Computers in Biology and Medicine 175 (2024) 108546

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

3D reconstruction from endoscopy images: A survey

Zhuoyue Yang [a,b], Ju Dai [b], Junjun Pan [a,b,*]

[a] State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, 37 Xueyuan Road, Haidian District, Beijing, 100191, China
[b] Peng Cheng Lab, 2 Xingke 1st Street, Nanshan District, Shenzhen, Guangdong Province, 518000, China

ARTICLE INFO

Keywords:
3D reconstruction
Endoscopy
SLAM
Depth estimation
Feature matching
Scene representation

ABSTRACT

Three-dimensional reconstruction of images acquired through endoscopes is playing a vital role in an increasing number of medical applications. Endoscopes used in the clinic are commonly classified as monocular endoscopes and binocular endoscopes. We have reviewed the classification of methods for depth estimation according to the type of endoscope. Basically, depth estimation relies on feature matching of images and multi-view geometry theory. However, these traditional techniques have many problems in the endoscopic environment. With the increasing development of deep learning techniques, there is a growing number of works based on learning methods to address challenges such as inconsistent illumination and texture sparsity. We have reviewed over 170 papers published in the 10 years from 2013 to 2023. The commonly used public datasets and performance metrics are summarized. We also give a taxonomy of methods and analyze the advantages and drawbacks of algorithms. Summary tables and result atlas are listed to facilitate the comparison of qualitative and quantitative performance of different methods in each category. In addition, we summarize commonly used scene representation methods in endoscopy and speculate on the prospects of deep estimation research in medical applications. We also compare the robustness performance, processing time, and scene representation of the methods to facilitate doctors and researchers in selecting appropriate methods based on surgical applications.

1. Introduction

An increasing number of medical applications, such as surgical navigation, medical image segmentation, preoperative registration, surgical simulation, surgical robotics, intelligent diagnostics, etc., can benefit from depth estimation and 3D reconstruction tasks. The problem of depth estimation has been studied in natural scenes for decades, but obtaining depth information using endoscopic images or videos is still an emerging problem. Many types of endoscopes are frequently used in minimally invasive procedures, including oral endoscope [1], bronchoscope [2], gastroscope [3], enteroscope [4], arthroscope [5], nasoscopes [6], stomatoscopes [7], colonoscopes [4], hysteroscopes [8], knee arthroscopes [5], fetal endoscopes [9].

Depending on the number of lenses used, these endoscopes can be divided into monocular endoscopes and binocular endoscopes. Binocular endoscopes are similar to human binoculars, the relative position and the internal parameters of the two lenses are known. Therefore, the use of binocular endoscopes for depth estimation is more easily understood and more common. Monocular endoscopes have one moving camera and cannot obtain depth information directly. Therefore, there are more challenging problems in monocular depth estimation. Since the endoscope needs to pass through the natural cavities of the body for examination, the size of the device is relatively small. Therefore, among the endoscopes mentioned above, only the laparoscope can be a binocular endoscope; the others are all basically monocular endoscopes.

The first generation of geometry-based methods relied on matching pixels between multiple images. However, this is not applicable in endoscopic scenes where the soft tissue surface is smooth resulting in featureless and repetitive texture regions. Although methods exist to improve feature matching using machine learning methods, the results still need to be improved. The second-generation methods use learning-based techniques to improve the integrity of reconstruction and reduce time consumption. The learning-based method cloud has challenges due to its dependence on large amounts of data and lack of ground truth in the medical scene.

In this article, we present a comprehensive and structured review of recent advances in the use of endoscopic images for 3D reconstruction and depth estimation over the last 10 years. These methods mainly use monocular endoscopes or binocular endoscopes. We have collected over 170 papers that were published in leading medical-engineering journals and conferences from 2013 to 2023. In addition, we describe

* Corresponding author.State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, 37 Xueyuan Road, Haidian District, Beijing, 100191, China.
E-mail address: pan_junjun@buaa.edu.cn (J. Pan).

https://doi.org/10.1016/j.compbiomed.2024.108546
Received 15 November 2023; Received in revised form 5 January 2024; Accepted 28 April 2024
Available online 30 April 2024

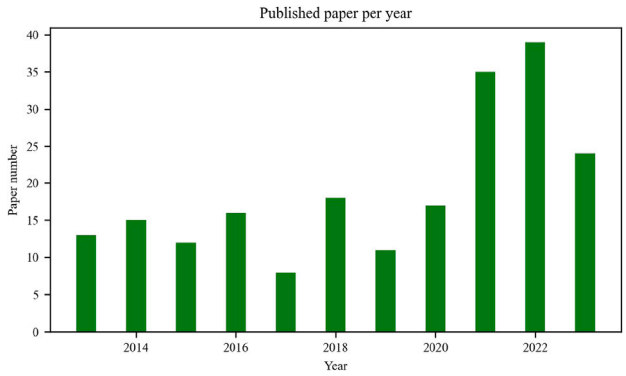


Fig. 1. The number of publications per year in endoscopy reconstruction.

the changes in technology, elaborate common pipelines, provide separate theoretical accounts for monocular and binocular endoscopes, and summarize each technique in a categorical manner. Moreover, we provide an in-depth analysis of the most popular learning-based approaches, including network architectures, training frameworks, and training strategies. We also collate publicly available datasets applied to depth estimation. The aim is to help readers navigate through this emerging field.

The remainder of this paper is organized as follows. Section 2 describes the definition and current classification of the problem, as well as the representation of the reconstruction. Section 3 investigates datasets and the metrics that can be used for depth estimation. Section 4 focuses on the theory and work on reconstruction using monocular endoscopes. Section 5 reviews methods for reconstruction methods using binocular endoscopes. Section 6 discusses the visualization and surgical navigation related to 3D reconstruction. Section 7 suggests potential future research directions and Section 8 concludes this survey.

2. Scope and taxonomy

2.1. Scope

The goal of depth estimation is to recover three-dimensional information about the endoscopic scene from a sequence of images or videos taken by the endoscope, where the internal or external reference of the camera is known or unknown. If a binocular endoscope is used, i.e. there are two cameras side by side, then it is possible to obtain two images taken at the same time. If a monocular endoscope is used, then one image can be obtained at one moment in time. The recovery of three-dimensional information about the scene based on two images or one image taken at a moment in time is generally represented as a depth map and is called a depth estimate. After obtaining multiple depth estimates, the multiple depth estimates are fused into a single representation of the scene.

In this survey, we focus on depth estimation and surface reconstruction under monocular and binocular endoscopy. Some papers on improvements to endoscopic hardware are outside the scope of our study. The related survey paper [10] is published. As shown in Fig. 1, in the past 10 years, a large number of works have emerged. Thus our survey is timely and necessary.

2.2. Taxonomy

Whether using a monocular endoscope or a binocular endoscope, two types of methods can be distinguished. According to the development of the technique, the geometry-based methods and the learning-based methods can be divided. **Geometry-based methods** include shape from shading (SfS) [11], structure from motion (SfM) [12], simultaneously localization and mapping (SLAM) [13], and multi-view stereo (MVS) [14]. The theoretical basis for the geometry-based approach is multi-view geometry. The main theories in multi-view geometry are epipolar geometry [15], triangulation [15], perspective-n-point (PnP) [16], iterative closest point (ICP) [17], etc. The steps that have the greatest impact on performance are feature extraction and feature matching.

Subsequently, **learning-based methods** have gained significant advances in natural scenes. As a result, many works use learning-based methods to improve the performance of feature matching in endoscopy. In conjunction with traditional pose estimation methods, learning-based methods obtain better results. In the last five years, direct depth estimation methods using deep learning network methods have achieved state-of-the-art results in quantitative and qualitative experiments. The learning-based method can be summarized as predicting the depth map (D) from the image set (I) through a prediction function (F). The depth map consists of the corresponding depth values of all pixels in the image. In natural scenes, the performance of learning-based methods mainly depends on a large number of high-quality training data. However, in the medical scene, it is difficult to obtain the ground truth of depth estimation due to the limitations of devices. Therefore, learning-based methods can be divided into two categories. The first category only uses endoscopic images for self-monitoring training, and the second category uses CT and other auxiliary data for training.

The classification of depth estimation methods and the types of scene representation is shown in Fig. 2. In summary, we divide depth estimation methods into two categories: geometry-based methods and learning-based methods, which are represented by different shades of green. The blue blocks indicate the different expressions used to display 3D information in the endoscopy scene. Scene representations are available as point clouds [18], surfel [19], and truncated signed distance functions (TSDF) [20].

Point cloud. A point cloud is a discrete set of points in 3D space. The point cloud is the common data form in 3D reconstruction tasks. Luo et al. [21] regard the point cloud as a result of intraoperative video 3D reconstruction and match the point cloud to a preoperative model that underwent downsampling. Most methods use point clouds as the representation of 3D scenes.

Surfel. Surfel is a real-time reconstruction method that supports dynamic scenes [19]. Surfel is designed based on the representation of flat and point. Each surfel contains a 3D point, a surface normal, radius, confidence, and timestamp. Several works also utilize surfel to represent a 3D scene. Turan et al. [22] combines electromagnetic positioning and visual positioning, and uses surfel for non-rigid map fusion. Li et al. [23] selects surfel as the scene representation and employs the embedded deform graph to track all surface sets. In the binocular endoscope, Wei et al. [24] also uses surfels as the expression of the whole surgical scene.

TSDF. TSDF (truncated signed distance functions) is a parameterized surface representation. The core idea of TSDF is to divide the selected 3D space into small pieces (voxels). Secondly, TSDF stores the distance between the current position and the nearest object surface in each voxel. A distance greater than 0 indicates that the position is in front of the surface; If the distance is less than 0, the position is behind the surface; Moreover, the position where the distance is zero is the object's surface. TSDF can be updated in a relatively straightforward way [20]. Recasens et al. [25] and Liu et al. [26] fuse depth information corresponding to multiple keyframes into a TSDF. Liu et al. [26] first use learning-based descriptions and SfM to generate dense point matching and camera trajectories. Then dense feature matching is utilized to provide supervision for the density estimation of each pixel, and finally, the depth maps are fused into a surface through TSDF.

NeRF. At present, implicit scene representation has shown great potential in 3D reconstruction. NeRF (Neural Radiance Field) represents the scene as the volume density and color value of any point

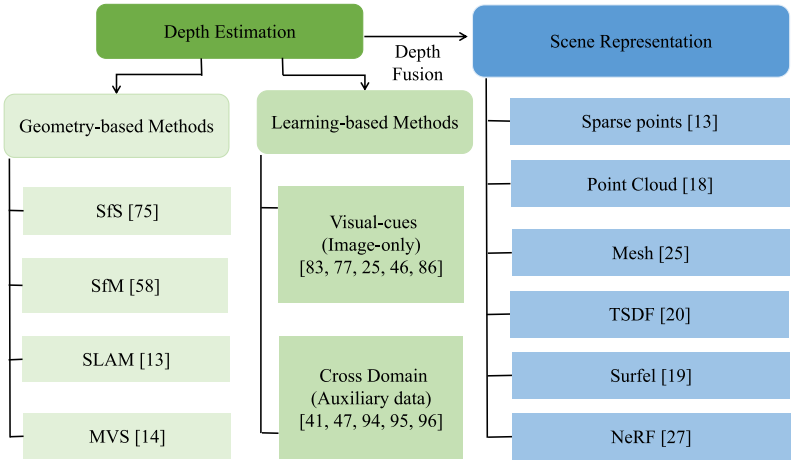


Fig. 2. The classification of depth estimation methods and the types of scene representation. Both the binocular depth estimation method and monocular depth estimation method can be divided into the geometry-based method and the learning-based method.

in space [27]. Wang et al. [28] and Yang et al. [29] use differentiable volume rendering to project scene representations onto 2D images and represent the deformable surgical scene in MLP (multilayer perceptrons).

We summarize the publicly available datasets in Table 1. We divide the performance of the robustness into 5 levels based on the theory, characteristics, and challenges of the methods, as shown in the Table 2. We summarize the accuracy, running time, robustness, and scene representation of methods in the following tables. Readers can choose the appropriate method based on the applied surgery.

3. Datasets and metrics

This section summarizes the datasets and metrics that can be utilized in the 3D reconstruction task.

3.1. Datasets

Few datasets can be used to monitor the depth estimation since it is very difficult to obtain the true value of the scene depth. Therefore, some works utilize VR-CAP [42] to create synthetic datasets. Some works create phantom models to collect truth values. Others use the results obtained from COLMAP [43] and surgery images as the ground truth. The images in the datasets are shown in Fig. 3. We have collated the public datasets used for depth estimation in the medical scene. Here we mainly introduce the size of the dataset and the method of obtaining the ground truth, as shown in Table 1.

3.1.1. Dataset size

The SERV-CT dataset [32] places the whole body of two miniature pigs in the endoscope field of vision, ensuring that the endoscope and the target anatomical structure can be seen in the CT scan, and uses CT reconstruction to obtain the true value. The SCARED benchmark [33] provides 7 training sets and 2 structured light data testing sets captured from pig carcasses for intensive depth estimation and camera poses. The EndoSLAM dataset [31] sews pig organs such as the colon and stomach onto the foam and utilizes a mechanical arm and a 3D scanner to record the true depth value and camera poses. The EndoMapper benchmark [30] provides 59 high-quality calibrated complete conventional endoscope records, corresponding sparse 3D reconstruction and pose estimation. Compared with other datasets, the video in this dataset takes a long time and comes from real records of the human body. The EndoAbs dataset [36] simulates the endoscopic stereoscopic images of abdominal organs and uses 3D organ surface references generated by laser scanners to provide camera calibration parameters, but only contains 120 binocular images.

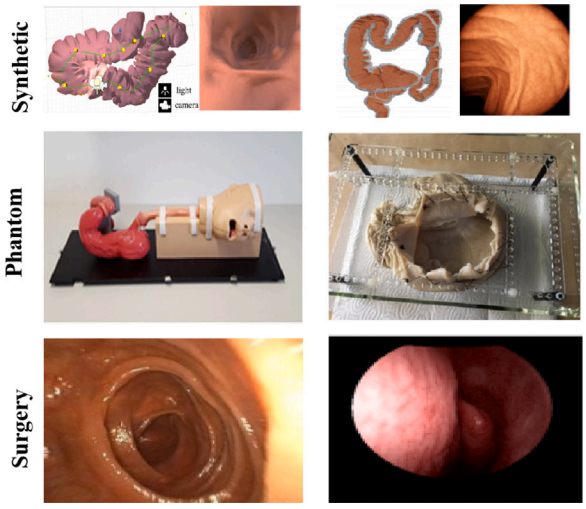


Fig. 3. Examples of synthetic, phantom, and surgery datasets.

3.1.2. Ground truth

The ground truth for the EndoSLAM dataset [31] and the SCARED dataset [33] are obtained from the 3D scanner, and each depth map is dense. the depth estimate for the first frame in each subsequence of the SCARED dataset is accurate, but subsequent dense depth maps are incomplete and cannot be updated for emerging scenes. The EndoSLAM dataset [31] includes a model stitched through the foam, and the anatomy is incomplete compared to the real scene. The ground truth of the EndoMapper benchmark [30] is derived from the sparse point cloud results of the SfM algorithm (COLMAP). The real values of the SERV-CT dataset [32] originated from the pre-operative CT reconstruction model, but there may be problems with intraoperative video deformation. The ground truth can be obtained from these tools, such as the CT-reconstructed model of a colon phantom molded from a real colon [41],[1] simulator used in [38],[2] slicer used in [44],[3] VR-caps

[1] https://www.thecgroup.com/product/colonoscopy-trainer-2003/.
[2] https://drive.google.com/drive/folders/1cypaTsHpi7TRVKI5cYvzk1UfpmdcOEts?usp=sharing.
[3] https://www.slicer.org/.

Table 1
Dataset for the depth in endoscopy. '–' means this item is not mentioned in the reference. '✓' represents this item is provided in the dataset. 'Int.' and 'Ext.' are the abbreviations of camera intrinsics and extrinsics.

