Computers in Biology and Medicine 163 (2023) 107121

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

FRSR: Framework for real-time scene reconstruction in robot-assisted minimally invasive surgery

Ziyang Chen [a,*], Aldo Marzullo [b], Davide Alberti [a], Elena Lievore [c], Matteo Fontana [c], Ottavio De Cobelli [c,d], Gennaro Musi [c,d], Giancarlo Ferrigno [a], Elena De Momi [a,c]

[a] Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, 20133, Italy
[b] Department of Mathematics and Computer Science, University of Calabria, Rende, 87036, Italy
[c] Department of Urology, European Institute of Oncology, IRCCS, Milan, 20141, Italy
[d] Department of Oncology and Onco-haematology, Faculty of Medicine and Surgery, University of Milan, Milan, 20122, Italy

ARTICLE INFO

Keywords:
3D reconstruction
Intra-operative scenes
Stereo endoscope
da Vinci Research Kit

ABSTRACT

3D reconstruction of the intra-operative scenes provides precise position information which is the foundation of various safety related applications in robot-assisted surgery, such as augmented reality. Herein, a framework integrated into a known surgical system is proposed to enhance the safety of robotic surgery. In this paper, we present a scene reconstruction framework to restore the 3D information of the surgical site in real time. In particular, a lightweight encoder–decoder network is designed to perform disparity estimation, which is the key component of the scene reconstruction framework. The stereo endoscope of da Vinci Research Kit (dVRK) is adopted to explore the feasibility of the proposed approach, and it provides the possibility for the migration to other Robot Operating System (ROS) based robot platforms due to the strong independence on hardware. The framework is evaluated using three different scenarios, including a public dataset (3018 pairs of endoscopic images), the scene from the dVRK endoscope in our lab as well as a self-made clinical dataset captured from an oncology hospital. Experimental results show that the proposed framework can reconstruct 3D surgical scenes in real time (25 FPS), and achieve high accuracy (2.69 ± 1.48 mm in MAE, 5.47 ± 1.34 mm in RMSE and 0.41 ± 0.23 in SRE, respectively). It demonstrates that our framework can reconstruct intra-operative scenes with high reliability of both accuracy and speed, and the validation of clinical data also shows its potential in surgery. This work enhances the state of art in 3D intra-operative scene reconstruction based on medical robot platforms. The clinical dataset has been released to promote the development of scene reconstruction in the medical image community.

1. Introduction

Robot-assisted minimally invasive surgery (RAMIS) can improve the performance of surgeons, because it enlarges the surgical vision and enhances the dexterity of instruments compared with the traditional open surgery [1]. More importantly, it opens the way for the integration of artificial intelligence in surgery [2–5]. Generally, surgeons can observe the surgical scene in real time through the transmission of stereo images using an endoscope. Then, the depth information of the scene is restored thanks to human inherent ability of binocular vision perception. However, this information only comes from the subjective consciousness of surgeons, due to the lack of accurate depth value calculation. It can be seen that the specific depth information is important, because it is the key step for intra-operative image guidance which is a popular assisted technology today [6,7], e.g., for the registration of pre-operative data, such as Computed Tomography (CT) and Magnetic Resonance Imaging (MRI), with the intra-operative scenes, to visualize the anatomical information of the patients during the operation. One of the main difficulties is that the surgical scene is always changing caused by the deformation of soft tissue and movement of instruments, which results in inaccurate registration. The real-time reconstruction quality of intra-operative surgical scenes directly affects the safety of operation.

The restoration of 3D scene information using a stereo camera has always been one of the hotspots in computer vision [8,9]. Given a rectified image pair, the task is to build stereo correspondence through the pixels or high-level features from left and right images. The 3D

* Corresponding author.
E-mail address: ziyang.chen@polimi.it (Z. Chen).

https://doi.org/10.1016/j.compbiomed.2023.107121
Received 14 February 2023; Received in revised form 12 May 2023; Accepted 30 May 2023
Available online 3 June 2023

point cloud can then be recovered based on these estimated matching values. Typical stereo correspondence always follows the process of cost calculation, cost aggregation, disparity calculation and refinement [10]. Two methods based on block matching and nonparametric census transform in [9] were developed to estimate disparity values of endoscopic heart images. Preprocessing approach including specularity removal and image equalization was implemented to remove specular highlights and enhance contrast, and linear iterative clustering (SLIC) super-pixels operation was added as a postprocessing way to refine the disparity map. This strategy could reconstruct intra-operative scenes, but the speed is not satisfactory due to the integration of many image processing steps. An optimized Quasi-Dense Matching [11] was proposed to perform a fast reconstruction, and it achieved the desired results through the validation of different endoscopic datasets by running on a GPU. Similarly, the authors in [12] first extracted initial 3D information from stereo optical videos, then use feature-based SLAM to mosaic the model, and effective post-processing steps, including outliers removal, hole filling and smoothing, were finally utilized to handle low-textured areas on soft tissue.

More recently deep learning based disparity estimation has been gradually explored, due to the expansion of big data and the enhancement of computing ability. An unsupervised learning approach was proposed in [13] to predict disparity maps of surgical scenes using a Generative Adversarial Network (GAN). The authors utilized a U-Net architecture [14] based generator to predict left and right disparity maps, and used a discriminator to judge the quality of prediction after reprojecting the disparity maps into RGB image pairs. Although it does not require annotated data, the reconstruction quality still needs to be improved. In [15], the authors also proposed an unsupervised model based on a GAN. A vertical correction module was designed for the compensation of imperfect image rectification, and the stereo image was reconstructed by fusing original images, estimated disparity maps and vertical correction maps, then a discriminator was adopted to distinguish the difference between the reconstructed images and the original images. This approach got satisfactory prediction accuracy by evaluating a public dataset. Supervised learning based neural networks for stereo correspondence are more commonly designed outside of MIS context, since the dataset with ground truth of disparity maps in the medical field is insufficient. To encode an image pair in a convolutional neural network, FlowNet [16] first proposed two possible strategies including stacking two images directly, or inputting them separately into two identical networks and then combining the feature maps using patch-based correlation operation. Through the alternate use of convolution and pooling, the high-level feature representation was generated and finally refinement was implemented to output the predicted optical flow with full image resolution. The second option, i.e., encoding the image pair separately using two identical branches, was subsequently widely utilized in the stereo matching field. The authors in [17] adopted two identical networks with Spatial Pyramid Pooling (SPP) modules to extract features of the image pair, and a stacked hourglass architecture consisting of two 3D convolution models and three small 3D U-Net modules was designed to perform the multi-scale disparity estimation. Similarly, a coarse to fine structure [18] for disparity estimation was proposed to refine the prediction by calculating four different scales of disparity estimation. The authors in [19] introduced a neural architecture search strategy to select the optimal architecture for those modules which contain trainable parameters, and the results showed that the accuracy can be improved while reducing computing resources. In [20], a novel cost volume was constructed based on the group-wise correlation and the authors trimmed the stacked hourglass architecture in their decoder to refine the model volume. It can be seen that constructing cost volume is an important step to regress disparity maps, so some researchers also designed different techniques to calculate cost volume in [21–23]. Considering that 3D convolutional operation is computationally expensive, the authors in [24] designed a hierarchically aggregated pyramid network to avoid the construction of cost volume, and it works in different surgical image pairs, including colon phantom, partial nephrectomy, and prostatectomy.

However, previous research has almost focused on stereo correspondence or point cloud generation, and these approaches are not integrated with real robotic platforms. It is possible to see that the performance of these methods on the physical platform may be inconsistent with separate test scenarios based on public datasets, due to the potential effects of signal interference, delay, etc. Hence, it is significant to perform the extended work, integrating the scene reconstruction approach into the physical robotic platform, in order to facilitate clinical application. In this paper, we built a Framework for Real-time Scene Reconstruction (FRSR) to visualize 3D surgical scenes interactively, and it was integrated into the da Vinci surgical system (Intuitive Surgical Inc., Sunnyvale, California) which is the most typical platform in RAMIS [25]. To the best of our knowledge, this is the first paper to demonstrate the effect of real-time scene reconstruction integrating into a known surgical robotic platform. The main contributions of this paper are summarized as follows:

(1) A real-time scene reconstruction framework integrated into da Vinci Research Kit (dVRK) is built to demonstrate 3D surgical scenes interactively.

