International Journal of Intelligent Robotics and Applications (2024) 8:866–877
https://doi.org/10.1007/s41315-024-00390-7

REGULAR PAPER

Binocular-based dense 3D reconstruction for robotic assisted minimally invasive laparoscopic surgery

Xin Sui[1] · Yang Zhang[1,2] · Xingwei Zhao[1] · Bo Tao[1]

Received: 16 May 2024 / Accepted: 20 September 2024 / Published online: 30 September 2024

Abstract

Dense 3D reconstruction of the abdominal environment for Minimally Invasive Surgery (MIS) is important for tasks in Computer Assisted Surgery (CAS), including the alignment with Computed Tomography (CT) and Magnetic Resonance Imaging (MRI), the autonomous navigation of surgical robots, and the application of augmented reality (AR). In this paper, we investigate the binocular laparoscopy-based stereo vision technology, and ultimately achieve fast and dense 3D reconstruction of the preoperative abdominal environment and intraoperative lesion localization based on visual guidance. We introduce binocular constraints and data looping combined to improve the hand–eye calibration algorithm based on binocular laparoscopy. As it is challenging to obtain the depth truth value from medical image data, we employ a binocular unsupervised learning algorithm based on the Parallax Attention Mechanism (PAM) for depth estimation, while a coarse-to-fine pyramid optimization method is used to minimize the photometric error to obtain the laparoscopic trajectory and reconstruct the abdominal environment by parallel processing. In order to confirm the effectiveness of the algorithm, we build a binocular laparoscope-based robot platform and conduct experiments on an abdominal phantom, and the results demonstrate that the simultaneous localization and mapping (SLAM) absolute pose error (APE) of our proposed method outperforms that of some other methods, and it can achieve precise intraoperative lesion localization based on visual guidance.

Keywords Binocular laparoscope · Abdominal environment · Hand–eye calibration · Depth estimation · Laparoscopic SLAM

1 Introduction

With the advantages of less trauma, low germ infection rate, fewer complications, less pain for patients, and faster recovery, MIS has been widely used in clinical practice over the past decade, promoting landmark advances in medicine. However, in comparison with traditional open surgery, there are several limitations, including restricted field of view, imprecise laparoscopic positioning, and a lack of information about the surrounding environment. The conventional monocular laparoscopy employed in MIS is constrained by its two-dimensional imaging capabilities, which leads to insufficient estimation of surgical instrument advancement distance and depth of the body cavity, as well as the 3D representation of the lesions, blood vessels, surrounding organs and tissues within the abdominal cavity. This significantly complicates surgical.

Furthermore, laparoscopes are unable to observe information beneath the surface of organs. Following the use of SLAM to reconstruct the abdominal environment in great detail, it is able to align and fuse with CT or MRI images from preoperative diagnosis, which allows the anatomical structure of the patient to be shown intraoperatively. Additionally, it can also be integrated with AR to superimpose supplementary information within the 3D scene, including annotations of target lesions, contour of organs, and

✉ Xingwei Zhao
zhaoxingwei@hust.edu.cn

Xin Sui
m202370692@hust.edu.cn

Yang Zhang
yangzhang@hust.edu.cn

Bo Tao
taobo@hust.edu.cn

[1] State Key Laboratory of Intelligent Manufacturing Equipment and Technology, Huazhong University of Science and Technology, Wuhan 430074, China

[2] Wuhan United Imaging Healthcare Surgical Technology Company Ltd., Wuhan 430074, China

tumor measurements, and also facilitates the expansion of the current laparoscopic field of view and promotes the development of autonomy for future laparoscopic robotic surgery.

In order to enhance laparoscopic surgical techniques, Mahmoud et al. (2016) enhance ORB-SLAM (Mur-Artal et al. 2015) using a monocular laparoscope by enlarging the search region for image keypoint matching, and in vivo pig experiment demonstrates that laparoscopic tracking remains robust even in the presence of organ deformation and partial instrument occlusion. However, this study uses a monocular laparoscope, which is unable to obtain depth-scale data and thus incapable of reconstructing accurate and usable environmental information. Song et al. (2018a, b) propose a real-time large-scale dense deformation SLAM system based on heterogeneous computing. Their enhanced ORB-SLAM runs on the CPU and transfers the ORB features and global pose to the GPU, thus enabling the generation and rendering of dynamic 3D shapes in real time for an autonomous surgical robot. However, the computational complexity and memory usage increase significantly with the growth of the model, and the depth estimation uses the ELAS algorithm, which exhibits a slight reduction in accuracy in medical scenarios. Qiu et al. (2020) introduce a CT based SLAM registration for laparoscopic navigation in oral surgery, where they utilize semi-dense contours with preoperative CT data for 3D point cloud registration, and the root mean square error (rmse) of the mapping is kept at 1mm. However, this study incorporates information other than visual data, increasing the time and cost of reconstruction.

In contrast to previous techniques, our study focuses on acquiring visual information exclusively by binocular laparoscope, without any other sensors, and the binocular laparoscope can obtain true depth map based on the principle of binocular disparity. We employ the Zhang's calibration method (Zhang 1999) to obtain the intrinsics of binocular laparoscopes and achieve sub-pixel level reprojection accuracy, which meets the requirements for medical. Additionally, we introduce binocular constraints and data loops to enhance the accuracy of the binocular laparoscope-based hand–eye calibration algorithm. The left and right images are rectified by intrinsics, and the depth information is obtained by finetuning the binocular unsupervised network based on the parallax attention mechanism, the pseudo RGBD sequence is obtained by aligning the timestamps, while the trajectory of the binocular laparoscope is obtained by using pyramid optimization method from coarse to fine to minimize the photometric error by parallel processing to achieve fast and dense 3D reconstruction of the abdominal environment. This provides surgeons or surgical robots with a full range of visual information in order to facilitate a more comprehensive understanding of the relationship between lesion and blood vessel positions and a reduction in intraoperative risks. In summary, our contributions of this paper include:

1. We introduce binocular constraints and data loops to improve hand–eye calibration algorithms, which is specifically designed for binocular laparoscope.
2. We introduce a novel SLAM method that generates pseudo-RGBD frames via an unsupervised PAM-based algorithm and uses parallel processing for trajectory tracking with coarse-to-fine pyramid optimization.
3. Our method enables fast and dense reconstruction of the preoperative abdominal environment and high-precision intraoperative localization of lesions.

2 Related work

2.1 Hand–eye calibration

Hand–eye calibration is the process of translating spatial information from complex environments to the robot base coordinate system, and proper hand–eye calibration is critical when sub-pixel perceptual accuracy is required (Enebuse et al. 2021). In robot-assisted surgery (RAS), excessive errors in hand–eye calibration may result in the damage of vital tissues and organs, thereby increasing the risk of surgical complications.

Krittin Pachtrachai et al. (2018) propose a hand–eye calibration method based on adjoint transformation of twist motions, which is solved iteratively by alternately estimating the rotation and translation matrices. Orhan Özgüner et al. (2020) add a Polaris Vicra optical tracking system to the da Vinci Surgical System for intraoperative real-time calibration of the remote-center-of-motion (RCM) based laparoscopic camera and the patient-side manipulator (PSM) arm through translation relations between coordinate systems. Lu et al. (2022) compute hand–eye calibration in a hierarchical manner based on monocular laparoscopy by manipulating PSM in a limited workspace, optimize globally to minimize the proposed aggregating sphere loss, and optimize locally based on beam adjustment model. Zhong et al. (2020) perform hand–eye calibration by fixing a monocular laparoscope and moving the surgical instrument without the use of a calibration target, they use only CAD model of the surgical instrument and a small amount of data to achieve calibration results with low error.