Dataset	Type	Organs	Tasks	Images		Depth		Pose	
				Size	Resolution	Type	Source	Int.	Ext.
EndoMapper [30] https://www.synapse.org/#!Synapse:syn26707219/wiki/615178	Surgery Synthetic	Colon	VSLAM	59 sequences at leat 6 sequences	320 × 240 –	Sparse Dense	COLMAP VR-Caps	✓ ✓	✓ ✓
Endo-SLAM [31] https://data.mendeley.com/datasets/cd2rtzm23r/1	Phantom Synthetic	Colon, small intestine, stomach	Disparity Disparity	42,700 35,900	640 × 480 320 × 320	Point cloud Dense-per-frame	3D scanner Unity	✓ ✓	✓ ✓
SERV-CT [32] https://www.ucl.ac.uk/interventional-surgical-sciences/weiss-open-research/weiss-open-data-server/serv-ct	Phantom	Torso cadavers	Disparity	16 stereo pairs	–	Depth map	CT	–	–
SCARED [33] https://endovissub2019-scared.grand-challenge.org/	Phantom	Abdominal cavity	Disparity	23,000	1280 × 1024	Point cloud	3D scanner	✓	✓
Hamlyn [34] https://hamlyn.doc.ic.ac.uk/vision/	Surgical	Stomach, colon, abdomen	Poly detection, tracking and retargeting disparity	37G	Multi-resolutions	Disparity map	CT	✓	–
UCL [4,35] http://cmic.cs.ucl.ac.uk/ColonoscopyDepth/	Synthetic	Colon	Depth	16,016	256 × 256	Depth map	CT	–	–
EndoAbs [36] https://zenodo.org/record/60593	Phantom	Spleen, liver, kidney	Disparity	120	640 × 480	Point cloud	Laser scanner	✓	–
Stereo surgical dataset used in [37] https://doi.org/10.5281/zenodo.7385603	Surgery	Lymph	Radical prostatectomy with lymphadenectomy	128G	1920 × 1080	–	–	–	–
Simulation platform used in [38] https://studentutsedu-my.sharepoint.com/	Synthetic	Colon	Pose estimation	15 cases	–	–	–	✓	✓
Colon10k used in [39] https://endoscopography.web.unc.edu/place-recognition-in-colonoscopy/	Surgical	Colon	Place recognition	10,126	270 × 216	–	–	–	–
Sinus Surgery used in [40] https://github.com/SURA23/Sinus-Surgery-Endoscopic-Image-Datasets	Surgical Phantom	Sinus	Instrument segmentation	9003	256 × 256	–	–	–	–
CVC-ClinicDB used in [41] https://polyp.grand-challenge.org/CVCClinicDB/	Surgical	Colon	Annotated poly	612	576 × 768	–	–	–	–
ASU-Mayo https://polyp.grand-challenge.org/AsuMayo/	Surgical	Colon	Annotated poly	18,902	–	–	–	–	–
LDPolypVideo used in [4] https://github.com/dashishi/LDPolypVideo-Benchmark	Surgical	Colon	Annotated poly	4,200,000	560 × 480	–	–	–	–

Table 2
The levels of the robustness performance.

Symbol	Explanation
Rigid	Clear limitations in theory, with refinement
Non-rigid	No limitations in theory, tolerance for small deformation of soft tissue during small time splits
Instrument	No limitations in theory, good tolerance for small deformation of soft tissue and instrument movement during small time splits
Dynamic	No limitations in theory, ideal results achieved when there is any deformation of soft tissue, movement of instruments within time slices.
Full cycle	No limitations in theory, ideal results achieved when there is any deformation of soft tissue, movement of instruments during surgery.

used in [42],[4] maya used in [45],[5] Blender used in [46,47][6] and 3D Systems GI Mentor Platform [46].

3.2. Metrics

According to different tasks and objectives, researchers utilize different metrics to evaluate the effectiveness and accuracy of the method. The main tasks are depth estimation and pose estimation. For the methods based on deep learning, it is also necessary to evaluate the performance of the network.

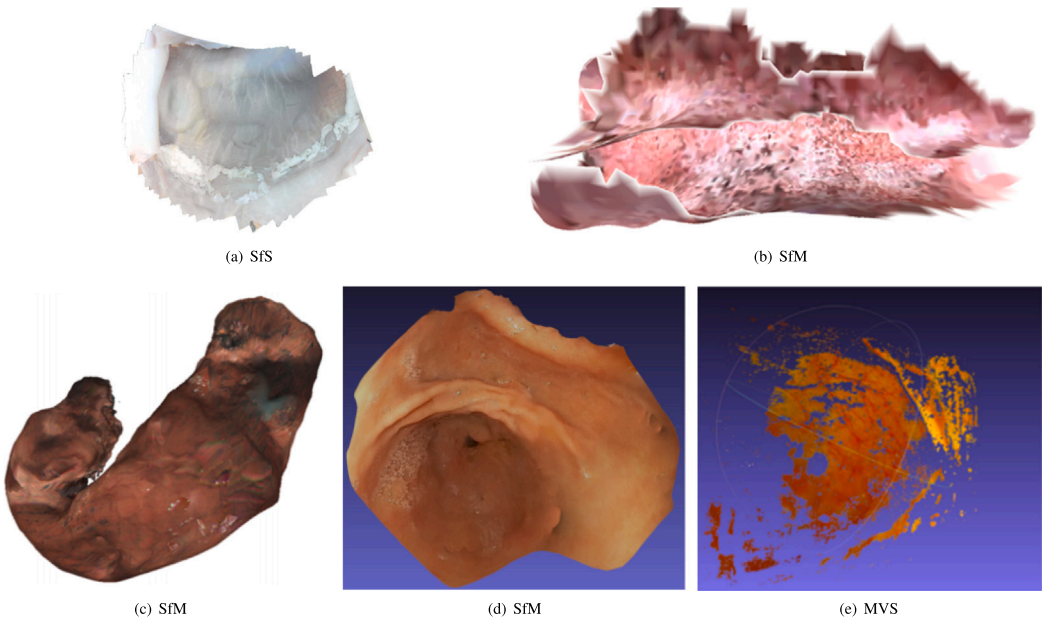
3.2.1. Depth evaluation metrics

The mostly used metric is the Root Mean Squared Error (RMSE), as shown in Eq. (1),

$$RMSE = \sqrt{\frac{1}{|\mathbf{D}|}\sum_{d\in\mathbf{D}}|d^* - d|^2}, \tag{1}$$

where d and d^* denote the predicted depth and the ground truth, respectively. $\mathbf{D}$ is the set of predicted depth. The other metrics, such as Eqs. (A.1), (A.2), and (A.3) are used to measure the error for depth

[4] https://github.com/CapsuleEndoscope/VirtualCapsuleEndoscopy.
[5] https://www.autodesk.com/products/maya/.
[6] https://www.blender.org/.



(a) SfS (b) SfM (c) SfM (d) SfM (e) MVS

Fig. 4. Results of SfS, SfM, and MVS using the monocular endoscopy. (a) is the result of stomach phantom model [11]; (b) shows the reconstructed model from nasal endoscopy images [48]; (c) [3], (d) [49], and (e) [50] represent the result of the stomach.

evaluation in Appendix. Eq. (A.4) represents the accuracy for depth estimation in Appendix.

3.2.2. Ego-motion evaluation metrics

Absolute trajectory error (ATE) [51] and relative pose error (RPE) are commonly used metrics. ATE mainly evaluates the global consistency and RPE measures the local accuracy of the trajectory [52]. 5-frame pose evaluation [53], breaks down the trajectory of the full sequence into 5-frame snippets with the reference coordinate frame adjusted to the central frame of each snippet. The trajectories are aligned by using the method [54]. Given this transformation S, the absolute trajectory error at the time step i is computed as $F_i = Q_i^{-1} S P_i$, between the estimated P_i and the ground truth Q_i. The root mean squared error over all time of the translational components ($trans$) in [52], i.e.,

$$RMSE(F_{1:n}) = \left(\frac{1}{n} \sum_{i=1}^{n} \|trans(F_i)\|^2 \right)^{1/2}. \quad (2)$$

3.2.3. Network evaluation metrics

There are some metrics used for evaluating the proposed network, including parameters (M), FLOPs, and speed (ms) [55]. Researchers utilize this metric to evaluate the consumption of memory resources. FLOPs, floating point operations, is used to measure the time complexity of the algorithm. The speed of the model means the number of reasoning times that the model can execute in one second.

4. Depth estimation from an endoscope

Methods for depth estimation from image sequences acquired by monocular endoscopes can be divided into two main categories: the first category is based on multi-view geometry, and the second category is based on deep learning. The classification of depth estimation methods is shown in Fig. 2. Each major category can be divided into specific sub-categories according to different theoretical methods.

4.1. Geometry-based methods

Geometry-based methods rely intensively on feature extraction and feature matching. The traditional features often used in endoscopic scenarios are SIFT [56], SURF [57], ORB [58], BRIEF [59] and DISP-SIFT [60]. In order to improve the speed, the fuzzy theory is used to optimize the matching strategy [48]. [61] evaluate the performance of traditional eight key point detectors and six feature descriptors in knee arthroscopic images. [62] improve the extraction process of ORB feature points based on motion vectors. [63] use dense contour clues to achieve target tracking and augmented reality navigation in endoscopic surgery. Geometry-based methods contains SfS [11], SfM [48], SLAM [13] and MVS [14]. The results of SLAM methods are shown in Fig. 5. The results of SfS, SfM, and MVS are shown in Fig. 4. The performance of surface estimation, pose estimation, processing time, robustness, and the scene representation of each method are shown in Table 3. For simplicity, we use different abbreviations in the table to represent different metrics.

4.1.1. SfS

Shape from shading (SfS) deals with the recovery of 3D shape which can be in terms of depth Z, the surface norm (nx,ny,nz), or surface gradient (p,q) from a single monocular image [75]. If the scene satisfies several assumptions (for example, there is only one light source in the scene, and its intensity and pose relative to the camera are known. In addition, the light reflected by the object conforms to the Lambert rule, and the object surface has a constant albedo), then SfS can use the brightness of the pixel to estimate the angle between the camera and the normal at that pixel [11]. Turan et al. [11] first filter the keyframes in the endoscopic image sequence and then use SfS to restore the three-dimensional structure, as shown in Fig. 4(a).

4.1.2. SfM

Structure from motion (SfM) refers to the phenomenon that 3D structures of the scene can be recovered from the projected 2D (retinal) motion field of a moving object. The pipeline of SfM is feature extraction and matching, pose estimation, 3D point triangulation, and bundle adjustment. In SfM, determining the correspondence between images plays a key role in reconstruction performance. To solve the problem of slow traditional feature matching, Rattanalappaiboon et al. [48] propose a fuzzy zoning method, which creates an adaptive matching region for each feature point according to the brightness, thus limiting the search space (Fig. 4(b)). Feature matching of endoscopic images

Table 3
Survey of the geometry-based methods mentioned in this paper. '–' represents the item that is not reported in the reference paper. The numbers in [] are the minimum and maximum values that appear in the reference. 'RMSE' is the root mean square error. 'RMSD' is the abbreviation for root mean square distance, which calculates the error of two sampled surfaces on the Z axis. Only medical-related public datasets are listed in the data availability column. '×' indicates that medical data is not available. '✓' indicates that the data can be downloaded.

Reference	Category	Surface estimation			Processing time	Robustness	Scene represent
		Data	Availability	Metric			
[13]	SLAM	Synthetic	×	Median error: 0.36 mm	18 ms per frame	Rigid	Sparse points (50–100)
[48]	SfM	Synthetic	×	Average error: [0.029 0.041]	Feature matching: 55.59 s per image pair	Rigid	Point cloud Poisson surface
[64]	SLAM	Phantom	×	RMSE: 4.1 mm	Tracking (average): 25 ms per frame	Rigid	Point cloud
[65]	SfS+SLAM	Phantom	×	RMSE: [0.023, 0.044]	55 ms	Non-rigid	Surfel
[11]	SfS	Phantom	×	RMSE: [2.14 cm, 4.45 cm]	919.15 ms per frame pair	Rigid	Point cloud
[66]	SLAM	Synthetic	✓	RMSD: [2.54 mm, 3.66 mm]	Sparse tracking: 25 ms (max) Dense pipeline: 600 ms	Rigid	Point clouod Mesh
[49]	SfM	Phantom	×	Diameter ratio: 99.33%	–	Rigid	Point cloud
[67]	SLAM	Phantom	✓	Average RMSE: [0.3 mm, 6.1 mm] stereo methods as GT	Sparse tracking: 730 ms Dense tracking: 16 280 ms	Rigid	Point cloud
[68]	SLAM+MVS	Phantom, cadavery	✓	RMSE: [0.2 mm, 0.5 mm]	–	Rigid	Point cloud
[69]	SLAM	Phantom	×	–	Sparse tracking: 80 ms per frame	Rigid	Point cloud
[70]	SLAM	Phantom	×	RMSE: 4.86 mm	Tracking: 13 ms	Non-rigid	3D points
[71][a]	SLAM	Phantom	✓	RMSE: [5 mm, 17 mm]	Deformable tracking: 50 ms Deformable mapping: 400 ms	Non-rigid	3D Points
[72][b]	SLAM	Phantom	✓	RMSE: [3 mm, 12.56 mm]	–	Non-rigid	3D points
[73][c]	SfM Learning	Phantom	×	Number of sparse points: [6763, 45 654] per case	37 ms per image pair	Rigid	Point cloud
[74]	SfS	Phantom	×	Mean absolute error: [0.534, 1.757]	–	Rigid	–

[a] https://github.com/UZ-SLAMLab/DefSLAM.
[b] https://github.com/UZ-SLAMLab/SD-DefSLAM.
[c] https://github.com/lppllppl920/DenseDescriptorLearning-Pytorch.

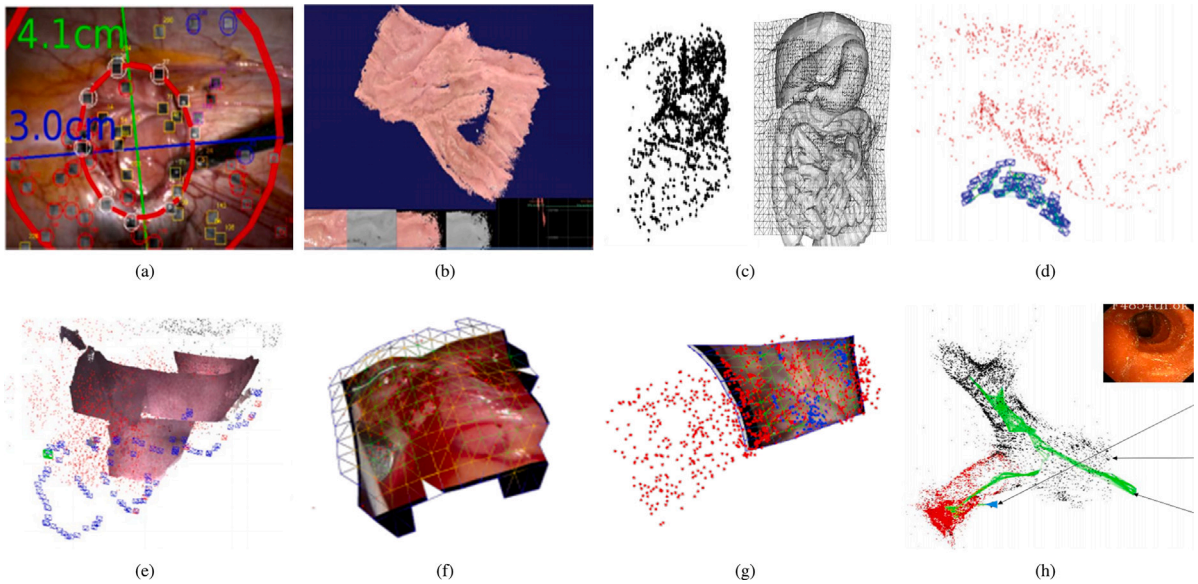
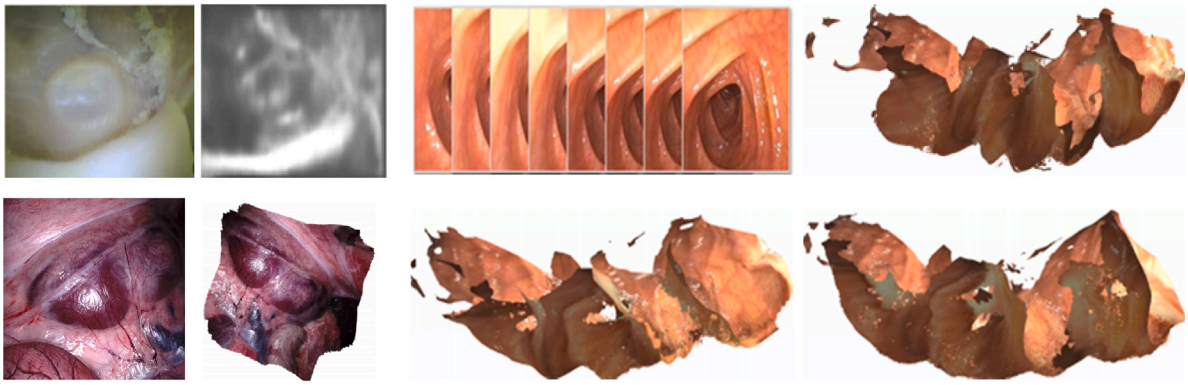


Fig. 5. Results of SLAM methods using the monocular endoscopy. (a) shows the image from laparoscopic surgery [13]; (b) is the reconstructed stomach phantom model [65]; (c) [66] shows the sparse result of abdominal organs; (d) represents the sparse point cloud of the phantom model [64]; (e) [67] is the dense result of abdominal scenes; (f) shows the 3D reconstruction of the heart sequences [71]; (g) shows the result from Hamlyn dataset [72]; (h) is the sparse point cloud reconstructed from the bronchoscopic images [69].

with weak structure and texture can only obtain sparse correspondence. To solve this problem, dense optical flow is used in [49] is used to determine the dense correspondence between the point sets of image pairs with small displacement, as shown in Fig. 4(d). Yang et al. [76] combine the contour feature and SIFT feature to improve the effect of SfM results. Widya et al. [3] collect endoscopic images after spraying indigo carmine (IC) on the stomach, and then use SfM to perform global reconstruction of the stomach. The result can be found in Fig. 4(c). Most of the follow-up works leverage SfM to provide supervision information [73,77].