(2) A lightweight deep learning based model, consisting of the U-Net based encoder and consecutive 3D residual modules based decoder, is designed for disparity estimation.

(3) A clinical surgical dataset, captured from an oncology hospital, is made and released online for the development of the medical scene reconstruction community.[1]

The remainder of this paper is structured as follows. Section 2 describes the framework proposed in this paper, and it also introduces the architecture of the neural network as a key component in this pipeline. In Section 3, it presents the training strategy of the designed network and comprehensive analysis for the reconstruction performance using three different scenarios. Section 4 discusses the findings and limitations of this work, and the conclusion is drawn in Section 5.

2. Methodology

This section presents the details of our real-time scene reconstruction framework, which is integrated into the dVRK. In particular, a lightweight neural network is introduced to perform disparity estimation since it closely affects the performance of the framework.

2.1. FRSR framework description

As the standard first generation of da Vinci surgical system, dVRK is currently developed to integrate various computer-assisted technologies as an open platform [26]. Thanks to the high integration of control hardware and software, we choose to extend it for the experimental study. A standard dVRK consists of the leader side and follower side: the stereoscopic endoscope mounted on an Endoscopic Camera Manipulator (ECM) captures the surgical image pairs and transmits them to the master control terminal. Then, the surgeon can observe the procedure through a High Resolution Stereo Viewer (HRSV) at the leader side and adjust surgical viewpoints by teleoperating ECM. Users can intuitively perceive the depth of scenes through different viewpoints from HRSV. Our experimental setup adopted the stereo endoscope of dVRK to extend the vision component.

The framework for scene reconstruction is managed by the Robot Operating System (ROS), an open source software development kit for robotic applications. Using ROS correspondence provides more flexibility, such as the enhanced security for signal transmission in the tele-surgery [27]. For the whole framework, the stereo endoscope

[1] Download link: https://doi.org/10.5281/zenodo.7385603

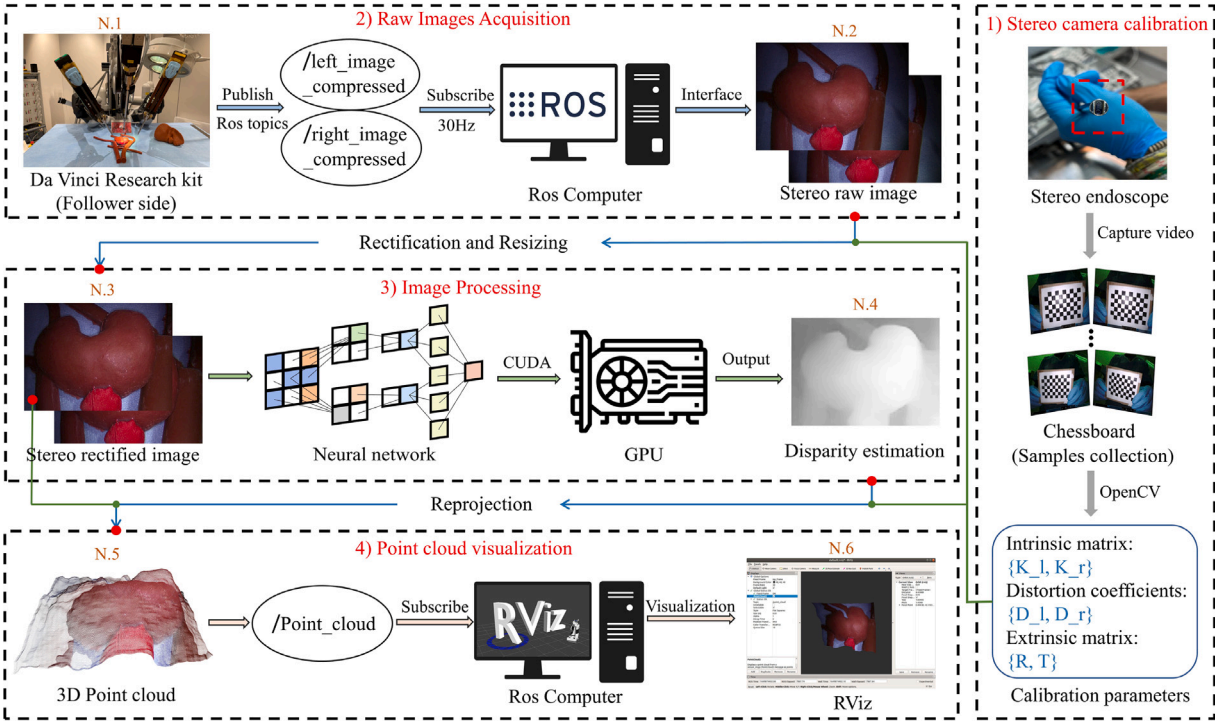

Fig. 1. Schematic representation of the surgical scene reconstruction framework. The name in the circles represents the ROS topics, including the compressed image pair and the generated point cloud. The blue solid arrow indicates the workflow, which flows from the end of the previous box to the first component of the next box through specific operations, and the green solid arrow indicates additional components necessary for these operations. The right block named stereo camera calibration is the preliminary step to generate calibration parameters for our framework, which means that it only needs to be executed once in advance. The main components consisting of three blocks on the left side directly affect the real-time performance of our framework, including the acquisition of raw compressed image pairs, stereo image processing and visualization of generated point clouds.

keeps capturing surgical scenes at the follower side and publishing them through the form of ROS topics, then the computer connected to dVRK through ROS correspondence receives these messages and processes surgical images based on the neural network, and finally the specific visualizer subscribes (meaning that it receives ROS messages) to the generated topic of point cloud and visualizes 3D surgical scenes interactively. Fig. 1 presents the construction details of FRSR, and it is basically composed of four parts,

(1) "Stereo camera calibration" is a preliminary step, and it involves acquiring image pairs of a chessboard with known dimensions and relating object properties with their representation in the acquired images. A 9×6 calibration black–white chessboard with 10 mm squares is used to collect calibration images. Calibration parameters, including intrinsic matrix, distortion coefficients and extrinsic matrix, are then generated using the OpenCV library. These parameters are fixed for the usage in later processes, so this preliminary step is performed once if the stereo endoscope is not changed.

(2) "Raw images acquisition" is the first part in the framework of this cycle. They are captured by the stereo endoscope and published as ROS topics both in a compressed and uncompressed format. The compressed format was chosen in FRSR because it takes less time for image transmission without losing useful information. The computer which hosts the visualization is connected to the ROS network and it subscribes to these topics at 30 Hz. Then, the left and right images are displayed separately through different windows.

(3) "Image processing" is the second part of our FRSR setup. To perform the accurate disparity estimation, the image rectification is required to be conducted using *stereoRectify* and *remap* OpenCV functions, since it facilitates the matching of pixels on the left and right images when they are in the same epipolar line. Then, the rectified image pairs are resized to the required resolution and input to the neural network running on GPU for enhancing the speed, and the final disparity values are predicted later. Considering that the neural network architecture directly affects the performance of the framework, a lightweight model will be introduced in the next section.

(4) Finally, "point cloud visualization" is implemented as the last part of this cycle. The 3D surface is reconstructed based on the estimated disparity values. Then, the generated point cloud is converted to a newly created ROS topic, and the color information extracted from the rectified left image is added to the message for the better visualization purpose. Here, RViz, an interactive 3D visualizer for ROS framework, is selected to visualize the generated 3D scene since it allows to visualize and manipulate the point cloud in real time.

2.2. Network architecture

A lightweight disparity estimation network is proposed in this work, consisting of two parts: a U-Net based 2D encoder to extract the high-level features of stereo images, and consecutive 3D residual blocks [28] based decoder to perform final disparity prediction. The overall architecture of the designed prediction network is shown in Fig. 2, and specific parameters are described in Table 1. Stereo image pairs after rectification are extracted with high-level feature maps by sharing encoder weights. The basic 2D convolutional blocks, composed of convolutional operation, batch norm (BN) and leaky rectified linear unit (Relu) activation, are utilized to extract the initial features of RGB stereo images. Then, four strided 2D convolutional blocks are added to further generate feature maps in a lower dimension. Similar to the U-Net architecture, 2D convolutional blocks with upsampling layers are implemented to enlarge the size of feature maps, and skip connection is used to concatenate different level feature maps on the channel.

