Stereo Dense Scene Reconstruction and Accurate Localization for Learning-Based Navigation of Laparoscope in Minimally Invasive Surgery

Ruofeng Wei®, Graduate Student Member, IEEE, Bin Li®, Graduate Student Member, IEEE, Hangjie Mo®, Bo Lu®, Yonghao Long®, Bohan Yang®, Qi Dou, Yunhui Liu®, Fellow, IEEE, and Dong Sun®, Fellow, IEEE

Abstract—Objective: The computation of anatomical information and laparoscope position is a fundamental block of surgical navigation in Minimally Invasive Surgery (MIS). Recovering a dense 3D structure of surgical scene using visual cues remains a challenge, and the online laparoscopic tracking primarily relies on external sensors, which increases system complexity. Methods: Here, we propose a learning-driven framework, in which an image-guided laparoscopic localization with 3D reconstructions of anatomical structures is obtained. To reconstruct the structure of the whole surgical environment, we first fine-tune a learning-based stereoscopic depth perception method, which is robust to texture-less and variant soft tissues, for depth estimation. Then, we develop a dense reconstruction algorithm to represent the scene by surfels, estimate the laparoscope poses and fuse the depth into a unified reference coordinate for tissue reconstruction. To estimate poses of new laparoscope views, we achieve a coarse-to-fine localization method, which incorporates our reconstructed model. Results: We evaluate the reconstruction method and the localization module on three datasets, namely, the stereo correspondence and reconstruction of endoscopic data (SCARED), the ex-vivo data collected with the Universal Robot (UR) and Karl Storz Laparoscope, and the in-vivo DaVinci robotic surgery dataset, where the reconstructed structures have rich details of surface texture with an error under 1.71 mm and the localization module can accurately track the laparoscope with images as input. Conclusions: Experimental results demonstrate the superior performance of the proposed method in anatomy reconstruction and laparoscopic localization. Significance: The proposed framework can be potentially extended to the current surgical navigation system.

Index Terms—Surgical navigation, tissue reconstruction, laparoscope localization, endoscope.

Manuscript received 24 December 2021; revised 22 March 2022; accepted 23 July 2022. Date of publication 29 July 2022; date of current version 20 January 2023. This work was supported by the Research Grants Council of the Hong Kong Special Administrative Region, China under Grants T42-409/18-R and 11211421. (Ruofeng Wei and Bin Li contributed equally to this work.) (Corresponding author: Dong Sun.)

Ruofeng Wei and Hangjie Mo are with the Department of Biomedical Engineering, City University of Hong Kong, Hong Kong.
Bin Li, Bohan Yang, and Yunhui Liu are with the Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong, Hong Kong.
Bo Lu was with the Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong, Hong Kong. He is now with the Robotics and Microsystems Center, School of Mechanical and Electric Engineering, Soochow University, China.
Yonghao Long and Qi Dou are with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong.
Dong Sun is with the Department of Biomedical Engineering, City University of Hong Kong, Hong Kong 518057, SAR China (e-mail: medsun@cityu.edu.hk).

This article has supplementary downloadable material available at https://doi.org/10.1109/TBME.2022.3195027, provided by the authors.

Digital Object Identifier 10.1109/TBME.2022.3195027

INTRODUCTION

MINIMALLY invasive surgery (MIS) has flourished over the past decade due to its small surgical trauma, less pain and shorter recovery [1]. In MIS, a laparoscope is utilized and inserted through a trocar into the human body to provide surgeons with visual information about surgical scene [2]. To assist the laparoscopic surgery, surgical navigation system that generally offers internal structural information for intra-operative planning and employs external trackers for laparoscope localization is popularly integrated into the existing platforms [3]. However, compared to conventional open surgeries, laparoscopic images observed in MIS are usually two-dimensional (2D) and the view of surgical field provided by laparoscope is commonly limited [4], which significantly decreases the understanding of the internal anatomy and negatively affects the practical operations. Moreover, extra tracking sensors may add complexity to surgical systems used in Operating Room (OR).

To improve the visualization of surgeons during surgery, depth information of the tissue surface needs to be extracted from the 2D stereo laparoscope. During the past decade, numerous depth estimation algorithms have been presented to provide depth measurements by establishing the correspondence between rectified left and right images’ pixels, and the result can be adopted for 3D reconstruction [5]. Considering the characteristics of tissue surface [6], Stoyanov et al. [7] used salient points based on Lucas-Kanade to establish sparse feature matching. However,
this method can only be operated at 10 fps estimation speed for images with $360 \times 288$ resolution. Zhou et al. [8] presented post-processing refinement steps, such as removing outliers, hole filling and disparities smoothing, to address low texture problems. However, the zero-mean normalized cross-correlation (ZNCC)-based local matching part only considered 100 candidate disparity values. Recently, the learning-based stereo depth estimation method was proposed. Liang et al. [9] used the convolution neural network (CNN) to extract features and compute similarity of each pixel for feature matching. Li et al. [10] proposed a transformer-based method that considered the sequential nature of videos in performing feature matching, running at 2 fps for $640 \times 512$ image pairs. However, these methods produce suboptimal depth information because of either poor texture and unique color of tissues or insufficient disparity candidates. In medicine, Huang et al. [11] proposed a self-supervised adversarial depth estimation method for laparoscopic binocular images. However, the estimation accuracy of this method is relatively low (e.g., 17 mm on the SCARED data), so the method cannot be used for further reconstructions. Karaoglou et al. [12] proposed a two-step domain-adaptive approach to estimate the depth of a bronchoscopy scene to overcome the lack of labeled data in surgical settings. But it constructs synthetic bronchoscopy images with perfect ground-truth depths for training depth estimation networks, which are not available in other scenarios. Shao et al. [13] considered the brightness variation among laparoscopic streams to aid in the task of depth estimation. However, this work yields scale-free depths that are arbitrarily scaled relative to the real world.

Furthermore, to provide extensive views of surgical site for surgeons, a simultaneous localization and mapping (SLAM)-based reconstruction module is utilized, which can enlarge the portion of reconstructed cavity by dynamically moving the laparoscope and fusing the 3D surfaces reconstructed at different time. Chen et al. [14] extended the SLAM algorithm to recover the sparse point clouds of the tissue surface. However, this method required the use of Poisson surface reconstruction method to fit the sparse point for inferring the tissue surface. Mahmoud et al. [15] embedded a multi-view reconstruction approach into the existing SLAM system, but the reconstruction results were not smooth and dense enough for surgical visualization and navigation. Marmol et al. [16] also combined the multi-view stereo method and SLAM module for 3D reconstruction, but it required an external camera and odometry data in surgical robot to calculate the arthroscope’s localization. In this paper, we propose a reconstruction method, which can estimate the online depth of the surgical scene and reconstruct large-scale, smooth, and dense 3D surface anatomical structures of tissues among the view only based on stereo images from the laparoscope.