(a) (b)

Fig. 6. 3D visualization results from learning-based methods that only use visual cues. There are 4 sub-figures in (a) and (b) respectively. The first row in (a) [78] shows the result of the stomach phantom model, and the second row [79] is the result from abdominal images. The first figure in (b) is the input image. The other figures in (b) are the reconstructed point cloud of the colon from [80,81], and [38].

4.1.3. SLAM

Simultaneous localization and mapping (SLAM) is the computational problem of constructing or updating a map of an unknown environment while simultaneously keeping track of an agent's location within it. Grasa et al. [13] propose a monocular vision SLAM algorithm combining joint compatibility branch and bound (JCBB) and EKF-SLAM for localization and reconstruction in the endoscope, but the reconstruction result is only a sparse point set (Fig. 5(a)).

ORB-SLAM [51] is a widely applied method and has stable performance in the natural environment. Mahmoud et al. [64] use ORB-SLAM [51] for the first time to estimate the endoscopic position and 3D structure of the surgical scene, and search for unmatched ORB features in adjacent keyframes to increase the number of reconstruction points. The result is shown in Fig. 5(d). Ye et al. [82] realize online tracking and localization of gastrointestinal endoscopic optical biopsy based on Haar-like random binary descriptor of the local region. Although this method can obtain a certain amount of feature points, it is still relatively sparse. Turan et al. [65] propose a non-rigid dense direct SLAM method. First, the initial depth is estimated according to the shape from shading algorithm. For each new incoming frame, the pose is estimated according to photometric and geometric constraints, and it is fused with the initial surfel model (Fig. 5(b)). Wang et al. [83] leverage dense contour clues to achieve target tracking and augmented reality navigation in endoscopic surgery. Chen et al. [66] retrain the bag of words model for surgical data and utilize the moving least squares (MLS) smoothing algorithm and Poisson surface reconstruction framework to process sparse point cloud data sets in real-time to build dense surfaces (Fig. 5(c)). Mahmoud et al. [67] propose a real-time tracking and intensive reconstruction method of a hand-held monocular endoscope based on ORB-SLAM, as shown in Fig. 5(e). Zhou et al. [44] put forward an improved ORB-SLAM-based video soft tissue surface real-time intensive reconstruction method, which still utilizes ORB as its feature point matching method.

To meet the challenge of large soft tissue deformations and long time-consuming, researchers have attempted to design novel SLAM algorithms. Lamarca et al. [70,71] present a deformable SLAM method called DefSLAM. The whole system includes two threads: deformation tracking and deformation mapping. The deformation tracking uses shape from template (SfT) for the camera pose and the deformation mapping leverages non-rigid structure from motion (NR-SfM) to estimate the surface and update the template used in the tracking thread. The deformation tracking thread recovers the camera pose and observed map deformation at frame rate by SfT processing on the template of static scene shape modeling. The key idea of SfT is to use the pixel point position and its first-order differential structure at the same time, which can easily extract features from the distortion between the template and the input image. The deformation mapping thread runs in parallel with tracking and updates the template at the key frame rate by processing a batch of full perspective keyframes of isometric NR-SfM. Gómez-Rodríguez et al. [72] combine direct and indirect methods in SD-DefSLAM, including enhanced illumination invariant Lucas Kanade tracker for data association, geometric bundle adjustment for pose and deformable map estimation, and feature descriptor for camera relocation. The results are shown in Fig. 5(f) and (g), respectively. Convolutional Neural Network (CNN) is trained for specific application fields to detect and segment dynamic objects. Dense-ArthroSLAM [68] is a system used in arthroscopy that combines SLAM and MVS for the dense reconstruction of selected keyframes. In [69], SLAM with more strict restrictions can also be used for bronchial navigation (Fig. 5(h)). Ma et al. [81] present a method combining DSO [84] and dense depth estimation based on depth learning is proposed, which achieves better results in colonoscopy reconstruction.

4.1.4. MVS

Multi-view stereo (MVS) is the general term given to a group of techniques that use stereo correspondence as their main cue and use more than two images [14]. [50] proposed the most advanced multi-view stereo (MVS) algorithm based on image patch similarity, which usually fails to obtain dense reconstruction from endoscopic images with weak texture. The result can be found in Fig. 4(e). Overall this method still has limitations in computational efficiency, reconstruction sparsity, depth estimation accuracy, and other issues.

In summary, geometry-based methods demonstrate the feasibility of 3D reconstruction based on endoscopic images. These methods have great advantages in the interpretability of the algorithm compared with the learning-based methods. Since it is in the early stage of development, the data utilized by each method is basically self-designed. The evaluation process and metrics are not completely unified.

4.2. Learning-based methods

Deep learning (DL) based approaches can be divided into methods that only use visual clues and methods that utilize auxiliary data. The method of visual clues and auxiliary data are summarized in Table 4.

4.2.1. Visual-cues methods

Depth estimation methods in natural scenes have been studied to some extent and typically leverage real depth values as supervised signals to model the problem as a regression or classification problem. However, true depth values are difficult to obtain in an endoscopic environment. It is not until after unsupervised methods are put forward, those deep learning methods are formally applied to endoscopic depth estimation tasks. The visual results are shown in Fig. 6. Zhou et al. [53] propose an unsupervised training method using

Table 4
Survey of learning-based methods with images mentioned in this paper. '–' represents the item that is not reported in the reference paper. 'RMSE' is the root mean square error. The numbers in [] represent the minimum and maximum values reported in the references. Only medical-related public datasets are listed in the source column. '×' indicates that medical data is not available. '✓' indicates that the data can be downloaded according to the reference. '†' means that the results are extracted from [79].

Reference	Theory	Depth estimation			Processing time	Robustness	Scene Represent.
		Data	Availability	Results			
[78]	Visual	Phantom	×	Qualitative	–	Non-rigid	–
[77][a]	Visual	Surgery	✓	Average residual error: 0.38 ± 0.13 mm	–	Non-rigid	Point cloud Poisson surface
[53]†	Visual	Phantom	✓	RMSE: 6.896	–	Non-rigid	TSDF Mesh
[80]†	Visual	Phantom	✓	RMSE: 5.606	3.8 ms	Non-rigid	TSDF Mesh
[85]†	Visual	Phantom	✓	RMSE: 5.988	–	Non-rigid	TSDF Mesh
[46]	Visual	Phantom	✓	Mean Relative Error (MRE): 0.168	17 ms	Non-rigid	–
[25][b]	Visual	Phantom	✓	RMSE: [1.428, 24.026]	15 ms	Non-rigid	TSDF Mesh
[86]	Visual	Phantom	×	RMSE: 14.195	–	Non-rigid	Point cloud
[31][c]	Visual	Phantom	✓	RMSE: [0.1785 ± 0.0265, 0.2966 ± 0.0622]	4.2 ms	Non-rigid	Point cloud Mesh
[81][d]	Visual+SLAM	Synthetic Real+SfM Synthetic+SfM	× × ×	Pearson correlation coefficient (CORR): 0.3909 CORR: 0.5853 CORR: 0.3260	84 ms	Non-rigid	Point cloud Surfel
[87]	Visual	Synthetic	✓	RMSE: 0.701	28 ms	Non-rigid	Mesh
[39]	Visual+SLAM	Synthetic	✓	RMSE: 0.521	–	Non-rigid	TSDF Mesh
[26][e]	Visual	Surgery GT from CT model	×	Average residual error: 0.69 ± 0.14 mm	Average runtime: 127 min per sequences	Non-rigid	Point cloud Mesh
[45,88][f]	Visual	Phantom	✓	Chamfer distance between point clouds: [0.075, 0.545]	–	Non-rigid	Point cloud
[4][g]	Visual	Synthetic	✓	RMSE: 0.057	–	Non-rigid	–
[79][h]	Visual	Phantom	✓	RMSE: 4.925	3.8 ms	Non-rigid	TSDF Mesh
[89]	Visual	Phantom	✓	RMSE: [5.975, 11.732]	–	Non-rigid	Surfel
[90]	Visual+SLAM	Synthetic	✓	RMSE: 0.9375 mm	–	Non-rigid	TSDF
[91][i]	Visual	Phantom	✓	RMSE: 13.405 ± 5.20	–	Non-rigid	–
[92]	Visual	Synthetic Phantom	✓ ✓	RMSE: 0.043 RMSE: 0.047	–	Non-rigid	Point cloud
[93]	Visual	Synthetic	✓	RMSE: 0.650	–	Non-rigid	–
[94][j]	Real to Synthetic	Phantom	✓	Mean accuracy: 1.5 mm	–	Non-rigid	–
[41]	GAN	Phantom	×	SSIM: 0.59	–	Non-rigid	–
[47]	Real to Synthetic CNN+CRF	Synthetic Phantom	✓	RMSE: 0.612 RMSE: 0.973	–	Non-rigid	Point cloud
[95]	Conditional GAN	Phantom	×	RMSE: 0.054	–	Non-rigid	Surfel
[35]	Conditional GAN	Synthetic Phantom	✓	RMSE: 0.175 cm RMSE: 1.655 cm	–	Non-rigid	Point cloud
[96]	Cycle GAN	Surgery	×	Qualitative	–	Non-rigid	Point cloud
[97]	Cycle-Consistent GAN	Phantom	×	RMSE: 10.6 ± 3.0 mm	–	Non-rigid	–
[2]	GAN	Phantom	✓	RMSE: 7.532 mm	–	Non-rigid	Point cloud
[98][k]	Transfer Training	Phantom	✓	RMSE: [0.029, 13.893]	–	Non-rigid	Point cloud

[a] https://github.com/lppllppl920/EndoscopyDepthEstimation-Pytorch.
[b] https://github.com/UZ-SLAMLab/Endo-Depth-and-Motion.
[c] https://github.com/CapsuleEndoscope/EndoSLAM.
[d] https://github.com/RicardoEspinosaLoera/RNN-SLAM.
[e] https://github.com/lppllppl920/DenseReconstruction-Pytorch.
[f] https://github.com/LONG-XI/Endoscopic-3D-Point-Clouds-Datasets/.
[g] https://github.com/ckLibra/Self-Supervised-Depth-Estimation-for-Colonoscopy.
[h] https://github.com/ShuweiShao/AF-SfMLearner.
[i] https://github.com/EndoluminalSurgicalVision-IMR/TCL.
[j] http://www.marcovs.com/bronchoscopy-navigation.
[k] https://github.com/YYM-SIA/LINGMI-MR.

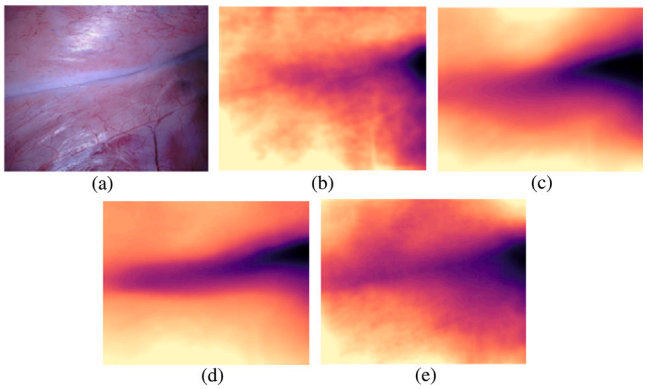


Fig. 7. Qualitative depth comparison on the SCARED dataset. (a) is the input image. (b), (c), (d), and (e) are the results of [31,79,80,85].

only monocular video sequences. The method uses the computed depth and poses as mediators, and warps nearby views to the target view as supervised information. Godard et al. [99] leverage binocular videos instead of depth truth to train the fully convolutional network. The authors hypothesize that, given a pair of calibrated binocular images, the network can fit 3D information about a scene if one image can be reconstructed from the other. In this deep neural network, the disparity map is generated by using the epipolar geometry constraint, left and right image consistency, and image reconstruction loss.

The first article [78] applies unsupervised depth estimation to endoscopy. The authors use a fully convolutional depth estimation approach with a similar structure to the method in [53]. Two encoder–decoder structures are used in this method, the first for depth estimation and the second for pose estimation and mask generation. The feature maps of the encoder in the depth estimation network are involved in the decoder process via short connections. The last few convolutional layers of the decoder are followed by a prediction layer, and each prediction is used as input to the next convolution. The pose and mask generation networks use the same encoding structure. There are two branches after the encoder, one generating the rotation and translation matrices and the other using the decoder structure to generate the masks. The results are shown in the first row of Fig. 6(a).

Based on the methods mentioned above, the researchers found that a network structure containing two branches could achieve better performance, and this structure became the baseline for subsequent methods. Godard et al. [80] propose the baseline method that on the basis of [99]'s network framework. The predictor behind the decoder in the depth estimation network and the decoder in the pose estimation network is deleted. A separate network is used to predict the pixels of scenes violating the static assumption. Because of the good performance, most researchers regard unsupervised depth estimation as an image reconstruction problem. The framework includes a depth network and a separate pose estimation network. To deal with edge conditions, such as object motion and occlusion, predictive interpretable masks are used. Li et al. [77] propose a self-monitoring method to train convolutional neural networks for intensive depth estimation from monocular endoscopic data. Supervised signals are derived from the positional and sparse point clouds of the motion recovery structure. Recasens et al. [25] leverage monodepth2 [80] in this work to train an endoscopic depth estimation network to obtain the depth corresponding to each image. Ozyoruk et al. [31] put forward EndoSfMLearner, which is an unsupervised monocular depth and pose estimation method. This method combines residual networks and a spatial attention module to focus on highly textured tissue areas. Li et al. [86] add the LSTM module in the pose estimation network to model time information, thus improving the accuracy of pose estimation. Shao et al. [79] joint use optical flow appearance flow to deal with the brightness inconsistency problem, the point cloud is shown in the second row of Fig. 6(a). Zhang et al. [87] propose a network that shares an encoder and contains two branches in the decoder. The two branches estimate the depth information and normal information respectively. At the same time, they improve and design a set of training loss functions to solve the challenges such as illumination inconsistency. It may be the first deep-learning network that can be utilized in clinical applications.

In summary, for monocular depth estimation networks using only visual cues, there are basically two components, the first component is the -pose estimation network and the second component is the depth estimation network. The training strategy is mainly to reconstruct the image for supervision by means of the predicted poses and depth predictions. The backbone of the current popular network structure is based on ResNet [100]. Most depth estimation networks follow a U-Net structure. Previously, we introduce the development of methods in chronological order. We also summarize the paper from the perspectives of interest to researchers, such as the selection of loss function, feature matching, and lighting transformation.

Loss function in DL. The commonly used loss function is the reconstructed image loss. It contains photometric error (L_p) and smooth term (L_s). Photometric error function [80], i.e.,

$$L_p = \sum_{t'} \frac{\alpha}{2}(1 - SSIM(I_t, I_{t'})) + (1-\alpha)\|I_t - I_{t'}\|_1, \tag{3}$$

where $\alpha = 0.85$, I_t is the target frame and the $I_{t'}$ is the synthesized frame warped according to ego-motion estimation. Structure similarity index measure (SSIM) is defined in [101]. The L_p minimizes the photometric reprojection error between the target frame and the transformed source frame. The smoothness loss (L_s) is defined in [99] as follows:

$$L_s = |\partial_x d_t^*| e^{-|\partial_x I_t|} + |\partial_y d_t^* \partial| e^{-|\partial_y I_t|}, \tag{4}$$

where d_t^* is the mean-normalized inverse depth. ∂_x and ∂_y are gradients in the x and y directions of the image, respectively. The L_s is designed to obtain locally smooth surfaces and suppose that the discontinuities occur at image gradients.

Feature matching with DL. Feature matching is a key step in depth estimation. It is a good idea to employ the deep-learning method to improve the performance of feature matching. The key problem is that the data available for training in the medical environment is very limited. Liu et al. [73] apply the output of SfM as the supervision signal to transform the feature matching problem into the point location problem. Yang et al. [76] employ contour cues of the colonoscope images to improve the performance of SfM and provide a better-supervised signal for the learning methods. Liu et al. [26] also use the feature matching results to recover the 3D surface.

Inconsistency of illumination. The main challenge in the medical scene is that the light source in the human cavity will change violently. Aiming at the problems of reflection and dark light in endoscope images, an image restoration and enhancement system based on depth learning is proposed in [102]. A recurrent neural network (RNN) [103] is designed and trained to adjust the gamma value in the process of gamma correction. In 2021, Ozyoruk et al. [31] use a brightness-aware photometric loss to improve the fast frame-to-frame illumination changes common in the endoscopic video. Shao et al. [79] joint leverage optical flow and appearance flow to deal with the brightness inconsistency problem. The depth estimation and point clouds are shown in Fig. 7(e) and the second row in Fig. 6(a).

4.2.2. Cross-domain methods

Networks also use the auxiliary data for supervised learning and then perform unsupervised learning on real endoscopic data.

