run in real time on a single CPU and handle large-scale environments and a variety of lighting conditions. However, it requires high-quality feature detection and feature matching for camera locating and scene mapping.

https://doi.org/10.1017/S026357472300125X Published online by Cambridge University Press

The aforementioned methods are generally based on monocular endoscopes, where it is difficult to process endoscopic images with small viewing angles and rapid frame transitions. Lin et al. [23] extended the application scope of PTAM to stereo endoscopy, which allows for simultaneous stereoscopic tracking, 3D reconstruction, detection of deformation points in the MIS setting and can generate denser 3D maps compared to EKF-SLAM methods. However, this stereo system suffers from time-consuming feature point matching. Later, Lin et al. [24] improved texture feature selection, green channel selection, and reflective area processing of the endoscopic images and proposed a revised VSLAM method to restore the surface structure of a 3D scene of abdominal surgery based on SLAM. However, the proposed method relies heavily on tissue surface vascular texture. In cases where the tissue being imaged has little or no vascularity, this method may not be effective in detecting unique features. Recently, Mur-Artal [25] provided an ORBSLAMsystem constructed via a robust camera tracking and mapping estimator with remarkable camera relocation capabilities. Mahmoud [26] applied the ORBSLAM algorithm to track the position of the endoscope without additional tracking elements and provide 3D reconstruction in real time. This extended the ORBSLAM to reconstruct semi-dense maps of soft organs. However, although the above two ORBSLAM methods based on feature point approaches reduce computational complexity, the reduction in the amount of information compared to the original graph also implies that some useful information is lost. While the two ORBSLAM methods reduce computational complexity, the reduction in useful information can lead to inaccurate camera location and visceral surface texture mapping.

Feature point detection is a fundamental and important processing step in Visual Odometry (VO) or VSLAM. Local features, such as the Scale Invariant Feature Transform (SIFT), Speed Up Robust Feature (SUFT), Oriented FAST, and Rotated BRIEF (ORB), for camera pose estimation are commonly hand-crafted by calling OpenCV algorithms from a third-party function library. However, the feature points extracted by these algorithms are often unevenly distributed, with large amounts of useful data lost, resulting in inaccurate camera positioning and scene mapping [27–29]. Moreover, the surface of the human viscera often has poor texture. Endoscope images often have a small field of view and are commonly taken with lighting changes and specular reflection, Fig. 1. Weak textures and specular reflections pose challenges to VSLAM [27], making many SfM or SLAM frameworks such as ORB-SLAM3 [30] ineffective in these situations. In this paper, a self-supervised feature extraction method "SuperPoint" [31] and a matching feature technique "SuperGlue" are applied to address challenges such as illumination changes, weak textures, and specular reflections in the human viscera. Moreover, this approach accelerates convolutional Neural Network (CNN) computations to enable real-time endoscopic pose estimation and viscera surface map construction.

Feature matching is another critical step in feature-based VO or SLAM techniques. This involves finding the same features in two images and establishing correspondences between them to achieve camera pose estimation and map updates. The performance of the feature-matching process directly affects the accuracy and stability of the VO or SLAM system. Chang et al. [32] used feature matching to perform heart surface sparse reconstruction through structural propagation. The algorithm obtained parallax data between point pairs to estimate stereo parallax of each frame and motion information between consecutive frames. However, the method obtained a sparse parallax field, and further complex interpolation calculations were required to obtain a denser reconstructed scene of the heart surface. Lin et al. [33] utilized a vessel-based line-matching approach based on block-matching geometry to avoid pixel-wise matching. However, the application of local characteristics of image features of the viscera can lead to mismatched point pairs and thus incorrect camera location. Direct methods such as DSO [34] or DSM [35] and hybrid methods such as SVO [36] assume that ambient illumination remains constant, which is difficult to ensure due to severe illumination variations of endoscopic images. The Self-Supervised Learning (SSL) approach can match images by using image content itself as supervision, without requiring explicit labels or annotations. SSL methods have shown promising performance in image-matching tasks such as stereo matching and optical flow estimation of real-life scenarios and have enhanced robustness to local illumination changes [37]. However, the performance of SSL in endoscopic image matching is unknown and remains to be studied. This paper proposes an

https://doi.org/10.1017/S026357472300125X Published online by Cambridge University Press

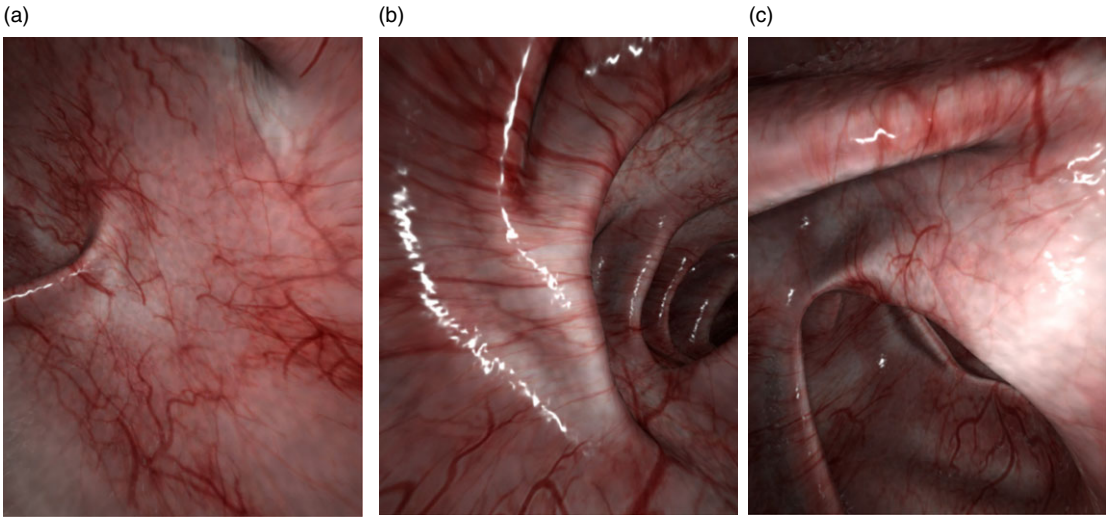

Figure 1. *Frames from "colon_reconstruction_dataset." (a) a Small field of view, (b) specular reflections, (c) lighting changes.*

improved SSL method with adaptive deep learning to address data association between endoscopic images.

This paper introduces SPSVO, a self-supervised surgical perception stereo visual odometer for endoscope pose (position and rotation) estimation and scene reconstruction. The proposed method overcomes adverse effects of endoscopic images on feature extraction and tracking, such as irregular illumination, poor surface texture, low contrast and extensive, and similar areas. The main contributions of this paper are as follows:

- A VO system is proposed that integrates a SuperPoint feature extraction method based on CNN and a SuperGlue feature-matching network. The SPSVO system enables extraction of enriched feature points compared with common hand-crafted local feature-detecting methods, such as ORB, SIFT, and SUFT.
- An image illumination pre-processing technique is proposed to address mirror reflection and illumination variations of endoscopic images.
- The SPSVO system includes image pre-processing, feature extraction, stereo matching, feature tracking, keyframe selection, and pose graph optimization.
- The performance of the proposed system is evaluated based on a public dataset: "colon_reconstruction_dataset" [38]. Results indicate that the proposed system outperforms ORB-SLAM2 [39] and ORB_SLAM2_Endoscopy [40] methods in feature detection and tracking. ORB-SLAM2 cannot extract sufficient feature points to initialize the scene map of viscera and thus results in loss of the endoscope track.
- The proposed system is capable of accurate and rapid operation within the human viscera; the computation speed of the SPSVO system is as fast as 131ms per frame, enabling real-time surgical navigation.

The rest of this paper is organized as follows: Section 2 presents related work on endoscopic VSLAM methods. Section 3 presents the proposed SPSVO system. Section 4 presents experimental results and analysis. Finally, conclusions are drawn in Section 5.

https://doi.org/10.1017/S026357472300125X Published online by Cambridge University Press

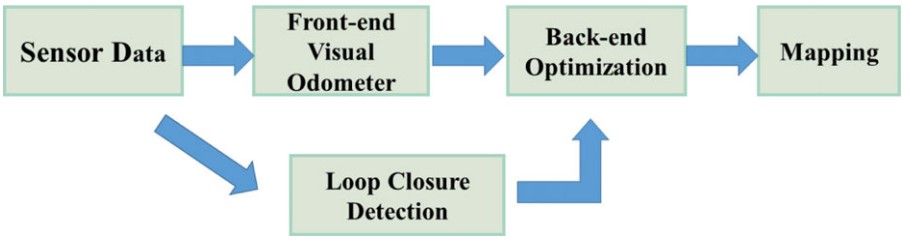

Figure 2. *Classic VSLAM framework.*

2. Related work

2.1. VSLAM and VO for endoscopy

VSLAM is a technique that uses camera vision for simultaneous robot self-locating and scene map construction [41]. It enables autonomous robot exploration in unknown or partially unknown environments. The architecture of a classical VSLAM system typically includes a front-end visual odometer, backend optimization, loop closure detection, and finally mapping, as shown in Fig. 2.

VSLAM has the potential to estimate the relative pose of the endoscope camera and construct a viscera surface texture map, which is important for lesion localization and surgical navigation. However, complicated intraoperative scenes (e.g., deformable targets, surface texture, sparsity of visual features, viscera specular reflection, etc.) and strict accuracy requirements have posed challenges to the application of VSLAM to minimally invasive surgery. Recently, Lamarca et al. [42] proposed a monocular non-rigid SLAM method that combines shape from template (SfT) and non-rigid structure from motion (NRSfM) methods for non-rigid environment scene map construction. However, this method is susceptible to variations in illumination and does not perform well under poor visual texture conditions, rendering it unsuitable for reconstruction of viscera with non-isometric deformations. Later, Gong et al. [43] constructed an online tracking and relocation framework which employs a rotation invariant Haar-like descriptor and a simplified random forest discriminator to select and track the target region for gastrointestinal biopsy images. Song [44] constructed a real-time SLAM system to address scope-tracking problems through an improved localization technique. Much work has focused on adapting VSLAM to enable application to an endoscopic scene, addressing problems such as poor texture [45, 46], narrow field of view [11], and specular reflections [47]. Still, the variable illumination problem remains unaddressed. Intraoperative scenarios require accurate camera localization; complex viscera images can lead to mismatched point pairs and thus incorrect camera location. Data association also remains a challenging problem for VSLAM systems in MIS scenarios [48]. This paper focuses on addressing the problems of variable illumination and data association for intraoperative scenes.