After constructing the 3D structure of surgical scene, the surgeons can navigate in the environment and automatically localize the laparoscope within a given view. Traditional methods using external trackers such as optical tracking system [17] and electromagnetic tracking system [18] may increase the system complexity when tracking the position and orientation of the camera, whereas the direct positional relationship between the laparoscope and the surgical scene cannot be provided [15]. Given the recent advances in autonomous driving, several learning-based visual localization methods, which can recover environments and camera poses, were proposed [19] [20]. However, estimating the pose of laparoscope using only the images is scarce in surgical navigation because of the texture-less and complex geometry nature of surgical tissues. To locate the laparoscope only using images, we creatively combine the dense 3D model from our reconstruction module with the laparoscopic localization method.

In this paper, we propose a learning-driven framework to recover the dense 3D structure of the surgical scene and estimate the laparoscopic pose of the new view. The main contribution of our work is fourfold.

First, we fine-tune a learning-based depth estimation module for dense depth computation per single frame using supervised and unsupervised methods from surgical data. It can be applied to challenging surgical scenarios such as tissues with textureless and monochromatic surfaces.

Second, to reconstruct the entire surgical scene, we propose a dense visual reconstruction algorithm that utilizes surfeels to efficiently represent 3D structures and can simultaneously compute camera poses. It utilizes only the stereoscopic images from laparoscopes, thus completing the entire processes from online depth estimation to reconstruction of dense surgical scenes.

Third, based on the reconstructed dense 3D structure, we propose a laparoscopic localization module to achieve coarse-to-fine pose estimation, where a knowledge distillation strategy is used to train an efficient feature extraction network.

Finally, based on the SCARED dataset, our in-vivo Da Vinci robotic surgery dataset, as well as self-collected ex-vivo phantom and tissue-based data with their 3D anatomy ground truths obtained using structure light techniques, we performed quantitative and qualitative experiments. The corresponding results demonstrate the accuracy and effectiveness of our proposed reconstruction and localization module, showing its potential application in surgical navigation systems.

The remainder of this paper is organized as follows. Section II introduces the proposed framework systematically. Section III presents the experimental procedures. Section IV evaluates the proposed method through experiments based on different datasets. Discussions of some key issues are provided in Section V. Finally, conclusions and future works are given in Section VI.

**II. METHODOLOGY**

**A. Overview of Framework**

Fig. 1(a) shows an overview of the proposed stereo-dense reconstruction and laparoscopic tracking framework. The rectified left $(L)$ and right $(R)$ RGB image at timestamp $t \in [0, T]$ are defined as $L_t$ and $R_t$, respectively. In Section II-B, by fine-tuning a Hierarchical Deep Stereo Matching (HSM) network [21], the disparity map $P_t$ is first computed using $L_t$ and $R_t$, which is then converted to a depth image $D_t$. In Section II-C, the estimated depth $D_t$ and the corresponding left frame $L_t$ from $t = 0$ to $t = T$ are input into a dense visual reconstruction algorithm that
Fig. 1. Overview of the proposed method. (a) Workflow of our proposed stereo dense reconstruction method and its application to laparoscope tracking. (b) The process of the learning-based depth estimation module. (c) The training workflow of the feature extraction network in laparoscopic localization.

Fig. 2. Examples of the SACRED dataset, our ex-vivo phantom data and the in-vivo DaVinci robotic surgery dataset.

recovers the entire 3D structure of the tissue surface. Notably, both the depth estimation network and reconstruction algorithm are designed to achieve real-time performance. Finally, by combining the scale-aware 3D model of the surgical scene with a visual tracking method, a laparoscopic localization module can be formulated to estimate the laparoscopic pose of a new given frame, which is presented in Section II-D.

B. Learning-Based Depth Estimation for Single Frame

Considering the poor and homogeneous textures and unique color of tissue appearance shown in Fig. 2, we find that learned features with large receptive fields and multi-scale properties will help establish accurate pixel-level correspondence between left and right binocular images. Then, given that generating high-resolution textures is important to help clinicians make a diagnosis, a large number of candidate disparity values are required; thus, a high-resolution cost volume must be handled. In this case, we choose the Hierarchical Deep Stereo Matching (HSM) network as the single-frame disparity estimation module, and then transfer the estimated disparity map to the depth image. The HSM network uses a U-Net (encoder-decoder) architecture to efficiently extract features with different levels, the encoder part of which is followed by 4 spatial pyramid pooling (SPP) [22] layers to broaden the receptive fields. After feature extraction, a 3D convolution-based feature decoder is utilized to process high resolution cost volumes more efficiently. Considering that HSM is designed for high-resolution images, it can estimate depth information more accurately by providing more candidate disparity values in computing feature volume. The detailed structure is shown in Fig. 1(b).

Given that publicly available datasets with depth annotations are scarce in surgical scene, an expert model \( f_{pre} \) of HSM can be pretrained by using the autonomous driving dataset KITTI [23], which is a commonly used training dataset for stereo depth estimation network. To alleviate the domain gap between driving data KITTI and surgical scene, we first use the SERV-CT [24] tissue data and a small amount of SACRED data, including endoscopic images and corresponding depth annotations, to supervised fine-tune the expert model, and then try the unsupervised method [25] to continue building a refined depth estimation network \( f_{unsup} \) by using the warping-based view generation loss.

For supervised fine-tuning, we employ a multiscale L1 disparity regression loss to refine the pretrained model. The fine-tuning loss \( \mathcal{L}_{sup} \) is defined as:

\[
\mathcal{L}_{sup} = \sum_{m=0}^{M-1} w_m \cdot \mathcal{L} \left( P^m_{sup} - P_{gt} \right)
\]

where \( M \) is the number of scale, \( w_m \) denotes the weighting factor for the \( m \)-th scale, \( \mathcal{L} \) is the smooth L1 loss, \( P^m_{sup} \) represents the \( m \)-th disparity map predicted by the supervised fine-tuning HSM model, and \( P_{gt} \) is the ground truth disparity.
Then, the objective of our unsupervised fine-tuning task can be formulated as the optimization of the following surgical scene-aware loss function:

\[
L_{\text{unsup}} = L_{\text{photo}} + \alpha_1 L_{\text{dis}} + \alpha_2 L_s
\]  

(2)

where \( \alpha_1 \) and \( \alpha_2 \) are weighting factors.

In (2), the unsupervised loss has three main terms expressed as follows.

**Term 1:** \( L_{\text{photo}} \) denotes the photometric loss computed from the new warping-based view synthesis with the predicted disparity map. The rectified right image \( R \) is first warped to the left view using the disparity estimated by bilinear sampling to obtain the warped right image \( R^w \), which is the calculation of the left image except for the occlusion area. Then, the photometric loss is given by:

\[
L_{\text{photo}} = \frac{1}{K} \sum_{i,j} O_{(i,j)} \cdot |L_{(i,j)} - R_{(i,j)}^w|
\]

(3)

where \( K \) is the total number of pixels, the subscript \( i \) and \( j \) denote the value at the \( i \)-th row and the \( j \)-th column of the image or disparity map, respectively, \( O \) is the ground truth occlusion mask, \( L \) is the corresponding left image, and \( P_{\text{unsup}} \) represents the disparity map estimated by \( f_{\text{unsup}} \).