Due to the inability to obtain large-scale and accurate RGB-D data sets, fully supervised depth regression is challenging in the endoscopic scene. Visentini et al. [94] and Chen et al. [95] try to overcome this challenge by training synthetic data. Given the input bronchoscope

frame and its corresponding rendered dataset, Visentini et al. [94] create an RGB transformer network and map the input frame to a texture-free representation similar to computed tomography (CT) rendering. Then, the depth estimation network is proposed to predict the depth of the pseudo-rendered image. Among them, the depth estimation and conversion network are both CNN-structured. The result is shown in Fig. 8(a). Due to the size limitation of endoscopy and deformed mucosa, it is difficult to image tissue topographic maps during colonoscopy. Most existing methods make geometric assumptions or contain prior information, which limits accuracy and sensitivity. A method to avoid these limitations is proposed, using the joint depth convolution neural network conditional random field (CNN-CRF) framework. The estimated depth is employed to reconstruct the topography of the colon surface from a single image. Mahmood et al. [47] train the univariate and pairwise potential energy functions of CRF on CNN. These potential energy functions are generated by developing an endoscope camera model and drawing more than 100 000 anatomical and realistic colon images. The result can be found in Fig. 8(c). However, the gap between the real domain and the composite domain can hardly be bridged by imitating the appearance, which may lead to performance degradation.

The use of adversarial training makes real medical images more like synthetic images. Mahmood et al. [41] preserve related features through self-regularization. With the adversarial training, the generator generates a similar composite representation of the real endoscope image. The loss function of the generator includes a discriminator that classifies the endoscopic image into a real image or a composite image and a self-regularization term that penalizes large deviations from the real image. Rau et al. [35] also leverage the simulation environment, as shown in Fig. 8(d). The difference is that the generator and discriminator of the model are additionally trained on the unlabeled real video frames to adapt to the real colonoscopy environment. The synthetic image of a simple colon model is applied in [95] to train the depth network, and then the region-randomized and realistic images rendered from the computed tomography measurement of the human colon are used to fine-tune. The generator in the generative adversarial network (GAN) is used to predict the depth, and the discriminator is used to provide supervision information. Their method uses the rendering of color images and depth maps to train a fully supervised depth network. In the evaluation, the appearance conversion network is utilized to convert the real endoscope image to the analog image or convert the analog image to the real endoscope image for depth prediction (Fig. 8(e)).

The proposal of CycleGAN further promotes the development of generative learning methods in the field of endoscopes. Widya et al. [96] utilize CycleGAN [104] to convert the real gastroscopic image to the image sprayed with dye to reconstruct the stomach and achieved good results, as shown in Fig. 8(f). Let A and B be two different image fields. CycleGAN consists of two generator and discriminator pairs (G_A, D_A) and (G_B, D_B). The task of the generator is to generate a virtual image by converting the input image from one domain to another and fool the discriminator of its relative domain. On the other hand, the task of the discriminator is to distinguish between the generated image and the real image. In the endoscope scene, A is the endoscope image and B is the composite image with real values. A new GAN network is proposed in [105] to convert colonoscopy images into virtual renderings, and the annotations of Haustral folds are displayed in the renderings. Rivoir et al. [106] propose a new method, which combines unpaired image translation with neural rendering to transform the simulated abdominal surgery scene into a realistic surgery scene. Pfeiffer et al. [107] extend the MUNIT framework and introduce an additional multi-scale structure similarity loss to complete the transformation and construction from virtual image to real image. Wang et al. [108] utilize the prepared real bronchoscope image and virtual depth image as the input of CycleGAN and find the mapping relationship between the real bronchoscope image and virtual depth directly through CycleGAN. The depth image is generated by the virtual bronchoscope system from preoperative CT images. Banach et al. [97] extend and validate an unsupervised learning method, which makes use of the three-loop consistent generation countermeasure network (3cGAN) to generate depth maps directly from bronchoscopic images and registered the depth maps to preoperative CT. The depth map can be found in Fig. 8(b). The authors in [109] train the teacher network on labeled data and then provide pseudo labels for all images to guide the student network. To improve the quality of pseudo tags, they enhance the consistency based on anti-learning compulsion, and they also use confidence to filter noise. Cheng et al. [4] employ synthetic data and real data for depth estimation, in which the GAN mode is leveraged for synthetic data and the self-monitoring mode is used for real data. Karaoglu et al. [2] propose a joint method. Firstly, the depth estimation network is trained with labeled composite images in a supervised manner. Then an unsupervised adaptive scheme against domain features is adopted to improve the performance of real images. A dual network architecture proposed in [46] realizes the colonoscopy coverage method, which is mainly used in the depth estimation method of colonoscopy without calibration and supervision. Without calibration, it is based on a camera sub-network to predict the camera's internal matrix. Widya et al. [110] complete the conversion with the dye in endoscopy.

4.3. Pose estimation

Like depth estimation methods, pose estimation for monocular endoscopic image sequences can be divided into geometric-based methods and learning-based methods. Geometry-based methods mainly contain steps such as feature extraction, feature matching, epipolar geometry, perspective-n-point (PnP), and iterative closest point (ICP). Geometry-based pose estimation methods rely on the accuracy of feature matching. The performance of feature matching in endoscopes is often unsatisfactory due to soft tissue deformation and illumination changes. Learning-based methods mainly include convolution-based feature encoders, which are then converted into pose matrices through convolution. Encoding images using convolution is more robust to photometric transformations in endoscopic scenes.

We summarize the performance of pose estimation in Table 5. We can observe that the first 6 rows are geometry-based methods. They are estimated on the synthetic datasets and then evaluated on phantom datasets. The errors of these methods are reduced. As shown in the middle of the table, the deep-learning-based method [111] firstly performs better than traditional ORB-SLAM. And the authors in [78] prove that the deep-learning-based method has a lower error when the trajectory gets longer. After that, many of the researchers focus on reducing the error of the pose estimation on the phantom and surgery datasets.

4.4. Summary

At present, the learning-based method obtains more dense disparity results, and the reconstruction results are closer to the real environment. The method of deep learning depends on the construction of large data sets. In order to alleviate the difficulty of obtaining data in the medical field, some methods use auxiliary data for training. However, there is a big gap between the auxiliary data and the real data. Other methods directly use endoscope images for training. But the real endoscope video makes it difficult to get the dense ground truth. Therefore, it is possible to obtain the experimental data closest to the human body through animal experiments.

By analyzing Tables 3 and 4, in terms of depth estimation, most of the work focuses on reducing the error rate of estimation and improving the accuracy. Feature matching took up a lot of time in early work based on SfM and SfS. In Table 3, the processing speed of most SLAM-based methods ranges from 13 ms to 730 ms for the tracking thread, and the processing speed for the mapping thread ranges from 400 to 16 280 ms. In Table 4, the deep inference speed of learning-based

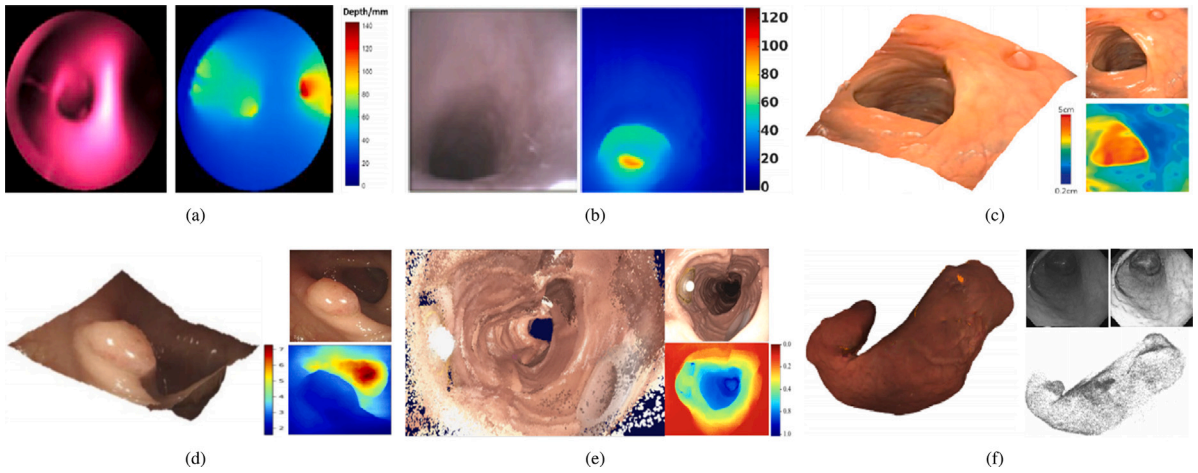


Fig. 8. Results of learning-based monocular methods with auxiliary data. (a) [94] and (b) [97] are the depth estimation from bronchoscopy images. (c) [47], (d) [35], and (e) [95] are the results from the colonoscopy images. (f) shows the reconstructed result of the stomach [96].

Table 5
Survey of pose estimation with images mentioned in this paper. '–' represents the item that is not reported in the reference paper. Type represents the type of endoscope, 'M' represents monocular, and 'S' represents binocular. 'L' means the learning-based method and 'G' means the geometry-based method. 'ATE' means the absolute trajectory error. 'RMSE' is the root mean square error. The numbers in [] represent the minimum and maximum values reported in the references.

Reference	Type	Theory	Pose estimation	
			Data	Results
[13]	M	G	Synthetic	Mean error of translation is 0.82 mm and mean error of rotation is 0.38 degrees
[65]	M	G	Phantom	RMSE: [0.015 cm, 0.049 cm]
[11]	M	G	Phantom	ATE RMSE: [4.10 cm, 8.32 cm]
[66]	M	G	Synthetic	RMSE: [1.24 mm, 4.39 mm]
[68]	M	G	Phantom	Relative pose error (RPE): [0.6, 2.9]
[69]	M	G	Phantom and CT (GT created by registration)	RMSE: 3.06 mm
[111]	M	L	–	ATE: Deep-learning based VO performs better than ORB-SLAM.
[78]	M	L	–	ATE: The trajectory length increase from 10 cm to 50 cm results a change of more than 4 cm in both ORB-SLAM and LSD-SLAM methods, whereas deep learning-based VO error increases around 1 cm.
[53]	M	L	–	–
[80]	M	L	Phantom	[0.0769, 0.0554]
[85]	M	L	Phantom	[0.0767, 0.0509]
[31]	M	L	Phantom	[0.0759, 0.0500]
[97]	M	L	Phantom	6.2 ± 2.9 mm
[103]	M	L	Surgery (GT from COLMAP)	RMSE of absolute pose error (APE): 0.680
[81]	M	L+SLAM	Surgery (GT from COLMAP)	RMSE of absolute pose error (APE): 0.335
[79]	M	L	Phantom	[0.0742, 0.0478]
[112]	M	L	Phantom	7.65 ± 2.99
[89]	M	L	Phantom	ATE: [1.68, 6.27]
[39]	M	L	Synthetic Phantom	RMSE: 28.9 cm RMSE: 7.88 cm
[91]	M	L	Phantom	ATE: [3.110, 4.601]
[38]	S	L	Phantom	Euler angle error (EAE): [−0.02, 0.02] rad
[113]	S	L	Phantom	ATE: [0.744, 4.070]
[114]	S	L	Phantom Surgery	ATE: [0.04, 0.19] ATE: 1.38 ± 0.93

methods ranges from 3.8 ms–84 ms. For real-time operations, it is more appropriate to use learning-based methods for depth estimation. According to Tables 3 and 4, we compare the robust performance of the methods and find that only a few works among the geometry-based methods have good tolerance to non-rigid soft tissues. In learning-based methods, almost all work has no theoretical restrictions on non-rigid soft tissues. Therefore, for surgical scenarios with a lot of soft tissue, it seems more appropriate to use learning-based methods.

5. Depth estimation from stereo endoscopes

Binocular and monocular depth estimation methods have similarities in principle but differences in processing. From classification, binocular depth estimation methods can also be divided into geometry-based methods and learning-based methods. From data sources, binocular endoscopes can acquire two images of the same scene at the same moment, while monocular endoscopes can only acquire one. In the feature-matching process for depth estimation, the monocular depth

Table 6
Survey of stereo methods mentioned in this paper. '–' represents the item that is not reported in the reference paper. 'L' means learning-based methods. 'G' means geometry-based methods. RMSE is the root mean square error. The numbers in [] represent the minimum and maximum values reported in the references. '×' indicates that medical data is not available. '✓' indicates that the data can be downloaded according to the reference.

Reference	Category	Depth estimation			Processing time	Robustness	Scene represent.
		Data	Availability	Metric	per frame		
[115]	G	Phantom	×	Qualitative	–	Non-rigid	–
[116]	G	Phantom	×	RMSE of registration to CT model: 1.08 ± 0.7 mm	–	Non-rigid	Point cloud
[117]	G	Phantom	✓	Average error: [0.46 mm 0.82 mm]	70 ms	Non-rigid	Point cloud
[21]	G	Phantom	✓	RMSE: [1.77 mm, 2.27 mm]	140,000 ms per case	Non-rigid	Point cloud
[44]	L	Phantom	×	RMSE: [1.0 mm, 2.0 mm]	76.3 ms	Non-rigid	TSDF
[118]	G	Phantom	✓	AEDE: [0.27 mm, 1.38 mm]	–	Non-rigid	Point cloud
[38][a]	G	Synthetic	✓	–	–	Non-rigid	Point cloud
[113]	L	Phantom	✓	RMSE: [0.714, 1.705]	83.3 ms	Non-rigid	Surfel
[23][b]	L	Phantom	✓	Qualitative	125 ms	Instrument	Surfel
[119]	L	Phantom	✓	SSIM: 42.41 ± 7.12	35 ms	Instrument	Surfel
[120][c]	L	Phantom	✓	RMSE: 1.119 mm	–	Instrument	3D points
[121]	L	Phantom	✓	MAE: 3.054 mm	–	Instrument	Point cloud
[109]	L	Phantom	✓	End-point disparity error (EPE): 0.77 ± 0.10 px	–	Non-rigid	–
[37]	L	Phantom	✓	RMSE: 5.47 ± 1.34	40 ms	Non-rigid	Point cloud
[122][d]	L	Phantom	✓	Disparity MAE (px): 0.74 ± 0.11	–	Non-rigid	–
[28][e]	L	Phantom	✓	SSIM: 0.921 ± 0.022	14 h per case	Dynamic	NeRF
[29][f]	L	Phantom	✓	SSIM: 0.901 ± 0.021	3 min per case	Dynamic	NeRF
[123][g]	L	Phantom	✓	SSIM: 0.953	9 h per case	Dynamic	NeRF

[a] https://drive.google.com/drive/folders/1cypaTsHpi7TRVKI5cYvzk1UfpmdcOEts.
[b] https://github.com/ucsdarclab/Python-SuPer.
[c] https://github.com/Ultraicee/tpsNet.
[d] https://github.com/HK-Shi/Bidirectional-SemiSupervised-Dual-branch-CNN.
[e] https://github.com/med-air/EndoNeR.
[f] https://github.com/Loping151/LerPlane.
[g] https://github.com/Ruyi-Zha/endosurf.git.

estimation method needs to search and match across the entire image because the pose estimation between the two images is unknown. In contrast, in the binocular depth estimation method, the search between two images can theoretically be narrowed down to one line by binocular correction. Learning-based methods for binocular depth estimation in natural scenes are summarized in detail in [124]. This paper focuses on methods that are often used in endoscopic environments. These methods are summarized in Table 6 and visual results are shown in Fig. 9. The abbreviations mentioned in the table are listed below. AEDE is the abbreviation of the mean of the absolute Euclidean distance errors. EAE represents the Euler angle error and TE represents the translation error. ATE means the absolute trajectory error. RMSE is the root mean square error.

5.1. Geometry-based methods

The common pipeline for geometry-based methods consists of the following steps, firstly extracting features and feature matching, secondly constructing cost volumes and performing parallax estimation, and finally optimization and post-processing. The key step in traditional geometry-based methods is feature matching, and the most commonly used framework is SLAM. Some papers also use other traditional techniques. Cao et al. [125] model the imaging process of binocular endoscope based on a shadow restoration framework and used an improved illumination reflection model. Kumar et al. [116] utilize SfS [126] method to reconstruct the 3D shape from one image.

5.1.1. Feature matching

The main problem in binocular feature matching is how to determine the similarity between pixels, i.e. where a pixel in the left image appears in the corresponding right image. Similarity matching costs based on proximity regions can be employed in the clinic, generally using metrics such as normalized cross-correlation (NCC) [115] and zero-normalized cross-correlation (ZNCC) [44]. In addition, SGM [127] and ELAS [128] are frequently used feature matching methods in endoscopy, such as in [129].

Bernhardt et al. [115] use normalized cross-correlation (NCC) as a metric to calculate the similarity between patches in the left and right images. This method can adaptively and accurately find the best corresponding relationship between each pair of images according to three strict confidence standards, so as to achieve dense matching between two stereo camera views and 3D scene reconstruction. Song et al. [117] leverage the Efficient Large scale Stereo (ELAS) algorithm to calculate the disparity map between pixels. Zhou et al. [44] utilize the zero mean normalized cross-correlation (ZNCC) metric to evaluate the similarity between local image blocks, and then present robust outlier removal and hole filling methods to refine the ZNCC matching results (Fig. 9(b)). Xia et al. [118] present a new feature point detection method based on the gradient change of the image, and a robust edge-preserving stereo matching of the laparoscopic image is proposed by combining light correction and variational theory. The result is shown in Fig. 9(g).