2.2. SLAM based on SuperPoint and SuperGlue

CNNs have made outstanding achievements in computer vision to aid lesion diagnosis or intraoperative scene reconstruction [48–52]. Researchers have studied and improved many aspects of VSLAM with learning-based feature extraction techniques to address variable illumination and poor visceral surface texture in complex surgical scenarios [53, 54]. Bruno et al. [49] presented a novel hybrid VSLAM algorithm based on a Learned Invariant Feature Transform network to perform feature extraction in a traditional backend based on an ORB-SLAM system. Li et al. [52] attempted to use an end-to-end deep CNN in VSLAM to extract local descriptors and global descriptors from endoscopic images for pose estimation. Schmidt et al. [50] proposed Real-Time Rotated descriptor (ReTRo), which was more effective than classical descriptors and allowed for the development of surgical tracking and mapping frameworks. However, the aforementioned methods are based on traditional Fast Library for Approximate Nearest Neighbors (FLANN) techniques to track keypoints and match extracted features. FLANN does not perform well at feature point matching of high-similarity images, resulting in mismatches between

https://doi.org/10.1017/S026357472300125X Published online by Cambridge University Press

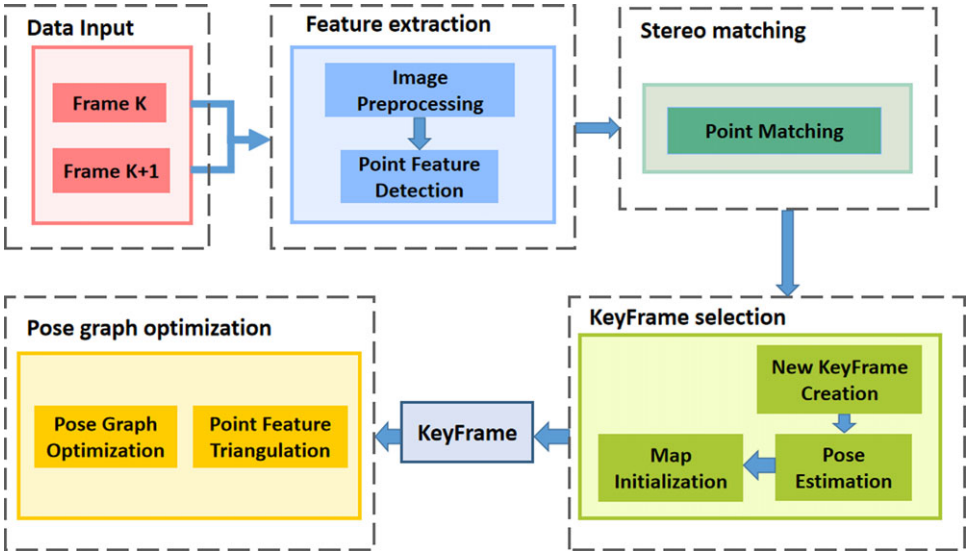

Figure 3. *Structure of the SPSVO system.*

extracted new features and potential features. Its performance is even worse under variable illumination; therefore, FLANN is not always applicable for MIS [55].

This paper proposes to apply a SuperPoint approach for keypoint detection and to utilize the SuperGlue technique to deal with complex data associations in intraoperative scenes. SuperPoint [31] is a self-supervised framework for detecting features and describing points of interest, while SuperGlue [37] is a network that can simultaneously filter outliers and match features. Recently, researchers have studied the effectiveness of SuperPoint and SuperGlue in VSLAM systems for MIS [56, 57]. Barbed et al. [56] demonstrated that SuperPoint delivers better feature detection in VSLAM than using hand-crafted local features. Laura et al. [58] applied SuperPoint to a monocular VSLAM system to estimate the pose of the ureteroscope tip. Sarlin et al. [57] proposed a Hierarchical Feature Network (HF-Net) algorithm based on SuperPoint and SuperGlue to predict local features and global descriptors for a 6-DoF localization of the camera. However, existing algorithms require substantial computing power to run in real time, which presents a significant obstacle to building maps in real time. In this work, a SPSVO algorithm is proposed to accelerate the CNN to realize real-time endoscopic pose estimation and viscera surface map construction.

3. Proposed SPSVO approach

3.1. System overview

The proposed SPSVO approach consists of four main modules: feature extraction, stereo matching, keyframe selection, and pose graph optimization, as shown in Fig. 3. The SPSVO can perform feature matching and keypoint tracking between stereo images and images in different frames, and can avoid incorrect data associations by using matching results of relevant key points. For real-time performance, the SPSVO performs feature tracking of images from only the left eye to reduce computation time. Nvidia TensorRT Toolkit is used to accelerate feature extraction and matching. On the backend, the SPSVO uses a traditional pose graph optimization framework for map construction. The above modules are designed to enable real-time application of the SPSVO within human enterococci and achieve accurate tracking by combining the efficiency of traditional optimization methods and the robustness of learning-based techniques.

https://doi.org/10.1017/S026357472300125X Published online by Cambridge University Press

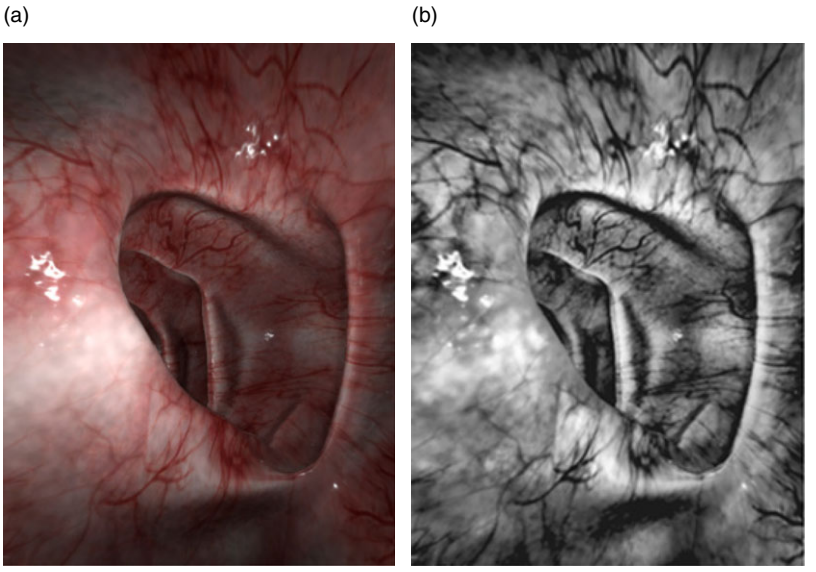

Figure 4. *Image pre-processing. (a) Original image and (b) image after using CLAHE.*

3.2. Image pre-processing

For image pre-processing, the SPSVO uses Contrast-Constrained Adaptive Histogram Equalization (CLAHE) [59] to enhance contrast, brightness, details, and texture of the input image. Due to severe variability in illumination in optical colonoscopy, some parts of the L-channel color space of the image are overexposed, resulting in image specular reflections, while some images are underexposed and lead to dark areas. In this work, pixels with a luminance greater than 50 are marked as reflective regions, and pixel values in the reflective region are set to the average of surrounding pixels. Possible noise is eliminated by a morphological closure operation. Performance of the CLAHE is demonstrated in Fig. 4. The proposed CLAHE effectively improves the uniformity of illumination and improves the contrast of endoscopic images. Due to the proximity of the endoscope light source to the inner wall of the organ and rapid movement of the endoscope, this pre-processing step allows the system to eliminate the effects of mismatches caused by specular reflections.

3.3. Proposed SuperPoint model

The SuperPoint network consists of four parts: encoding network, feature point detection network, descriptor detection network, and loss function. The encoder network converts the input image into a high-dimensional tensor representation for the decoder, making it easier to detect and describe key points. The feature point detection network is a decoding structure that calculates keypoint probability for each pixel and embeds a sub-pixel convolution algorithm to reduce computational effort. The descriptor detection network is also a decoding structure that extracts semi-dense descriptors first, performs a bicubic interpolation algorithm to obtain full descriptors, and uses L2-normalization to obtain unit-length descriptors. The loss function is a measure of the difference between the network output and ground truth label, guiding the network to optimize and improve its performance in detecting and describing key points of the input image. This provides better performance for related applications such as VSLAM, 3D reconstruction, and autonomous navigation. The SuperPoint network is trained in PyTorch. The input of the SuperPoint network is a single image I with $I \in R^{H \times W}$, where H is the height and W is the width of the image, in pixels. The output of the network is positions of key points extracted in each image and their corresponding descriptors.

https://doi.org/10.1017/S026357472300125X Published online by Cambridge University Press

Table I. *The parameters used in ORB-SLAM2 and SPSVO are presented. Default parameters indicate in italics, the optimal parameters indicate in bold. Parameters A, B, C, and D all use default values in order to follow the principle of variable control. A: Scale factor between levels in ORB scale pyramid. B: Number of levels in ORB scale pyramid. C: Initial response threshold of FAST detector. D: Minimum response threshold of FAST detector.*

	ORB-SLAM2					**SPSVO**
	Threshold of feature points	**A**	**B**	**C**	**D**	**Threshold of feature points**
Value 1	700	*1.20*	*8*	*20*	*7*	700
Value 2	*1200*	*1.20*	*8*	*20*	*7*	*1200*
Value 3	**1600**	*1.20*	*8*	*20*	*7*	**1600**

Based on Barbed [56], the loss function can be expressed as

$$L_{SP}\left(X, X', D, D'; Y, Y', S\right) = L_P\left(X, Y\right) + L_P\left(X', Y'\right) + \lambda L_d\left(D, D', S\right) \tag{1}$$

where the X and X ́are outputs of the original detection header of image I and warped image I,́respectively. The associated detection pseudo-labels are Y and Y.;́D and D ́are outputs of the raw description header. $S \in R^{H/8\times W/8\times H/8\times H/8}$ is the homography estimation matrix. L_P represents the loss of feature points during detection, which can be used to measure the difference between detected outputs and the pseudo-label. The L_d is the loss function of the descriptor; λ is a weight parameter used to balance the weight of L_p and L_d.