**Term 2:** \( L_{\text{dis}} \) is a regularization loss used to make the newly predicted disparity map close to the predictions of the supervised fine-tuning model. The term is derived as:

\[
L_{\text{dis}} = \frac{1}{K} \sum_{i,j} (1 - O_{(i,j)} + \alpha_3) \cdot |P_{\text{unsup}}(i,j) - P_{\text{sup}}(i,j)|
\]  

(4)

where \( \alpha_3 \) is the regularization coefficient.

**Term 3:** \( L_s \) is the edge-aware smoothness loss used to regularize the disparity smoothness, expressed as:

\[
L_s = \frac{1}{K} \sum_{i,j} |\nabla P_{\text{unsup}}(i,j) - \nabla P_{\text{sup}}(i,j)|
\]  

(5)

where \( \nabla \) denotes the gradients of disparity.

Using the supervised and unsupervised fine-tuning procedures described above, the HSM model can estimate textureless and monochromatic robust disparity maps of surgical scenes. Afterwards, the estimated disparity can be transformed into a depth image \( D \) using the stereo calibration information.

### C. Dense Visual Reconstruction of Whole Surgical Scene

In order to reconstruct the whole surgical scene, the estimated depth of single frame at different time will be gradually fused. We adopt an unordered list of surfels \( S \) [26] which is more memory efficient to represent the 3D structure of tissue surface, where each surfel \( s \) contains following attributes: a 3D point \( v = (x, y, z)^T \in \mathbb{R}^3 \), surface normal \( n = (n_x, n_y, n_z)^T \in \mathbb{R}^3 \), radius \( r \in \mathbb{R} \), confidence \( c \in \mathbb{R} \), and timestamp. When a pair \((L_t, D_t)\) is coming from depth estimation module, new surfels \( S_t \) under the current camera coordinates would be obtained. For a 2D pixel \( p = (i, j)^T \in \mathbb{R}^2 \) in depth image \( D_t \), we convert each depth sample \( D_t(p) \) into a 3D point \( v_{(i,j)}^t = D_t(p)K^{-1}(p^T, 1)^T \) of surfel, where the superscript \( (i,j) \) marks the position of the 3D point on image coordinate and \( K \) denotes the laparoscope intrinsic parameter. The process is presented in Fig. 3(b).

The normal in surfel \( s_t \) is expressed as:

\[
n_{t_{(i,j)}} = \frac{(v_{(i+1,j)} - v_{(i,j)}) \times (v_{(i,j+1)} - v_{(i,j)})}{\|v_{(i+1,j)} - v_{(i,j)}\| \cdot \|v_{(i,j+1)} - v_{(i,j)}\|}
\]  

(6)

The radius represents the local area around a point, i.e.:

\[
r_{t_{(i,j)}} = \frac{D_t(p)^2}{f |n_z|}
\]  

(7)

where \( f \) is the focal length part of \( K \). The surfel confidence is initialized as:

\[
c_{t_{(i,j)}} = e^{-\left(\frac{(v_{(i,j)x} - c_x)^2 + (v_{(i,j)y} - c_y)^2}{2\sigma^2}\right)}
\]  

(8)

where \( v_{[x]} \) and \( v_{[y]} \) represent the X and Y coordinates of the 3D point \( v \), \((c_x, c_y)\) are the center of camera, and \( \sigma = 0.6 \) in line with related work [27]. After calculating each surfel, \( S_t \) will be fused into the canonical surfels \( S_{\text{ref}} \) which are under the reference coordinates defined by the first frame based on the current laparoscope pose \( T_t \). The surfels \( S_{\text{ref}} \) are illustrated in Fig. 3(a).

For computing the current pose \( T_t \), reference surfels \( S_{\text{ref}} \) are initially transformed to \( S_{t_{-1}} \) under the camera coordinates of \( L_{t-1} \), and we then iteratively minimize the geometric and photometric reprojection errors between \( S_{t_{-1}} \) and \( S_t \). If the point distance and normal angle of the surfel between \( S_{t_{-1}} \) and \( S_t \), which are calculated according to (12)-(14), are smaller than the threshold [27], it can be added to the surfel set \( P \). Thus, the geometric reprojection error is expressed as:

\[
E_{\text{geo}} = \sum_{(S_{t_{-1}}, S_t) \in P} ((\Delta_T \cdot v_{t} - v_{t_{-1}}) \cdot n_{t_{-1}})^2
\]  

(9)

where \( \Delta_T \) is the transformation pose from the image \( L_{t-1} \) to \( L_t \). The photometric error, which is the image intensity difference,
is written as follows:

$$E_{\text{photo}} = \sum \left( L_t((i, j)) - L_{t-1} \left( K \cdot \Delta_{T}^{-1} \cdot v_{t}^{(i,j)} \right) \right)^2$$  \hspace{1cm} (10)

We define the minimization function as follows:

$$\min \left\{ E_{\text{geo}} + w_{\text{photo}} E_{\text{photo}} \right\}$$  \hspace{1cm} (11)

where $w_{\text{photo}} \in [0,10]$ is an adjustable parameter. Therefore, the laparoscope pose at time $t$ is calculated as $T_t = \Delta_T \cdot T_{t-1}$.

After calculating the current laparoscope pose, new surfels $S_t$ will be integrated into the $S_{ref}$ through surfel association and fusion. Each surfel $S_t$ is paired a corresponding $S_{ref}$ to find the association between $S_t$ and $S_{ref}$. First, surfels $S_{ref}$ are transformed to the current camera coordinates as $S_t'$ by using the camera pose $T_t$, and each point $v_t$ can be further projected onto the image plane to construct a depth image $D_t'$, the process of which is shown in Fig. 3(c). Second, for each pixel $p = (i, j)^T$ in $D_t$, we find a $3 \times 3$ neighborhood $I$ around the same position in $D_t'$, which is illustrated in Fig. 3(c). Then, three metrics are calculated as follows:

$$d_p = \| v_t^{(i,j)}[z] \cdot \| - v_t^{(i_u,j_u)}[z] \cdot \|$$  \hspace{1cm} (12)

$$\alpha_t = \left( \frac{v_t^{(i,j)}[x]}{v_t^{(i,j)}[x]} \right)^{\left( \frac{v_t^{(i,j)}[y]}{v_t^{(i,j)}[y]} \right)} \cdot \frac{1}{\sqrt{2}}$$  \hspace{1cm} (13)

$$\theta = \arccos \left( \frac{n_t^{(i,j)} \cdot n_t^{(i_u,j_u)}}{\| n_t^{(i,j)} \| \cdot \| n_t^{(i_u,j_u)} \|} \right)$$  \hspace{1cm} (14)