5.1.2. SLAM

The general framework includes two threads. One thread extracts sparse feature points for pose estimation, and the other thread runs dense binocular feature matching for depth estimation and then combines pose to fuse the point cloud. Song et al. [117] propose a SLAM algorithm based on embedded deformation nodes. This algorithm leverages images from stereoscopes to perform deformable intensive reconstruction of the surface, and estimate the deformation field by transforming the last updated model into the current real-time model (Fig. 9(a)). Wang et al. [130] improve the tracking quality of the ORB-SLAM [51] based laparoscope, which leverages motion vectors to improve the extraction process of ORB feature points. Song et al. [131] put forward MIS-SLAM, which is a complete real-time large-scale dense deformable SLAM system with a stereoscope. The CPU is used to execute ORB-SLAM to provide a robust global posture. The GPU performs depth recovery of binocular images to generate a complete 3D reconstruction scene. Zhou et al. [44] present effective post-processing steps for local stereo matching methods in a static environment using binocular endoscopes for common low-texture areas and changing lighting conditions on tissue surfaces. 3D information is extracted from

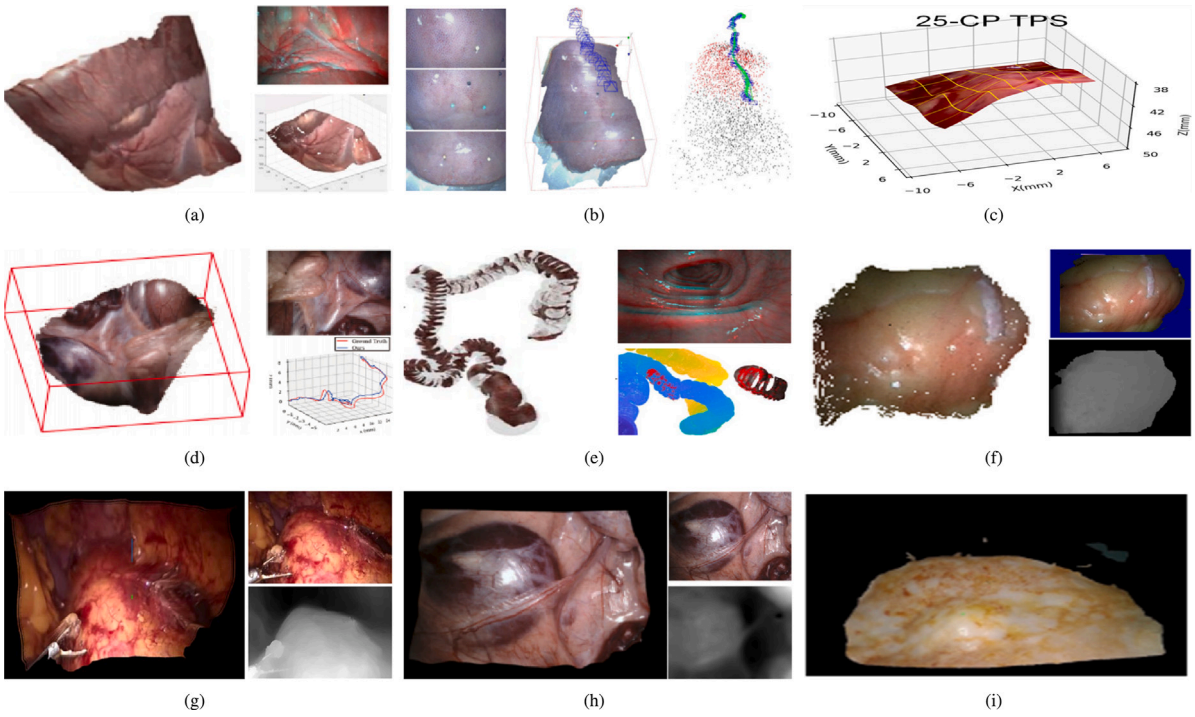


Fig. 9. Visualization of stereo methods. (a) [117] is the result from Hamlyn dataset; (b) shows the reconstruction of the phantom model [44]; (c) [21] represents the result of liver; (d) is the result from abdominal images [113]; (e) shows the result of colon [38]; (f) is the reconstructed point cloud of heart sequences [120]; (g) is the result from surgery images [118]; (h) is the reconstructed point from abdominal images [121]; (i) is the dynamic result of deformation of tissue [23].

video frames through stereo matching, and then 3D models are reconstructed through pose stitching provided by ORBSLAM [51,132]. The average processing time is 76.3 ms. The colon model segmented from CT scanning is used together with the colonoscopy image to achieve the high-precision 3D reconstruction of the colon in [38] (Fig. 9(e)). The specific process is to first estimate the parallax image using binocular images and SGM [127] algorithm, and then register the reconstructed point cloud with the CT model. Based on the established point correspondence, the barycenter-based mapping algorithm is utilized to extract the texture coordinates between the 2D color image and the colon model.

5.2. Learning-based methods

Because binocular endoscopes can provide two images of the same moment as input, most works are done without the aid of auxiliary data and focus mainly on improving performance and solving problems caused by instrument occlusion or soft tissue distortion.

5.2.1. Feature matching with DL

The stereo depth estimation depends on the best matching between the epipolar pixels in the left and right images to infer the depth. Li et al. [133] re-examine this problem from the perspective of sequence-to-sequence correspondence, and replace cost volume structure with dense pixel matching of location information and attention. Long et al. [119] utilize the network in [133] to obtain 3D information. Zhao et al. [134] propose a loss function for surface perception based on the work of [133] to improve the accuracy of the reconstruction.

5.2.2. Self-supervised network

The network framework for binocular depth estimation is similar to the network structures presented previously. The state-of-the-art approaches utilize a transformer structure to improve performance between binocular feature matching.

Luo et al. [21] use the unsupervised method in SfMLearner [99] to estimate the depth (Fig. 9(c)). [113] employ the U-Net framework to extract features at different resolutions, with the differences that firstly four spatial pyramid pooling (SPP) layers are added after the encoder and secondly a 3D convolutional decoder is used. The result is shown in Fig. 9(d). A new self-supervised superposition and siam encoding/decoding neural network is proposed in [135] to calculate the accurate disparity map of 3D laparoscopic depth estimation. An unsupervised optical flow-based depth estimation framework (END flow) is proposed in [136] to train uncorrected binocular video without calibrating camera parameters.

5.2.3. Dynamic reconstruction

For binocular reconstruction methods, researchers focus on reducing the impact of surgical instrument movements and occlusions on reconstruction. In addition, segmentation networks are added for processing dynamically changing surgical instruments, etc. in the surgical environment. Long et al. [119] design a lightweight tool splitter to deal with tool occlusion. This method can gather information according to the time series and be used for surgical scene reconstruction. The running speed is 28 fps. In terms of reconstructing the latest deformed shape of the soft tissue surface, Li et al. [23] propose a new surgical perception framework named SuPer for surgical robot control in the nasal cavity mirror, which tracks the deformable surgical region and the rigid instruments in the region at the same time. It selects the surface as the scene representation and uses the embedded deform graph to track all surface sets. The result can be found in Fig. 9(i). Yang et al. [120] put forward a one-way neural network equivalent to the thin plate spline (TPS) model (proved in this paper) to estimate the subsequent disparity map more accurately by combining the disparity of the previous binocular image. The result is shown in Fig. 9(f). Luo et al. [121] present the correction module to compensate for the imperfect stereo correction. The generating network creates the corresponding reconstruction map according to the disparity map and the original map, and the distinguishing network judges the reconstruction map and the original map (Fig. 9(h)).

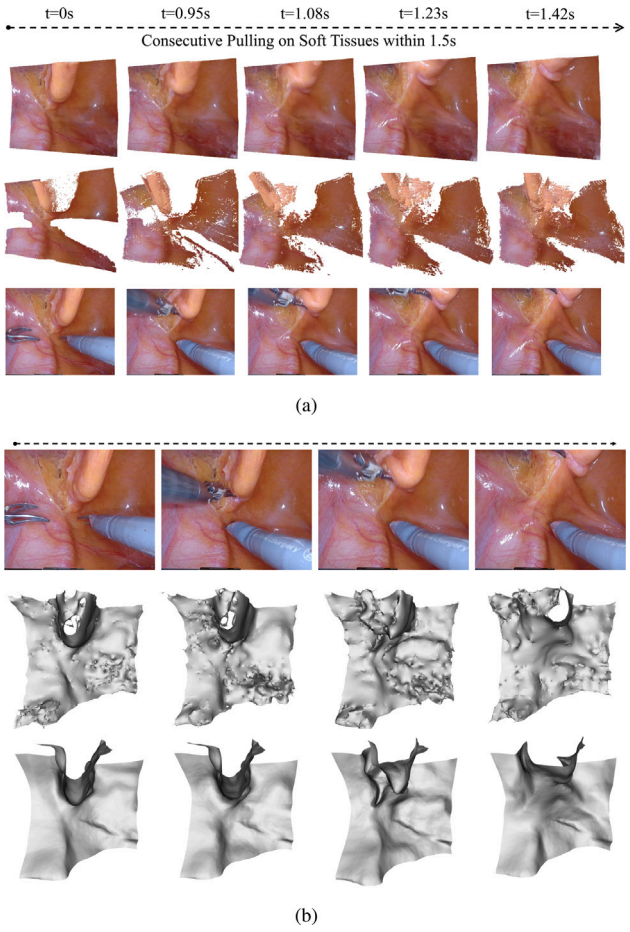


Fig. 10. Dynamic results with NeRF. (a) shows the results of the first paper using NeRF [28]. (b) displays soft tissue shape extraction from NeRF [123].

5.2.4. NeRF

Wang et al. [28] propose a dynamic neural radiation field to represent the deformable surgical scene in MLP (multilayer perceptrons). In order to overcome the limitations of tool occlusion and single view angle, tool masks are employed to guide ray casting, and 3D depth prompts ray travel. The results are shown in Fig. 10(a). This work shows great potential, but in the training process, an accurate disparity map is still needed as the monitoring information. Zha et al. [123] use three neural fields to model surface dynamics, shape, and texture respectively. As shown in Fig. 10(b), this method extracts smoother and more accurate geometric shapes. Yang et al. [29] represent the surgical 3D space and time axis as a 4D volume and accelerate the training process by decomposing the 4D volume into 2D planes.

5.3. Summary

In conclusion, the approaches for binocular endoscopes are the fastest growing in terms of practical surgical applications. In my opinion, this is due to three reasons. Firstly, the hardware of the binocular endoscope can generate depth information. Secondly, laparoscopic surgery provides a larger field of view and maneuvering space for the endoscope. Finally, deep learning methods are rapidly evolving for tasks such as anatomical structure recognition and instrument tracking. Many fields remain to be explored in order to bridge the gap between current methods and practical applications, such as soft tissue deformation, soft tissue hemorrhage, and the interaction of instruments and soft tissue.

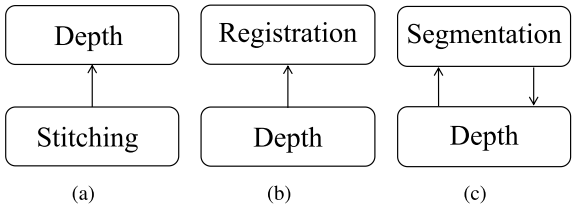


Fig. 11. Relationships between depth estimation and other tasks.

As shown in Table 6, the range of processing time for stereo-based methods is 35 ms–83 ms per frame. It is more convenient to obtain depth information with two parallel lenses in a binocular endoscope. Except for the laparoscope, monocular endoscopes are still used in most examinations. In terms of robust performance, most methods can handle non-rigid surgical scenarios. Some researchers have begun to explore combining segmentation networks to eliminate the impact of equipment movement. However, only a few methods currently deal with the deformation of surgical instruments and soft tissue interactions. By combining Tables 3, 4, and 6, we analyze the advantages and disadvantages of several scene expression methods. The detailed analysis is shown in Table 7. Point clouds have the widest range of applications in terms of robust performance. Surfel [19] seems more suitable for handling mobile equipment. TSDF [20] and Mesh can be used to display the three-dimensional geometric shape of soft tissue. For soft tissue deformation problems, NeRF [28] based methods can achieve better performance.

6. Discussion

Depth estimation and 3D reconstruction tasks are closely related to surgical visualization, surgical navigation, and surgical recognition. The relationship between the discussed tasks and the depth estimation is shown in Fig. 11. For surgical visualization, the depth estimation method can be combined with the mosaic algorithm to improve the visual effect, as shown in Fig. 11(a). 3D reconstruction is the key step of surgical navigation, which provides a good basis for the non-rigid registration steps, as shown in Fig. 11(b). The combination of 3D reconstruction and semantic clues can improve the performance of polyp recognition, segmentation, relocalization, and anatomical structure recognition, as shown in Fig. 11(c).

6.1. Mosaicing and stitching

There are some works that utilize mosaicing and stitching technology to extend vision. Bergen et al. [10] summarize a comprehensive overview of the technology of endoscopic image mosaics and surface reconstruction before 2014. Liu et al. [137] design the hardware of the endoscope with a positioning sensor. Image stitching is carried out through more accurate pose estimation and constraint model, and errors caused by stomach expansion and non-rigid deformation are handled to a certain extent. Gong et al. [138] propose a hybrid rigid registration combining feature matching and template matching to splice all frames into a full view using probe-based conformal laser endoscopy. To solve the problem that traditional stitching methods cannot work on non-rigid scenes in laparoscope, Zhou et al. [139] present a SLAM that can generate dense deformation fields from sparse feature matching and compensate for the deformation of pixels. Omar et al. [140] utilize a multispectral endoscope, build a panoramic view of the stomach based on video splicing technology and achieve registration with the endoscopic image based on optical biopsy targeting. Zhang et al. [141] improve ORB descriptors through Gaussian pyramids, and use the improved descriptors to improve image matching accuracy and perform endoscopic image mosaics.

Table 7
The analysis of several scene representations.

Scene Represent.	The range of application	Advantages	Disadvantages
3D points	The probability of using sparse 3D points in rigid, non-rigid, and instrument scenes is 9%, 6%, and 25%, respectively.	Easy to extraction	Sparse result
Point cloud	The probability of using point cloud in rigid, non-rigid, and instrument scenes is 81%, 38%, and 25%, respectively.	Basic 3D information Convenient conversion to other formats	Presence of noise and holes
Surfel	The probability of using Surfel in non-rigid and instrument scenes is 17% and 50%, respectively.	More geometry information (positions, normals, and radius) Suitable for curve surface	Moderate accuracy
Mesh	19% of articles related to non-rigid scenarios choose mesh.	More geometry information (positions, normals, and faces) Suitable for most surfaces High accuracy	High complexity
TSDF	17% of articles related to non-rigid scenarios choose TSDF.	Volume information Convenient conversion to mesh	Moderate accuracy
NeRF	SOTA methods use NeRF to deal with the large deformation of soft tissues.	Implicit information Efficient geometry representation Reasonable filling-in	Poor generalization performance

To solve the problem that traditional stitching methods cannot work stably in real time for laparoscopic images, Zhou et al. [139] propose a new two-dimensional non-rigid simultaneous positioning and mapping system, which can compensate for the deformation of pixels and perform image stitching in real-time. The image transformation matrix is estimated in [142] based on SIFT feature extraction and K-nearest neighbor feature point matching algorithm to achieve an endoscopic image mosaic.

6.2. Non-rigid registration and navigation

For surgical navigation tasks, 3D reconstruction of surgical scenes is a prerequisite. The registration between the results of endoscopic 3D reconstruction and the preoperative model is the core step to complete the navigation. For example, Raposo et al. [143] match the preoperative model with the bone by estimating the pose and reconstructing the contour through visual markers and guiding the drilling. However, it is difficult to put markers in the endoscope environment manually. In order to visualize the middle ear structure during transtympanic surgery, Hussain et al. [144] set reference points around the tympanic membrane, register CT reconstruction and video images, and then used SURF tracking endoscope and Kalman filter to track the instrument, so that the instrument and model can be displayed on the screen.

Therefore, in the endoscope environment, researchers try to leverage the unmarked method (SfM and SLAM) to reconstruct 3D scenes, and then complete the visualization through non-rigid registration. Qiu et al. [1] create a sparse map of the oral environment based on ORBSLAM [51,132]. In view of the fact that the key parts involved in functional endoscopic sinus surgery are at high risk, Leonard et al. [145] design an image-based enhanced endoscopic navigation method. In [146], the point cloud model of the nasal endoscope is reconstructed according to SfM and MVS methods, and the registration of the point cloud and CT reconstruction model is realized through fast point feature histogram (FPFH) [147] features. According to the surgical path and registration results of the reference planning, navigation systems provide doctors with navigation of the nasal. For the highly active uterus in the human abdominal cavity, authors in [8] fuse preoperative CT and laparoscopic video in real-time through initial registration and updated registration to provide assistance for doctors. In [148], the point cloud generated by SLAM is registered with the model reconstructed by CT to provide navigation for minimally invasive transient surgery.

Reconstruction based on deep learning is also common in surgical navigation. Bano et al. [149] train the improved deep image homography model to obtain the homography matrix between image pairs, so as to achieve the purpose of splicing on the fetal mirror images and expanding the field of vision. The convolutional neural network [150] is trained by the organ model with biomechanics to perform the corresponding search and non-rigid registration of the organ surface. Liu et al. [5] obtain depth estimation and semantic segmentation through joint training of arthroscopic images and convolution-based neural networks.