Next, a 3D cost volume is built in a combination approach. On the one hand, we compute the difference between the left and right feature maps at each disparity hypothesis value along the horizontal epipolar line [18]. On the other hand, we also introduce the group-wise

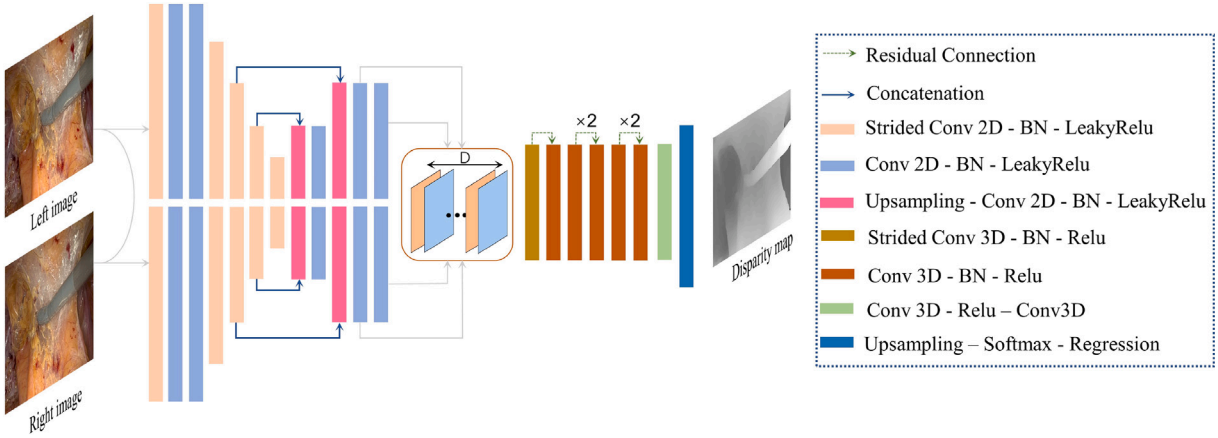

Fig. 2. Architecture of the proposed network. The blue solid arrow indicates the concatenation of different level feature maps on the channel, while the green dashed one denotes the sum operation at the end of the residual block. The 2D convolutional operation followed by Batch Norm (BN) and leaky rectified linear unit (Relu) activation function is adopted to form the basic module of the encoder. On the decoder side, the 3D convolutional operation is utilized with BN and Relu layers as basic modules. The disparity values are finally estimated through upsampling, softmax activation and regression operations.

correlation inspired by [20]. Hence, a combined cost volume V_{com} can be defined as:

$$\begin{aligned} V_{gwc}(d,x,y,g) &= \frac{1}{N_c/N_g}\left\langle f_l^g(x,y), f_r^g(x-d,y)\right\rangle \\ V_{dif}(d,x,y) &= \left|f_l(x,y) - f_r(x-d,y)\right| \\ V_{\text{com}} &= \left\|V_{gwc}, V_{dif}\right\| \end{aligned} \tag{1}$$

Where V_{gwc} is the cost volume generated by calculating group-wise correlation, while V_{dif} is the cost volume by considering the difference between left and right feature maps. d is the disparity hypothesis value, g denotes the feature group, and f_l and f_r represent left and right feature maps. N_c is the channel number of feature maps, while N_g is the divided group number. $\langle,\rangle$ denotes the inner product, $\|$ is the absolute value operation, and $\|,\|$ represents the vector concatenation.

3D convolutional blocks are utilized in the decoder due to the extended disparity channel, consisting of 3D convolution, BN, and Relu activation. Considering that 3D convolutional blocks consume high computing resources, we selected four consecutive residual blocks [17] without extra operations to squeeze the parameter size of our model. The disparity regression approach is implemented to estimate the final disparity values $\bar{d}$ by taking the sum of each disparity with its weighted probability [29]:

$$\bar{d} = \sum_{d=0}^{D_{\max}} d \times \sigma_{axis=1}\left(-c_d\right) \tag{2}$$

where D_{max} is the maximum disparity value, c_d represents the predicted cost, and $\sigma(\cdot)$ denotes the softmax operation. Here, $axis = 1$ means that the softmax operation is performed in axis 1 (i.e., the disparity dimension of the predicted cost c_d).

Regarding the loss function, we consider two different terms. On the one hand, to compare the absolute pixel difference between the ground truth and the predicted one, the "smooth L_1" is adopted due to its high robustness and low sensitivity to outliers [30], and the loss term L_{abs} is defined as:

$$L_{abs}(\hat{d},\bar{d}) = \frac{1}{N}\sum_{i=1}^{N} \text{smooth}_{L_1}\left(\hat{d}_i - \bar{d}_i\right) \tag{3}$$

in which

$$\text{smooth}_{L_1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases} \tag{4}$$

Where N is the total number of pixels, and $\hat{d}_i$ is the disparity value of ground truth. On the other hand, we introduce the loss term L_{gra} to pay more attention to the object boundaries by comparing image gradient [31]:

$$L_{gra}(\hat{d},\bar{d}) = \frac{1}{N}\sum_{i=1}^{N}\left|\boldsymbol{g}_x\left(\hat{d}_i,\bar{d}_i\right)\right| + \left|\boldsymbol{g}_y\left(\hat{d}_i,\bar{d}_i\right)\right| \tag{5}$$

Where $\boldsymbol{g}_x$ and $\boldsymbol{g}_y$ represent the gradient difference of images in the x and y directions, respectively. Hence, the final loss function L_{sum} can be calculated as:

$$L_{sum}(\hat{d},\bar{d}) = \lambda_1 L_{abs}(\hat{d},\bar{d}) + \lambda_2 L_{gra}(\hat{d},\bar{d}) \tag{6}$$

Where λ_1 and λ_2 denote the weights of different loss terms.

Table 1
Parameters of the proposed network. The encoder consisting of Layers 2 to 13 extracts the feature maps with the dimension of (Height, Width, Channel), while the features in the decoder from Layers 15 to 21 contain the dimension of (Height, Width, Disparity, and Channel). The input is the stereo image with a resolution of 512 × 384 here, and we present the dimension of a single map in the encoder part.

ID	Layer setting	Output dimension	Connected to
1	Input	512*384*3	2
2	Strided Conv2D, BN, Leaky Relu	256*192*16	3
3	Conv2D, BN, Leaky Relu	256*192*16	4
4	Conv2D, BN, Leaky Relu	256*192*32	5
5	Strided Conv2D, BN, Leaky Relu	128*96*32	6
6	Strided Conv2D, BN, Leaky Relu	64*48*64	7, 12
7	Strided Conv2D, BN, Leaky Relu	32*24*128	8, 10
8	Strided Conv2D, BN, Leaky Relu	16*12*128	9
9	Upsampling, Conv2D, BN, Leaky Relu	32*24*64	10
10	Conv2D, BN, Leaky Relu	32*24*128	11
11	Upsampling, Conv2D, BN, Leaky Relu	64*48*64	12
12	Conv2D, BN, Leaky Relu	64*48*120	13,14
13	Conv2D, BN, Leaky Relu	64*48*12	14
14	Cost Volume	64*48*48*32	15
15	Strided Conv3D, BN, Relu	64*48*24*32	16,17
16	Conv3D, BN, Relu	64*48*24*32	17
17	Residual 3D Block	64*48*24*32	18
18	Residual 3D Block	64*48*24*32	19
19	Residual 3D Block	64*48*24*32	20
20	Residual 3D Block	64*48*24*32	21
21	Conv3D, Relu, Conv3D	64*48*24*1	22
22	Upsampling, Softmax, Regression	512*384*1	Output

3. Experiment and demonstration

3.1. Network training strategy

Some common public stereo datasets were first utilized to train our model, including Scene Flow [30], KITTI 2015 [32], KITTI2012 [33],

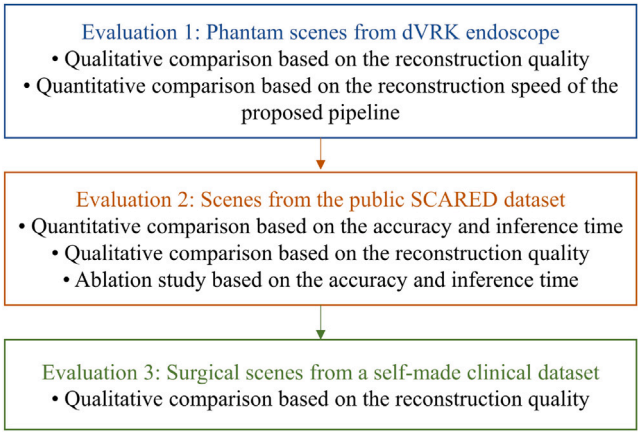

Fig. 3. The flowchart of our evaluation framework. It was performed based on three different surgical scenes.

and ETH3D [34]. Also, a synthetic dataset related to stereo endoscopic images was made by Blender [35]. Here, five thousand stereo pairs (640 × 480 resolution) with ground truth of disparity maps were generated using a moving camera based on five different phantom scenes. In addition, the SERV-CT dataset [36] imaged by two different ex vivo porcine samples was adopted for finetuning the model to enhance the generalization. Common operations of data augmentation were performed to enlarge the training data, including random scaling and cropping, and adjusting brightness, gamma and contrast. It can also help to improve the generalization capability of the model [37]. Finally, 98751 image pairs were utilized to train our network from the scratch, and finetuning contains 16 image pairs.