$$d_a = \frac{\| v_t^{(i,j)} \cdot \alpha_t \|}{\| \alpha_t \|}$$  \hspace{1cm} (15)

where $v[z]$ denotes the $Z$ coordinate of the 3D point $v$, and $(i_u, j_u)$ is a pixel within $I$. If $d_p$ and $\theta$ are lower than threshold $\gamma_{\text{depth}}$ and $\gamma_{\theta}$, then the pixel holding the smallest $d_a$ will be considered as the matching pixel in $D_t'$; thus, the corresponding surfel $S_{ref}$ can be found for surfel in $S_t$. When the association between $S_{ref}$ and $S_t$ is established, we use following rules to update the reference surfels $S_{ref}$:

$$v_{\text{ref}} \leftarrow \frac{c_{\text{ref}} \cdot v_{\text{ref}} + c_t \cdot v_t}{c_{\text{ref}} + c_t}$$  \hspace{1cm} (16)

$$n_{\text{ref}} \leftarrow \frac{c_{\text{ref}} \cdot n_{\text{ref}} + c_t \cdot n_t}{c_{\text{ref}} + c_t}$$  \hspace{1cm} (17)

$$r_{\text{ref}} \leftarrow \frac{c_{\text{ref}} \cdot r_{\text{ref}} + c_t \cdot r_t}{c_{\text{ref}} + c_t}$$  \hspace{1cm} (18)

$$c_{\text{ref}} \leftarrow c_{\text{ref}} + c_t$$  \hspace{1cm} (19)

The corresponding pseudo codes of the surfel association and fusion algorithm are summarized in Algorithm 1.

**D. Accurate Laparoscopic Localization for Navigation**

Based on the computed 3D structure of the whole surgical scene, we aim at localizing the camera of a given intra-operative view using the coarse-to-fine laparoscopic localization module. The process is shown in intra-operative laparoscopic localization part of Fig. 1(a). First, a global map is established to combine the 3D structure and input images. Second, images from global map with similar location to the query frames can be recognized by a learning-based image retrieval system. After that, we cluster the retrieved images based on the observed 3D points. An iterative estimation process is then used to compute the fine pose of the laparoscope.

**Map Building:** We build a global map shown in Fig. 1(a) by using the input pre-operative images, the estimated laparoscope poses, and reconstructed 3D structure of the tissue surface from the proposed reconstruction framework. First, we combine the input images into an image database. Second, 3D points in the reconstructed structure are projected onto the image plane of the camera coordinates, which is defined on the basis of each estimated laparoscope pose. Then, we regard the 2D pixels projected from the 3D points as key points of the image, and the correspondence between the 3D structure and the input images can be stored by the coordinates of the corresponding pixels on the image plane.

**Coarse Retrieval:** Based on the NetVLAD network [28], we utilize knowledge distillation [29] to train an efficient and smaller student feature extraction network $f_c$ at learning global features predicted by the teacher (NetVLAD). The student net is composed of an encoder and a smaller VLAD layer [19]. Using the $f_c$, global features are computed and indexed for every image in the image database. For each intra-operative query frame, we initially extract the global features. Then, we employ the KNN algorithm to find the nearest images, which have the shortest distance on feature space in the image database. These nearest images are then clustered by the 3D points they co-observe.

NetVLAD network has been applied to mobile robotic applications for place recognition because it integrates the robust feature extraction capability of traditional retrieval methods into CNNs. The advantages of NetVLAD, namely its remarkable robustness to partial scene occlusion, illumination changes,
camera translation and rotation, and great scale-invariant capability, facilitate stable and efficient feature extraction of laparoscopic image sequences. However, the original network is too expensive to generate a large number of global features of surgical images [20]. Therefore, we distill the feature representations from the off-the-shelf trained teacher network into a smaller student model \( f_e \). The training process is shown in Fig. 1(c). We train our feature extraction network \( f_e \) based on the mean square error (MSE) loss, which is written as:

\[
L_{\text{feature}} = \frac{1}{K} \sum_{k=1}^{K} \left( \mathbf{d}_i^k - \mathbf{d}_e^k \right)^2
\]  

where \( K \) denotes the number of extracted features, \( \mathbf{d}_i^k \) is the feature descriptor vector estimated by the teacher NetVLAD, and \( \mathbf{d}_e^k \) is the descriptor predicted by the student model \( f_e \).

Using the trained \( f_e \) model, we can efficiently extract stable features from the laparoscopic images. Then, images in the database with similar localization as query frames can be recognized by feature matching and the KNN algorithm. However, the retrieved images may correspond to different regions of the reconstructed 3D structure. Therefore, we need to cluster the images based on the observed 3D points. If two frames see some 3D points in common, they correspond to the same place and can be grouped. This local search process can reduce the number of false matches and increase the probability of successful localization [19]. Fig. 4 shows an example of clustering process. By retrieving a list of nearest images in global feature space using the KNN, the laparoscope pose can be roughly calculated.

**Fine Localization:** The cluster with the most images is initially used to estimate a fine laparoscopic pose by utilizing a perspective-n-point (PnP) geometric consistency check. We first extract hand-crafted ORB features [30] from the query image and retrieved nearest images and then calculate the feature matches between them. Therefore, the corresponding 3D points in the reconstructed structure for the 2D keypoints of the query image can be selected. After outlier rejection within a RANSAC scheme, we can estimate a global laparoscopic pose from n 3D-2D matches using the PnP method. If a valid pose is calculated, then the process will terminate, and the image of query laparoscopic view is successfully localized.

**III. EXPERIMENTAL PROCEDURES**

**A. Datasets**

**SCARED:** The public SCARED data [5] consists of seven training datasets and two test datasets captured by a da Vinci Xi surgical robot. Each dataset corresponds to a porcine subject, and it has four or five keyframes. A keyframe contains a 1280 \( \times \) 1024-res stereo video, relative endoscope poses and a 3D point cloud of the scene computed by a structured light-based reconstruction method. The data selected in the experiment is called dx_ky, where x and y represent the number of dataset and keyframes, respectively. Here, we also convert all nine stereo endoscopic videos to image streams, which include 26831 frames.

**Ex-Vivo Data:** Our ex-vivo phantoms and tissues data are collected by a Karl Storz laparoscope attached at the end-effector of UR5 robot, each consists of 640 \( \times \) 480-res calibrated stereo videos, laparoscope poses \( b^T_e \), calculated by using the pre-calibrated transformation from end-effector to laparoscope \( c^T_e \) and the end-effector pose \( b^T_{ee} \): \( b^T_e = b^T_{ee} \cdot c^T_e \), and corresponding ground truths of 3D point cloud reconstructed by an active stereo surface reconstruction method assisted with the structured light (SL) [31], of which the accuracy is 45.4 \( \mu m \).