2.2 Depth estimation

Depth estimation algorithm is the foundation for binocular stereo vision technology, which is a crucial component of laparoscopic surgery. Scharstein et al. (2001) have proposed a four-step framework for traditional stereo matching, including matching cost computation, cost aggregation, parallax computation, and parallax optimization. The accuracy and speed of binocular-based parallax estimation in complex environment have been significantly enhanced with the advent of deep learning algorithms, which can be classified into two distinct categories: supervised and unsupervised.

To address the challenges of laparoscopic surgery, Chang et al. (2013) propose a stereoscopic depth estimation algorithm that constructs a 3D cost volume per pixel based on disparity, followed by Huber-L convex optimization. This method efficiently achieves dense reconstruction of smooth surfaces, such as the heart, even in texture-poor regions or regions occluded by surgical instruments, and it preserves depth discontinuity and real time on GPU. Godard et al. (2017) introduce a fully convolutional deep neural network with differentiable training losses, including left–right consistency checks, to enhance the quality of synthesized depth maps by minimizing reprojection errors. Huang et al. (2018) presented a multi-view convolutional neural network that aggregates information from an unordered set of images, which integrates multi-layer feature activations from a pre-trained VGG-19 network on a real dataset, demonstrating superior reconstruction results in low-texture regions. Wang et al. (2020) use Blender to simulate binocular laparoscopy data in gastrointestinal environment. They propose a 23-layer convolutional neural network for real-time parallax map generation, and introduce a scale-invariant loss function to improve depth estimation accuracy with minimal training data.

2.3 Visual SLAM

Robotic SLAM systems are distinguished by their core functional modules, which can be broadly categorized into two forms: laser SLAM and vision SLAM(VSLAM) (Zaffar et al. 2018). These can also be further subdivided into sparse and dense reconstruction. VSLAM includes monocular, stereo, event-based, omnidirectional, and RGBD cameras (Tourani et al. 2022). VSLAM is relatively straightforward to install, utilizes inexpensive sensors, enables dense reconstruction, and is more suitable for medical applications in abdominal environments. In recent years, the use of light model correction and highly robust feature points have become prevalent in the field of VSLAM research, with encouraging results.

ORB-SLAM2 (Mur-Artal et al. 2017) incorporates loop closure, reposition and map reuse to enhance accuracy with bundled adjustment. Additionally, it includes a lightweight positioning mode that utilizes visual odometer for unmapped areas and matches map points to permit zero drift positioning. DSO (Engel et al. 2018) enhance the direct pose estimation model by incorporating affine luminance transformation, photometric calibration and depth optimization, yet lack loop closure detection. Mahjourian et al. (2018) presented a novel unsupervised learning approach for depth and ego motion estimation in monocular video, which is validated on the KITTI dataset and on a dataset of videos captured by calibrated mobile phones in micro-landscapes. Static Fusion (Scona et al. 2018) is an element-based RGB-D SLAM system designed for dynamic environments, but initialization uncertainty arises in the presence of numerous dynamic objects. In this paper, we utilize binocular laparoscopy and separate the depth and SLAM reconstruction processes.

3 Method

3.1 Framework

In this paper, we propose a fast and dense preoperative 3D reconstruction of the abdominal tissue and intraoperative precise lesion localization using only the visual information provided by binocular laparoscope. The algorithmic flow of this paper is shown in Fig. 1, which illustrates that intrinsic calibration and hand–eye calibration of the binocular laparoscope is required only once and a set of left and right images of the scene are acquired to finetune the unsupervised depth estimation network. The video captured by the binocular endoscope is uploaded to the GPU side of the computer in real time through the capture card. The pseudo-RGBD frames are created after acquiring the depth map through the trained depth estimation network, and then sequentially passed to the SLAM algorithm for parallel processing, estimating the trajectory of the binocular endoscope and completing the dense reconstruction of the abdominal surface. Intraoperatively, the pixel points of the lesion are extracted through threshold segmentation and contour extraction, and the spatial position in the base of robotic coordinate system is calculated according to the depth map combined with the hand–eye calibration matrix, and the surgical instrument is visually guided to move to the lesion.

3.2 Binocular-based hand–eye calibration

A series of checkerboard images and the corresponding angular data for each joint angle of the robot are collected. The Perspective-n-Point (PnP) algorithm is then used to solve the transformation matrix of the checkerboard origin to the left laparoscopy. We construct the AX = XB hand–eye calibration equation, and in order to minimize the quantity of data collected and ensure the X iteration solution remained accurate, a data loop processing approach rather than a sequential one is implemented, and where A is the product of any two robot ends to bases transformation matrices, rather than the i + 1th to the ith, B is the product of any two checkerboard origins to left laparoscopies transformation matrices and X is the desired hand–eye calibration matrix.

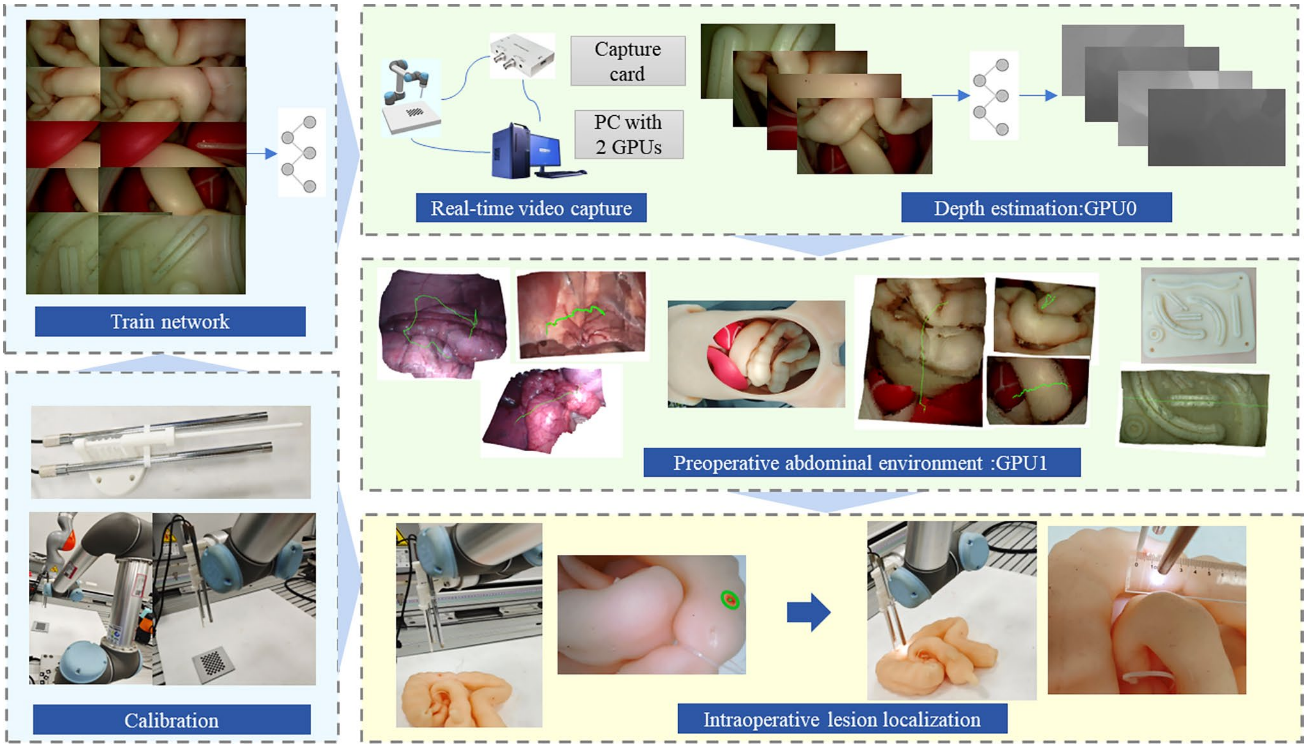
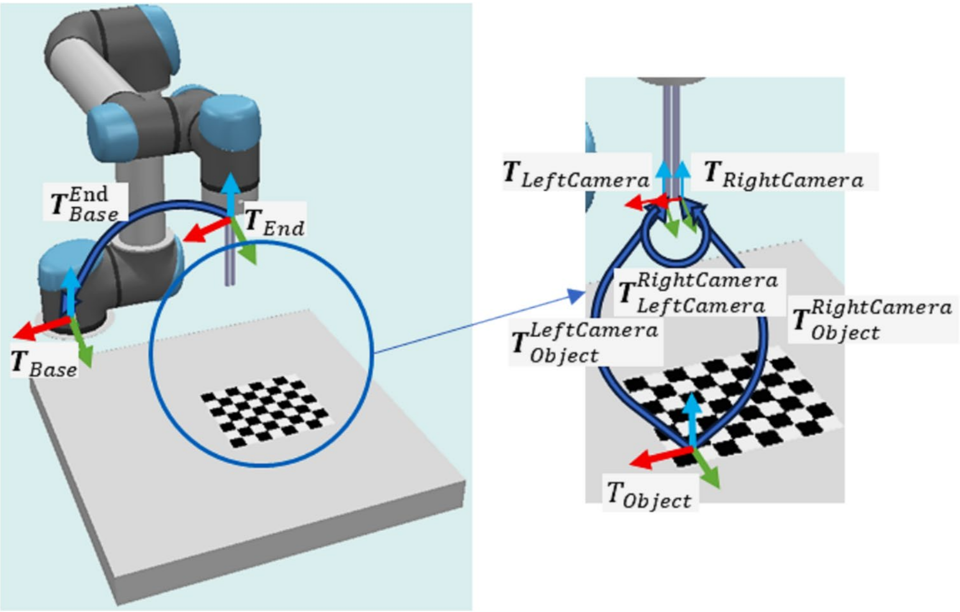