Images were randomly cropped to size 512 × 384 for the model training, and the maximum disparity value D_{max} was set to 384. Here, we checked the images and we noticed that a few images have the maximum disparity values which are close to 384, so we set D_{max} to 384 in our case. The network was trained with Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$). It was pre-trained for 9 epochs with a learning rate of 0.001, and we set the rate as 0.0001 in the last epoch. Then, we finetuned the network with a constant learning rate of 0.0001 for 1000 epochs. The whole training process was performed on an Ubuntu server with an NVIDIA A100 GPU.

Fig. 3 presents the specific components of our evaluation framework. Three different surgical scenes were utilized to evaluate the reconstruction performance. Six state-of-the-art approaches for disparity estimation were adopted to conduct the comparison study. Among them, [10] is an optimization based method, and the other five methods [17,20–23] are based on neural networks. Hence, we retrained the deep learning based methods using the same training datasets with the original training configurations, and finetuned them using the same strategy as ours for a fair comparison.

3.2. System performance evaluation using dVRK endoscope

To evaluate the scene reconstruction performance of the proposed FRSR framework, a low-textured liver phantom model, was 3D printed to test the whole pipeline. Fig. 5 shows the mean computing time comparison of the proposed framework for scene reconstruction. 100 consecutive samples were collected for each evaluation. The duration of computing time starts with subscribing to raw image topics and ends with interactive visualization using RViz. It shows that our approach is the fastest (0.0402 ± 0.0021 s), significantly less than the other methods by an order of magnitude. In particular, we divided the running time of the framework in each continuous frame into four stages: Stage 1 (from N.1 to N.3 as annotated in Fig. 1) is the time of image preprocessing, including the time to subscribe the topics of one compressed image pair, rectify and reshape the images (640×360); Stage 2 (N.3–N.4 in Fig. 1) denotes the time to estimate final disparity values; Stage 3 (from N.4 to N.5) shows the time to reproject the disparity map into the 3D surface; and Stage 4 (N.5–N.6 as annotated in Fig. 1) is the time for the interactive visualization using RViz. From the distribution of computing time, it presents that the disparity estimation (Stage 2) always occupies the most time, image preprocessing (Stage 1) and point cloud visualization (Stage 4) take a similar time, and the time of the reprojection operation (Stage 3) is minimal. Fig. 4 demonstrates the reconstruction of 3D surfaces using the proposed framework in a 3D printed liver phantom. It can be seen that our approach performs a smoother surface with fewer outliers than the advanced methods.

3.3. Quantitative comparison on ex vivo dataset

To conduct the quantitative accuracy evaluation, the SCARED dataset [38], captured using porcine cadavers, was adopted to perform 3D surgical scene reconstruction. It contains 2 test datasets with sparse ground truth. After checking the dataset manually, we chose the frames with useful points that are more than thirty percent, so 3018 frames were used for our quantitative evaluation. Image rectification was performed using OpenCV functions, and the rectification error is 1.32 ± 0.46 pixels by calculating all image pairs with 30 random feature points generated by SIFT algorithm [39] on each image pair. Three accuracy-related metrics were chosen to evaluate the reconstruction error, comparing the estimated depth with the provided ground truth, both expressed in millimeters. The metrics include Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Squared Relative Error (SRE), as well as the inference time in single frame [40,41].

$$\mathrm{MAE} = \frac{1}{|D|}\sum_{(x,y)} |d(x,y) - d'(x,y)| \tag{7}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{|D|}\sum_{(x,y)} |d(x,y) - d'(x,y)|^2} \tag{8}$$

$$\mathrm{SRE} = \frac{1}{|D|}\sum_{(x,y)} \frac{|d(x,y) - d'(x,y)|^2}{d'(x,y)} \tag{9}$$

Where D is the set of predicted depth values for each frame, $d(x,y)$ is the predicted depth value related to pixel in position (x,y) and $d'(x,y)$

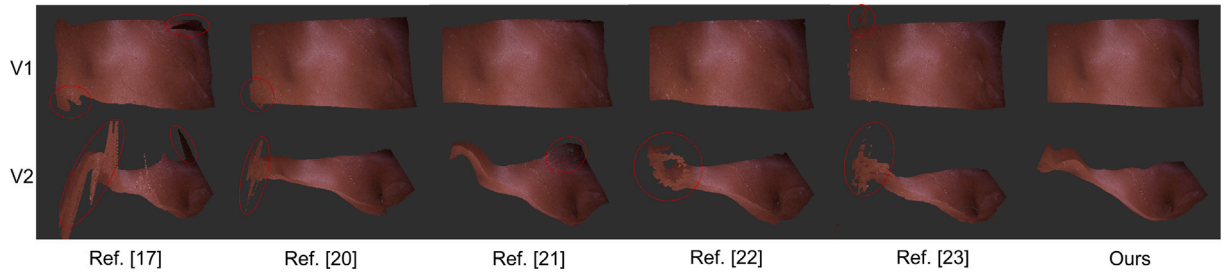

Fig. 4. 3D surfaces of a low-textured liver phantom based on RViz. "V1" represents the front view, while "V2" is the tangential view. Red ellipses mark some unsatisfactory reconstruction areas.

Table 2
Quantitative evaluation of 3D surgical scene reconstruction based on the SCARED dataset. Two test datasets 'D1' and 'D2' are utilized for the comparison. "ALL" means the whole dataset consisting of 3018 frames, "SD" means the Wilcoxon rank-Sum test to compute the significant differences between different state-of-the-art methods and our model, and the result is shown as ns : $0.05 < p \leq 1$, $*$: $0.01 < p \leq 0.05$, $**$: $0.001 < p \leq 0.01$, $***$: $0.0001 < p \leq 0.001$, and $****$: $p \leq 0.0001$. Parameters and FLOPs of the deep learning based models are also provided.

		Ref. [10]	Ref. [17]	Ref. [20]	Ref. [21]	Ref. [22]	Ref. [23]	Ours
Parameters (M)	ALL	N/A	3.68	6.91	10.51	5.31	139.67	**1.06**
FLOPs (G)	ALL	N/A	2330.79	2424.81	2226.82	2590.95	1840.35	**116.01**
MAE (mm)	D1	3.13 ± 2.17	2.63 ± 1.49	2.57 ± 1.49	2.63 ± 1.34	**2.55 ± 1.42**	2.63 ± 1.50	2.57 ± 1.46
	D2	3.03 ± 1.38	3.03 ± 1.45	**2.97 ± 1.51**	3.33 ± 1.60	3.06 ± 1.55	3.13 ± 1.38	2.98 ± 1.50
	ALL	3.10 ± 1.98	2.74 ± 1.49	**2.68 ± 1.50**	2.82 ± 1.45	2.70 ± 1.48	2.77 ± 1.49	2.69 ± 1.48
	SD	****	ns	ns	****	ns	**	
RMSE (mm)	D1	9.88 ± 3.89	5.64 ± 1.38	5.57 ± 1.39	5.51 ± 1.17	5.49 ± 1.24	5.80 ± 1.26	**5.48 ± 1.26**
	D2	9.47 ± 1.78	5.64 ± 1.47	5.54 ± 1.61	6.00 ± 2.05	5.99 ± 2.26	6.67 ± 1.70	**5.42 ± 1.52**
	ALL	9.77 ± 3.44	5.64 ± 1.41	5.56 ± 1.46	5.64 ± 1.49	5.63 ± 1.61	6.04 ± 1.45	**5.47 ± 1.34**
	SD	****	****	ns	****	**	****	
SRE	D1	1.54 ± 1.76	0.41 ± 0.23	0.40 ± 0.25	**0.38 ± 0.19**	0.38 ± 0.21	0.44 ± 0.24	0.38 ± 0.21
	D2	1.60 ± 0.88	0.59 ± 0.44	0.57 ± 0.58	0.79 ± 1.18	0.86 ± 1.73	1.09 ± 1.27	**0.50 ± 0.26**
	ALL	1.56 ± 1.57	0.46 ± 0.31	0.45 ± 0.38	0.50 ± 0.67	0.52 ± 0.96	0.62 ± 0.76	**0.41 ± 0.23**
	SD	****	****	ns	****	*	****	
Inference time (s)	ALL	0.39 ± 0.04	0.78 ± 0.02	0.78 ± 0.03	0.55 ± 0.00	0.92 ± 0.03	0.61 ± 0.01	**0.04 ± 0.00**
	SD	****	****	****	****	****	****	