**In-Vivo Data:** Our in-vivo DaVinci robotic surgery dataset from HKPWH contains six cases of 1280 \( \times \) 1024-res stereoscopic videos documenting the entire procedure of robotic prostatectomy. Since the laparoscope and surgical instruments cannot be operated simultaneously in the DaVinci surgical system [32], it is assumed that the surgical scene remains relatively stationary while the laparoscope moves. Therefore, in our study, we manually collected 95 high-quality video clips from six surgical cases in which the environment was kept quasi-static. Each clip lasts about 1 s (20 \( \sim \) 60 frames) and the camera moves rapidly. The environment of these clips contains complex tissue surfaces, small tissue deformations, and slight instrument movements.

For depth estimation, we fine-tuned the HSM model on the SCARED and SERV-CT datasets, and then directly applied it to ex-vivo and in-vivo data to verify the generalization ability. In unsupervised fine-tuning, the SCARED dataset was organized into 20924, 4721 and 1186 frames for training, validation and test sets according to the original data organization. To evaluate dense reconstructions, we utilized eight video sequences in the SCARED test datasets, one in the training dataset, two cases (\( \alpha \), \( \beta \)) in our ex-vivo data, and four clips in the in-vivo data. Furthermore, to examine the performance of the visual localization method, we employed three types of endoscopic motions commonly found in robotic surgery, namely zoom-in, zoom-out, and follow, and random camera motions to generate the test dataset. Typical examples of these three movements are shown in Fig. 5. In each SCARED test dataset, we sampled 70% of the images to build the map, and then picked frames with three endoscopic motion types from the remaining 30% of images. Additionally, we collected validation data on our robotic platform and split it for map building and camera localization according to the above sampling rules in the SCARED test datasets. Ultimately, ten sets...
of data, each with 100 ~ 200 frames, were generated to validate our laparoscopic motion estimation.

B. Implementation

Unsupervised Fine-Tuning of HSM: The HSM model was implemented in PyTorch and fine-tuned for 20 epochs on two GPUs of NVIDIA TITAN XP. The Adam optimizer was used here, where $\beta_1 = 0.9$, $\beta_2 = 0.99$, and the batch size was 4. The following hyperparameters were chosen for training: $w_{\text{trans}} = (0.75, 0.19, 0.05, 0.01)$, $\alpha_1 = 0.025$, $\alpha_2 = 0.05$, and $\alpha_3 = 0.1$. The original image was cropped to 640 × 512 as network input. The training started with an initial learning rate of $2 \times 10^{-5}$, and then degraded by half at epochs 5, 10, and 15.

Dense Visual Reconstruction: In the 3D reconstruction of the entire surgical scene, we computed the pose for each frame by optimizing geometric and photometric errors. In the experiments, the parameter $w_{\text{photo}}$ used to adjust the proportion of the photometric error in the overall errors was set to 1. In surfel association and fusion, the distance threshold $\gamma_{\text{depth}}$ and angle threshold $\gamma_{\text{angle}}$ were set to 0.05 and 0.5, respectively. The code for the dense visual reconstruction algorithm run in CUDA C++.

Laparoscope Localization: To distill the ability of feature extraction from the teacher NetVLAD network, we chose the Google Landmarks dataset [33] which has 100 k images, all 26831 SCARED images, and 95 video clips from the in-vivo dataset. The datasets chosen for knowledge distillation contain general features in natural scenes, such as edges, lines, points, and special characters possessed by surgical scenes. These rich data are beneficial to train the student model $f_e$. The $f_e$ model is able to extract robust features from medical images for endoscope localization. Furthermore, we employed pretrained NetVLAD to generate pseudo-descriptor labels for training the $f_e$ model. All feature extraction networks were implemented in the TensorFlow library and used RMSProp as the optimizer. We trained the model for 85000 epochs with a batch size of 16. The initial learning rate was 0.001 and was divided by 10 at the 60,000th and 80,000th epochs.

C. Performance Metrics

Table I shows the depth evaluation metrics used in our experiments [34], where $d$ and $d^*$ are the predicted depth value and the corresponding ground truth, respectively, $D$ denotes a set of predicted depth values, and $v \in \{1,25^1,1,25^2,1,25^3\}$. Then, we utilized the Root Mean Squared Error (RMSE) to validate the quantified accuracy of the reconstructed 3D model. The RMSE is computed as follows. The 3D reconstructed structure is initially registered with the ground truth 3D point cloud by manually selecting landmarks such as edge points. Then, the registration is refined by the ICP method. In addition, we adopted three metrics, namely, absolute trajectory error (ATE), relative translation error (RTE) and relative rotation error (RRE) [35], to estimate the precision of the laparoscope pose, and the three metrics are defined as follows:

$$ATE = \frac{1}{T} \sum_{t=1}^{T} || \text{trans} \left( Q_{t}^{-1} \Delta S P_{t} \right)||$$

$$E_t = \left( Q_{t}^{-1} Q_{t+1} \right)^{-1} \left( P_{t}^{-1} P_{t+1} \right)$$

$$RTE = \frac{1}{T} \sum_{t=1}^{T} || \text{trans} \left( E_t \right)||$$

$$RRE = \frac{1}{T} \sum_{t=1}^{T} || \text{rot} \left( E_t \right)||$$

where $Q_t$ is the ground truth camera pose; $P_t$ denotes the estimated pose, and $\Delta S$ is the rigid transformation between $Q_t$ and $P_t$.

IV. RESULTS

Extensive experiments were conducted to verify the performance of the proposed framework in terms of the stereo depth estimation accuracy, 3D dense reconstruction results, and the laparoscopic localization ability.

A. Evaluation of Depth Accuracy

We compared the depth estimation accuracy of the fine-tuned HSM model with several stereo-based methods that employed binocular images as training data, including the original HSM [21], AANet [36], STTR in E-DSSR [37], Monodepth2 [34], and AF-SMILearner [13]. For HSM, AANet, and E-DSSR, the pretrained models were directly utilized to estimate the depth information. For Monodepth2, we used the binocular pair to calculate the photometric error, which was then used to optimize the network, and the model trained by this process was
TABLE II

| Methods               | Abs Rel ↓ | Sq Rel ↓ | RMSE ↓ | RMSElog ↓ | δ < 1.25 1↑ | δ < 1.25 2↑ | δ < 1.25 3↑ | Time (ms) |
|-----------------------|-----------|----------|--------|------------|-------------|-------------|-------------|-----------|
| MD+Stereo [34]        | 0.062     | 0.466    | 4.280  | 0.076      | 0.946       | 0.993       | 1.000       | 94.21     |
| AF-SL+Stereo [13]     | 0.062     | 0.532    | 4.259  | 0.077      | 0.940       | 0.990       | 0.999       | 98.75     |
| AANet [36]            | 0.045     | 0.508    | 4.553  | 0.070      | 0.983       | 0.995       | 0.998       | 143.91    |
| E-DSSR [37]           | 0.073     | 2.450    | 11.062 | 0.350      | 0.957       | 0.971       | 0.980       | 406.18    |
| Original HSM [21]     | 0.035     | 0.169    | 3.130  | 0.072      | 0.993       | 0.999       | 0.999       | 48.51     |
| Ours                  | 0.029     | 0.124    | 2.959  | 0.042      | 1.000       | 1.000       | 1.000       | 50.19     |