Fig. 1 Pipeline Overview. The process includes a one-time intrinsic and hand–eye calibration of the binocular laparoscope, followed by fine-tuning of the depth estimation network using rectified images. Real-time video is processed into pseudo-RGBD frames for preoperative abdominal environment reconstruction and intraoperative lesion localization

Coordinate system employed in the process of hand–eye calibration is shown in Fig. 2.

A binocular laparoscope is employed in this procedure, and it should be noted that some other calibration methods may result in the loss of information from one of the laparoscopes. The transformation matrix $\mathbf{T}_{LeftCamera}^{RightCamera}$ of the left to the right laparoscope can be derived from the intrinsic calibration and remains constant throughout the calibration procedure. In order to achieve accurate

Fig. 2 All coordinates employed in the process of hand–eye calibration. The symbol $\mathbf{T}_B^A$ denotes the transformation relationship between the B coordinate to the A coordinate, the direction in which the arrow is pointing


T_{Base}^{End}
T_{End}
T_{Base}
$T_{LeftCamera}$
$T_{RightCamera}$
$T_{LeftCamera}^{RightCamera}$
$T_{Object}^{LeftCamera}$
$T_{Object}^{RightCamera}$
T_{Object}

calibration results, we use both the left and right images and incorporate binocular constraints.

$$
\begin{aligned}
&A = T_{Base_i}^{End^{-1}} \cdot T_{Base_j}^{End}\\
&B_{l_1} = T_{Object_i}^{LeftCamera^{-1}} \cdot T_{Object_j}^{LeftCamera}\\
&B_{r_1} = T_{Object_i}^{RightCamera^{-1}} \cdot T_{Object_j}^{RightCamera}\\
&B_{l_2} = \left(T_{LeftCamera}^{RightCamera^{-1}} \cdot T_{Object_i}^{RightCamera}\right)^{-1} \cdot T_{Object_j}^{LeftCamera}\\
&B_{r_2} = \left(T_{LeftCamera}^{RightCamera} \cdot T_{Object_i}^{LeftCamera}\right)^{-1} \cdot T_{Object_j}^{RightCamera}\\
&B_{l_3} = T_{Object_i}^{LeftCamera^{-1}} \cdot \left(T_{LeftCamera}^{RightCamera^{-1}} \cdot T_{Object_j}^{RightCamera}\right)\\
&B_{r_3} = T_{Object_i}^{RightCamera^{-1}} \cdot \left(T_{LeftCamera}^{RightCamera} \cdot T_{Object_j}^{LeftCamera}\right)\\
&B_{l_4} = \left(T_{LeftCamera}^{RightCamera^{-1}} \cdot T_{Object_i}^{RightCamera}\right)^{-1}\\
&\cdot \left(T_{LeftCamera}^{RightCamera^{-1}} \cdot T_{Object_j}^{RightCamera}\right)\\
&B_{r_4} = \left(T_{LeftCamera}^{RightCamera} \cdot T_{Object_i}^{LeftCamera}\right)^{-1}\\
&\cdot \left(T_{LeftCamera}^{RightCamera} \cdot T_{Object_j}^{LeftCamera}\right)
\end{aligned}
\tag{1}
$$

where the range of i is the number of calibrated images captured, while the range of j is the number of calibrated images captured minus 1, and for each i, the loop iterates j times. Finally, the classical Tsai method (Tsai et al. 1989) is employed to solve the hand–eye calibration equation for AX = XB, where subscript l represents the left camera and r represents the right camera, B_{l1},B_{l2},B_{l3},B_{l4} are stacked as B_l matrix, and B_{r1},B_{r2},B_{r3},B_{r4} are stacked as B_r matrix, and the transformation matrices of the left and right camera with respect to the robot base are calculated respectively.

3.3 Unsupervised binocular-based depth estimation

There are few medical open binocular datasets with depth groundtruth, so we use parallax attention for unsupervised stereo correspondence learning network (PASMnet) (Wang et al. 2022) architecture and training procedures, which achieves the state-of-the art performance, and we generalize it to medical images. The PASMnet feeds the rectified left and right images into an hourglass-type feature extraction network respectively, and the resulting feature maps are fed into a cascaded PAM, which uses coarse-to-fine matching cost regression, and the parallax map is output after the hourglass-type parallax refinement module. The general flow of the algorithm for depth estimation is shown in Fig. 3.