As shown in Fig. 1(b), there are generally multiple specular reflection areas (white spot areas) that exist in an endoscopic image. Most existing feature detection methods tend to detect many feature points around contour areas or specular reflection areas [56]. For VSLAM, the more evenly the feature points are distributed in the image, the more accurately feature matching can estimate spatial pose relation. To make feature points extracted by SuperPoint evenly distributed in the region of interest, the specularity loss (L_S), which reconsiders weights of all extracted key points in specular regions, is proposed. The revised loss function is defined as

$$L_{ESP}\left(I, I', X, X', D, D'; Y, Y', S\right) = L_{SP}\left(\ldots\right) + \lambda_S L_S\left(X, I\right) + \lambda_S L_S\left(X, I'\right), \tag{2}$$

in which λ_S is a scale weighting factor determined by characteristics of the dataset and contribution of each objective function to the model performance. In this work, $\lambda_s = 100$. The L_S is defined as

$$L_S\left(X, I\right) = \frac{\sum_{h,w=1}^{H,W}\left[m\left(I\right)_{hw} \cdot d2s\left(softmd\left(X\right)\right)_{hw}\right]}{\varepsilon + \sum_{h,w=1}^{H,W} m\left(I\right)_{hw}}, \tag{3}$$

where *softmd()* and *d2s()* are SoftMax functions. The ε is a constant with $\varepsilon = 10^{-10}$ [31, 56]. The $m(I)_{hw}$ is a weighting mask, where $m(I)_{hw} > 0$ for pixels near a specularity and 0 otherwise. The value of L_S is close to zero when there is no key point at that location.

The default thresholds of the parameters of the ORB-SLAM2 and SPSVO are determined based on [30, 51], as shown in Table 1. The algorithms were run with default thresholds at the beginning and calibrated by comparing with the results of the ground truth values through increasing or decreasing the thresholds. In this work, ±40% variations were made with respect to the default thresholds. Figure 5 shows the comparison of the number of keypoints matched per keyframe with feature points threshold of 1600, as can be observed that the proposed SPSVO outperforms the ORB-SLAM2 in terms of matched feature points (approximately 700 points versus 500 points).

Comparison of the distribution of feature points extracted by the SPSVO algorithm and ORB-SLAM2 on the "colon_reconstruction_dataset" [38] is shown in Fig. 6; the image resolution is 480 × 640.

https://doi.org/10.1017/S026357472300125X Published online by Cambridge University Press

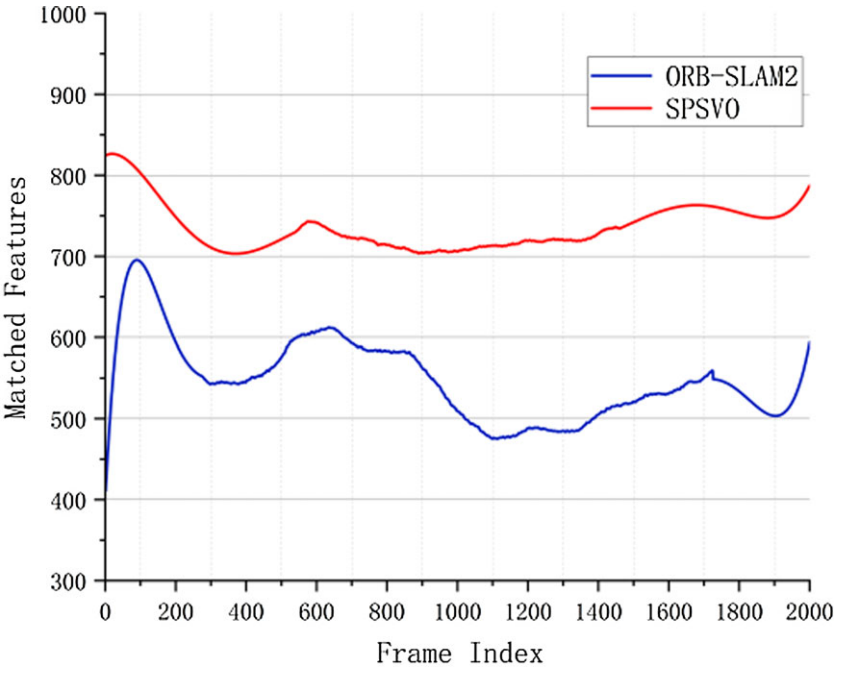

Figure 5. *Comparison of number of keypoints matched per keyframe with feature points threshold of 1600.*

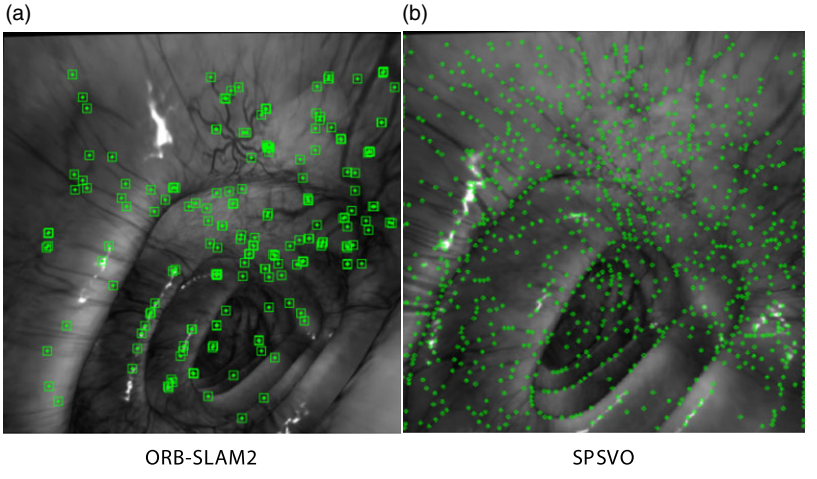

Figure 6. *Comparison of feature extraction results of the ORB-SLAM2 and SPSVO methods. (a) ORB-SLAM2 and (b) SPSVO.*

According to the results of Table 1, the upper threshold of feature point extraction is set to 1600 to ensure that both algorithms have the potential to obtain perfect system performance in most scenarios. It can be seen that the SPSVO extracts more effective features than ORB-SLAM2. The large number and even distribution of feature points will provide more scene information, thus improving the accuracy of camera localization. Furthermore, the feature points extracted by SPSVO are evenly distributed and located in textured areas, which is beneficial for subsequent VSLAM tasks such as keypoint matching, camera localization, map construction, and path planning.

https://doi.org/10.1017/S026357472300125X Published online by Cambridge University Press

Algorithm 1. SuperGlue Stereo Matching

Input: features_0 features_0
Output: the matched feature points
foreach exists(features_0 [idx]) || exists(features_1 [idx] **do**
norm0 = Normalize (features_0, image_width, image_height);
norm1 = Normalize (features_1, image_width, image_height); **then**
superglue.infer(norm_features0, norm_features1, indices0, indices1);
for(size_t i = 0; i < indices0.size(); i++) do
d = 1.0 - (mscores0[i] + mscores1[indices0[i]]) / 2.0; **then**
reject outliers
end
end

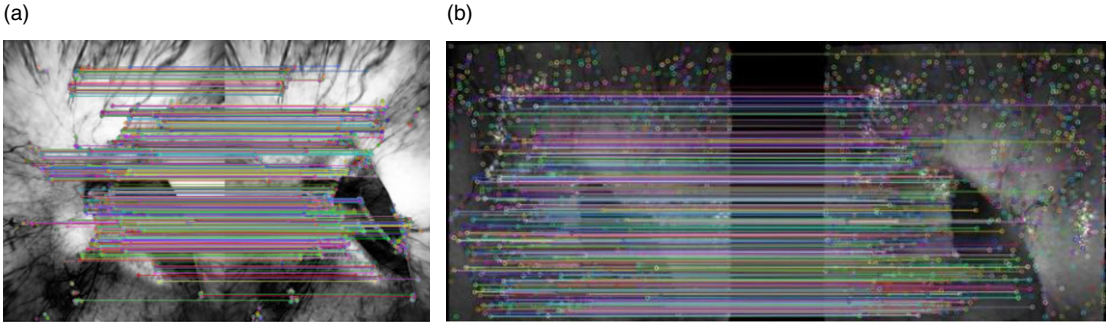

Figure 7. *Comparison of feature matching. (a) ORB_SLAM2 and (b) SPSVO.*

3.4. Feature matching

The SuperGlue algorithm is commonly applied to simultaneously address feature matching and outlier filtering for real-time pose estimation in indoor and outdoor environments [60–62]. SuperGlue needs to be trained on the true value of the trajectory in the abdominal cavity to achieve an adaptive intra-abdominal environment. A bi-directional brute force matching algorithm is utilized to establish correspondence between features in consecutive frames of an image sequence. Additionally, SPSVO uses the Random Sample Consensus algorithm to remove false matches of feature points for robust geometric estimation, see Algorithm 1. Figure 7 shows the results of the proposed algorithm for stereo matching. The successfully matched feature pairs are connected by lines. It can be seen that the SPSVO can accurately match a large number of key points. Moreover, the SPSVO has good consistency in feature matching between frames, where a feature point can be consistently matched across multiple frames. Consistent matching indicates that the proposed SPSVO can effectively estimate camera position.