Fig. 6. Qualitative comparison of depths estimated by MD+Stereo [34], AF-SL+Stereo [13], AANet [36], E-DSSR [37], and ours, and the ground truth (GT) depths. (a) Typical depth estimation results of SCARED. (b) Depth estimation on in-vivo data where ground truth depth does not exist in surgical data. The colorbar on the right represents the distance scale and the unit is mm.

called MD+Stereo. To train AF-SfMLearner, an additional error computed from stereo images was added to the original network. The AF-SfMLearner improved by this stereo-based error was named AF-SL+Stereo. Table II lists the quantitative depth comparison results. Since the SCARED dataset has ground truth depth per frame, quantitative depth evaluation was performed on it. The HSM network achieved the best performance among all depth evaluation metrics. It is worth noting that the fine-tuned HSM model shows significant improvement compared to the original model due to the use of supervised and unsupervised fine-tuning strategies. The model achieved a low RMSE on the test dataset with an error of about 2.959 mm, indicating that the depth of the tissue surface can be estimated with high accuracy. Furthermore, our method took 50.19 ms per frame for depth estimation, which was only a little more than the original HSM.

Furthermore, we selected several typical images from different datasets for qualitative depth comparison. SCARED trained models were directly used to estimate the depth of in-vivo data without any fine-tuning. As shown in Fig. 6, our fine-tuned HSM network can provide stable and clear depth estimates for medical images compared to other methods. Thanks to the use of fine-tuning strategies and HSM networks, our proposed method showed good performance in complex scenes, such as tissues with complex geometries and the edges and tips of surgical instruments.

B. Performance Assessment of 3D Dense Reconstruction

Based on the estimated depth for each frame, we can perform a 3D reconstruction of the entire scene. In the study, we quantitatively validated the accuracy of the reconstruction method on SCARED and ex-vivo datasets, and compared the method to the well-known open-source SLAM approach ORB-SLAM2 [14],
Fig. 7. Reconstruction results and rendering comparisons. (a) 3D reconstruction results of the SCARED and ex-vivo dataset. (b) The input laparoscope images and corresponding rendering images. In each dataset, the left column is the input images, and the right column is the rendering image.

TABLE III

| Dataset | SL-Points ($10^6$) | SR-Points ($10^6$) | RMSE (mm) |
|---------|-------------------|-------------------|-----------|
| d1_k2   | 1.08              | 0.76              | 1.027     |
| d8_k0   | 1.04              | 1.11              | 0.938     |
| d8_k1   | 1.18              | 1.26              | 1.308     |
| d8_k2   | 1.04              | 1.80              | 1.068     |
| d8_k3   | 1.31              | 1.46              | 0.351     |
| d9_k0   | 1.06              | 1.53              | 0.967     |
| d9_k1   | 1.02              | 1.41              | 1.339     |
| d9_k2   | 0.94              | 1.55              | 1.362     |
| d9_k3   | 0.84              | 1.21              | 0.714     |
| α       | 1.71              | 2.47              | 1.705     |
| β       | 0.70              | 1.59              | 1.220     |

TABLE IV

| Dataset | d1_k2 | d8_k0 | d8_k1 | d8_k2 | d8_k3 | d9_k0 | d9_k1 | d9_k2 | d9_k3 | α | β |
|---------|-------|-------|-------|-------|-------|-------|-------|-------|-------|---|---|
| Number of Frames | 280 | 637 | 590 | 309 | 212 | 160 |
| Trajectory Length (mm) | 42.156 | 178.791 | 129.393 | 81.412 | 36.999 | 49.425 |
| ATE (mm) | 0.744 | 2.464 | 4.079 | 1.539 | 3.708 | 2.182 |
| RTE (mm) | 0.053 | 0.093 | 0.077 | 0.076 | 0.247 | 0.155 |
| RRE(deg) | 0.097 | 0.122 | 0.109 | 0.143 | 0.155 | 0.143 |

[38]. Considering that acquiring the ground truth of the tissue’s 3D model in surgery is currently impractical because of clinical regulation, we qualitatively tested our approach on in-vivo DaVinci robotic surgery dataset.

**Quantitative Evaluation of 3D Reconstruction**: As shown in Fig. 7(a), the obtained 3D tissue models usually contained millions of points, which can provide rich details of the surface texture. Furthermore, a surgical simulator was established for rendering color images generated by the estimated camera pose and the 3D structure. We compared the rendering images with corresponding input images, and the results are presented in Fig. 7(b). Our reconstructed 3D surfaces of tissues and the textures on their re-projections both matched those observed from the input images. As for the quantitative results concerning the reconstruction, we compared the SL-Points and SR-Points which accordingly refer to the numbers of points in surface geometry calculated by using the structure light and our stereo reconstruction method. As can be noticed in Table III, the results of RMSE are under the level of 1.71 mm in all testing sets, which to a certain extend demonstrates the high accuracy of our reconstruction framework.

We simultaneously estimated the laparoscope pose in surfel fusion. Since the precision of the camera pose estimation can also affect the accuracy of our reconstruction outcomes, we hence validated the poses by comparing the calculated results with the ground truth camera poses using ATE, RTE and RRE metrics. Table IV shows the quantitative comparisons, and the result illustrates that the estimated camera pose matches closely with the ground truth poses, thereby proving the effectiveness of the proposed reconstruction framework.

**Qualitative Evaluation on SCARED and In-Vivo Data**: To handle low texture in medical images, we set the number of features per image extracted by ORB-SLAM2 to 12000 and the minimum threshold to detect FAST corners to 1. As shown in Fig. 8(a), the 3D structure reconstructed by ORB-SLAM2 is sparse compared to our method, which makes it difficult to observe texture from ORB-SLAM2 results. Additionally, in Fig. 8(b), we only presented our reconstruction results because ORB-SLAM2 cannot be initialized with fewer images. Although the laparoscope moved quickly and the surgical scene was complicated with slight deformations, a potential 3D point cloud and smooth laparoscope poses can be estimated, which qualitatively proves that the proposed method is accurate.
C. Performance of Laparoscopic Localization for Navigation

With dense and accurate reconstruction of the tissue surface, we subsequently performed experiments to validate the performance of the laparoscopic localization module. Since there are no image-based laparoscopic localization methods reported in the literature, we compared our method with MapNet [39], a CNN-based end-to-end camera pose estimation method commonly used in autonomous driving. As described in Section III-A, we sampled 70% of the dataset to build maps in each dataset for dense visual reconstruction, so here we used the same number of images to train MapNet. The remaining images in the data were used to test our visual localization method and MapNet.