PAM employs 1 × 1 convolution to extract the left and right image feature maps $\mathbf{I}_{left}$ and $\mathbf{I}_{right}$ respectively, with dimensions H × W × C. H a batch of matrix multiplication, with each matrix comprising W points, and C represents the feature dimension of each point. The geometrical-aware matrix multiplication and SoftMax function operations are conducted on these two feature maps to encode the feature similarity of any two pixels along the pole line into the PAM $\mathbf{M}_{right\rightarrow left}$ and $\mathbf{M}_{right\rightarrow left}$ respectively, with dimensions H × W × W. The masked pixels are removed according to the point pair matching correlation to obtain the valid mask. Employing PAM instead of the cost volume reduces the amount of computation and memory occupancy. Furthermore, there is no need to set a fixed parallax maximum value. In order to obtain reliable matching relationships, both left–right consistency and cycle consistency are introduced, left–right consistency is calculated as follows, and torch.matmul is the tensor's multiplication method for multiplying two tensor matrices:

$$
\begin{cases}
\mathbf{I}_{left} = torch.matmul(\mathbf{M}_{right\rightarrow left}, \mathbf{I}_{right})\\
\mathbf{I}_{right} = torch.matmul(\mathbf{M}_{left\rightarrow right}, \mathbf{I}_{left})
\end{cases}
\tag{2}
$$

Cycle consistency is calculated as follows:

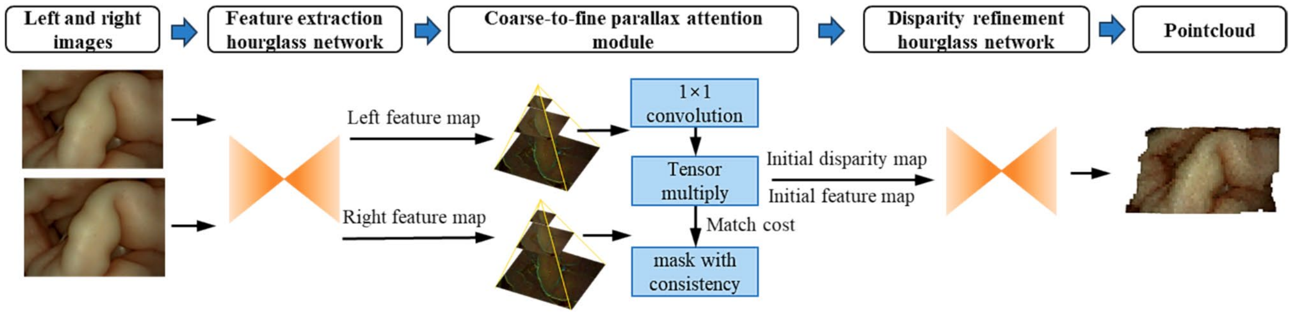


Fig. 3 The process of depth estimation. The net feeds the rectified left and right images into an hourglass-type feature extraction network respectively, and the resulting feature maps are fed into a cascaded PAM, which uses coarse-to-fine matching cost regression, and the disparity map is output after the hourglass-type parallax refinement module, which can be transformed into a 3D pointcloud

$$\left\{\begin{array}{l} \mathbf{I}_{left} = torch.matmul(\mathbf{M}_{left\rightarrow right\rightarrow left}, \mathbf{I}_{left}) \\ \mathbf{I}_{right} = torch.matmul(\mathbf{M}_{right\rightarrow left\rightarrow right}, \mathbf{I}_{right}) \end{array}\right. \tag{3}$$

The loss function of this unsupervised network incorporates three components: photometric loss $\mathcal{L}_p$, smoothness loss $\mathcal{L}_s$ and PAM loss $\mathcal{L}_{PAM}$, and λ_p, λ_s, λ_{PAM} are their respective weights:

$$\begin{aligned} &\mathcal{L}_{total} = \lambda_p \mathcal{L}_p + \lambda_s \mathcal{L}_s \\ &+\lambda_{PAM}\left(0.2\mathcal{L}^s_{PAM-p} + 0.3\mathcal{L}^s_{PAM-s} + 0.5\mathcal{L}^s_{PAM-c}\right) \end{aligned} \tag{4}$$

The photometric loss comprises two components: the mean absolute error and the structural similarity index (SSIM). The edge-aware smoothness loss enhances the local smoothness of parallax. PAM loss introduces three terms of varying scales to regulate the process, thereby ensuring the generation of accurate and consistent stereo correlation, which includes PAM photometric loss $\mathcal{L}^s_{PAM-p}$:

$$\begin{aligned} &\mathcal{L}^s_{PAM-p} = \frac{1}{N^s_{left}} \sum_{p \int V^s_{left}} \\ &(\left\| I^s_{left}(p) - torch.matmul\left(M^s_{right\rightarrow left}, I^s_{right}\right)(p)\right\|_1 \\ &+\frac{1}{N^s_{right}} \sum_{p \int V^s_{right}} \\ &(\left\| I^s_{right}(p) - torch.matmul\left(M^s_{left\rightarrow right}, I^s_{left}\right)(p)\right\|_1 \end{aligned} \tag{5}$$

PAM smoothness loss $\mathcal{L}^s_{PAM-s}$: $\mathcal{L}^s_{PAM-s} = \frac{1}{N^s_{right\rightarrow lef\mathrm{t}}} \sum_{i,j,k}$

$$\begin{aligned} &(||\mathbf{M}^s_{right\rightarrow left}(i.j.k) - \mathbf{M}^s_{right\rightarrow left}(i+1.j.k)||_1 \\ &+|\mathbf{M}^s_{right\rightarrow left}(i.j.k) - \mathbf{M}^s_{right\rightarrow left}(i.j+1.k+1)||_1) \\ &+\frac{1}{N^s_{left\rightarrow \mathrm{right}}} \sum_{i,j,k} \\ &(||\mathbf{M}^s_{left\rightarrow \mathrm{right}}(i.j.k) - \mathbf{M}^s_{left\rightarrow \mathrm{right}}(i+1.j.k)||_1 \\ &+|\mathbf{M}^s_{left\rightarrow \mathrm{right}}(i.j.k) - \mathbf{M}^s_{left\rightarrow \mathrm{right}t}(i.j+1.k+1)||_1) \end{aligned} \tag{6}$$

PAM cycle consistency loss $\mathcal{L}^s_{PAM-c}$:

$$\begin{aligned} &\mathcal{L}^s_{PAM-c} = \frac{1}{N^s_{left}} \sum_{p \int V^s_{left}} (\left\|(M^s_{left\rightarrow right\rightarrow left} - I^s)(p)\right\|_1 \\ &+\frac{1}{N^s_{right}} \sum_{p \int V^s_{right}} (\left\|(M^s_{right\rightarrow left\rightarrow right} - I^s)(p)\right\|_1 \end{aligned} \tag{7}$$

where N^s_{left} and N^s_{right} are the number of valid pixels in V^s_{left} and V^s_{right} respectively, $N^s_{right\rightarrow lef\mathrm{t}}$ and $N^s_{left\rightarrow \mathrm{right}}$ are the number of pixels in $\mathbf{M}^s_{right\rightarrow left}$ and $\mathbf{M}^s_{left\rightarrow \mathrm{right}}$ respectively. Take $\mathbf{M}^s_{right\rightarrow left}(i.j.k)$ as an example, $\mathbf{M}^s_{right\rightarrow left}(i.j.k)$ measures the contribution of (i, k) in $\mathbf{I}^s_{right}$ to (i, j) in $\mathbf{I}^s_{left}$, $\mathbf{I}^s$ is a stack of identity matrices.

3.4 SLAM in the abdominal environment

VSLAM for laparoscopic surgery consists of three parts: image acquisition and feature information extraction for the environment in which the binocular laparoscope itself is located firstly, and then positioning of the binocular laparoscope according to the visual information, reconstruction of the spatial information of the abdominal environment finally. We use the dense direct method, which does not compute feature points and descriptors and is suitable for the abdominal environment that lacks texture. Based on the assumption that the same spatial point has the same grey value under different camera positions, we focus on the gradient of image pixel grey values.

The issue of discontinuities between successive images, which are the result of rapid motion, is addressed by the construction of an image pyramid comprising four levels. This approach entails a progressive reduction in the resolution of the original image by a factor of four, which is intended to facilitate the process of feature matching and stabilize the optimization process by initiating the process with a coarser representation of the scene. In this four-layer pyramid, the first layer is the coarsest and represents the lowest resolution image. Subsequently, the resolution of the layers in question increases gradually until it reaches that of the fourth and final layer, which corresponds to the resolution of the original image. Subsequently, each layer of the pyramid is optimized in a coarse-to-fine manner. The coarse-to-fine strategy allows the optimization process to converge in a more seamless and dependable manner, as it starts with a global approximation and gradually refines the solution in detail.