To solve the problem of deformable registration of preoperative models and intraoperative images, Min et al. [151] propose a two-stage method. Kokko et al. [152] use particle filter to estimate the best similarity transformation of patient kidney model. Modrzejewski et al. [153] provide the dataset of non-rigid fit and flexible body collision reasoning through the live pig model. Aiming at the problem of registration of stereo endoscopic images and CT models in laparoscopic surgery, a two-step method is proposed [154]. koeda et al. [155] utilize a point-to-plane ICP method to register the kidney organ model and the point cloud of endoscopic reconstruction, so as to improve the accuracy of pose estimation. Zhang et al. [156] improve 3D reconstruction with CT and non-rigid registration.

For surgical navigation tasks, the inaccuracy of depth estimation also needs to be considered. The uncertainty of depth estimation should also be considered for robot navigation in [157]. The unreliable depth measurement is eliminated through confidence measurement to improve the stereo accuracy [158].

6.3. Semantic cues

In the endoscope environment, the results of depth estimation can promote some semantic tasks, such as polyp detection, segmentation, tracking, and annotation. We elaborate on the relationship between different tasks and depth estimation.

6.3.1. Polyp detection

Szczypinski et al. [159] utilize color and texture information, and polyp detection of wireless capsule endoscope image is realized through feature extraction, relevant feature subset selection, and classification. Itoh et al. [160] propose a method to improve the accuracy of polyp classification by using depth estimation information and conducting quantitative and qualitative evaluation through different types of polyps.

6.3.2. Segmentation

Jonmohamadi et al. [161] propose the first knee arthroscope 3D semantic mapping system, but the supervision network and depth estimation network are separated, and the segmentation results are directly mapped to the depth estimation. Celik et al. [162] use an unsupervised adaptive technology, which can further improve the performance of gastrointestinal polyps. Transunet [163] method, which has the advantages of both transformer and U-Net, is proposed to enhance more

detailed details by restoring local spatial information and has achieved many excellent performances in multiple organ segmentation and heart segmentation. SegFormer [164] combines transformer. This structure effectively uses local attention and global attention to effectively improve the performance of segmentation networks in natural scenes. Psychogyios et al. [165] propose a learning framework for joint disparity estimation and device segmentation is proposed. A shared feature encoder and two prediction heads are used to perform segmentation and disparity estimation tasks respectively. The accuracy of disparity estimation is improved by supervising the segmentation task.

6.3.3. Relocalization and tracking

Schmidt et al. [166] present a current implicit neural graph to track any number of points in a specified region. Ye et al. [167] propose a matching method based on a multi-scale color histogram feature descriptor and a random forest to relocate the image during the inspection. A method based on optical flow to estimate colonoscopy motion is proposed in [168], which combines CT data with colonoscopy to display the corresponding patient's anatomical structure. Jia et al. [169] put forward the tracking method based on depth learning to eliminate the motion interference, and then the motion is estimated based on ORBSLAM [51,132]. In [170], the poses of the instrument relative to the camera are estimated using the tracking method of the tracking tip and the tracking method of the Hoff Kalman tracking instrument axis.

6.3.4. 3D annotation

Tong et al. [171] convert the nasal endoscope image into a virtual endoscope image for depth estimation, and combine the camera parameters and the pose obtained by the electromagnetic sensor. Such 3D annotation can be firmly and stably anchored to the target structure when the camera moves.

7. Future directions

There have been significant improvements and advancements in endoscopy in recent years, but a few opening problems are still waiting to be solved according to the need for medical applications and technological development. This section suggests the potential research directions of 3D reconstruction and navigation in endoscopy.

7.1. Multi-tasking navigation for the entire procedure

For current surgical navigation, the most important thing is the lack of a fully automated navigation system for multiple tasks. The surgery often lasts for several hours. However, most researchers focus on edited video clips with a duration of about ten minutes or a few seconds. For the depth estimation task, how to save and update the long-period 3D model is also a key issue. Therefore, for long-term surgeries, it is necessary to fully automate the identification of surgical stages, surgical processes, anatomical positions, and other information. The combination of multi-task output results allows doctors to interact more efficiently and comfortably with the navigation system. In [172], the estimated depth image can be applied to the recognition of anatomical position. Later work can consider the surgical stage and anatomical structure as a priori information to help the depth estimation task. In addition, different surgical instruments will be used at different stages of the operation. Making full use of the prior information on the instruments and organs in the operation will greatly improve the dynamic environment reconstruction.

7.2. Network architecture

Improving the robustness of the network is necessary for applications in medical environments. In real surgeries, various situations, such as reflections, blood fog, and burning, affect the performance of depth estimation, detection, and recognition tasks. Also, the complexity of existing models is a key consideration during deployment. In deep learning methods, the performance of the network is greatly affected by the architecture. Thus, the improvement of the backbone is essential for depth estimation. Nowadays, the transformer mechanism has made great progress in improving its performance by using the global processing of information [173]. A transformer-based approach is proposed for the 3D reconstruction in nature scenes instead of the convolutional network as the backbone in [174]. Compared with the fully convolutional network, these characteristics allow the dense prediction network to provide a more fine-grained and globally consistent prediction. When there is a large amount of available data, the depth estimation task of natural scenes has been improved by 28%. However, labeled data is limited in medical scenes and transformer-based networks still remain to be developed.

7.3. Deformable soft tissue

The expression of soft tissue in the surgical scene is an important content that has not been structured and carefully studied. Some researchers consider combining the biomechanical model with the reconstruction results. A method combining the finite element method and deep neural network to learn complex elastic deformation is proposed in [175], which can only use sparse local surface displacement data to solve the deformation state of organs. The current work also mainly focuses on the validation of synthetic and external data. It is difficult to obtain the corresponding ground truth for the deformation caused by the patient's own breathing and heartbeat movement.

7.4. Multi-sensor data fusion

Some works [176,177] have been devoted to the combination of new sensors and endoscopes in hardwares. On the basis of hardware improvement, multi-sensor data fusion is a very potential direction. These works combine multi-camera [178], ultrasound [179], Time-of-Flight [180], VIO [22], etc. However, the comprehensive use of different sensors puts forward higher requirements for the robustness of the method.

8. Conclusion

This paper provides a detailed summary of recent advances in 3D reconstruction and depth estimation in the field of endoscopy over the last 10 years. This decade has seen landmark events, such as the first introduction of deep learning methods to endoscopic depth estimation tasks. We have already observed an almost two-fold increase in the number of papers published in conferences and journals on computer vision, medical engineering, and more in 2021. Many more papers are being published in 2022, indicating that an increasing number of researchers are focusing on this area. The advantages and disadvantages of these methods are summarized and performance comparison can be found in tables and figures. Those who are new to the field can read this paper to get an overview of the technical processes, and to understand the key technologies and core frameworks. The paper also identifies future research directions in this area.

CRediT authorship contribution statement

Zhuoyue Yang: Conceptualization, Formal analysis, Methodology, Visualization, Writing – original draft. **Ju Dai:** Conceptualization, Methodology, Writing – review & editing. **Junjun Pan:** Conceptualization, Methodology, Writing – review & editing.

Declaration of competing interest

None Declared

Acknowledgments

This research is supported by National Key Research and Development Program of China (No. 2022ZD0115902), National Natural Science Foundation of China (Nos. U20A20195, 62272017, 62172437, 62102208), Beijing Municipal Science and Technology Commission (No. Z221100007422005).

Appendix

These metrics are also used to measure the error for depth evaluation.

$$Abs\ Rel = \frac{1}{|\mathbf{D}|}\sum_{d\in\mathbf{D}} |d^* - d|/d^* \tag{A.1}$$

$$Sq\ Rel = \frac{1}{|\mathbf{D}|}\sum_{d\in\mathbf{D}} |d^* - d|^2/d^* \tag{A.2}$$

$$RMSE\ log = \sqrt{\frac{1}{|\mathbf{D}|}\sum_{d\in\mathbf{D}} |\log d^* - \log d|^2} \tag{A.3}$$

$$\delta = \frac{1}{|\mathbf{D}|}\left|\left\{d \in \mathbf{D} | max(\frac{d^*}{d}, \frac{d}{d^*} < 1.25)\right\}\right| \times 100\% \tag{A.4}$$

where $\mathbf{D}$ is the set of predicted depth. d and d^* denote the predicted depth and the ground truth, respectively.

References

[1] L. Qiu, H. Ren, Endoscope navigation and 3D reconstruction of oral cavity by visual SLAM with mitigated data scarcity, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 2197–2204.
[2] M.A. Karaoglu, N. Brasch, M. Stollenga, W. Wein, N. Navab, F. Tombari, A. Ladikos, Adversarial domain feature adaptation for bronchoscopic depth estimation, in: Int. Conf. Med. Image Comput. and Computer-Assisted Intervention, Springer, 2021, pp. 300–310.
[3] A.R. Widya, Y. Monno, K. Imahori, M. Okutomi, S. Suzuki, T. Gotoda, K. Miki, 3D reconstruction of whole stomach from endoscope video using structure-from-motion, in: Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, 2019, pp. 3900–3904.
[4] K. Cheng, Y. Ma, B. Sun, Y. Li, X. Chen, Depth estimation for colonoscopy images with self-supervised learning from videos, in: Int. Conf. Med. Image Comput. and Computer-Assisted Intervention, Springer, 2021, pp. 119–128.
[5] F. Liu, Y. Jonmohamadi, G. Maicas, A.K. Pandey, G. Carneiro, Self-supervised depth estimation to regularise semantic segmentation in knee arthroscopy, in: Int. Conf. Med. Image Comput. and Computer-Assisted Intervention, Springer, 2020, pp. 594–603.
[6] S. Liu, X. Meng, Y. Chu, J. Fan, J. Yang, Surface and volume fusion rendering for augmented reality based functional endoscopic sinus surgery, in: International Conference on Digital Signal Processing, 2021, pp. 103–108.
[7] T. Jia, X. Chen, P. Dong, X. Chen, Monocular endoscope video-based augmented reality for transoral laryngeal tumor resection surgery, in: International Conference on Mechatronics and Machine Vision in Practice, IEEE, 2021, pp. 750–754.
[8] T. Collins, D. Pizarro, S. Gasparini, N. Bourdel, P. Chauvet, M. Canis, L. Calvet, A. Bartoli, Augmented reality guided laparoscopic surgery of the uterus, IEEE Trans. Med. Imaging 40 (1) (2020) 371–380.
[9] P. Sadda, J.A. Onofrey, M.O. Bahtiyar, X. Papademetris, Better feature matching for placental panorama construction, in: Data Driven Treatment Response Assessment and Preterm, Perinatal, and Paediatric Image Analysis, Springer, 2018, pp. 128–137.
[10] T. Bergen, T. Wittenberg, Stitching and surface reconstruction from endoscopic image sequences: a review of applications and methods, IEEE J. Biomed. Health Inform. 20 (1) (2014) 304–321.
[11] M. Turan, Y.Y. Pilavci, I. Ganiyusufoglu, H. Araujo, E. Konukoglu, M. Sitti, Sparse-then-dense alignment-based 3D map reconstruction method for endoscopic capsule robots, Mach. Vis. Appl. 29 (2) (2018) 345–359.
[12] J.L. Schonberger, J.-M. Frahm, Structure-from-motion revisited, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4104–4113.
[13] O.G. Grasa, E. Bernal, S. Casado, I. Gil, J. Montiel, Visual SLAM for handheld monocular endoscope, IEEE Trans. Med. Imaging 33 (1) (2013) 135–146.
[14] S.M. Seitz, B. Curless, J. Diebel, D. Scharstein, R. Szeliski, A comparison and evaluation of multi-view stereo reconstruction algorithms, in: IEEE Conf. on Comput. Vision and Pattern Recog., Vol. 1, IEEE, 2006, pp. 519–528.
[15] R. Hartley, A. Zisserman, Multiple View Geometry in Computer Vision, Cambridge University Press, 2003.
[16] V. Lepetit, F. Moreno-Noguer, P. Fua, Epnp: An accurate o (n) solution to the pnp problem, Int. J. Comput. Vis. 81 (2) (2009) 155–166.
[17] P.J. Besl, N.D. McKay, Method for registration of 3-D shapes, in: Sensor Fusion IV: Control Paradigms and Data Structures, vol. 1611, SPIE, 1992, pp. 586–606.
[18] R.B. Rusu, S. Cousins, 3D is here: Point cloud library (pcl), in: 2011 IEEE International Conference on Robotics and Automation, IEEE, 2011, pp. 1–4.
[19] M. Keller, D. Lefloch, M. Lambers, S. Izadi, T. Weyrich, A. Kolb, Real-time 3d reconstruction in dynamic scenes using point-based fusion, in: International Conference on 3D Vision, IEEE, 2013, pp. 1–8.
[20] B. Curless, M. Levoy, A volumetric method for building complex models from range images, in: Annual Conference on Computer Graphics and Interactive Techniques, 1996, pp. 303–312.
[21] 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.
[22] M. Turan, Y. Almalioglu, E.P. Ornek, H. Araujo, M.F. Yanik, M. Sitti, Magnetic-visual sensor fusion-based dense 3d reconstruction and localization for endoscopic capsule robots, in: IEEE International Conference on Intelligent Robots and Systems, IEEE, 2018, pp. 1283–1289.
[23] Y. Li, F. Richter, J. Lu, E.K. Funk, R.K. Orosco, J. Zhu, M.C. Yip, Super: A surgical perception framework for endoscopic tissue manipulation with surgical robotics, IEEE Robot. Autom. Lett. 5 (2) (2020) 2294–2301.
[24] R. Wei, B. Li, H. Mo, B. Lu, Y. Long, B. Yang, Q. Dou, Y. Liu, D. Sun, Stereo dense scene reconstruction and accurate localization for learning-based navigation of laparoscope in minimally invasive surgery, IEEE. Trans. Biomed. Eng. (2022).
[25] D. Recasens, J. Lamarca, J.M. Fácil, J.M.M. Montiel, J. Civera, Endo-depth-and-motion: Reconstruction and tracking in endoscopic videos using depth networks and photometric constraints, IEEE Robot. Autom. Lett. 6 (4) (2021) 7225–7232.
[26] X. Liu, M. Stiber, J. Huang, M. Ishii, G.D. Hager, R.H. Taylor, M. Unberath, Reconstructing sinus anatomy from endoscopic video – towards a radiation-free approach for quantitative longitudinal assessment, in: Medical Image Computing and Computer Assisted Intervention, 2020, pp. 3–13.
[27] B. Mildenhall, P.P. Srinivasan, M. Tancik, J.T. Barron, R. Ramamoorthi, R. Ng, NeRF: Representing scenes as neural radiance fields for view synthesis, in: European Conference on Computer Vision, 2020.
[28] Y. Wang, Y. Long, S.H. Fan, Q. Dou, Neural rendering for stereo 3d reconstruction of deformable tissues in robotic surgery, in: Int. Conf. Med. Image Comput. and Computer-Assisted Intervention, Springer, 2022, pp. 431–441.
[29] C. Yang, K. Wang, Y. Wang, X. Yang, W. Shen, Neural LerPlane representations for fast 4D reconstruction of deformable tissues, in: Int. Conf. Med. Image Comput. and Computer-Assisted Intervention, 2023, pp. 46–56.
[30] P. Azagra, C. Sostres, Á. Ferrandez, L. Riazuelo, C. Tomasini, O.L. Barbed, J. Morlana, D. Recasens, V.M. Batlle, J.J. Gómez-Rodríguez, et al., EndoMapper dataset of complete calibrated endoscopy procedures, 2022, arXiv preprint arXiv:2204.14240.
[31] K.B. Ozyoruk, G.I. Gokceler, T.L. Bobrow, G. Coskun, K. Incetan, Y. Almalioglu, F. Mahmood, E. Curto, L. Perdigoto, M. Oliveira, et al., EndoSLAM dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos, Med. Image Anal. 71 (2021) 102058.
[32] 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.
[33] 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:2101.01133.
[34] M. Ye, E. Johns, A. Handa, L. Zhang, P. Pratt, G.-Z. Yang, Self-supervised siamese learning on stereo image pairs for depth estimation in robotic surgery, 2017, arXiv preprint arXiv:1705.08260.
[35] A. Rau, P. Edwards, O.F. Ahmad, P. Riordan, M. Janatka, L.B. Lovat, D. Stoyanov, Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy, Int. J. Comput. Assist. Radiol. Surg. 14 (7) (2019) 1167–1176.
[36] V. Penza, A.S. Ciullo, S. Moccia, L.S. Mattos, E. De Momi, Endoabs dataset: Endoscopic abdominal stereo image dataset for benchmarking 3d stereo reconstruction algorithms, Int. J. Med. Robot. Comput. Assist. Surg. 14 (5) (2018) e1926.
[37] Z. Chen, A. Marzullo, D. Alberti, E. Lievore, M. Fontana, O. De Cobelli, G. Musi, G. Ferrigno, E. De Momi, FRSR: Framework for real-time scene reconstruction in robot-assisted minimally invasive surgery, Comput. Biol. Med. (2023) 107121.
[38] S. Zhang, L. Zhao, S. Huang, M. Ye, Q. Hao, A template-based 3d reconstruction of colon structures and textures from stereo colonoscopic images, IEEE Trans. Med. Robot. Bionics 3 (1) (2020) 85–95.
[39] E. Posner, A. Zholkover, N. Frank, M. Bouhnik, C 3 fusion: consistent contrastive colon fusion, towards deep slam in colonoscopy, in: International Workshop on Shape in Medical Imaging, Springer, 2023, pp. 15–34.
[40] F. Qin, S. Lin, Y. Li, R.A. Bly, K.S. Moe, B. Hannaford, Towards better surgical instrument segmentation in endoscopic vision: Multi-angle feature aggregation and contour supervision, IEEE Robot. Autom. Lett. 5 (4) (2020) 6639–6646.