Given that the ground truths of the camera poses can be obtained in each data, we can quantitatively evaluate the accuracy of the calculated laparoscope pose. As reported in Table V, translation and rotation errors concerning the camera pose estimation were presented. It is worth noticing that the average errors in translation and rotation were only 2.17 mm and 2.23°, showing that our localization method can track the camera in real laparoscopic views and simulated new views. However, MapNet lacked the localization ability in new scenes. Therefore, our visual localization module has the preliminary ability to track the laparoscope in complicated surgical scenes with only images as input.

Fig. 9 shows typical examples of a comparison between estimated poses and ground truth poses. For each type of motion, the black wireframe represents the origin of the camera motion, while the red and blue wireframes represent the ground truth of the camera pose and those computed by our visual localization module accordingly. These experimental results show that the estimated pose is qualitatively similar to the ground truth in both the rotation and translation parts.

Ablation Study: The number of images retrieved in coarse retrieval significantly affects the accuracy and runtime of laparoscopic pose estimation, so we analyzed its impact through an experimental study. Pose recall at position and orientation errors is reported as test data, measured as follows. First, the position and orientation errors between the estimated pose and the true pose were calculated. Second, the error percentage within 2.0 mm and 1.5° was calculated, named recall@\((2.0 \text{ mm}, 1.5°)\). As shown in Table VI, when the number of retrieved images...
was changed from $N = 3$ to $N = 5$, the recall improved while the runtime increased only slightly. However, increasing $N$ from 5 to 10 increased the overall runtime of camera localization, but did not change the accuracy. Therefore, we chose 5 for the number of images retrieved in laparoscopic localization.

### D. Runtime

As shown in Fig. 10, we run our holistic stereo reconstruction framework on the platform composed of a UR5 robot, a Karl Storz laparoscope system, a joystick, and a uterus phantom. All of the source code of the proposed method were executed on an ASUS desktop with an Intel Xeon E5-2640 v4 2.40GHZ CPU and two GPUs of NVIDIA TITAN XP. We utilized the joystick to control the movement of laparoscope to collect stereo image sequences while achieving live reconstruction of the tissue surface. The flow of image data and the estimated depth between different parts in our proposed framework was based on Robot Operating System. For stereo dense reconstruction method in the framework, the average runtime of the two parts is shown in Table VII, which is the average results of 2000 frames on 640 × 480 laparoscopy videos. The computational time to process one image in reconstruction is 81.35 ms ($\sim$12 fps), which demonstrates that the reconstruction method is real time. Here, we used the 640 × 480-res to compute the runtime of depth estimation per frame, so it consumed less time compared with the results in Section IV-A. Besides, since the computing process contains the stereo image sequences reading, data flow in computer and the robot control, the runtime of the reconstruction method is little more than the actual consuming time. We then calculated the average runtime of the laparoscopic localization of all ten sets of data, where each image was 640 × 480 in size. As shown in the table, it takes approximately 349 ms to estimate the laparoscopic pose per query image. Although the speed of the camera localization module is only 2.8 fps, we can successfully track the laparoscope with the images.

### V. DISCUSSION

Laparoscopic surgery has been widely accepted as a minimally invasive surgical procedure. In the current clinical routine, surgeons make a small incision in the patient’s belly button and insert the laparoscope. For further diagnosis, the doctor will observe the abdominal cavity based on the laparoscopic images. If surgery is required, additional incisions will be made and instruments will be inserted through these holes. Then, the operation is performed using the laparoscope as a guide. In this paper, we propose a new learning-based framework to reconstruct the dense 3D structure of a surgical environment and compute the laparoscopic pose within a given view. The dense reconstruction module of the proposed method can potentially integrate into the diagnosis stage of clinical routine. When the surgeon moves the laparoscope during diagnosis, the entire 3D structure of the internal anatomy can be reconstructed online. After that, the image-based laparoscope localization module is introduced to help the doctor locate the camera and understand the relationship between the laparoscope and the surgical site in the early stage of the surgery.

To train the stereo depth estimation model, we only use the binocular images from the SCARED dataset and not the corresponding ground truth depth. The reason is that the ground truth depth map for each frame in the SCARED data is relatively poor, as shown in Fig. 11. Most uninformative depth maps affect supervised training. In future work, we will improve our depth perception capability by extracting more information from the ground truth. Notably, a large number of stereoscopic images of robotic surgery without ground-truth depth are available from hospitals. Based on our unsupervised fine-tuning strategy, we can make full use of these real surgical data, thus solving the data-hungry problem of deep learning-based algorithms to a certain extent. Furthermore, we will design more robust training losses for surgical scenarios in the future. To train the feature extraction network in the laparoscopic localization module, we can directly use the pretrained NetVLAD to generate pseudo-descriptor labels. While the trained student net $f_e$ can extract stable features for coarse retrieval, we will further leverage
traditional methods to obtain ground-truth descriptor labels for surgical images, and then train our NetVLAD model for more robust \( f_e \).

Although the proposed framework can only reconstruct static or slightly deformed tissue surfaces, it is capable of estimating depth information for surgical scenes with large deformations, thus still providing surgeons with 3D depth data. The laparoscopic localization module assumes that the surgical scene is slightly deformed, so it is more suitable for the early stage of the operation. When the surgeon moves the laparoscope to the tumor or lesion, it is important for the surgeon to understand the relationship between the camera and the surgical site. An example of applying the proposed method is functional endoscopic sinus surgery (FESS). In FESS, the sinus environment is rigid and static, and the endoscope-tumor relationship should be provided to the surgeon. Furthermore, the image-based laparoscope localization module is now time consuming, so we will design a more efficient network for image retrieval in the future.

Fig. 12 shows some examples of relatively low-confidence dense 3D reconstruction results. The reconstruction on the left presents a 3D structure with many noisy point clouds. This was caused by the rapid movement of the laparoscope due to the jitter of UR5 arm. However, we may design some motion criteria in the future to filter out those abnormal motions of the robot. The reconstructed 3D structure on the right shows some cracks in the tissue surface. These cracks were caused by the incorrect calculation of the camera pose during the dense reconstruction. The reason of wrong laparoscopic pose estimation is the jitter of UR5, but it is rare in all experiments.

VI. CONCLUSION

In this paper, we propose an efficient learning-driven framework, which can achieve an image-only 3D reconstruction of surgical scenes and preliminary laparoscopic localization. A fine-tuned learning-based stereo estimation network and a dense visual reconstruction algorithm are proposed to recover the 3D structure of tissue surface. In addition, a visual localization module that incorporates our reconstructed 3D structure is presented to achieve coarse-to-fine laparoscopic tracking using only image as input. We also evaluate our framework qualitatively and quantitatively in three datasets to demonstrate its accuracy and efficiency.

This work assumes a surgical scene with small deformation for the reconstruction and localization framework. In the future, we will apply our stereo dense reconstruction and camera localization framework to ENT surgery.