At each stage of the optimization process, a nonlinear optimization technique is employed to estimate the camera's rotation matrix **R** and translation vector **t**, which are fundamental for determining the camera's pose with respect to the scene. However, this process is susceptible to the bias of reprojection error, which refers to the discrepancy between the observed image features and their expected positions based on the estimated camera pose. In order to minimize this bias and refine the estimates of **R** and **t**, a nonlinear iterative optimization approach, which is the Gauss–Newton method is employed, represents an adaptation of the Newton method for addressing nonlinear systems. This method employs an iterative adjustment of the estimates of **R** and **t** with the objective of reducing the reprojection error and thus minimizing the bias. The Gauss–Newton method employs second-order derivatives (Hessian matrices) to approximate the optimized surface, thereby facilitating a more rapid convergence to a more accurate solution in comparison to first-order methods. This iterative process is continued until the bias is sufficiently reduced to result in a negligible reprojection error. The transformation relationship between the two frames and the schematic of the solution are shown in Fig. 4.

The laparoscope left intrinsic matrix is defined as **A**, the coordinates of the pixel points on the image are defined as u, v, and the depth of the spatial point P to the laparoscope imaging plane is defined as Z. The relative motion of the camera between the k-1th and kth frame is estimated as the rotation matrix **R** and the translation vector **t**, corresponding to the Lie algebra ξ. This can be obtained from the projection equation based on the principle of small-aperture imaging:

$$\begin{cases} p_1 = \begin{bmatrix} \mathbf{u}1 \\ \mathbf{v}1 \\ 1 \end{bmatrix} = \frac{1}{Z_1}\mathbf{A}P \\ p_2 = \begin{bmatrix} \mathbf{u}2 \\ \mathbf{v}2 \\ 1 \end{bmatrix} = \frac{1}{Z_2}\mathbf{A}(\mathbf{R}P + \mathbf{t}) = \frac{1}{Z_2}\mathbf{A}\left(\exp\left(\xi^{\wedge}\right)P\right) \end{cases} \tag{8}$$

For all spatial points P, I(p) is photometry of the pixel of p, the problem of minimizing the photometric error e is formulated:

$$\min_{\xi} J(\xi) = \sum_{i=1}^{N} e_i^T e_i,\ e_i = I\left(p_{i.1}\right) - I\left(p_{i,2}\right) \tag{9}$$

The Lie group is multiplied by a small quantity, and the perturbation model is employed to ascertain the rate of change of the Lie algebra with respect to the aforementioned quantity,

set $q = \delta\xi^{\wedge} \exp\left(\xi^{\wedge}\right)P$, $u = \frac{1}{Z_2}\mathbf{A}q$, and it is able to derive the following formula:

$$\begin{aligned}(\xi \oplus \delta\xi) = I_1\left(\frac{1}{Z_1}\mathbf{A}P\right) \\ - I_2\left(\frac{1}{Z_2}\mathbf{A}\exp\left(\delta\xi^{\wedge}\right)\exp\left(\xi^{\wedge}\right)P\right) \\ = e(\xi) - \frac{\partial I2}{\partial u}\frac{\partial u}{\partial q}\frac{\partial q}{\partial \delta\xi}\delta\xi\end{aligned} \tag{10}$$

where $\frac{\partial I_2}{\partial u}$ is the partial derivative of grey scale with respect to pixel points, $\frac{\partial u}{\partial q}$ is the partial derivative of pixel points with respect to spatial points, and is the partial derivative of spatial points with respect to the corresponding Lie algebra of the transformation matrix Lie group. Then the Jacobi matrix of the error with respect to the Lie algebra is derived. Subsequently, the increments are calculated using the Gaussian Newton method to solve iteratively.

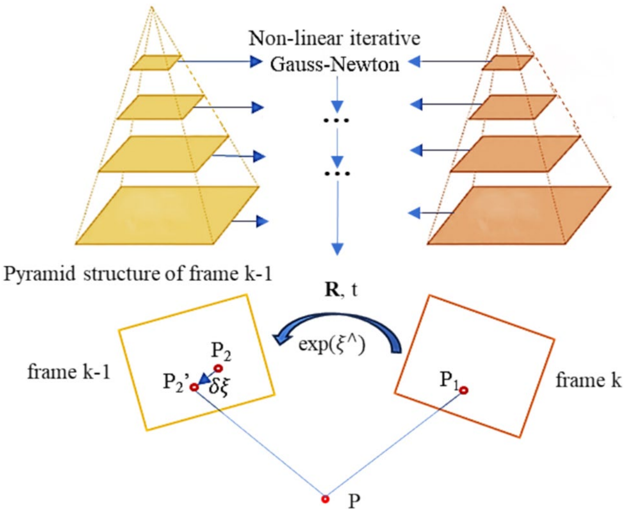


Fig. 4 A four-layer image pyramid is constructed and non-linear optimization is performed on each layer of the pyramid, from coarse to fine, in order to determine the camera rotation matrix **R** and translation vector **t**

4 Experiment

4.1 Datasets and implementation details

In this paper, the hardware environments are Intel(R) Core (TM) i5-8300H CPU @ 2.30GHz and 2 NVIDIA GeForce GTX 1060 GPUs.

A 3D reconstruction system of abdominal cavity is developed based on a binocular laparoscope and a Universal Robot 5 (UR5). The imaging resolution of both the left and right image of the laparoscope is 1920×1080, and two video capture cards are used to connect to the computer via a USB interface with a transmission speed of up to 60 frames per second. The binocular image data is acquired and processed by Python and OpenCV function, the RTDE drive toolkit is utilize for UR5 trajectory control by Ubuntu 18.04 operating system.

One of the image sequences datasets of the binocular abdominal environment we use are derived from an open-source video dataset provided by the Hamlyn Medical Center. Additionally, another image sequences datasets of the abdominal phantom and the suture practice model are captured using our own binocular laparoscopes.

Firstly, the binocular laparoscope is fixed to the end of UR5 by a 3D-printed fixture, and the checkerboard calibration board is fixed on the robot platform. Different positions and attitudes of the end of UR5 are transformed to record 30 sets of joint angles and the corresponding binocular checkerboard images. The intrinsics of the laparoscope, including the focal length, optical center, distortion coefficients, and left and right transformation matrix, are calibrated by the MATLAB toolbox, and the re-projection error is 0.2 pixels, which meets the medical use requirement at the sub-pixel level, and the hand–eye calibration matrix is calculated using our proposed method. It is important to note that images acquired by binocular laparoscopes must be rectified using the intrinsics parameters obtained from the calibration process before they can be input to the depth estimation network for training and obtaining parallax maps.

4.2 Preoperative SLAM accuracy evaluation

The image should be center cropping and the values of u0 and v0, which represent the position of the optical center pixel in the laparoscope's intrinsics parameters, should be modified, this operation will result in a reduction of SLAM mapping anomalies, with a corresponding improvement in the effect. The video stream is transmitted to the computer in real time, where it is used to run the depth estimation and SLAM algorithms in parallel process. This can be achieved at an average speed of 500 ms per frame, resulting in the generation of the laparoscopic trajectory and dense 3D reconstruction of abdominal environment. The process of 3D reconstruction of the abdominal cavity by SLAM is shown in Fig. 5.