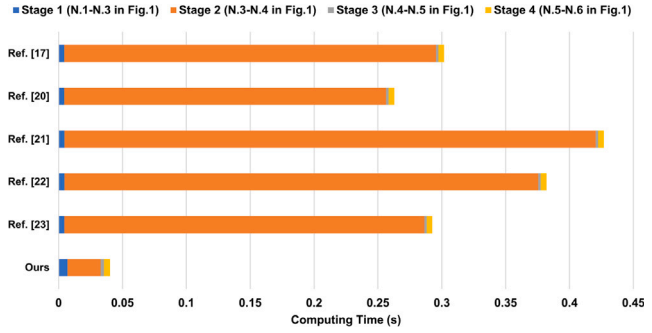

Fig. 5. Mean computing time distribution of our proposed framework in one frame. Stages are defined as the preprocessing, including the time to subscribe topics, rectify and reshape the images (Stage 1, corresponding to N.1–N.3 in Fig. 1), time to estimate disparity values (Stage 2, corresponding to N.3–N.4 in Fig. 1), time to reproject to the 3D surface (Stage 3, corresponding to N.4–N.5 in Fig. 1), and time to visualize using RViz (Stage 4, corresponding to N.5–N.6 in Fig. 1).

is the ground truth of depth value. It can be seen that RMSE is more sensitive to outliers compared with MAE, and SRE is utilized to measure the relative difference.

Table 2 presents the quantitative comparison results of the reconstructed 3D scenes using two SCARED test datasets that consist of 3018 surgical frames, and parameters and FLOPs of the deep learning based methods are also provided. Our model has the smallest volume when referencing the model parameter, and the FLOPs show the lightweight advantage of our model. Furthermore, our approach achieves promising reconstruction accuracy compared with the state-of-the-art methods [10,17,20–23] when calculating the whole dataset (2.69 ± 1.48 mm in MAE, 5.47 ± 1.34 mm in RMSE, and 0.41 ± 0.23 in SRE, respectively). Specifically, our method performs best when considering both RMSE and SRE. Also, the MEA error shows that our method can reach a high accuracy even though the MEA error of [20] is slightly lower than ours. More importantly, our model is one order of magnitude smaller than other methods in the inference time, which makes it possible to run the framework in real time. Next, the statistical results based on WILCOXON Rank-Sum Test show that our method is significantly different from the state-of-the-art methods in the reconstruction accuracy and speed. Lastly, the qualitative reconstruction results of 5 keyframes extracted from different scenes were shown in Fig. 6. To give a comprehensive demonstration of the error distribution, we also provided the 3D error maps at the end. Our method could generate smoother soft tissue surfaces in 3D space compared with other approaches.

Then, we conducted the ablation study for the proposed model to find out the best configuration. Particularly, we divided the results of the ablation study into 3 groups, and the rank-sum test was also performed to compare other possible configurations with our final version,

- In group 1, we explored the effect of the loss function. Different weight combinations of the loss function were set, and it can be noticed that $\lambda_2 = 0$ means that the loss term L_{gra} is removed.
- In group 2, different cost volumes were designed to search for the best construction approach. In our work, we considered both V_{gwc} and V_{dif} to form a combination cost volume. Differently, some previous models chose to generate the cost volume V_{concat} by concatenating the left and right feature maps [17,29], or combining V_{gwc} and V_{concat} [20]. Hence, different architectures of the cost volume were built to explore the effect in this group.
- To show the simplicity of our lightweight model, we modified or added some components in group 3 to check if the extra modules could improve the performance: (1) We noticed that some models [17,18] added Spatial Pyramid Pooling (SPP) module to expand the perceptive field when extracting high-level features, so we implemented this module in the encoder part; (2) We did not downsample the disparity channels after forming the cost volume to keep the larger disparity hypothesis range; (3) We added an upsampling layer in the decoder to generate bigger feature maps; (4) We replaced the LeakyRelu activation with the Relu activation; (5) We built the 3D U-Net modules consisting of three 3×3 convolutional layers with downsampling and upsampling operation to replace the residual blocks.

The results of the ablation study in Table 3 show that the loss function with proper weight setting (λ_1, λ_2 are 1 and 15, respectively) could improve the accuracy compared with the single common loss term L_{abs}. Furthermore, the combination cost volume ($V_{gwc} + V_{dif}$) proposed by us could provide higher accuracy compared with other existing architectures. Finally, the results of group 3 present that complexing our model by introducing extra components does not enhance the reconstruction accuracy and it slows down the speed.

3.4. Qualitative evaluation on in vivo clinical data

To measure the potential of the proposed approach in clinical practice, a complete surgical operation named a radical prostatectomy

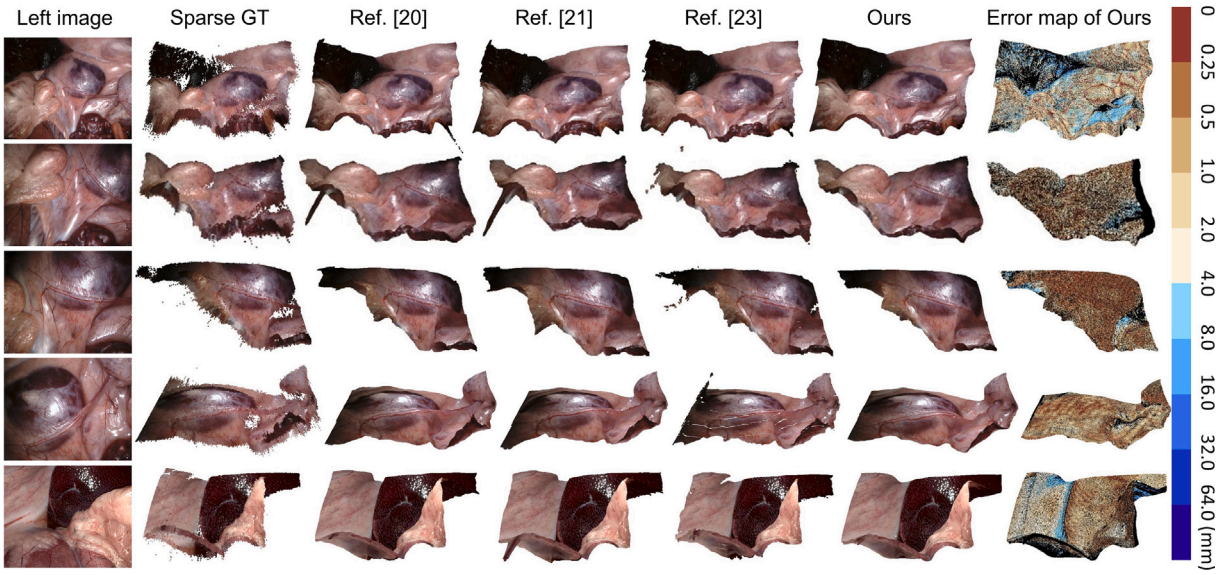

Fig. 6. Qualitative comparison of reconstruction results using SCARED test frames. "Sparse GT" means the sparse ground truth of point clouds provided in the dataset, and the last column gives the 3D error maps of our reconstructed scenes compared with the ground truth.

Table 3
Ablation study based on the SCARED dataset. "SD" means the Wilcoxon rank-sum test to check the statistical differences between different configurations and our final model, and the result is annotated as ns : $0.05 < p \leq 1$, $*$: $0.01 < p \leq 0.05$, $**$: $0.001 < p \leq 0.01$, $***$: $0.0001 < p \leq 0.001$, and $****$: $p \leq 0.0001$. The layer ID in group 3 can be found in Table 1.