REFERENCES

[1] T. Bergen and T. Wittenberg, “Stitching and surface reconstruction from endoscopic image sequences: A review of applications and methods,” IEEE J. Biomed. Health Inform., vol. 20, no. 1, pp. 304–321, Jan. 2016.
[2] L. Maier-Hein et al., “Optical techniques for 3D surface reconstruction in computer-assisted laparoscopic surgery,” Med. Image Anal., vol. 17, no. 8, pp. 974–996, 2013.
[3] Y. Hayashi et al., “Clinical application of a surgical navigation system based on virtual laparoscopy in laparoscopic gastrectomy for gastric cancer,” Int. J. Comput. Assist. Radiol. Surg., vol. 11, no. 5, pp. 827–836, 2016.
[4] C.-H. Wu, Y.-N. Sun, and C.-C. Chang, “Three-dimensional modeling from endoscopic video using geometric constraints via feature positioning,” IEEE Trans. Biomed. Eng., vol. 54, no. 7, pp. 1199–1211, Jul. 2007.
[5] M. Nazeri et al., “Stereo correspondence and reconstruction of endoscopic data challenge,” 2021, arXiv:2101.01133.
[6] B. Lin, Y. Sun, and X. Qian, “Dense surface reconstruction with shadows in MIS,” IEEE Trans. Biomed. Eng., vol. 60, no. 9, pp. 2411–2420, Sep. 2013.
[7] D. Stoyanov et al., “Real-time stereo reconstruction in robotically assisted minimally invasive surgery,” in Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Interv., 2010, pp. 275–282.
[8] H. Zhou and J. Jagadeesan, “Real-time dense reconstruction of tissue surface from stereo optical video,” IEEE Trans. Med. Imag., vol. 39, no. 2, pp. 400–412, Feb. 2020.
[9] Z. Li et al., “Learning for disparity estimation through feature consistency,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 2811–2820.
[10] Z. Li et al., “Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers,” in Proc. IEEE Int. Conf. Comput. Vis., 2021, pp. 6197–6206.
[11] B. Huang et al., “Self-supervised generative adversarial network for depth estimation in laparoscopic images,” in Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Interv., 2021, pp. 227–237.
[12] M. A. Karaoglu et al., “Adversarial domain feature adaptation for bronchoscopic depth estimation,” in Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Interv., 2021, pp. 300–310.
[13] S. Shao et al., “Self-supervised monocular depth and ego-motion estimation in endoscopy: Appearance flow to the rescue,” Med. Image Anal., vol. 61, 2021, Art. no. 102338.
[14] L. Chen et al., “Slam-based dense surface reconstruction in monocular minimally invasive surgery and its application to augmented reality,” Comput. Methods Programs Biomed., vol. 158, pp. 135–146, 2018.
[15] N. Mahmoud, T. Collins, A. Hostettler, L. Soler, C. Doignon, and J. M. M. Montiel, “Live tracking and dense reconstruction for handheld monocular endoscopy,” IEEE Trans. Med. Imag., vol. 38, no. 1, pp. 79–89, Jan. 2019.
[16] A. Marmol, A. Banach, and T. Peynot, “Dense-arthroslam: Dense intra-articular 3-D reconstruction with robust localization prior for arthroscopy,” IEEE Robot. Autom. Lett., vol. 4, no. 2, pp. 918–925, Apr. 2019.
[17] J. Wang et al., “Augmented reality navigation with automatic marker-free image registration using 3D image overlay for dental surgery,” IEEE Trans. Biomed. Eng., vol. 61, no. 4, pp. 1295–1304, Apr. 2014.
[18] S. Leonard et al., “Evaluation and stability analysis of video-based navigation system for functional endoscopic sinus surgery on in vivo clinical data,” IEEE Trans. Med. Imag., vol. 37, no. 10, pp. 2185–2195, Oct. 2018.
[19] P.-E. Sarlin et al., “Leveraging deep visual descriptors for hierarchical efficient localization,” in Proc. Conf. Rob. Learn., 2018, pp. 456–465.
[20] P.-E. Sarlin et al., “From coarse to fine: Robust hierarchical localization at large scale,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 12716–12725.
[21] G. Yang et al., “Hierarchical deep stereo matching on high-resolution images,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 5515–5524.
[22] H. Zhao et al., “Pyramid scene parsing network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 2881–2890.
[23] M. Menze and A. Geiger, “Object scene flow for autonomous vehicles,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 3061–3070.
[24] P. Edwards et al., “SERV-CT: A disparity dataset from CT for validation of endoscopic 3D reconstruction,” Med. Image Anal., vol. 76, 2020, Art. no. 102302.
[25] F. Liu et al., “Self-supervised depth estimation to regularise semantic segmentation in knee arthroscopy,” in Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Interv., 2020, pp. 594–603.

[26] M. Keller, D. Lefloch, M. Lambers, S. Izadi, T. Weyrich, and A. Kolb, “Real-time 3D reconstruction in dynamic scenes using point-based fusion,” in Proc. Int. Conf. 3D Vis., 2013, pp. 1–8.

[27] T. Whelan et al., “ElasticFusion: Real-time dense SLAM and light source estimation,” Int. J. Rob. Res., vol. 35, no. 14, pp. 1697–1716, 2016.

[28] R. Arandjelovic et al., “Netvlad: CNN architecture for weakly supervised place recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 5297–5307.

[29] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” in Proc. 28th Int. Conf. Neural Inf. Process. Syst. Workshop, 2014.

[30] E. Rublee et al., “ORB: An efficient alternative to sift or surf,” in Proc. IEEE Int. Conf. Comput. Vis., 2011, pp. 2564–2571.

[31] C. Sui et al., “Active stereo 3-D surface reconstruction using multistep matching,” IEEE Trans. Autom. Sci. Eng., vol. 17, no. 4, pp. 2130–2144, Oct. 2020.

[32] S. DiMaio, M. Hanuschik, and U. Kreaden, “The Da Vinci surgical system,” in Proc. Surg. Robot., 2011, pp. 199–217.

[33] H. Noh et al., “Large-scale image retrieval with attentive deep local features,” in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 3456–3465.

[34] D. Recasens et al., “Endo-depth-and-motion: Reconstruction and tracking in endoscopic videos using depth networks and photometric constraints,” IEEE Robot. Autom. Lett., vol. 6, no. 4, pp. 7225–7232, Oct. 2021.

[35] K. B. Ozyoruk et al., “Endoslam dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos,” Med. Image Anal., vol. 71, 2021, Art. no. 102058.

[36] H. Xu and J. Zhang, “AANet: Adaptive aggregation network for efficient stereo matching,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 1959–1968.

[37] Y. Long et al., “E-dssr: Efficient dynamic surgical scene reconstruction with transformer-based stereoscopic depth perception,” in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv., 2021, pp. 415–425.

[38] R. Mur-Artal and J. D. Tardós, “Orb-slam2: An open-source SLAM system for monocular, stereo, and rgb-d cameras,” IEEE Trans. Robot., vol. 33, no. 5, pp. 1255–1262, Oct. 2017.

[39] S. Brahmbhatt et al., “Geometry-aware learning of maps for camera localization,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 2616–2625.