3.5. Keyframe selection

Keyframe selection plays an important role in reducing computational cost, decreasing redundant information, and improving accuracy of VSLAM [22, 25, 63]. The general criteria for keyframe selection are (1) distribution of the keyframes should not be too dense or too sparse; (2) the number of keyframes should generate sufficient local map points [54]. Unlike other SLAM or VO systems, SPSVO integrates a learning-based matching method that can effectively match frames with large differences in baseline

https://doi.org/10.1017/S026357472300125X Published online by Cambridge University Press

Algorithm 2. The keyframe selection

```
Input: last_keyframe current_frame num_match
Output: bool(is_new_keyframe);
foreach exists(current_frame [idx]) do
        Matrix4d current_pose = get_current_pose(); then
        Matrix4d last_keyframe_pose = get_last_keyframe_pose(); then
        if (num_feature_matches < min_num_feature_matches ||
                delta_rotation_angle > max_rotation_angle ||
                delta_translation_distance > max_translation_distance ||
                num_frames_since_last_keyframe > max_frames_since_last_keyframe){
            is_new_keyframe = true;
        }
        return is_new_keyframe;
        end
end
```

length. Therefore, during feature-matching SPSVO only matches the current frame with keyframes, which can reduce tracking error. The keyframe selection criteria should take into account the movement between frames, information gain, tracking stability, and previous experience. Based on the key frame selection principle [30, 51], the keyframe selection criteria corresponding to the matching process of SPSVO are defined as:

- The distance between the current frame and the nearest keyframe (L) satisfies the condition of $L > D_f$;
- The angle between the current frame and the nearest keyframe (θ) satisfies the condition of $\theta > \theta_f$;
- The number of map points (N_A) tracked by the current frame satisfies the condition $N_1^u < N_A < N_2^l$;
- The number of the map points (N_B) tracked by the current frame satisfies the condition $N_B < N_3$;
- The number of frames since the last keyframe inserted (N_C) satisfies the condition of $N_C > N_4$.

in which, $D_f, \theta_f, N_1^u, N_2^l, N_3, N_4$ are preset thresholds. A frame is selected as a keyframe if it meets any of the above conditions, see Algorithm 2. The proposed keyframe selection criteria consider both image quality and keypoint quality. These can play an important role in filtering useless or incorrect information and avoiding adverse impacts on endoscope localization and scene mapping.

3.6. Keyframe selection

The Levenberg Marquardt (LM) algorithm is used as the optimization solver in the backend of the proposed SPSVO to construct the Covisibility Graph. For each optimizing iterative loop, when LM optimization converges, both inputs and outputs of the optimization process are set as inputs of the loss function for decoding network training. The optimization variables are keyframes and map points, and the corresponding constraints are the monocular and stereo constraints.

3.6.1 The monocular constraint

If a 3D map point ${}^wP_i$ is observed by the left eye camera, the reprojection error $e_{k,i}$ of the i-th point in the k-th frame is defined as

$$e_{k,i} = \hat{p}_i - \pi_i \left({}^cR_w {}^wP_i + {}^ct_w \right), \tag{4}$$

https://doi.org/10.1017/S026357472300125X Published online by Cambridge University Press

where ${}^{w}P_i$ is the *i-th* point observed by frame k, w is the world coordinate system and c is the camera coordinate system. R and t are the rotation and translation of the camera. $\hat{p}_i = (\hat{u}_i, \hat{v}_i)$ is the observation data of the map point on the frame, and $\pi_i(\cdot)$ is the camera projection model representing coordinates of the 3D map point projection on the left eye image, expressed as

$$\pi_i \left(\begin{bmatrix} x_i \\ y_i \\ z_i \end{bmatrix} \right) = \begin{bmatrix} f_x \frac{x_i}{z_i} + c_x \\ f_y \frac{y_i}{z_i} + c_y \end{bmatrix}, \tag{5}$$

where $[\, x_i \; y_i \; z_i \,]^{\mathrm{T}}$ are the world coordinates of point ${}^{w}P_i$, and f_x, f_y, c_x, c_y are the intrinsic parameters of camera.

3.6.2 The stereo constraint

If a 3D map point ${}^{w}P_j$ is observed by both left and right cameras at the same time, the reprojection error is defined as

$$e_{k,j} = \hat{p}_j - \pi_j \left({}^{c}R\, {}^{w}P_j + {}^{c}t \right), \tag{6}$$

where $\hat{p}_j = (\hat{u}_j, \hat{v}_j, \hat{r}_j)$ is the observation data of the map point on the k-th frame of the right image, and $\hat{r}_j$ is the horizontal coordinate of the right image. $\pi_j(\cdot)$ is the camera projection model representing the 3D map point projection on the stereo image and defined as

$$\pi_j \left(\begin{bmatrix} x_j \\ y_j \\ z_j \end{bmatrix} \right) = \begin{bmatrix} f_x \frac{x_j}{z_j} + c_x \\ f_y \frac{y}{z_j} + c_y \\ f_x \frac{x_j - b}{z_j} + c_x \end{bmatrix}, \tag{7}$$

where b represents the baseline of the stereo camera. $[\, x_j \; y_j \; z_j \,]^{\mathrm{T}}$ are the world coordinates of point ${}^{w}P_j$.

3.6.3 Graph optimization

Assuming that the distribution of key points satisfies a Gaussian distribution [64], the final cost function of the proposed SPSVO can be defined as

$$J = \sum_{k,i} \rho_{k,i} \left(\left(e_{k,i} \right)^{\mathrm{T}} \left(\Sigma_{k,i} \right)^{-1} \left(e_{k,i} \right) \right) + \sum_{k,j} \rho_{k,j} \left(\left(e_{k,j} \right)^{\mathrm{T}} \left(\Sigma_{k,j} \right)^{-1} \left(e_{k,j} \right) \right), \tag{8}$$

where $\rho_{k,i}$ and $\rho_{k,j}$ are robust kernel functions to further reduce the impact of any possible outliers. $(e_{k,i})^{\mathrm{T}}$ and $(e_{k,j})^{\mathrm{T}}$ are the transpose of matrix $e_{k,i}$ and $e_{k,j}$, respectively. $\Sigma_{k,i}$ and $\Sigma_{k,j}$ are covariance matrices, and $(\Sigma_{k,i})^{-1}$, $(\Sigma_{k,j})^{-1}$ are the inverse of these covariance matrices, respectively.

4. Experimental validation of the proposed SPSVO method

In this section, the performance of the proposed SPSVO is evaluated based on the "colon_reconstruction_dataset" [38] and compared with ORB-SLAM2. SPSVO is a stereo VO system without loop closure detection module. Furthermore, the colon_reconstruction_dataset does not involve scene re-identification or map closure situations, so the impact of loop closure detection module on algorithm comparison is very limited. Therefore, to ensure fair and accurate comparison, loop closure detection is turned off in ORB-SLAM2. Frame threshold is defined as the number of times a map point is observed by a keyframe for monocular and stereo constraints in graph optimization.

https://doi.org/10.1017/S026357472300125X Published online by Cambridge University Press

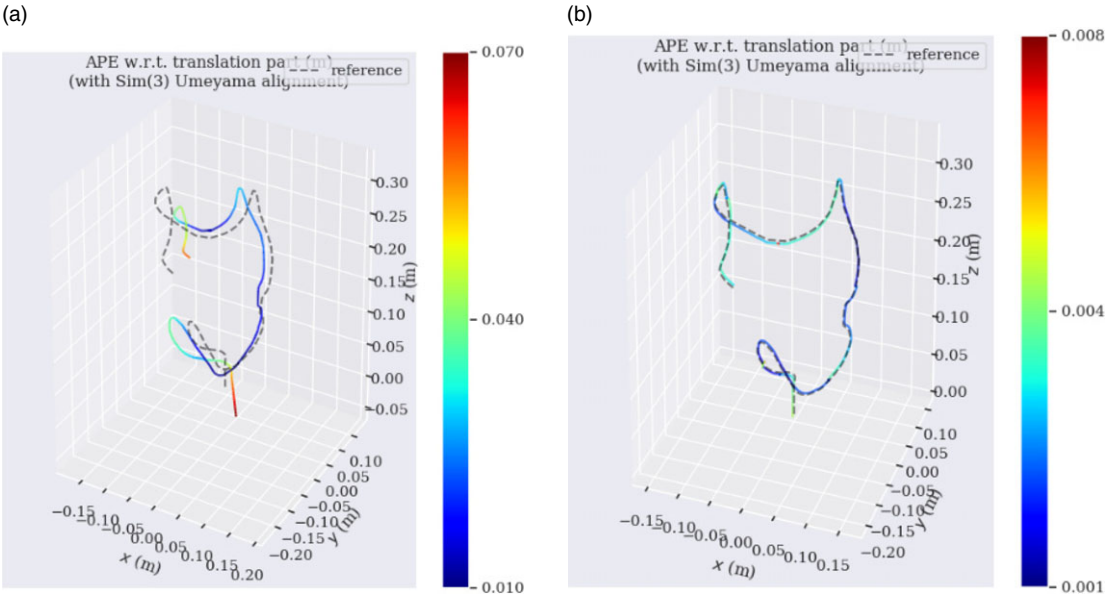

Figure 8. *Comparison between the endoscope trajectories estimated by ORB-SLAM2 and SPSVO, and the true trajectories. (a) ORB-SLAM2 and (b) SPSVO.*

4.1. Dataset

The "colon_reconstruction_dataset" contains 16 stereo colonoscope sequences (named as Case 0–Case 15, there are total of 17,362 frames) with corresponding depth and ego-motion ground truth.

4.2. Implementation details

The proposed SPSVO algorithm runs in a C++ environment on a laptop with an i7-10750H CPU and NVIDIA GTX1650Ti. SPSVO uses Nvidia TensorRT Toolkit to accelerate feature extraction and matching networks and uses the LM algorithm of the g2o library for nonlinear squared optimization. OpenCV and the Ceres library are applied to implement computer vision functions and statistical estimation, respectively.

4.3. Results on the colon reconstruction dataset

The performances of the ORB-SLAM2 and SPSVO were tested with the "colon_reconstruction_dataset"; however, the ORB-SLAM2 could only successfully obtain the endoscope trajectories of "Case 0," and results are shown in Figs. 8, 9 and Table 2. The data sequences of "Case 0" contain 4751 frames of images for each left and right camera and have slower camera motion speed and smaller translation and rotation amplitude compared to "Case 1" to "Case 10." It can be observed from Fig. 8 that ORB-SLAM2 has larger drift error compared to the proposed SPSVO method.