Group 1: Different weights of the loss function				
	MAE (mm)/SD	RMSE (mm)/SD	SRE/SD	Inference time (s)/SD
$\lambda_1 = 1, \quad \lambda_2 = 0$	2.80 ± 1.49/***	5.53 ± 1.34/*	0.42 ± 0.23/ns	0.04 ± 0.00/ns
$\lambda_1 = 1, \quad \lambda_2 = 5$	2.81 ± 1.49/***	5.54 ± 1.34/*	0.43 ± 0.27/*	0.04 ± 0.00/ns
$\lambda_1 = 1, \quad \lambda_2 = 10$	2.87 ± 1.51/****	5.57 ± 1.36/***	0.43 ± 0.23/**	0.04 ± 0.00/ns
$\lambda_1 = 1, \quad \lambda_2 = 20$	2.77 ± 1.46/**	5.52 ± 1.31/*	0.43 ± 0.24/*	0.04 ± 0.00/ns
$\lambda_1 = 1, \quad \lambda_2 = 15$ (Ours)	**2.69 ± 1.48**	**5.47 ± 1.34**	**0.41 ± 0.23**	**0.04 ± 0.00**
Group 2: Different combination modes of the cost volume				
	MAE (mm)/SD	RMSE (mm)/SD	SRE/SD	Inference time (s)/SD
V_{gwc}	2.81 ± 1.46/****	5.57 ± 1.31/***	0.44 ± 0.27/***	0.04 ± 0.00/ns
V_{dif}	2.74 ± 1.47/ *	5.50 ± 1.32/ns	0.42 ± 0.23/ns	0.04 ± 0.00/ns
V_{concat}	2.91 ± 1.49/****	5.61 ± 1.34/****	0.44 ± 0.25/****	0.04 ± 0.00/ns
$V_{gwc} + V_{concat}$	2.77 ± 1.47/**	5.51 ± 1.33/*	0.42 ± 0.24/ns	0.04 ± 0.00/ns
$V_{gwc} + V_{dif}$ (Ours)	**2.69 ± 1.48**	**5.47 ± 1.34**	**0.41 ± 0.23**	**0.04 ± 0.00**
Group 3: Modify the different components of the proposed architecture				
	MAE (mm)/SD	RMSE (mm)/SD	SRE/SD	Inference time (s)/SD
Add SPP between layers 8 and 9	2.85 ± 1.48/****	5.70 ± 1.45/****	0.52 ± 0.61/****	0.04 ± 0.00/ns
No striding in layer 15	2.71 ± 1.53/ns	5.48 ± 1.37/ns	0.42 ± 0.23/ns	0.07 ± 0.00/****
Upsampling before layer 19	2.85 ± 1.48/****	5.57 ± 1.31/***	0.43 ± 0.23/****	0.13 ± 0.00/****
Adopt Relu activation in the encoder	2.69 ± 1.50/ns	5.50 ± 1.33/ns	0.43 ± 0.26/ns	0.04 ± 0.00/ns
Adopt 3D U-Net modules in the decoder	2.78 ± 1.47/**	5.49 ± 1.34/ns	0.41 ± 0.22/ns	0.09 ± 0.00/****
Simplest configuration (Ours)	**2.69 ± 1.48**	**5.47 ± 1.34**	**0.41 ± 0.23**	**0.04 ± 0.00**

with lymphadenectomy was recorded with a 3D HD video recorder (HVO-3300MT, SONY, Tokyo) based on the da Vinci Xi surgical system at European Institute of Oncology (IEO, Milan, Italy). The endoscope calibration process was also performed at the hospital to obtain the intrinsic and extrinsic parameters. The keyframes from four different surgical phases divided by a senior surgeon were extracted and performed the scene reconstruction using our pipeline, as shown in Fig. 7. We demonstrated the qualitative evaluation using the comparison study with five deep learning based advanced methods. Furthermore, we also tested the reconstruction quality in the challenging surgical scenes, including smoke, specularity and blur. It can be seen that our approach could get smoother 3D surfaces compared with other existing methods in different phases as well as the challenging scenes. Also, other methods are prone to outliers from boundary regions while ours is not. The scene with smoke is more challenging than that with specularity and blur in our observation.

To promote the development of the medical scene reconstruction community, we further created a clinical dataset with annotated surgical context information (phases, steps, and types of instruments). This dataset contains 145,694 image pairs with a resolution of 1920 × 1080, and calibration parameters are also provided for image rectification and triangulation. It could be utilized for the evaluation of surgical scene reconstruction models, also the training of unsupervised learning based networks. Moreover, it provides the possibility for surgical workflow recognition based on stereo images or point clouds. This dataset has been released online.

The collection of data was in accordance with the ethical standards of the Istituto Europeo di Oncologia and with the 1964 Helsinki declaration, revised in 2000. No personal data was recorded. All the subjects involved in this research were informed and agreed to data treatment before the intervention.

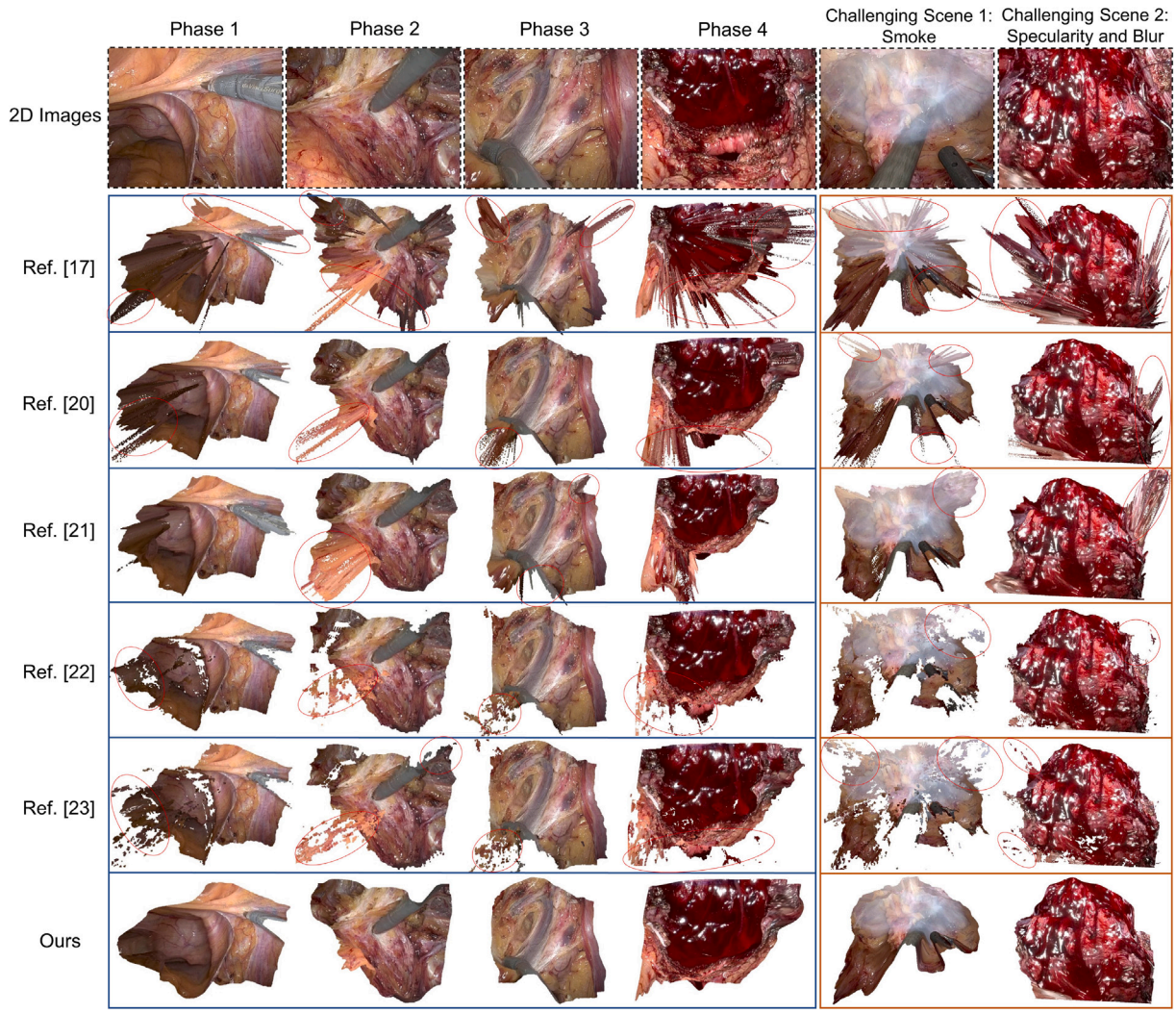

Fig. 7. Reconstruction demonstration of surgical scenes in a radical prostatectomy with lymphadenectomy. The first row represents the rectified left images photoed from the stereo endoscope. Examples from four phases (Phase 1: collapse of the peritoneum, Phase 2: prostate removal, Phase 3: lymphadenectomy, Phase 4: anastomosis) are extracted in sequence, and the challenging surgical scenes including smoke, specularity and blur are also provided. The red ellipses mark some undesirable reconstruction areas.