Based on EVO (Sturm et al. 2012), an open-source tool used for evaluating the performance of SLAM systems, providing a suite of metrics and visualization tools to assess accuracy and reliability. We select the absolute pose error (APE) metric, which is a metric within EVO that directly calculates the discrepancy between the groundtruth of the laparoscopic trajectory and the estimated value of the SLAM system, and is essential for understanding the long-term precision of SLAM systems. It involves aligning the estimated trajectory with the ground truth, computing the Euclidean distance errors for each pose, and averaging these to obtain the APE, which provides a highly intuitive reflection of the algorithmic accuracy and the global consistency of the trajectory. The groundtruth of the laparoscopic trajectory mentioned above is achieved by employing UR5 to execute a specified trajectory at a low velocity, and the position of the UR5 end to the base is obtained real-time via communication with the RTDE drive toolkit, ensuring that the frequency of position acquisition is consistent with the frame rate of the video captured by the laparoscope. The trajectory groundtruth and the trajectory computed by the SLAM algorithm are converted into the KITTI pattern required by EVO and aligned using the timestamps, and then input them into EVO to compute the spatial position error of XYZ.

Three sets of abdominal phantom image sequences acquired by ourselves are chosen to acquire laparoscopic trajectories using the RGBD method in ORB-SLAM2 (Mur-Artal et al. 2017), Endo-depth-and-motion (Recasens et al. 2021) and our proposed method respectively, and APE is compared with the trajectory acquired by UR5 as groundtruth. Rmse, mean, median, Standard Error(std), minimum, maximum and sum of squares due to error (sse) of our method are all smaller than the other two methods.

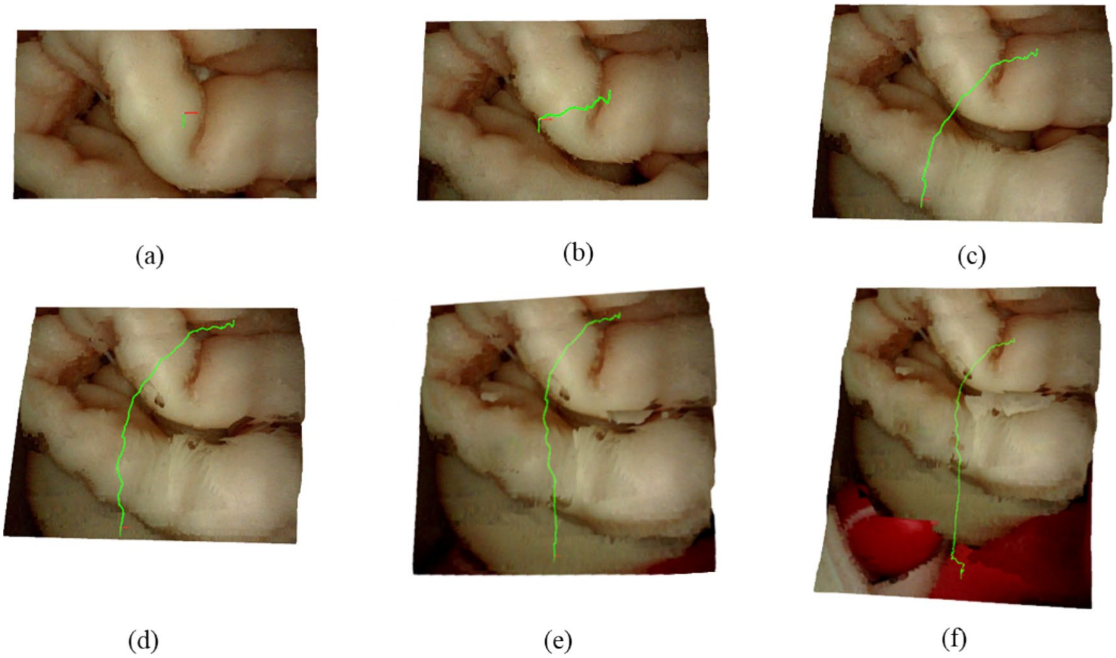


Fig. 5 abcdef are screenshots of the gradual reconstruction of the intraoperative abdominal environment in the order of increasing frames

Among these, the mean error of our proposed method is 0.6mm for dataset1, but the ORB-SLAM2 reaches 2mm, the Endo-depth-and-motion reaches 5mm. Further, ORB-SLAM2 cannot get dense depth to reconstruct the environment. Therefore, our method is superior. Specific comparison results are shown in Table 1 and Fig. 6, of which the optimal indicator is in bold.

4.3 Intraoperative lesion localization

The binocular laparoscopic fixture we use has a needle fixed to it, and the transformation relationship of the tool coordinate system to laparoscopy coordinate system is known by the designed fixture. Intraoperatively, the lesion area and center point are extracted by color space transformation, threshold segmentation, and contour extraction fitting. The parallax map is then obtained and combined with the hand–eye calibration matrix to calculate the spatial position of the lesion point under the robot base coordinate system. and visually guided the movement of the surgical

Table 1 ATE assessment results. Dataset1, dataset2 and dataset3 are the abdominal phantom we collect ourselves using the UR5, we compare the accuracy of the SLAM trajectories using ORB-SLAM2, Endo-depth and our proposed method using metrics such as Rmse, mean, median, Standard Error(std), minimum, maximum and sum of squares due to error (sse). Data in the table are in meters, and optimal indicators are in bold

Image sequence	Method	RMSE	Mean	Median	STD	Min	Max	SSE
Dataset1	ORB_SLAM2	0.00223	0.00208	0.002085	0.000802	0.000364	0.00424	0.00086
	Endo_depth	0.006115	0.005502	0.005468	0.002668	0.001368	0.012641	0.006469
	Ours	**0.000745**	**0.000658**	**0.000582**	**0.00035**	**8.41E-05**	**0.002362**	**9.61E-05**
Dataset2	ORB_SLAM2	0.010383	0.009036	0.008163	0.005114	0.002552	0.024925	0.01714
	Endo_depth	0.00823	0.008053	0.007872	0.001695	0.003339	0.013433	0.010769
	Ours	**0.001012**	**0.000872**	**0.000763**	**0.000512**	**9.82E-05**	**0.002289**	**0.000163**
Dataset3	ORB_SLAM2	0.00896	0.007597	0.007456	0.00475	2.11E-05	0.016084	0.010837
	Endo_depth	0.002452	0.002152	0.001763	0.001174	0.000412	0.005228	0.000812
	Ours	**0.001518**	**0.001412**	**0.001408**	**0.000557**	**0.000262**	**0.002609**	**0.000311**