Comparisons between estimated trajectories and true trajectories of the endoscope are shown in Fig. 10. Colored solid lines represent estimated trajectories of the SPSVO. gray dotted lines represent real motion trajectories of the endoscope corresponding to the "colon_reconstruction_dataset" [38]. Statistics for SPSVO are shown in Table 3. The average measurement error of SPSVO for the

https://doi.org/10.1017/S026357472300125X Published online by Cambridge University Press

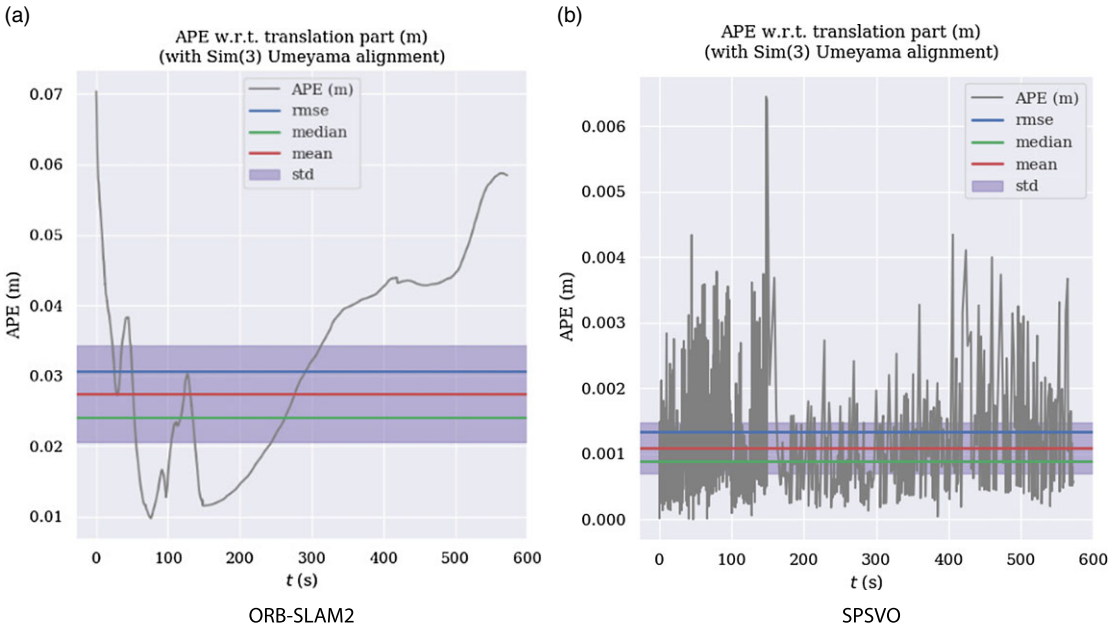

Figure 9. *Variation of the absolute pose error (APE) between the estimated trajectories of ORB-SLAM2 and SPSVO and the true trajectories. (a) ORB-SLAM2 and (b) SPSVO.*

Table II. *Statistical error analysis of estimated trajectories of ORB-SLAM2 and SPSVO on Case0 sequence (unit: mm).*

	Indicators									
	Max		**Mean**		**Min**		**RMSE**		**STD**	
	ORB-SLAM2	**SPSVO**	**ORB-SLAM2**	**SPSVO**	**ORB-SLAM2**	**SPSVO**	**ORB-SLAM2**	**SPSVO**	**ORB-SLAM2**	**SPSVO**
Case 0	70.379	6.447	27.455	1.089	9.81	0.003	30.643	1.332	13.610	0.767

10 cases is between 0.058 and 0.740 mm, with the RMSE between 0.278 and 0.690 mm. This indicates that the proposed SPSVO method can accurately track the true trajectory of the endoscope. Figure 11 shows the variation of the absolute pose error between estimated and true trajectories with respect to time. It can be observed that the proposed SPSVO method has high accuracy and reliability for endoscope trajectory estimation. ORB-SLAM2 cannot extract enough feature points to initialize the viscera scene map, resulting in a loss of feature tracking and failure to construct endoscopic trajectories. Therefore, quantitative results for ORB-SLAM2 on Case1-Case10 are not presented.

4.4. Computational cost

Computational time of the SPSVO and ORB-SLAM2 on Case 0 sequence for one frame of the "colon_reconstruction_dataset" [38] is shown in Table 4. For fair comparison, 1000 points were extracted in this experiment, loop closure, relocalization, and visualization parts were disabled. Keypoint

https://doi.org/10.1017/S026357472300125X Published online by Cambridge University Press

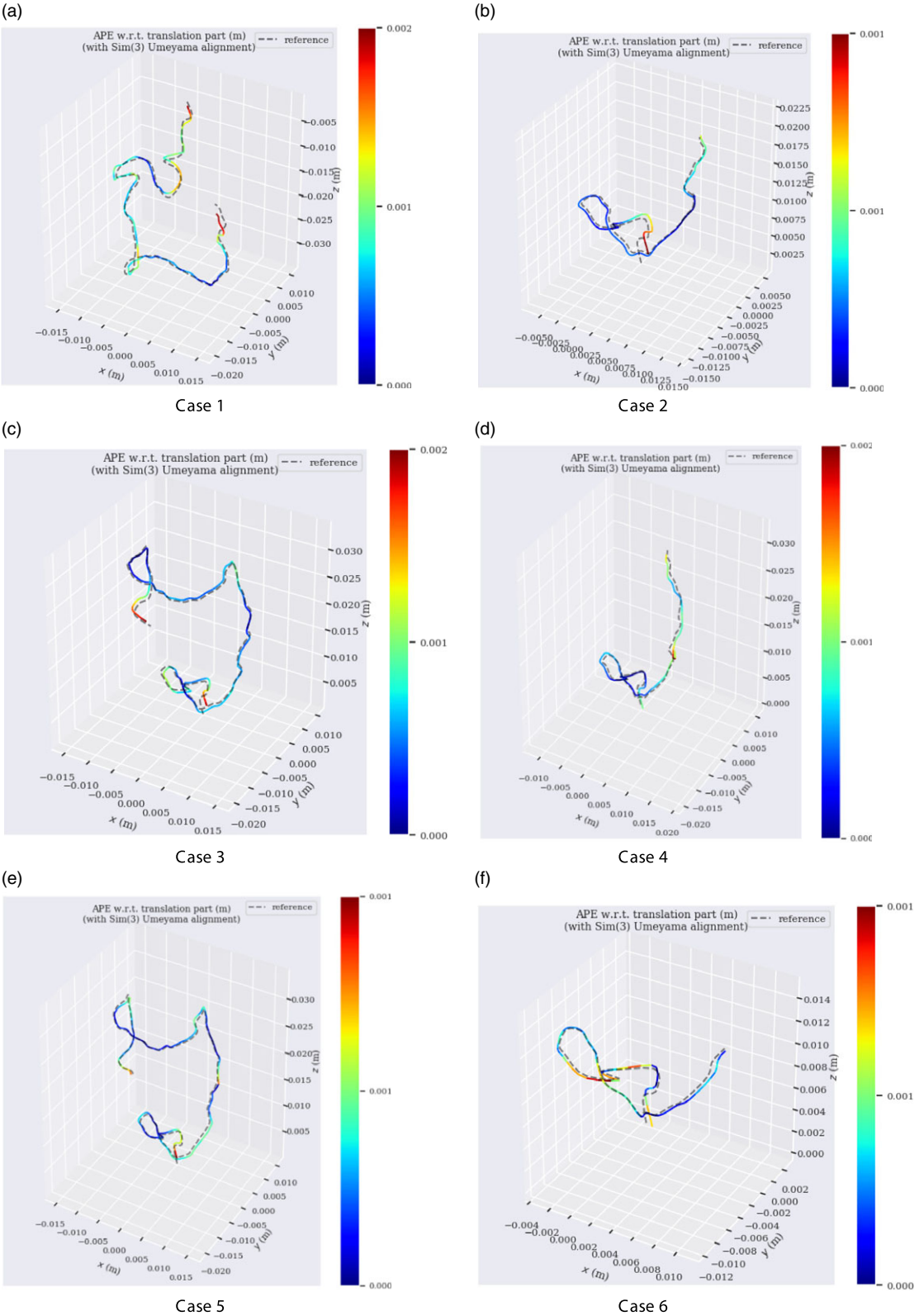

Figure 10. *Comparison between SPSVO-estimated and true trajectories of the endoscope. (a) Case 1, (b) Case 2, (c) Case 3, (d) Case 4, (e) Case 5, (f) Case 6, (g) Case 7, (h) Case 8, (i) Case 9, and (j) Case 10.*

https://doi.org/10.1017/S026357472300125X Published online by Cambridge University Press