[41] F. Mahmood, R. Chen, N.J. Durr, Unsupervised reverse domain adaptation for synthetic medical images via adversarial training, IEEE Trans. Med. Imaging 37 (12) (2018) 2572–2581.

[42] K. İncetan, I.O. Celik, A. Obeid, G.I. Gokceler, K.B. Ozyoruk, Y. Almalioglu, R.J. Chen, F. Mahmood, H. Gilbert, N.J. Durr, et al., VR-caps: a virtual environment for capsule endoscopy, Med. Image Anal. 70 (2021) 101990.

[43] J.L. Schönberger, J.-M. Frahm, Structure-from-motion revisited, in: Conference on Computer Vision and Pattern Recognition, CVPR, 2016.

[44] H. Zhou, J. Jagadeesan, Real-time dense reconstruction of tissue surface from stereo optical video, IEEE Trans. Med. Imaging 39 (2) (2019) 400–412.

[45] L. Xi, Y. Zhao, L. Chen, Q.H. Gao, W. Tang, T.R. Wan, T. Xue, Recovering dense 3D point clouds from single endoscopic image, Comput. Methods Programs Biomed. 205 (2021) 106077.

[46] D. Freedman, Y. Blau, L. Katzir, A. Aides, I. Shimshoni, D. Veikherman, T. Golany, A. Gordon, G. Corrado, Y. Matias, et al., Detecting deficient coverage in colonoscopies, IEEE Trans. Med. Imaging 39 (11) (2020) 3451–3462.

[47] F. Mahmood, N.J. Durr, Deep learning and conditional random fields-based depth estimation and topographical reconstruction from conventional endoscopy, Med. Image Anal. 48 (2018) 230–243.

[48] S. Rattanalappaiboon, T. Bhongmakapat, P. Ritthipravat, Fuzzy zoning for feature matching technique in 3D reconstruction of nasal endoscopic images, Comput. Biol. Med. 67 (2015) 83–94.

[49] T.-B. Phan, D.-H. Trinh, D. Lamarque, D. Wolf, C. Daul, Dense optical flow for the reconstruction of weakly textured and structured surfaces: Application to endoscopy, in: IEEE International Conference on Image Processing, 2019, pp. 310–314.

[50] G. Bae, I. Budvytis, C.-K. Yeung, R. Cipolla, Deep multi-view stereo for dense 3D reconstruction from monocular endoscopic video, in: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention, Springer, 2020, pp. 774–783.

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

[52] J. Sturm, N. Engelhard, F. Endres, W. Burgard, D. Cremers, A benchmark for the evaluation of RGB-d SLAM systems, in: IEEE International Conference on Intelligent Robots and Systems, 2012, pp. 573–580.

[53] T. Zhou, M. Brown, N. Snavely, D.G. Lowe, Unsupervised learning of depth and ego-motion from video, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1851–1858.

[54] B.K. Horn, Closed-form solution of absolute orientation using unit quaternions, JOSA A 4 (4) (1987) 629–642.

[55] Z. Zhou, X. Fan, P. Shi, Y. Xin, R-MSFM: Recurrent multi-scale feature modulation for monocular depth estimating, in: IEEE Int. Conf. on Comput. Vis., 2021, pp. 12757–12766.

[56] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.

[57] H. Bay, T. Tuytelaars, L.V. Gool, Surf: Speeded up robust features, in: European Conference on Computer Vision, Springer, 2006, pp. 404–417.

[58] E. Rublee, V. Rabaud, K. Konolige, G. Bradski, ORB: An efficient alternative to SIFT or SURF, in: International Conference on Computer Vision, IEEE, 2011, pp. 2564–2571.

[59] M. Calonder, V. Lepetit, C. Strecha, P. Fua, Brief: Binary robust independent elementary features, in: European Conference on Computer Vision, Springer, 2010, pp. 778–792.

[60] J. Dong, S. Soatto, Domain-size pooling in local descriptors: DSP-SIFT, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5097–5106.

[61] A. Marmol, T. Peynot, A. Eriksson, A. Jaiprakash, J. Roberts, R. Crawford, Evaluation of keypoint detectors and descriptors in arthroscopic images for feature-based matching applications, IEEE Robot. Autom. Lett. 2 (4) (2017) 2135–2142.

[62] Y. Chu, H. Li, X. Li, Y. Ding, X. Yang, D. Ai, X. Chen, Y. Wang, J. Yang, Endoscopic image feature matching via motion consensus and global bilateral regression, Comput. Methods Programs Biomed. 190 (2020) 105370.

[63] R. Wang, M. Zhang, X. Meng, Z. Geng, F.-Y. Wang, 3-D tracking for augmented reality using combined region and dense cues in endoscopic surgery, IEEE J. Biomed. Health Inform. 22 (5) (2017) 1540–1551.

[64] N. Mahmoud, I. Cirauqui, A. Hostettler, C. Doignon, L. Soler, J. Marescaux, J.M.M. Montiel, ORBSLAM-based endoscope tracking and 3D reconstruction, in: International Workshop on Computer-Assisted and Robotic Endoscopy, Springer, 2016, pp. 72–83.

[65] M. Turan, Y. Almalioglu, H. Araujo, E. Konukoglu, M. Sitti, A non-rigid map fusion-based direct SLAM method for endoscopic capsule robots, Int. J. Intell. Robot. Appl. 1 (4) (2017) 399–409.

[66] L. Chen, W. Tang, N.W. John, T.R. Wan, J.J. Zhang, SLAM-based dense surface reconstruction in monocular minimally invasive surgery and its application to augmented reality, Comput. Methods Programs Biomed. 158 (2018) 135–146.

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

[68] A. Marmol, A. Banach, T. Peynot, Dense-ArthroSLAM: Dense intra-articular 3-D reconstruction with robust localization prior for arthroscopy, IEEE Robot. Autom. Lett. 4 (2) (2019) 918–925.

[69] C. Wang, M. Oda, Y. Hayashi, B. Villard, T. Kitasaka, H. Takabatake, M. Mori, H. Honma, H. Natori, K. Mori, A visual SLAM-based bronchoscope tracking scheme for bronchoscopic navigation, Int. J. Comput. Assist. Radiol. Surg. 15 (10) (2020) 1619–1630.

[70] J. Lamarca, J.M.M. Montiel, Camera tracking for SLAM in deformable maps, in: European Conference on Computer Vision Workshops, 2019, pp. 730–737.

[71] J. Lamarca, S. Parashar, A. Bartoli, J.M.M. Montiel, DefSLAM: Tracking and mapping of deforming scenes from monocular sequences, IEEE Trans. Robot. 37 (1) (2021) 291–303.

[72] J.J. Gómez-Rodríguez, J. Lamarca, J. Morlana, J.D. Tardós, J.M.M. Montiel, SD-defslam: Semi-direct monocular SLAM for deformable and intracorporeal scenes, in: IEEE International Conference on Robotics and Automation, 2021, pp. 5170–5177.

[73] X. Liu, Y. Zheng, B. Killeen, M. Ishii, G.D. Hager, R.H. Taylor, M. Unberath, Extremely dense point correspondences using a learned feature descriptor, in: IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 4846–4855.

[74] J. Fan, Y. Feng, J. Mo, S. Wang, Q. Liang, Texture-less surface reconstruction using shape-based image augmentation, Comput. Biol. Med. 150 (2022) 106114.

[75] T. Ping-Sing, M. Shah, Shape from shading using linear approximation, Image Vis. Comput. 12 (8) (1994) 487–498.

[76] Z. Yang, J. Pan, R. Li, H. Qin, Scene-graph-driven semantic feature matching for monocular digestive endoscopy, Comput. Biol. Med. (2022) 105616.

[77] X. Liu, A. Sinha, M. Ishii, G.D. Hager, A. Reiter, R.H. Taylor, M. Unberath, Dense depth estimation in monocular endoscopy with self-supervised learning methods, IEEE Trans. Med. Imaging 39 (5) (2020) 1438–1447.

[78] M. Turan, E.P. Ornek, N. Ibrahimli, C. Giracoglu, Y. Almalioglu, M.F. Yanik, M. Sitti, Unsupervised odometry and depth learning for endoscopic capsule robots, in: IEEE International Conference on Intelligent Robots and Systems, 2018, pp. 1801–1807.

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

[80] C. Godard, O.M. Aodha, M. Firman, G. Brostow, Digging into self-supervised monocular DepthEstimation, in: IEEE International Conference on Computer Vision, 2019, pp. 3827–3837.

[81] R. Ma, R. Wang, Y. Zhang, S. Pizer, S.K. McGill, J. Rosenman, J.-M. Frahm, RNNSLAM: Reconstructing the 3D colon to visualize missing regions during a colonoscopy, Med. Image Anal. 72 (2021) 102100.

[82] M. Ye, S. Giannarou, A. Meining, G.-Z. Yang, Online tracking and retargeting with applications to optical biopsy in gastrointestinal endoscopic examinations, Med. Image Anal. 30 (2016) 144–157.

[83] R. Wang, M. Zhang, X. Meng, Z. Geng, F.-Y. Wang, 3-D tracking for augmented reality Using Combined Region and dense cues in endoscopic surgery, IEEE J. Biomed. Health Inform. 22 (5) (2018) 1540–1551.

[84] J. Engel, V. Koltun, D. Cremers, Direct sparse odometry, IEEE Trans. Pattern Anal. Mach. Intell. 40 (3) (2017) 611–625.

[85] J. Bian, Z. Li, N. Wang, H. Zhan, C. Shen, M.-M. Cheng, I. Reid, Unsupervised Scale-Consistent Depth and Ego-Motion Learning from Monocular Video, Vol. 32, 2019.

[86] L. Li, X. Li, S. Yang, S. Ding, A. Jolfaei, X. Zheng, Unsupervised-learning-based continuous depth and motion estimation with monocular endoscopy for virtual reality minimally invasive surgery, IEEE Trans. Ind. Inform. 17 (6) (2021) 3920–3928.

[87] Y. Zhang, J.-M. Frahm, S. Ehrenstein, S.K. McGill, J.G. Rosenman, S. Wang, S.M. Pizer, ColDE: A depth estimation framework for colonoscopy reconstruction, 2021, arXiv preprint arXiv:2111.10371.

[88] L. Chen, W. Tang, T.R. Wan, N.W. John, Self-supervised monocular image depth learning and confidence estimation, Neurocomputing 381 (2020) 272–281.

[89] R. Wei, B. Li, H. Mo, F. Zhong, Y. Long, Q. Dou, Y.-H. Liu, D. Sun, Distilled visual and robot kinematics embeddings for metric depth estimation in monocular scene reconstruction, in: IEEE Int. Conf. Intell. Rob. Syst., IEEE, 2022, pp. 8072–8077.

[90] R. Liu, Z. Liu, J. Lu, G. Zhang, Z. Zuo, B. Sun, J. Zhang, W. Sheng, R. Guo, L. Zhang, et al., Sparse-to-dense coarse-to-fine depth estimation for colonoscopy, Comput. Biol. Med. 160 (2023) 106983.

[91] H. Yue, Y. Gu, TCL: Triplet consistent learning for odometry estimation of monocular endoscope, in: Int. Conf. Med. Image Comput. and Computer-Assisted Intervention, 2023, pp. 144–153.

[92] Y. Liu, S. Zuo, Self-supervised monocular depth estimation for gastrointestinal endoscopy, Comput. Methods Programs Biomed. (2023) 107619.

[93] A. Mathew, L. Magerand, E. Trucco, L. Manfredi, SoftEnNet: Symbiotic monocular depth estimation and lumen segmentation for colonoscopy endorobots, 2023, arXiv preprint arXiv:2301.08157.

[94] M. Visentini-Scarzanella, T. Sugiura, T. Kaneko, S. Koto, Deep monocular 3D reconstruction for assisted navigation in bronchoscopy, Int. J. Comput. Assist. Radiol. Surg. 12 (7) (2017) 1089–1099.

[95] R.J. Chen, T.L. Bobrow, T. Athey, F. Mahmood, N.J. Durr, Slam endoscopy enhanced by adversarial depth prediction, in: KDD Workshop on Applied Data Science for Healthcare, 2019.