4. Discussion

Integrating scene reconstruction into surgical robot platforms is significant to promote clinical application. However, we found that there is a lack of details on the integrated approach when surveying the relevant literature. Hence, a real-time surgical scene reconstruction framework was proposed in this paper to enhance the visualization of intra-operative scenes for the safety of surgery. It was integrated into da Vinci Research Kit, a popular surgical system in RAMIS today. This framework is developed based on ROS, which ensures the safety of signal transmission and its high possibility of migration to other robot platforms. The specific time distribution of the 3D reconstruction pipeline can be seen in Fig. 5, and it shows that the speed of our method is significantly faster than other methods. The reasons could be explained by that on the one hand, to squeeze the model volume and speed up the inference time, we adopted four consecutive residual blocks in our decoder, which is lighter than other decoders such as the stacked 3D hourglass network which is more time-consuming [17,20]; On the other hand, we adopted the strided convolution operation to downsample the dimension of cost volume at the beginning of our decoder, which can further accelerate the time because of the smaller resolution. Additionally, Fig. 5 also presents that disparity estimation is the most time-consuming step in the framework. To perform a real-time reconstruction, there are two proper approaches to accelerate the pipeline, one is to downsample the resolution of the raw images, and another one is to adopt a lightweight model especially in the 3D decoder part since 3D convolution takes much time.

The SCARED dataset was adopted to conduct the quantitative comparison study between our network and other advanced methods since it contains sufficient surgical images with ground truth. The comparison result in Table 2 shows that our model can provide reliable reconstruction quality in surgical scenes, and real-time performance is also a highlight. The rank-sum test presents the statistical differences between our approach and others when calculating different metrics. Furthermore, the ablation study in Table 3 shows the feature of our lightweight model: simplification of the model will not deteriorate the reconstruction quality. More specifically, the contour information of the scene can be beneficial to the accuracy of the model to a certain extent if the proper weight was set in the loss function, while it may also impair the performance with an inappropriate weight. Next, considering both the feature correlation and feature difference when establishing the cost volume could enhance the estimation performance, while concatenating the feature maps directly as the cost volume presents the worst result. We could also notice that the SPP module cannot improve

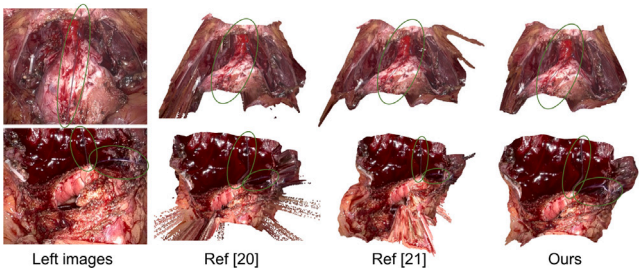

Fig. 8. An instance to show the limitation of our work. All methods fail to separate the thin surgical sutures from the background. The green ellipses mark the position of the sutures.

the accuracy of our network. Moreover, using the feature maps with higher resolution not only does not help the accuracy of the model but also slows down the inference speed. Also, adopting the 3D U-Net modules with the downsampling and upsampling operations does not present a promising result compared to using the 3D residual blocks in our case.

Insufficient stereo endoscopic datasets hinder the evaluation of 3D reconstruction methods in the medical field, so we made a clinical dataset by capturing a complete surgical operation at an oncology hospital. The qualitative evaluation based on this clinical dataset in Fig. 7 shows the potential of our approach in the real surgical environment. We noticed that other advanced methods can perform better when evaluating the public SCARED test datasets, while the reconstruction quality of the clinical dataset is not satisfactory. Some apparent outliers can be observed from the results generated by other methods, especially in scenes with smoke. As a comparison, our method can still keep a high reconstruction quality with a smoother surface and fewer outliers in these clinical scenes.

Nevertheless, a limitation of the proposed framework can be found in the evaluation of the clinical dataset. The environment inside the human body is always complex and challenging. For instance, we noticed that all methods occasionally failed to extract some small instruments (such as the thin surgical suture) from the background, as shown in Fig. 8. It can be seen that the reconstructed sutures were attached to the background, while there should be a certain distance between the sutures and the background in the realistic scene. On the one hand, we think that there is a lack of annotation for such small targets in the training set, so the model has unsatisfactory adaptability when such targets appear in the test set. Adding annotations containing small targets in the training set may enhance the performance of the model. On the other hand, we could also consider enhancing the fine feature extraction ability of the model in a complex surgical environment, and a possible solution is to concatenate special layers focusing on extracting local fine features when constructing the cost volume, although it will slow down the prediction speed of the model.

Another limitation comes from the stereo matching itself in surgical scenes. The disparity calculation relies on searching for the corresponding pixels on the stereo images. However, human organs do not have apparent texture in some areas, which increases the difficulty to estimate the accurate disparity values in these textureless areas. One possible solution is to augment our training images by introducing some textureless scenes to enhance the adaptability of our network in dealing with such regions. Also, more depth estimation approaches, such as the feature points based monocular estimation, could be introduced to compare the applicability in real surgical applications.

5. Conclusion

To conclude, the comprehensive evaluation results show that the proposed framework can not only achieve promising results in the endoscopic scene reconstruction quality, but also real-time performance is a significant highlight in this work. It provides the possibility and potential to be implemented in the clinical procedure. The next step is to extract interesting soft tissues from the reconstructed 3D surgical scenes. A promising strategy is to segment the region of interest from the endoscopic images and match the generated masks with the predicted disparity maps. Both optimization based segmentation method [42] and deep learning based method [43] will be compared and explored to implement a high-performance segmentation in accuracy and speed. In this way, we can perform the registration between pre-operative models and intra-operative soft tissues for image-guided surgery [44–46].

Declaration of competing interest

None declared.