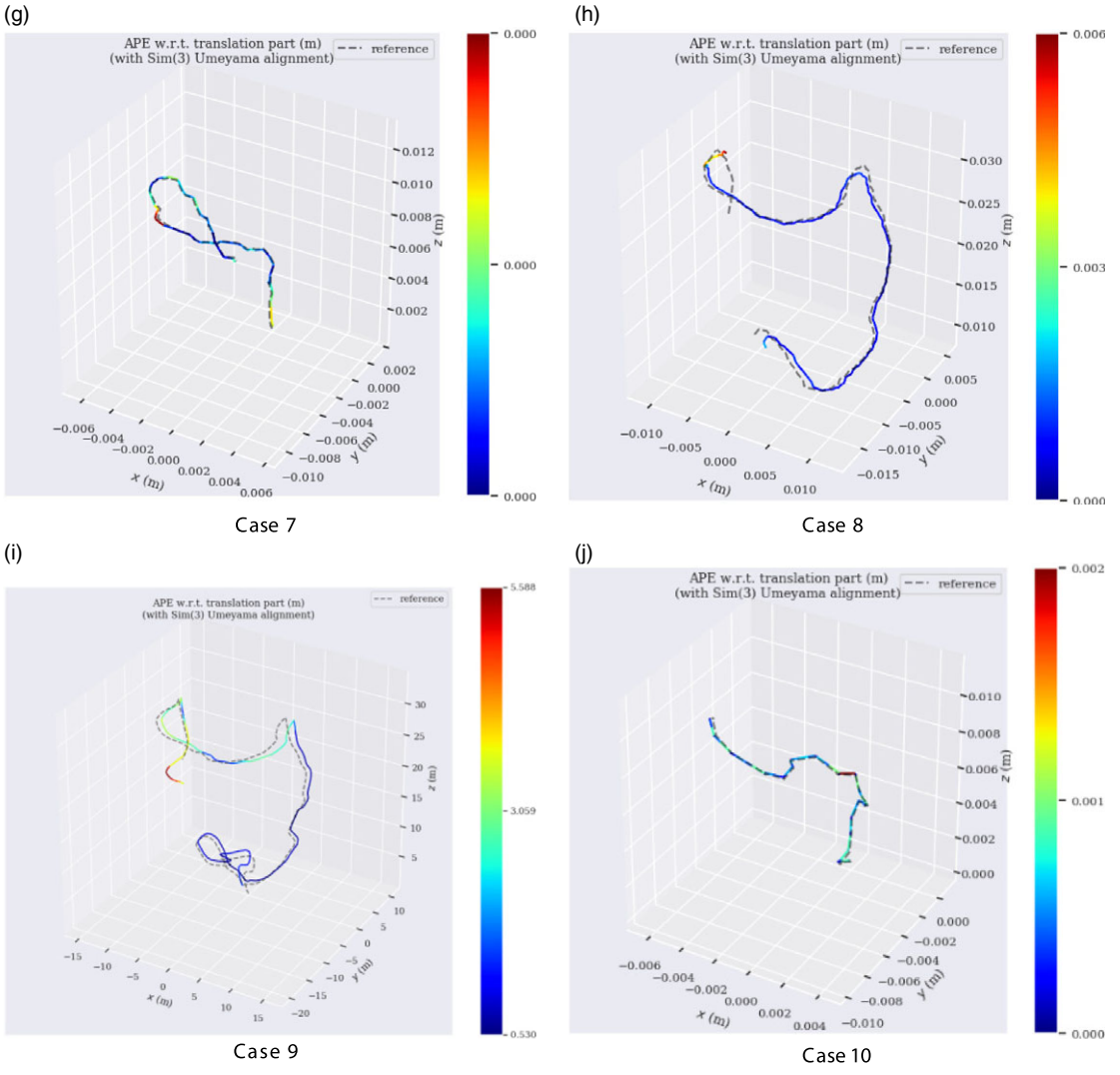

Figure 10. *continue.*

detection takes 25 ms for keypoint extraction of one stereo image. 29 ms are required for stereo matching and feature tracking between frames. Pose estimation is fast and only costs 8ms for one image. Therefore, SPSVO can operate at 14 fps; this speed can be further boosted by parallel implementation. It can be observed that the proposed SPSVO method has faster processing speed compared to ORB-SLAM2.

5. Conclusions

An important goal in VSLAM for medical applications is accurate estimation of endoscopic pose to better assist surgeons in locating and diagnosing lesions. Extreme illumination variations and weak texture of endoscopy images result in difficulties for accurate estimation of camera motion and scene reconstruction. This paper proposed a novel self-supervised Surgical Perception Stereo Visual Odometer

https://doi.org/10.1017/S026357472300125X Published online by Cambridge University Press

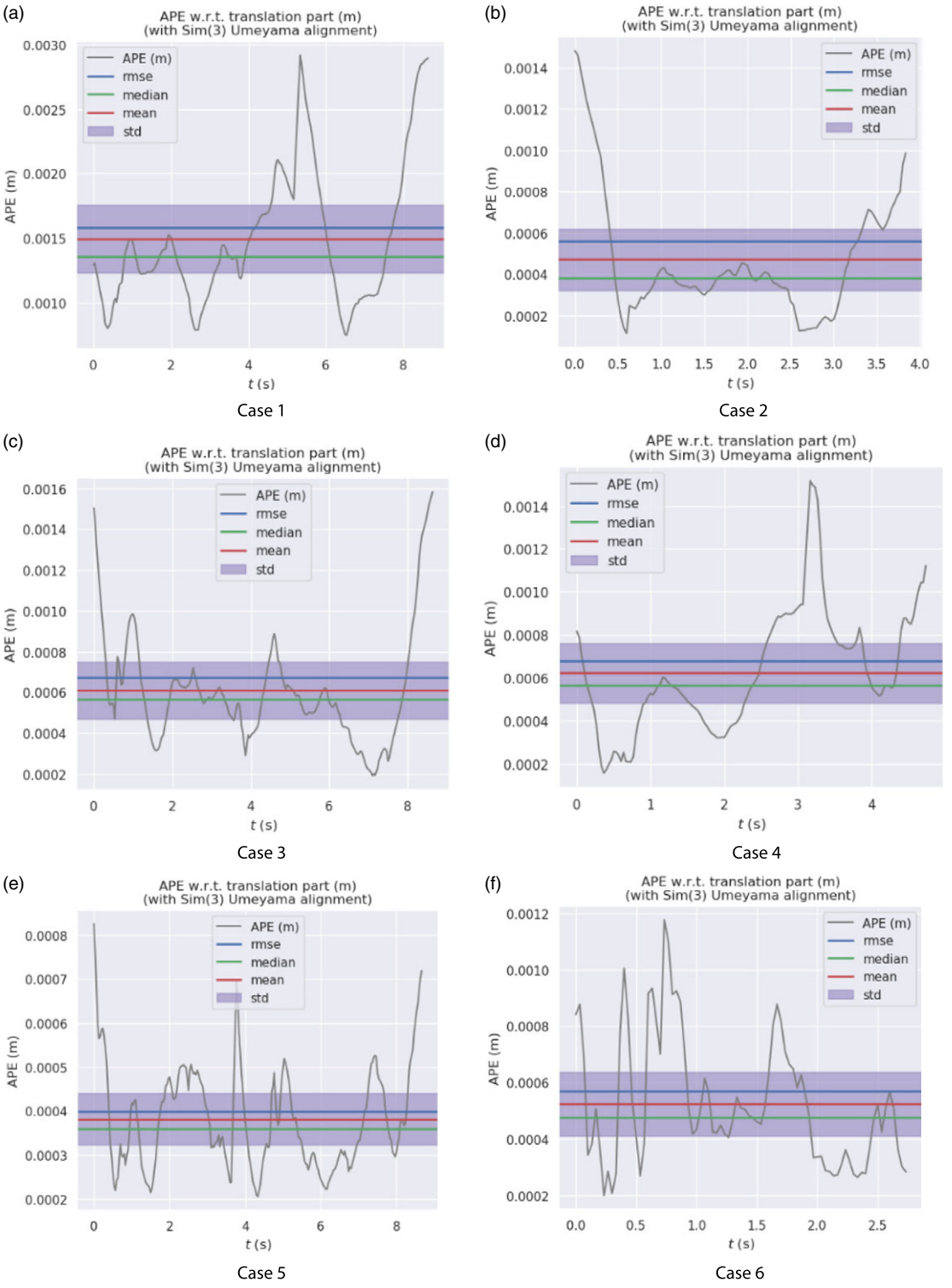

Figure 11. *Variation of the absolute pose error (APE) between SPSVO-estimated and true trajectories. (a) Case 1, (b) Case 2, (c) Case 3, (d) Case 4, (e) Case 5, (f) Case 6, (g) Case 7, (h) Case 8, (i) Case 9, and (j) Case 10.*

https://doi.org/10.1017/S026357472300125X Published online by Cambridge University Press

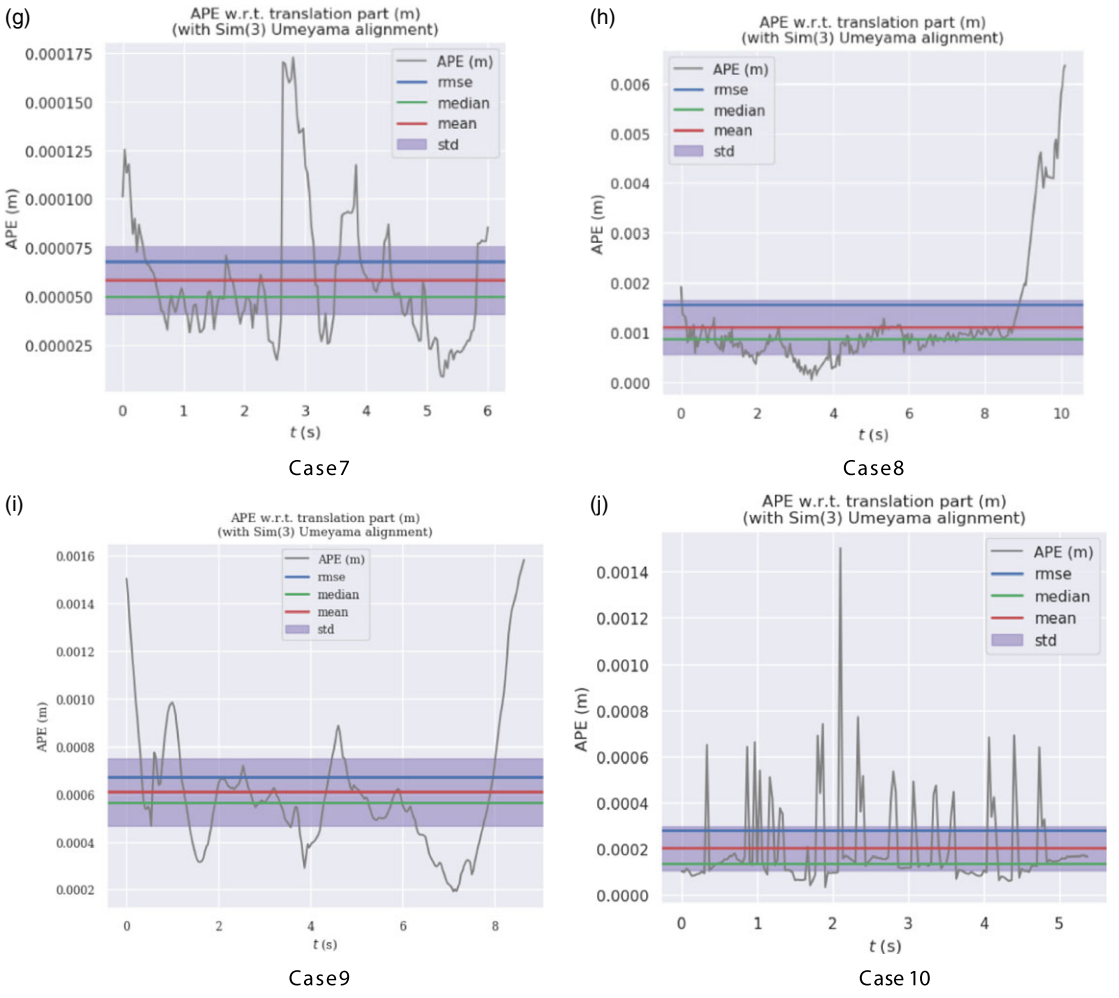

Figure 11. *continue.*

(SPSVO) framework for real-time endoscopic pose estimation and viscera surface map construction. The proposed SPSVO method reduced adverse effects of local illumination variability and specular reflections by using a self-supervised learning (SSL) approach for feature extraction and matching, as well as image illumination pre-processing. In the proposed SPSVO, keyframe selection strategies and the Nvidia TensorRT Toolkit were applied to accelerate computation speed for real-time lesion localization and surgical navigation. Comparison between estimated and the ground truth trajectories of the endoscope were obtained from the colon_reconstruction_dataset. Through experimental tests, the following conclusions are made:

1. The proposed SPSVO system achieves superior performance in variable illumination environments and can track key points in human enterococci with intraperitoneal cavities. Simulation results show that SPSVO has average tracking error of 0.058–0.704 mm with respect to true camera trajectories in the given dataset. Comparison with existing methods also indicates that the proposed method outperforms ORB-SLAM2.

https://doi.org/10.1017/S026357472300125X Published online by Cambridge University Press

Table III. *Statistical error analysis of SPSVO-estimated trajectories (unit: mm). RMSE is the Root Mean Square Error, STD stands for the Standard Deviation. SSE refers to the Sum of Squared Errors.*

	Indicators					
Sequence	**Max**	**Mean**	**Min**	**RMSE**	**SSE**	**STD**
Case 1	1.729	0.704	0.038	0.680	0.153	0.347
Case 2	1.482	0.470	0.111	0.558	0.035	0.300
Case 3	3.047	0.363	0.025	0.493	0.060	0.333
Case 4	1.516	0.621	0.155	0.630	0.065	0.276
Case 5	0.826	0.381	0.205	0.398	0.040	0.117
Case 6	1.178	0.0524	0.199	0.570	0.027	0.224
Case 7	0.173	0.058	0.009	0.068	0.001	0.034
Case 8	3.699	0.440	0.018	0.058	0.076	0.375
Case 9	3.588	1.743	0.529	0.704	0.113	0.346
Case 10	1.504	0.201	0.032	0.278	0.012	0.192

Table IV. *Computational cost of ORB-SLAM2 and SPSVO on Case 0.*

Time	**Keypoint detection**	**Feature tracking**	**Pose estimation**	**Total**
ORB_SLAM2	33 ms	37 ms	7 ms	87 ms
SPSVO	25 ms	29 ms	8 ms	71 ms

2. The proposed SPSVO system combines advantages of traditional optimization and learning-based methods and demonstrates an operating speed of 14 frames per second on a normal computer. This is adequate for real-time navigation in surgical procedures.
3. The proposed method can effectively eliminate effects of irregular illumination and specular reflections and can accurately estimate the position of the endoscope.

Acknowledgments. I would first like to thank Dr Qimin Li and Dr Yang Luo, whose expertise was invaluable in formulating the research questions and methodology. Your insightful feedback pushed me to sharpen my thinking and brought my work to a higher level.

Author contribution. Junjie Zhao: Conceptualization, Methodology, Software. Yang Luo: Data curation, Writing – Original draft preparation. Qimin Li: Supervision. Natalie Baddour: Writing – Reviewing and Editing. Md Sulayman Hossen: Writing-Reviewing and Editing.

Financial support. This research is funded by the Open Fund of Guangdong Provincial Key Laboratory of Precision Gear Digital Manufacturing Equipment Technology Enterprises (Grant No. 2021B1212050012-04), with contributions from Zhongshan MLTOR Numerical Control Technology Co., LTD and South China University of Technology, as well as the Innovation Group Science Fund of Chongqing Natural Science Foundation (No. cstc2019jcyj-cxttX0003).

Competing interests. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Ethical approval. Not applicable.

Data availability statement. The datasets "colon_reconstruction_dataset" for this study can be found at the https://github.com/zsustc/colon_reconstruction_dataset.

https://doi.org/10.1017/S026357472300125X Published online by Cambridge University Press

References

[1] S. Bernhardt, S. A. Nicolau, L. Soler and C. Doignon, " The status of augmented reality in laparoscopic surgery as of 2016," *Med. Image Anal.* **37**, 66–90 (2017).

[2] S. Shao, Z. Pei, W. Chen, W. Zhu, X. Wu, D. Sun and B. Zhang, " Self-supervised monocular depth and ego-motion estimation in endoscopy: Appearance flow to the rescue," *Med. Image Anal.* **77**, 102338 (2022).

[3] M. Feuerstein. *Augmented Reality in Laparoscopic Surgery* (Vdm Verlag Dr.mller Aktiengesellschaft & Co.kg, 2007).

[4] P. K. Lim, G. S. Stephenson, T. W. Keown, C. Byrne, C. C. Lin, G. S. Marecek and J. A. Scolaro, " Use of 3D printed models in resident education for the classification of acetabulum fractures," *J. Surg. Educ.* **75**(6), 1679–1684 (2018).

[5] Z. Zhang, Y. Xie, F Xing, M. McGough, L. Yang, "MDNet: A semantically and visually interpretable medical image diagnosis network," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017) pp. 3549–3557.

[6] C. M. Low, J. M. Morris, J. S. Matsumoto, J. K. Stokken, E. K. O'Brien and G. Choby, " Use of 3D-printed and 2D-illustrated international frontal sinus anatomy classification anatomic models for resident education," *Otolaryngol. Head Neck Surg.* **161**(4), 705–713 (2019).

[7] A. Afifi, C. Takada, Y. Yoshimura and T. Nakaguchi, " Real-time expanded field-of-view for minimally invasive surgery using multi-camera visual simultaneous localization and mapping," *Sensors* **21**(6), 2106 (2021).

[8] F. Tatar, J. R. Mollinger, R. C. Den Dulk, W. A. van Duyl, J. F. L. Goosen and A. Bossche, "Ultrasonic Sensor System for Measuring Position and Orientation of Laproscopic Instruments in Minimal Invasive Surgery," *2nd Annual International IEEE-EMBS Special Topic Conference on Microtechnologies in Medicine and Biology. Proceedings (Cat. No. 02EX578)*, (2002) pp. 301–304.

[9] P. Lamata, T. Morvan, M. Reimers, E. Samset and J. Declerck, "Addressing Shading-based Laparoscopic Registration," *World Congress on Medical Physics and Biomedical Engineering, September 7-12, 2009, Munich, Germany: Vol. 25/6 Surgery, Nimimal Invasive Interventions, Endoscopy and Image Guided Therapy*, (2009) pp. 189–192.

[10] C.-H. Wu, Y.-N. Sun and C.-C. Chang, "Three-dimensional modeling from endoscopic video using geometric co Qax' nstraints via feature positioning," *IEEE Trans. Biomed. Eng.* **54**(7), 1199–1211 (2007).

[11] S. Seshamani, W. Lau and G. Hager. Real-time endoscopic mosaicking. **In:** *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2006: 9th International Conference, , Copenhagen, Denmark, October 1-6, 2006. Proceedings, Part I 9*, (2006) pp. 355–363,

[12] T. Thormahlen, H. Broszio and P. N. Meier, "Three-dimensional Endoscopy," *Falk Symposium*, (2002), 2002-01.

[13] D. Koppel, C.-I. Chen, Y.-F. Wang, H. Lee, J. Gu, A. Poirson and R. Wolters, "Toward Automated Model Building from Video in Computer-assisted Diagnoses in Colonoscopy," *Medical Imaging 2007: Visualization and Image-Guided Procedures*, (2007) pp. 567–575.

[14] D. Mirota, H. Wang, R. H. Taylor, M. Ishii and G. D. Hager, "Toward Video-based Navigation for Endoscopic Endonasal Skull Base Surgery," *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2009: 12th International Conference, London, UK, September 20-24, 2009, Proceedings, Part I 12*, (2009) pp. 91–99.

[15] A. Kaufman and J. Wang, "3D surface reconstruction from endoscopic videos," *Math. Visual.*, 61–74 (2008).

[16] D. Hong, W. Tavanapong, J. Wong, J. Oh, P.-C. De Groen, "3D reconstruction of colon segments from colonoscopy images, "*2009 Ninth IEEE International Conference on Bioinformatics and BioEngineering*, (2009) pp. 53–60.

[17] J. Y. Jang, H.-S. Han, Y.-S. Yoon, Y. Jai and Y. Choi, "Retrospective comparison of outcomes of laparoscopic and open surgery for T2 gallbladder cancer - thirteen-year experience," *Surg. Oncol.* **29**, 29–147 (2019).

[18] H. Wu, J. Zhao, K. Xu, Y. Zhang, R. Xu, A. Wang and Y. Iwahori, " Semantic SLAM based on deep learning in endocavity environment," *Symmetry-Basel* **14**(3), 614 (2022).

[19] C. Xie, T. Yao, J. Wang and Q. Liu, " Endoscope localization and gastrointestinal feature map construction based on monocular SLAM technology," *J. Infect Public Health* **13**(9), 1314–1321 (2020).

[20] P. Mountney, D. Stoyanov, A. Davison, and G.-Z. Yang, "Simultaneous Stereoscope Localization and Soft-tissue Mapping for Minimal Invasive Surgery, " *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2006: 9th International Conference, Copenhagen, Denmark, October 1-6, 2006. Proceedings, Part I 9*, (2006) pp. 347–354.