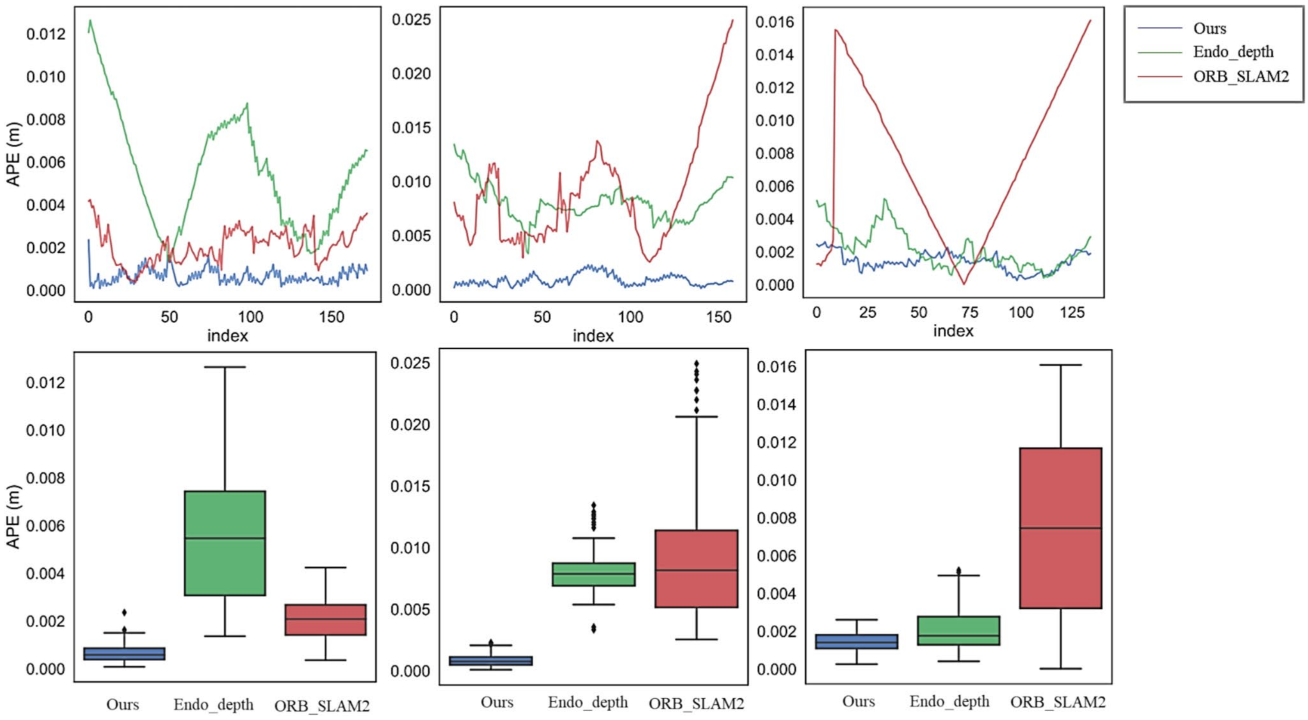


Fig. 6 ATE assessment results. Data in the diagrams are in meters

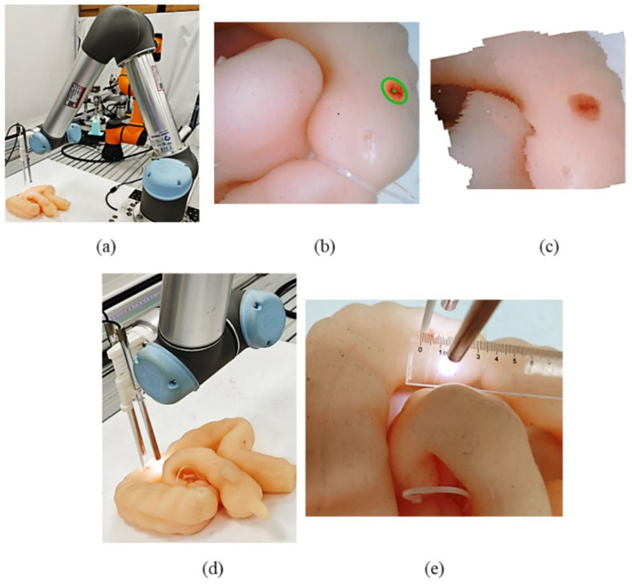


Fig. 7 Lesion localization experiments. (**a**) The entire image of experimental scene (**b**) lesion is extracted through threshold segmentation and contour extraction (**c**) 3D pointcloud of the lesion (**d**) Detailed image of the experimental scene (**e**) Experimental result and error assessment

instrument to the lesion point. Following the experiment, the positioning accuracy is found to be approximately 5mm. The experimental scenarios and localization results are shown in Fig. 7.

5 Conclusion

In this paper, we investigate the binocular laparoscopic stereo imaging technique, complete sub-pixel intrinsic calibration to meet the requirements of medical use, combine the binocular constraints and data loops to complete a accurate hand–eye calibration specifically for binocular laparoscopic, and acquire the depth map through the binocular unsupervised depth estimation algorithm based on PAM. The pseudo-RGBD frames are established, while the SLAM algorithm run based on the pyramid optimization method to minimize the luminance error from coarse to fine by parallel process. The trajectory accuracy of our proposed algorithm on our dataset also reaches 1mm. We complete fast and dense 3D reconstruction of the preoperative abdominal environment and intraoperative lesion localization.

The abdominal cavity environment presents a number of challenges for the accurate estimation of depth, which include lack of texture, specular luminescence on the tissue surface and smoke. Additionally, the deformation of tissues due to heart beating and occlusion of surgical instruments affects SLAM construction. Consequently, 3D reconstruction of abdominal cavity is a complex and challenging process, with numerous issues that require further investigation. In future research, we intend to further investigate the potential for improvement of unsupervised depth estimation algorithms to address the specific challenges posed by smoke and mirror reflection, and to study SLAM algorithms suitable for dynamic environments, with the objective of achieving higher speed and more robust 3D reconstruction of abdominal tissues. Furthermore, we intend to combine the Nerf (Mildenhall et al. 2020) and 3D Gaussian splatting (Kerbl et al. 2023) algorithms to obtain a more realistic and reconstructed view of the region of instrument occlusion. This paper focuses on the system overall integrity, with a lack of in-depth research on MIS, such as RCM constraints and the support of clinical invivo data (Liu et al. 2024). The proposed algorithm will be further validated with the clinical data of our group's project.

Acknowledgements This work is supported by Major Program (JD) of Hubei Province (2023BAA016); the National Natural Science Foundation of China under Grant 52188102.

Author Contribution Sui wrote the first draft of the manuscript text, Zhang improved the experimental setup, Zhao and Tao directed and supervised the research. All authors reviewed and approved the final manuscript.

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest All authors declare no conflicts of interest. The authors declare no competing interests.

References

Chang, P.L., Stoyanov, D., Davison, A.J., Edwards, P.E..: Real-time dense stereo reconstruction using convex optimization with a cost-volume for image-guided robotic surgery. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013. MICCAI 2013. Lecture Notes in Computer Science, vol 8149. Springer, Berlin, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40811-3_6

Enebuse, I., Foo, M., Ibrahim, B.S.K.K., Ahmed, H., Supmak, F., Eyobu, O.S.: A comparative review of hand–eye calibration techniques for vision guided robots. IEEE Access **9**, 113143–113155 (2021). https://doi.org/10.1109/ACCESS.2021.3104514

Engel, J., Koltun, V., Cremers, D.: Direct sparse odometry. IEEE Trans. Pattern Anal. Mach. Intell. **40**(3), 611–625 (2018). https://doi.org/10.1109/TPAMI.2017.2658577

Godard, C., Aodha, O.M., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6602–6611 (2017). https://doi.org/10.1109/CVPR.2017.699

Huang, P.H., Matzen, K., Kopf, J., Ahuja, N., Huang, J.B.: DeepMVS: Learning Multi-view Stereopsis. In: Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2821–2830 (2018). https://doi.org/10.1109/CVPR.2018.00298

Kerbl, B., Kopanas, G., Leimkuehler, T., Drettakis, G.: 3D Gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. (TOG). **42**, 1–14 (2023). https://doi.org/10.1145/3592433

Liu, P., Qian, L., Zhao, X., Tao, B.: Joint knowledge graph and large language model for fault diagnosis and its application in aviation assembly. IEEE Trans. Ind. Inf. 20(6):8160–8169 (2024). https://doi.org/10.1109/TII.2024.3366977

Lu, B., Li, B., Dou, Q., Liu, Y.: A unified monocular camera-based and pattern-free hand-to-eye calibration algorithm for surgical robots with RCM constraints. IEEE/ASME Trans. Mechatron. **27**(6), 5124–5135 (2022). https://doi.org/10.1109/TMECH.2022.3166522