[96] A.R. Widya, Y. Monno, M. Okutomi, S. Suzuki, T. Gotoda, K. Miki, Stomach 3D reconstruction using virtual chromoendoscopic images, IEEE J. Transl. Eng. Health Med. 9 (2021) 1–11.
[97] A. Banach, F. King, F. Masaki, H. Tsukada, N. Hata, Visually navigated bronchoscopy using three cycle-consistent generative adversarial network for depth estimation, Med. Image Anal. 73 (2021) 102164.
[98] Y. Yang, S. Shao, T. Yang, P. Wang, Z. Yang, C. Wu, H. Liu, A geometry-aware deep network for depth estimation in monocular endoscopy, Eng. Appl. Artif. Intell. 122 (2023) 105989.
[99] C. Godard, O.M. Aodha, G.J. Brostow, Unsupervised monocular depth estimation with left-right consistency, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6602–6611.
[100] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[101] Z. Wang, A. Bovik, H. Sheikh, E. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.
[102] M. Asif, L. Chen, H. Song, J. Yang, A.F. Frangi, An automatic framework for endoscopic image restoration and enhancement, Appl. Intell. 51 (4) (2021) 1959–1971.
[103] Y. Zhang, S. Wang, R. Ma, S.K. McGill, J.G. Rosenman, S.M. Pizer, Lighting enhancement aids reconstruction of colonoscopic surfaces, in: International Conference on Information Processing in Medical Imaging, Springer, 2021, pp. 559–570.
[104] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 2223–2232.
[105] S. Mathew, S. Nadeem, A. Kaufman, FoldIt: Haustral folds detection and segmentation in colonoscopy videos, in: Int. Conf. Med. Image Comput. and Computer-Assisted Intervention, Springer, 2021, pp. 221–230.
[106] D. Rivoir, M. Pfeiffer, R. Docea, F. Kolbinger, C. Riediger, J. Weitz, S. Speidel, Long-term temporally consistent unpaired video translation from simulated surgical 3d data, in: IEEE International Conference on Computer Vision, 2021, pp. 3343–3353.
[107] M. Pfeiffer, I. Funke, M.R. Robu, S. Bodenstedt, L. Strenger, S. Engelhardt, T. Roß, M.J. Clarkson, K. Gurusamy, B.R. Davidson, et al., Generating large labeled data sets for laparoscopic image processing tasks using unpaired image-to-image translation, in: Int. Conf. Med. Image Comput. and Computer-Assisted Intervention, Springer, 2019, pp. 119–127.
[108] C. Wang, Y. Hayashi, M. Oda, T. Kitasaka, H. Takabatake, M. Mori, H. Honma, H. Natori, K. Mori, Depth-based branching level estimation for bronchoscopic navigation, Int. J. Comput. Assist. Radiol. Surg. 16 (10) (2021) 1795–1804.
[109] H. Shi, Z. Wang, J. Lv, Y. Wang, P. Zhang, F. Zhu, Q. Li, Semi-supervised learning via improved teacher-student network for robust 3D reconstruction of stereo endoscopic image, in: Proceedings the ACM International Conference on Multimedia, 2021, pp. 4661–4669.
[110] A.R. Widya, Y. Monno, M. Okutomi, S. Suzuki, T. Gotoda, K. Miki, Learning-based depth and pose estimation for monocular endoscope with loss generalization, in: Annual International Conference of the IEEE Engineering in Medicine & Biology Society, IEEE, 2021, pp. 3547–3552.
[111] M. Turan, Y. Almalioglu, H. Araujo, E. Konukoglu, M. Sitti, Deep endovo: A recurrent convolutional neural network (rcnn) based visual odometry approach for endoscopic capsule robots, Neurocomputing 275 (2018) 1861–1870.
[112] Y. Gu, C. Gu, J. Yang, J. Sun, G.-Z. Yang, Vision-kinematics-interaction for robotic-assisted bronchoscopy navigation, IEEE Trans. Med. Imaging (2022).
[113] R. Wei, B. Li, H. Mo, B. Lu, Y. Long, B. Yang, Q. Dou, Y. Liu, D. Sun, Stereo dense scene reconstruction and accurate laparoscope localization for learning-based navigation in robot-assisted surgery, IEEE Trans. Biomed. Eng. 70 (2) (2022) 488–500.
[114] M. Hayoz, C. Hahne, M. Gallardo, D. Candinas, T. Kurmann, M. Allan, R. Sznitman, Learning how to robustly estimate camera pose in endoscopic videos, Int. J. Comput. Assist. Radiol. Surg. (2023) 1–8.
[115] S. Bernhardt, J. Abi-Nahed, R. Abugharbieh, Robust dense endoscopic stereo reconstruction for minimally invasive surgery, in: International MICCAI Workshop on Medical Computer Vision, Springer, 2012, pp. 254–262.
[116] A. Kumar, Y.-Y. Wang, C.-J. Wu, K.-C. Liu, H.-S. Wu, Stereoscopic visualization of laparoscope image using depth information from 3D model, Comput. Methods Programs Biomed. 113 (3) (2014) 862–868.
[117] J. Song, J. Wang, L. Zhao, S. Huang, G. Dissanayake, Dynamic reconstruction of deformable soft-tissue with stereo scope in minimal invasive surgery, IEEE Robot. Autom. Lett. 3 (1) (2017) 155–162.
[118] W. Xia, E.C. Chen, S. Pautler, T.M. Peters, A robust edge-preserving stereo matching method for laparoscopic images, IEEE Trans. Med. Imaging (2022).
[119] Y. Long, Z. Li, C.H. Yee, C.F. Ng, R.H. Taylor, M. Unberath, Q. Dou, E-dssr: efficient dynamic surgical scene reconstruction with transformer-based stereoscopic depth perception, in: Int. Conf. Med. Image Comput. and Computer-Assisted Intervention, 2021, pp. 415–425.
[120] B. Yang, S. Xu, H. Chen, W. Zheng, C. Liu, Reconstruct dynamic soft-tissue with stereo endoscope based on a single-layer network, IEEE Trans. Image Process. 31 (2022) 5828–5840.
[121] 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.
[122] H. Shi, Z. Wang, Y. Zhou, D. Li, X. Yang, Q. Li, Bidirectional semi-supervised dual-branch CNN for robust 3D reconstruction of stereo endoscopic images via adaptive cross and parallel supervisions, IEEE Trans. Med. Imaging 42 (11) (2023) 3269–3282.
[123] R. Zha, X. Cheng, H. Li, M. Harandi, Z. Ge, EndoSurf: Neural surface reconstruction of deformable tissues with stereo endoscope videos, in: Int. Conf. Med. Image Comput. and Computer-Assisted Intervention, 2023, pp. 13–23.
[124] H. Laga, L.V. Jospin, F. Boussaid, M. Bennamoun, A survey on deep learning techniques for stereo-based depth estimation, IEEE Trans. Pattern Anal. Mach. Intell. 44 (4) (2022) 1738–1764.
[125] Z. Cao, Y. Wang, W. Zheng, L. Yin, Y. Tang, W. Miao, S. Liu, B. Yang, The algorithm of stereo vision and shape from shading based on endoscope imaging, Biomed. Signal Process. Control 76 (2022) 103658.
[126] M. Visentini-Scarzanella, D. Stoyanov, G.-Z. Yang, Metric depth recovery from monocular images using shape-from-shading and specularities, in: IEEE International Conference on Image Processing, IEEE, 2012, pp. 25–28.
[127] H. Hirschmuller, Stereo processing by semiglobal matching and mutual information, IEEE Trans. Pattern Anal. Mach. Intell. 30 (2) (2008) 328–341.
[128] A. Geiger, M. Roser, R. Urtasun, Efficient large-scale stereo matching, in: Asian Conference on Computer Vision, Berlin, Heidelberg, 2011, pp. 25–38.
[129] M. Mikamo, R. Furukawa, S. Oka, T. Kotachi, Y. Okamoto, S. Tanaka, R. Sagawa, H. Kawasaki, Active stereo method for 3D endoscopes using deep-layer GCN and graph representation with proximity information, in: Annual International Conference of the IEEE Engineering in Medicine & Biology Society, IEEE, 2021, pp. 7551–7555.
[130] C. Wang, M. Oda, Y. Hayashi, K. Misawa, H. Roth, K. Mori, Motion vector for outlier elimination in feature matching and its application in SLAM based laparoscopic tracking, in: Computer Assisted and Robotic Endoscopy and Clinical Image-Based Procedures, 2017, pp. 60–69.
[131] J. Song, J. Wang, L. Zhao, S. Huang, G. Dissanayake, Mis-slam: Real-time large-scale dense deformable slam system in minimal invasive surgery based on heterogeneous computing, IEEE Robot. Autom. Lett. 3 (4) (2018) 4068–4075.
[132] R. Mur-Artal, J.D. Tardós, Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras, IEEE Trans. Robot. 33 (5) (2017) 1255–1262.
[133] Z. Li, X. Liu, N. Drenkow, A. Ding, F.X. Creighton, R.H. Taylor, M. Unberath, Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers, in: IEEE International Conference on Computer Vision, 2021, pp. 6197–6206.
[134] S. Zhao, C. Wang, Q. Wang, Y. Liu, S.K. Zhou, 3D endoscopic depth estimation using 3D surface-aware constraints, 2022, arXiv preprint arXiv:2203.02131.
[135] F. Bardozzo, T. Collins, A. Forgione, A. Hostettler, R. Tagliaferri, StaSiS-Net: A stacked and siamese disparity estimation network for depth reconstruction in modern 3D laparoscopy, Med. Image Anal. 77 (2022) 102380.
[136] Z. Yang, R. Simon, Y. Li, C.A. Linte, Dense depth estimation from stereo endoscopy videos using unsupervised optical flow methods, in: Annual Conference on Medical Image Understanding and Analysis, Springer, 2021, pp. 337–349.
[137] J. Liu, B. Wang, W. Hu, P. Sun, J. Li, H. Duan, J. Si, Global and local panoramic views for gastroscopy: an assisted method of gastroscopic lesion surveillance, IEEE Trans. Biomed. Eng. 62 (9) (2015) 2296–2307.
[138] L. Gong, J. Zheng, Z. Ping, Y. Wang, S. Wang, S. Zuo, Robust mosaicing of endomicroscopic videos via context-weighted correlation ratio, IEEE Trans. Biomed. Eng. 68 (2) (2020) 579–591.
[139] H. Zhou, J. Jayender, Real-time nonrigid mosaicking of laparoscopy images, IEEE Trans. Med. Imaging 40 (6) (2021) 1726–1736.
[140] O. Zenteno, D.-H. Trinh, S. Treuillet, Y. Lucas, T. Bazin, D. Lamarque, C. Daul, Optical biopsy mapping on endoscopic image mosaics with a marker-free probe, Comput. Biol. Med. 143 (2022) 105234.
[141] Z. Zhang, L. Wang, W. Zheng, L. Yin, R. Hu, B. Yang, Endoscope image mosaic based on pyramid ORB, Biomed. Signal Process. 71 (2022) 103261.
[142] yan Liu, J. Tian, R. Hu, B. Yang, S. Liu, L. Yin, W. Zheng, Front. Neurorobot. 16 (2022) 840594.
[143] C. Raposo, J.P. Barreto, C. Sousa, L. Ribeiro, R. Melo, J.P. Oliveira, P. Marques, F. Fonseca, D. Barrett, Video-based computer navigation in knee arthroscopy for patient-specific ACL reconstruction, Int. J. Comput. Assist. Radiol. Surg. 14 (9) (2019) 1529–1539.
[144] R. Hussain, A. Lalande, R. Marroquin, K.B. Girum, C. Guigou, A.B. Grayeli, Real-time augmented reality for ear surgery, in: Int. Conf. Med. Image Comput. and Computer-Assisted Intervention, 2018, pp. 324–331.
[145] S. Leonard, A. Sinha, A. Reiter, M. Ishii, G.L. Gallia, R.H. Taylor, G.D. Hager, Evaluation and stability analysis of video-based navigation system for functional endoscopic sinus surgery on in vivo clinical data, IEEE Trans. Med. Imaging 37 (10) (2018) 2185–2195.
[146] Z. Cui, Y. He, P. Zhang, Y. Hu, H. Jin, S. Liu, Virtual reality navigation system of nasal endoscopy with real surface texture information, in: IEEE International Conference on Real-Time Computing and Robotics, IEEE, 2021, pp. 135–140.

[147] R.B. Rusu, N. Blodow, M. Beetz, Fast point feature histograms (FPFH) for 3D registration, in: IEEE International Conference on Robotics and Automation, IEEE, 2009, pp. 3212–3217.
[148] L. Qiu, H. Ren, Endoscope navigation with SLAM-based registration to computed tomography for transoral surgery, Int. J. Intell. Robot. Appl. 4 (2) (2020) 252–263.
[149] S. Bano, F. Vasconcelos, M. Tella Amo, G. Dwyer, C. Gruijthuijsen, J. Deprest, S. Ourselin, E.V. Poorten, T. Vercauteren, D. Stoyanov, Deep sequential mosaicking of fetoscopic videos, in: Int. Conf. Med. Image Comput. and Computer-Assisted Intervention, Springer, 2019, pp. 311–319.
[150] M. Pfeiffer, C. Riediger, S. Leger, J.-P. Kühn, D. Seppelt, R.-T. Hoffmann, J. Weitz, S. Speidel, Non-rigid volume to surface registration using a data-driven biomechanical model, in: Int. Conf. Med. Image Comput. and Computer-Assisted Intervention, Springer, 2020, pp. 724–734.
[151] Z. Min, L. Liu, M.Q.-H. Meng, Generalized non-rigid point set registration with hybrid mixture models considering anisotropic positional uncertainties, in: Int. Conf. Med. Image Comput. and Computer-Assisted Intervention, Springer, 2019, pp. 547–555.
[152] M.A. Kokko, D.W. Van Citters, J.D. Seigne, R.J. Halter, A particle filter approach to dynamic kidney pose estimation in robotic surgical exposure, Int. J. Comput. Assist. Radiol. Surg. 17 (6) (2022) 1079–1089.
[153] R. Modrzejewski, T. Collins, B. Seeliger, A. Bartoli, A. Hostettler, J. Marescaux, An in vivo porcine dataset and evaluation methodology to measure soft-body laparoscopic liver registration accuracy with an extended algorithm that handles collisions, Int. J. Comput. Assist. Radiol. Surg. 14 (7) (2019) 1237–1245.
[154] D. Reichard, D. Häntsch, S. Bodenstedt, S. Suwelack, M. Wagner, H. Kenngott, B. Müller-Stich, L. Maier-Hein, R. Dillmann, S. Speidel, Projective biomechanical depth matching for soft tissue registration in laparoscopic surgery, Int. J. Comput. Assist. Radiol. Surg. 12 (7) (2017) 1101–1110.
[155] M. Koeda, N. Maeda, A. Hamada, A. Sawada, T. Magaribuchi, O. Ogawa, K. Onishi, H. Noborio, Position and orientation registration of intra-abdominal point cloud generated from stereo endoscopic images and organ 3D model using Open3D, in: International Conference on Human-Computer Interaction, Springer, 2022, pp. 52–65.
[156] S. Zhang, L. Zhao, S. Huang, R. Ma, B. Hu, Q. Hao, 3D reconstruction of deformable colon structures based on preoperative model and deep neural network, in: IEEE International Conference on Robotics and Automation, IEEE, 2021, pp. 1875–1881.
[157] J. Rodriguez-Puigvert, D. Recasens, J. Civera, R. Martinez-Cantin, On the uncertain single-view depths in colonoscopies, in: Int. Conf. Med. Image Comput. and Computer-Assisted Intervention, Springer, 2022, pp. 130–140.
[158] H. Luo, Q. Hu, F. Jia, Details preserved unsupervised depth estimation by fusing traditional stereo knowledge from laparoscopic images, Healthc. Technol. Lett. 6 (6) (2019) 154–158.
[159] P. Szczypiński, A. Klepaczko, M. Pazurek, P. Daniel, Texture and color based image segmentation and pathology detection in capsule endoscopy videos, Comput. Methods Programs Biomed. 113 (1) (2014) 396–411.
[160] H. Itoh, M. Oda, K. Jiang, Y. Mori, M. Misawa, S.-E. Kudo, K. Imai, S. Ito, K. Hotta, K. Mori, Binary polyp-size classification based on deep-learned spatial information, Int. J. Comput. Assist. Radiol. Surg. 16 (10) (2021) 1817–1828.
[161] Y. Jonmohamadi, S. Ali, F. Liu, J. Roberts, R. Crawford, G. Carneiro, A.K. Pandey, 3D semantic mapping from arthroscopy using out-of-distribution pose and depth and in-distribution segmentation training, in: Int. Conf. Med. Image Comput. and Computer-Assisted Intervention, Springer, 2021, pp. 383–393.
[162] N. Celik, S. Ali, S. Gupta, B. Braden, J. Rittscher, Endouda: a modality independent segmentation approach for endoscopy imaging, in: Int. Conf. Med. Image Comput. and Computer-Assisted Intervention, Springer, 2021, pp. 303–312.
[163] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A.L. Yuille, Y. Zhou, Transunet: Transformers make strong encoders for medical image segmentation, 2021, arXiv preprint arXiv:2102.04306.
[164] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J.M. Alvarez, P. Luo, SegFormer: Simple and efficient design for semantic segmentation with transformers, Adv. Neural Inf. Process. Syst. 34 (2021) 12077–12090.
[165] D. Psychogyios, E. Mazomenos, F. Vasconcelos, D. Stoyanov, MSDESIS: Multitask stereo disparity estimation and surgical instrument segmentation, IEEE Trans. Med. Imaging 41 (11) (2022) 3218–3230.
[166] A. Schmidt, O. Mohareri, S. DiMaio, S.E. Salcudean, Recurrent implicit neural graph for deformable tracking in endoscopic videos, in: Int. Conf. Med. Image Comput. and Computer-Assisted Intervention, Springer, 2022, pp. 478–488.
[167] M. Ye, E. Johns, B. Walter, A. Meining, G.-Z. Yang, An image retrieval framework for real-time endoscopic image retargeting, Int. J. Comput. Assist. Radiol. Surg. 12 (8) (2017) 1281–1292.
[168] J. Liu, K.R. Subramanian, T.S. Yoo, An optical flow approach to tracking colonoscopy video, Comput. Med. Imaging Graph. 37 (3) (2013) 207–223.
[169] T. Jia, Z.A. Taylor, X. Chen, Long term and robust 6dof motion tracking for highly dynamic stereo endoscopy videos, Comput. Med. Imaging Graph. 94 (2021) 101995.
[170] C. Loukas, V. Lahanas, E. Georgiou, An integrated approach to endoscopic instrument tracking for augmented reality applications in surgical simulation training, Int. J. Comput. Assist. Radiol. Surg. 9 (4) (2013) e34–e51.
[171] H.-S. Tong, Y.-L. Ng, Z. Liu, J.D. Ho, P.-L. Chan, J.Y. Chan, K.-W. Kwok, Real-to-virtual domain transfer-based depth estimation for real-time 3D annotation in transnasal surgery: a study of annotation accuracy and stability, Int. J. Comput. Assist. Radiol. Surg. 16 (5) (2021) 731–739.
[172] M. Oda, H. Itoh, K. Tanaka, H. Takabatake, M. Mori, H. Natori, K. Mori, Depth estimation from single-shot monocular endoscope image using image domain adaptation and edge-aware depth estimation, Comput. Methods Biomech. Biomed. Eng.: Imaging Vis. 10 (3) (2022) 266–273.
[173] S.F. Bhat, I. Alhashim, P. Wonka, Adabins: Depth estimation using adaptive bins, in: IEEE Conference on Computer Vision and Pattern Recognition, 2021, pp. 4009–4018.
[174] R. Ranftl, A. Bochkovskiy, V. Koltun, Vision transformers for dense prediction, in: IEEE International Conference on Computer Vision, 2021, pp. 12179–12188.
[175] J.-N. Brunet, A. Mendizabal, A. Petit, N. Golse, E. Vibert, S. Cotin, Physics-based deep neural network for augmented reality during liver surgery, in: Int. Conf. Med. Image Comput. and Computer-Assisted Intervention, Springer, 2019, pp. 137–145.
[176] T. Köhler, S. Haase, S. Bauer, J. Wasza, T. Kilgus, L. Maier-Hein, C. Stock, J. Hornegger, H. Feußner, Multi-sensor super-resolution for hybrid range imaging with application to 3-D endoscopy and open surgery, Med. Image Anal. 24 (1) (2015) 220–234.
[177] J. Guo, F. Gu, Y. Ye, Z. Song, An accurate speckle 3D reconstruction system based on binocular endoscope, in: IEEE International Conference on Real-Time Computing and Robotics, IEEE, 2021, pp. 703–708.
[178] A. Wachter, J. Kost, W. Nahm, Simulation-based estimation of the number of cameras required for 3D reconstruction in a narrow-baseline multi-camera setup, J. Imaging 7 (5) (2021) 87.
[179] X. Luo, H.-Q. Zeng, Y.-P. Du, X. Cheng, A novel endoscopic navigation system: simultaneous endoscope and radial ultrasound probe tracking without external trackers, in: Int. Conf. Med. Image Comput. and Computer-Assisted Intervention, Springer, 2019, pp. 47–55.
[180] A. Roberti, N. Piccinelli, F. Falezza, G. De Rossi, S. Bonora, F. Setti, P. Fiorini, R. Muradore, A time-of-flight stereoscopic endoscope for anatomical 3D reconstruction, in: International Symposium on Medical Robotics, IEEE, 2021, pp. 1–7.