[21] P. Mountney and G.-Z. Yang, "Motion Compensated SLAM for Image Guided Surgery," *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2010: 13th International Conference, Beijing, China, September 20-24, 2010, Proceedings, Part II 13*, (2010) pp. 496–504.

[22] G. Klein, D. Murray, "Parallel Tracking and Mapping for Small AR Workspaces," *2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*, (2007) pp. 225–234.

[23] B. Lin, A. Johnson, X. Qian, J. Sanchez and Y. Sun, "Simultaneous Tracking, 3D Reconstruction and Deforming Point Detection for Stereoscope Guided Surgery," *Augmented Reality Environments for Medical Imaging and Computer-Assisted Interventions: 6th International Workshop, MIAR 2013 and 8th International Workshop, AE-CAI 2013, Held in Conjunction with MICCAI 2013, Nagoya, Japan, September 22, 2013. Proceedings*, (2013) pp. 35–44.

[24] B. Lin, Y. Sun, J. E. Sanchez and X. Qian, " Efficient vessel feature detection for endoscopic image analysis," *IEEE Trans. Biomed. Eng.* **62**(4), 1141–1150 (2014).

[25] R. Mur-Artal, M. Montiel J.M. and J. D. Tardos, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE Trans. Robot.* **31**(5), 1147–1163 (2015).

[26] N. Mahmoud, I. Cirauqui, A. Hostettler, C. Doignon, L. Soler, J. Marescaux and J. M. M. Montiel, "ORBSLAM-based Endoscope Tracking and 3D Reconstruction," *Computer-Assisted and Robotic Endoscopy: Third International Workshop, CARE 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 17, 2016, Revised Selected Papers 3*, (2017) pp. 72–83.

https://doi.org/10.1017/S026357472300125X Published online by Cambridge University Press

[27] N. Mahmoud, T. Collins, A. Hostettler, L. Soler, C. Doignon and J. M. M Montiel, "Live tracking and dense reconstruction for handheld monocular endoscopy," *IEEE Trans. Med. Imag*. **38**(1), 79–89 (2019).

[28] D. Recasens, J. Lamarca, J. M. Facil, J. M. M. Montiel and J. Civera, " Endo-depth-and-motion: Localization and reconstruction in endoscopic videos using depth networks and photometric constraints," *IEEE Robot. Automat. Lett*. **6**(4), 7225–7232 (2021).

[29] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An Efficient Alternative to SIFT or SURF," *IEEE International Conference on Computer Vision, ICCV 2011*, (2011).

[30] C. Campos, R. Elvira, J. Rodriguez, M. Montiel and J. D. Tardós, "ORB-SLAM3: An accurate open-source library for visual, visual-inertial, and multimap SLAM," *IEEE Trans. Robot. Publ. IEEE Robot. Automat. Soc* **37**(6), 1874–1890 (2021).

[31] D. Detone, T. Malisiewicz and A. Rabinovich, "SuperPoint: Self-supervised interest point detection and description (2017), arXiv: 1712.07629.

[32] P.-L. Chang, D. Stoyanov, A. J. Davison, and P. E. Edwards, "Real-time Dense Stereo Reconstruction Using Convex Optimisation with a Cost-volume for Image-guided Robotic Surgery," *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part I 16*, (2013) pp. 42–49.

[33] B. Lin, Y. Sun, J. Sanchez and X. Qian "Vesselness based Feature Extraction for Endoscopic Image Analysis, "*2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, (2014) pp. 1295–1298.

[34] J. Engel, V. Koltun and D. Cremers, "Direct sparse odometry," (2016): arXiv e-prints.

[35] J. Zubizarreta, I. Aguinaga and J. Montiel, "Direct sparse mapping," (2019): arXiv:1904.06577.

[36] C. Forster, M. Pizzoli and D. Scaramuzza, "SVO: Fast Semi-direct Monocular Visual Odometry," *IEEE International Conference on Robotics & Automation*, (2014).

[37] P. E. Sarlin, D. Detone, T. Malisiewicz, and A. Rabinovich, "SuperGlue: Learning Feature Matching With Graph Neural Networks," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2020).

[38] S. Zhang, L. Zhao, S. Huang and Q. Hao, "A template-based 3D reconstruction of colon structures and textures from stereo colonoscopic images," *IEEE Trans. Med. Robot. Bionics* **3**(1), 85–95 (2021).

[39] R. Mur-Artal and J. Tardós, "ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras," *IEEE Trans. Robot* **33**(5), 1255–1262 (2017).

[40] https://github.com/UZ-SLAMLab/ORB_SLAM2_Endoscopy.

[41] H. Durrant-Whyte and T. Bailey, "Simultaneous localization and mapping: Part I," *IEEE Robot. Automat. Magaz.* **13**(2), 99–110 (2006).

[42] A. Bartoli, J. Montiel, J. Lamarca, and Q. Hao, DefSLAM: Tracking and Mapping of Deforming Scenes from Monocular Sequences, (2019): arXiv: 1908.08918.

[43] H. Gong, L. Chen, C. Li, J. Zeng, X. Tao and Y. Wang, "Online tracking and relocation based on a new rotation-invariant haar-like statistical descriptor in endoscopic examination," *IEEE Access* **8**, 101867–101883 (2020).

[44] J. Song, J. Wang, L. Zhao, S. Huang and G. Dissanayake, "Mis-slam: Real-time large-scale dense deformable slam system in minimal invasive surgery based on heterogeneous computing," *IEEE Robot. Automat. Lett.* **3**(4), 4068–4075 (2018).

[45] G. Wei, G. Feng, H. Li, T. Chen, W. Shi and Z. Jiang, "A Novel SLAM Method for Laparoscopic Scene Reconstruction with Feature Patch Tracking," *2020 International Conference on Virtual Reality and Visualization (ICVRV)*, (2020) pp. 287–291.

[46] J. Song, Q. Zhu, J. Lin, and M. Ghaffari, "BDIS: Bayesian dense inverse searching method for real-time stereo surgical image matching," *IEEE Trans. Robot.*, **39**(2), 1388–1406 (2022).

[47] G. Wei, H. Yang, W. Shi, Z. Jiang, T. Chen and Y. Wang, "Laparoscopic Scene Reconstruction based on Multiscale Feature Patch Tracking Method," *International Conference on Electronic Information Engineering and Computer Science (EIECS)*, (2021) pp. 588–592.

[48] R. Yadav and R. Kala, "Fusion of visual odometry and place recognition for slam in extreme conditions," *Appl. Intell.* **52**(10), 11928–11947 (2022).

[49] S. Bruno H.M. and E. L. Colombini, "LIFT-SLAM: A deep-learning feature-based monocular visual SLAM method," *Neurocomputing* **455**, 97–110 (2021).

[50] A. Schmidt and S. E. Salcudean, "Real-time rotated convolutional descriptor for surgical environments," *Medical Image Computing and Computer Assisted Intervention-MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV 24*, (2021) pp. 279–289.

[51] K. Xu, Y. Hao, C. Wang, and L. Xie, "AirVO: An illumination-robust point-line visual odometry, (2022): arXiv preprint arXiv: 2212.07595.

[52] D. Li, X. Shi, Q. Long, S. Liu, W. Yang, F. Wang and F. Qiao, "DXSLAM: A Robust and Efficient Visual SLAM System with Deep Features," *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, (2020) pp. 4958–4965.

[53] X. Liu, Y. Zheng, B. Killeen, M. Ishii, G. D. Hager, R. H. Taylor and M. Unberath, "Extremely Dense Point Correspondences using a Learned Feature Descriptor," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2020) pp. 4847–4856.

[54] X. Liu, Z. Li, M. Ishii, G. D. Hager, R. H. Taylor and M. Unberath, "Sage: Slam with Appearance and Geometry Prior for Endoscopy," *2022 International Conference on Robotics and Automation (ICRA)*, (2022) pp. 5587–5593.

[55] M. Muja and D. G. Lowe, "Fast approximate nearest neighbors with automatic algorithm configuration," *Proceedings of the Fourth International Conference on Computer Vision Theory and Applications*, Lisboa, Portugal (February 5-8, 2009).

[56] O. L. Barbed, F. Chadebecq, J. Morlana, J. M. Montiel and A. C. Murillo, " SuperPoint Features in Endoscopy," *MICCAI Workshop on Imaging Systems for GI Endoscopy*, (2022) pp. 45–55.

https://doi.org/10.1017/S026357472300125X Published online by Cambridge University Press

[57] P.-E. Sarlin, C. Cadena, R. Siegwart and M. Dymczyk, "From Coarse to Fine: Robust Hierarchical Localization at Large Scale," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2019) pp. 12716–12725.
[58] L. Oliva Maza, F. Steidle, J. Klodmann, K. Strobl and R. Triebel, "An ORB-SLAM3-based approach for surgical navigation in ureteroscopy," *Comput. Methods Biomech. Biomed. Eng. Imag. Visual.*, **11**(4), 1005–1011 (2022).
[59] K. Zuiderveld, "Contrast limited adaptive histogram equalization," *Graphics Gems*, 474–485 (1994).
[60] H. Jang, S. Yoon and A. Kim, "Multi-session Underwater Pose-graph Slam using Inter-session Opti-acoustic Two-view Factor," *2021 IEEE International Conference on Robotics and Automation (ICRA)*, (2021) pp. 11668–11674.
[61] S. Rao, "SuperVO: A Monocular Visual Odometry based on Learned Feature Matching with GNN," *2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, 2021) pp. 18–26.
[62] Y. Su and L. Yu, "A dense RGB-D SLAM algorithm based on convolutional neural network of multi-layer image invariant feature," *Meas. Sci. Technol.* **33**(2), 025402 (2021).
[63] H. Strasdat, J. Montiel, A. J. Davison, "Real-time Monocular SLAM: Why Filter?," *IEEE International Conference on Robotics and Automation*, (2010) pp. 2657–2664.
[64] R. Szeliski. *Computer Vision: Algorithms and Applications* (Springer Nature, 2022).

Cite this article: J. Zhao, Y. Luo, Q. Li, N. Baddour and M. S. Hossen (2023). "SPSVO: a self-supervised surgical perception stereo visual odometer for endoscopy", Robotica **41**, 3724–3745. https://doi.org/10.1017/S026357472300125X