References

[1] J. Koskinen, M. Torkamani-Azar, A. Hussein, A. Huotarinen, R. Bednarik, Automated tool detection with deep learning for monitoring kinematics and eye-hand coordination in microsurgery, Comput. Biol. Med. 141 (2022) 105121.
[2] S. Moccia, L. Migliorelli, V. Carnielli, E. Frontoni, Preterm infants' pose estimation with spatio-temporal features, IEEE Trans. Biomed. Eng. 67 (8) (2019) 2370–2380.
[3] C. da Costa Rocha, N. Padoy, B. Rosa, Self-supervised surgical tool segmentation using kinematic information, in: 2019 International Conference on Robotics and Automation, ICRA, IEEE, 2019, pp. 8720–8726.
[4] A. Casella, S. Bano, F. Vasconcelos, A.L. David, D. Paladini, J. Deprest, E. De Momi, L.S. Mattos, S. Moccia, D. Stoyanov, Learning-based keypoint registration for fetoscopic mosaicking, 2022, arXiv preprint arXiv:2207.13185.
[5] Y. Dai, J. Wu, Y. Fan, J. Wang, J. Niu, F. Gu, S. Shen, MSEva: A musculoskeletal rehabilitation evaluation system based on EMG signals, ACM Trans. Sensor Netw. 19 (1) (2022) 1–23.
[6] J. Lee, A.C. Gordon, H. Kim, W. Park, S. Cho, B. Lee, A.C. Larson, E.A. Rozhkova, D.-H. Kim, Targeted multimodal nano-reporters for pre-procedural MRI and intra-operative image-guidance, Biomaterials 109 (2016) 69–77.
[7] Z. Guo, Z. Dong, K.-H. Lee, C.L. Cheung, H.-C. Fu, J.D. Ho, H. He, W.-S. Poon, D.T.-M. Chan, K.-W. Kwok, Compact design of a hydraulic driving robot for intraoperative MRI-guided bilateral stereotactic neurosurgery, IEEE Robot. Autom. Lett. 3 (3) (2018) 2515–2522.
[8] D. Stoyanov, M.V. Scarzanella, P. Pratt, G.-Z. Yang, Real-time stereo reconstruction in robotically assisted minimally invasive surgery, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2010, pp. 275–282.
[9] V. Penza, J. Ortiz, L.S. Mattos, A. Forgione, E. De Momi, Dense soft tissue 3D reconstruction refined with super-pixel segmentation for robotic abdominal surgery, Int. J. Comput. Assist. Radiol. Surg. 11 (2) (2016) 197–206.
[10] A. Geiger, M. Roser, R. Urtasun, Efficient large-scale stereo matching, in: Asian Conference on Computer Vision, Springer, 2010, pp. 25–38.
[11] G. Zampokas, K. Tsiolis, G. Peleka, I. Mariolis, S. Malasiotis, D. Tzovaras, Real-time 3D reconstruction in minimally invasive surgery with quasi-dense matching, in: 2018 IEEE International Conference on Imaging Systems and Techniques, IST, IEEE, 2018, pp. 1–6.
[12] H. Zhou, J. Jagadeesan, Real-time dense reconstruction of tissue surface from stereo optical video, IEEE Trans. Med. Imaging 39 (2) (2019) 400–412.
[13] B. Huang, J.-Q. Zheng, A. Nguyen, D. Tuch, K. Vyas, S. Giannarou, D.S. Elson, Self-supervised generative adversarial network for depth estimation in laparoscopic images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 227–237.
[14] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.
[15] H. Luo, C. Wang, X. Duan, H. Liu, P. Wang, Q. Hu, F. Jia, Unsupervised learning of depth estimation from imperfect rectified stereo laparoscopic images, Comput. Biol. Med. 140 (2022) 105109.
[16] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, T. Brox, Flownet: Learning optical flow with convolutional networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2758–2766.
[17] J.-R. Chang, Y.-S. Chen, Pyramid stereo matching network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5410–5418.
[18] G. Yang, J. Manela, M. Happold, D. Ramanan, Hierarchical deep stereo matching on high-resolution images, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5515–5524.

[19] X. Cheng, Y. Zhong, M. Harandi, Y. Dai, X. Chang, H. Li, T. Drummond, Z. Ge, Hierarchical neural architecture search for deep stereo matching, Adv. Neural Inf. Process. Syst. 33 (2020) 22158–22169.
[20] X. Guo, K. Yang, W. Yang, X. Wang, H. Li, Group-wise correlation stereo network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3273–3282.
[21] X. Gu, Z. Fan, S. Zhu, Z. Dai, F. Tan, P. Tan, Cascade cost volume for high-resolution multi-view stereo and stereo matching, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 2495–2504.
[22] D. Garg, Y. Wang, B. Hariharan, M. Campbell, K.Q. Weinberger, W.-L. Chao, Wasserstein distances for stereo disparity estimation, Adv. Neural Inf. Process. Syst. 33 (2020) 22517–22529.
[23] B. Liu, H. Yu, G. Qi, GraftNet: Towards domain generalized stereo matching with a broad-spectrum and task-oriented feature, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 13012–13021.
[24] P. Brandao, D. Psychogyios, E. Mazomenos, D. Stoyanov, M. Janatka, HAPNet: hierarchically aggregated pyramid network for real-time stereo matching, Comput. Methods Biomech. Biomed. Eng.: Imaging Vis. 9 (3) (2021) 219–224.
[25] Z. Chen, S. Terlizzi, T. Da Col, A. Marzullo, M. Catellani, G. Ferrigno, E. De Momi, Robot-assisted ex vivo neobladder reconstruction: preliminary results of surgical skill evaluation, Int. J. Comput. Assist. Radiol. Surg. 17 (12) (2022) 2315–2323.
[26] A. Mariani, G. Colaci, T. Da Col, N. Sanna, E. Vendrame, A. Menciassi, E. De Momi, An experimental comparison towards autonomous camera navigation to optimize training in robot assisted surgery, IEEE Robot. Autom. Lett. 5 (2) (2020) 1461–1467.
[27] M. Rajakumaran, S. Ramabalan, Security for the networked robot operating system for biomedical applications, J. Med. Imag. Health Inform. 11 (12) (2021) 2937–2949.
[28] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[29] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, A. Bry, End-to-end learning of geometry and context for deep stereo regression, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 66–75.
[30] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1440–1448.
[31] I. Alhashim, P. Wonka, High quality monocular depth estimation via transfer learning, 2018, arXiv preprint arXiv:1812.11941.
[32] M. Menze, C. Heipke, A. Geiger, Object scene flow, ISPRS J. Photogramm. Remote Sens. 140 (2018) 60–76.
[33] A. Geiger, P. Lenz, R. Urtasun, Are we ready for autonomous driving? The kitti vision benchmark suite, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 3354–3361.
[34] T. Schops, J.L. Schonberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, A. Geiger, A multi-view stereo benchmark with high-resolution images and multi-camera videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3260–3269.
[35] J. Cartucho, S. Tukra, Y. Li, D. S. Elson, S. Giannarou, VisionBlender: a tool to efficiently generate computer vision datasets for robotic surgery, Comput. Methods Biomech. Biomed. Eng.: Imaging Vis. 9 (4) (2021) 331–338.
[36] P.E. Edwards, D. Psychogyios, S. Speidel, L. Maier-Hein, D. Stoyanov, SERV-CT: A disparity dataset from cone-beam CT for validation of endoscopic 3D reconstruction, Med. Image Anal. 76 (2022) 102302.
[37] V. Venugopal, J. Joseph, M.V. Das, M.K. Nath, DTP-net: A convolutional neural network model to predict threshold for localizing the lesions on dermatological macro-images, Comput. Biol. Med. 148 (2022) 105852.
[38] M. Allan, J. Mcleod, C. Wang, J.C. Rosenthal, Z. Hu, N. Gard, P. Eisert, K.X. Fu, T. Zeffiro, W. Xia, et al., Stereo correspondence and reconstruction of endoscopic data challenge, 2021, arXiv preprint arXiv:2101.01133.
[39] D.G. Lowe, Object recognition from local scale-invariant features, in: Proceedings of the Seventh IEEE International Conference on Computer Vision, 2, Ieee, 1999, pp. 1150–1157.
[40] D. Eigen, C. Puhrsch, R. Fergus, Depth map prediction from a single image using a multi-scale deep network, Adv. Neural Inf. Process. Syst. 27 (2014).
[41] C. Zhao, Q. Sun, C. Zhang, Y. Tang, F. Qian, Monocular depth estimation based on deep learning: An overview, Sci. China Technol. Sci. 63 (9) (2020) 1612–1627.
[42] A. Qi, D. Zhao, F. Yu, A.A. Heidari, Z. Wu, Z. Cai, F. Alenezi, R.F. Mansour, H. Chen, M. Chen, Directional mutation and crossover boosted ant colony optimization with application to COVID-19 X-ray image segmentation, Comput. Biol. Med. 148 (2022) 105810.
[43] K. Hu, L. Zhao, S. Feng, S. Zhang, Q. Zhou, X. Gao, Y. Guo, Colorectal polyp region extraction using saliency detection network with neutrosophic enhancement, Comput. Biol. Med. 147 (2022) 105760.
[44] H. Luo, D. Yin, S. Zhang, D. Xiao, B. He, F. Meng, Y. Zhang, W. Cai, S. He, W. Zhang, et al., Augmented reality navigation for liver resection with a stereoscopic laparoscope, Comput. Methods Programs Biomed. 187 (2020) 105099.
[45] T. Zhu, S. Jiang, Z. Yang, Z. Zhou, Y. Li, S. Ma, J. Zhuo, A neuroendoscopic navigation system based on dual-mode augmented reality for minimally invasive surgical treatment of hypertensive intracerebral hemorrhage, Comput. Biol. Med. 140 (2022) 105091.
[46] F. Giannone, E. Felli, Z. Cherkaoui, P. Mascagni, P. Pessaux, Augmented reality and image-guided robotic liver surgery, Cancers 13 (24) (2021) 6268.