Mahjourian, R., Wicke, M., Angelova, A.: Unsupervised Learning of Depth and Ego-Motion from Monocular Video Using 3D Geometric Constraints. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), pp. 5667–5675 (2018). https://doi.org/10.1109/CVPR.2018.00594

Mahmoud, N., Cirauqui, I.: Orbslam-based endoscope tracking and 3d reconstruction. In: Peters, T., et al. Computer-Assisted and Robotic Endoscopy. CARE 2016, LNCS, vol 10170. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-54057-3_7

Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: representing scenes as neural radiance fields for view synthesis. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, JM. (eds) Computer Vision – ECCV 2020. ECCV 2020. Lecture Notes in Computer Science(), vol 12346. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58452-8_24

Mur-Artal, R., Tardós, J.D.: ORB-SLAM2: An Open-source SLAM system for monocular, stereo, and RGB-D cameras. IEEE Trans. Rob. **33**(5), 1255–1262 (2017). https://doi.org/10.1109/TRO.2017.2705103

Mur-Artal, R., Montiel, J.M.M., Tardós, J.D.: ORB-SLAM: a versatile and accurate monocular SLAM system. IEEE Trans. Rob. **31**(5), 1147–1163 (2015). https://doi.org/10.1109/TRO.2015.2463671

Özgüner, O., et al.: Camera-robot calibration for the da vinci robotic surgery system. IEEE Trans. Autom. Sci. Eng. **17**(4), 2154–2161 (2020). https://doi.org/10.1109/TASE.2020.2986503

Pachtrachai, K., Vasconcelos, F., Chadebecq, F., et al.: Adjoint transformation algorithm for hand–eye calibration with applications in robotic assisted surgery. Ann. Biomed. Eng. **46**, 1606–1620 (2018). https://doi.org/10.1007/s10439-018-2097-4

Qiu, L., Ren, H.: Endoscope navigation with slam-based registration to computed tomography for transoral surgery. Int. J. Intell. Robot. Appl. **4**(2), 252–263 (2020). https://doi.org/10.1007/s41315-020-00127-2

Recasens, D., Lamarca, J., Fácil, J.M., Montiel, J., Civera, J.: Endo-depth-and-motion: localization and reconstruction in endoscopic videos using depth networks and photometric constraints. IEEE Robot. Autom. Lett. **6**(4), 7225–7232 (2021). https://doi.org/10.1109/LRA.2021.3095528

Scharstein, D., Szeliski, R., Zabih, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. In: Proceedings of the IEEE Workshop on Stereo and Multi-Baseline Vision (SMBV 2001), pp. 131–140 (2001). https://doi.org/10.1109/SMBV.2001.988771

Scona, R., Jaimez, M., Petillot, Y.R., Fallon, M., Cremers D.: StaticFusion: Background Reconstruction for Dense RGB-D SLAM in Dynamic Environments. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), pp. 3849–3856 (2018). https://doi.org/10.1109/ICRA.2018.8460681

Song, J., Wang, J., Zhao, L., Huang, S., Dissanayake, G.: Mis-slam: real-time large-scale dense deformable slam system in minimal invasive surgery based on heterogeneous computing. IEEE Robot. Autom. Lett. **3**(4), 4068–4075 (2018a). https://doi.org/10.1109/LRA.2018.2856519

Song, J., Wang, J., Zhao, L., Huang, S., Dissanayake, G.: Dynamic reconstruction of deformable soft-tissue with stereo scope in minimal invasive surgery. IEEE Robot. Autom. Lett. **3**(1), 155–162 (2018b). https://doi.org/10.1109/LRA.2017.2735487

Sturm, J., Engelhard, N., Endres, F., Burgard, W., Cremers, D.: A benchmark for the evaluation of rgb-d slam systems. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 573–580 (2012). https://doi.org/10.1109/IROS.2012.6385773

Tourani, A., Bavle, H., Sanchez-Lopez, J.L., Voos, H.: Visual SLAM: what are the current trends and what to expect? Sensors **22**(23), 9297 (2022). https://doi.org/10.3390/s22239297

Tsai, R.Y., Lenz, R.K.: A new technique for fully autonomous and efficient 3D robotics hand/eye calibration. IEEE Trans. Robot. Autom. **5**(3), 345–358 (1989). https://doi.org/10.1109/70.34770

Wang, X.Z., Nie, Y., Lu, S.P., Zhang, J.: Deep convolutional network for stereo depth mapping in binocular endoscopy. IEEE Access **8**, 73241–73249 (2020). https://doi.org/10.1109/ACCESS.2020.2987767

Wang, L., Guo, Y., Wang, Y., Liang, Z., Lin, Z., Yang, J., An, W.: Parallax attention for unsupervised stereo correspondence learning. IEEE Trans. Pattern Anal. Mach. Intell. **44**(4), 2108–2125 (2022). https://doi.org/10.1109/TPAMI.2020.3026899

Zaffar, M., Ehsan, S., Stolkin, R., Maier, K.M.: Sensors, slam and long-term autonomy: a review. In 2018 NASA/ESA Conference on Adaptive Hardware and Systems (AHS), pp. 285–290 (2018). https://doi.org/10.1109/AHS.2018.8541483

Zhang, Z.Y.: Flexible camera calibration by viewing a plane from unknown orientations. In: Proceedings of the Seventh IEEE International Conference on Computer Vision, Kerkyra, Greece, vol.1, pp. 666–673 (1999). https://doi.org/10.1109/ICCV.1999.791289

Zhong, F., Wang, Z., Chen, W., He, K., Wang, Y., Liu, Y.H.: Hand–eye calibration of surgical instrument for robotic surgery using interactive manipulation. IEEE Robot. Autom. Lett. **5**(2), 1540–1547 (2020). https://doi.org/10.1109/LRA.2020.2967685

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Xin Sui received the B.S. degrees in mechanical engineering from Huazhong University of Science and Technology (HUST), Wuhan, China, in 2023. She is currently working toward a M.S. degree at the State Key Laboratory of Intelligent Manufacturing Equipment and Technology, Huazhong University of Science and Technology (HUST), Wuhan, China. Her current research interests mainly include visual perception and control of surgical robots.

Yang Zhang received the B.S. and M.S. degrees in mechanical engineering from Huazhong University of Science and Technology (HUST), Wuhan, China, in 2015 and 2018, respectively. He is currently working toward a Ph.D. degree at the State Key Laboratory of Intelligent Manufacturing Equipment and Technology, Huazhong University of Science and Technology (HUST), Wuhan, China. His current research interests mainly include perception, planning and control of surgical robots.

Xingwei Zhao received B.S. and M.S. degrees in mechanical engineering from University of Duisburg-Essen, Duisburg, Germany, in 2012 and 2013, respectively. He received the Ph.D. degree in mechanical engineering from the Technical University of Berlin, Berlin, Germany, in 2017. He is currently an associate research fellow with the State Key Laboratory of Intelligent Manufacturing Equipment and Technology, Huazhong University of Science and Technology (HUST), Wuhan, China. His research interests mainly include nonlinear dynamics, robot control and robotic manufacture.

Bo Tao received the B.S. and Ph.D. degrees in mechanical engineering from Huazhong University of Science and Technology (HUST) in 1999 and 2007 respectively. After being a post-doctor from June 2007 to June 2009, he has been an Associate Professor in 2009 and a Professor in 2013 at the School of Mechanical Science and Engineering, HUST. From June 2013 to June 2014, he was a visiting scholar at the Mechanical Engineering Department of UC Berkeley, USA. And now he is a Changjiang Scholar Chair Professor of HUST. He has published 1 monograph, more than 80 papers in international journals and conference. His research interests mainly include intelligent manufacturing and robotics technologies, IOT technologies and applications.