Zhang *et al.*
Health Information Science and Systems (2024) 12:4
https://doi.org/10.1007/s13755-023-00262-7

Health Information Science and Systems

RESEARCH

Self-supervised neural network-based endoscopic monocular 3D reconstruction method

Ziming Zhang[1,2], Wenjun Tan[1,2*], Yuhang Sun[1,2], Juntao Han[1,2], Zhe Wang[3*], Hongsheng Xue[3] and Ruoyu Wang[3*]

Abstract

Based on deep learning, monocular visual 3D reconstruction methods have been applied in various conventional fields. In the aspect of medical endoscopic imaging, due to the difficulty in obtaining real information, self-supervised deep learning has always been a focus of research. However, current research on endoscopic 3D reconstruction is mainly conducted in laboratory environments, lacking experience in dealing with complex clinical surgical environments. In this work, we use an optical flow-based neural network to address the problem of inconsistent brightness between frames. Additionally, attention modules and inter-layer losses are introduced to tackle the complexity of endoscopic scenes in clinical surgeries. The attention mechanism allows the network to better focus on pixel texture details and depth differences, while the inter-layer losses supervise the network at different scales. We have established a complete monocular endoscopic 3D reconstruction framework and conducted quantitative experiments on a clinical dataset using the cross-correlation coefficient as a metric. Compared with other self-supervised methods, our framework can better simulate the mapping relationship between adjacent frames during endoscope motion. To validate the generalization performance of our framework, we tested the model trained on the clinical dataset on the SCARED dataset and achieved equally excellent results.

Keywords: Self-supervised learning, Monocular depth estimation, Ego-motion, Three-dimensional reconstruction, Endoscopy

Introduction

Lung diseases seriously affect human health. Take lung cancer as an example; it is the leading cause of cancer-related deaths worldwide[1]. Video-assisted Thoracic Surgery (VATS) is a reliable, precise, and safe minimally invasive treatment method for lung cancer. Doctors use a single-lens scope to observe the patient's condition and provide visual information during surgery[2, 3]. However, VATS also has its disadvantages, such as limited visibility and an inability to accurately position the scope. Augmented reality navigation systems based on computer vision can help doctors address these issues. Still, due to problems like changes in lighting and sparse features, accurately and densely reconstructing lung structures is not a straightforward task.

Three-dimensional reconstruction from monocular video has been a long-standing research topic[4–6]. Currently, deep learning methods are the primary research direction for this issue. Eigen et al.[7], Xu et al.[8], Cao et al.[9], and Fu[10] have used fully supervised convolutional neural networks for deep estimation. However, fully supervised three-dimensional reconstruction is challenging for endoscopy since obtaining true depth maps corresponding to endoscopy images is difficult.

Therefore, self-supervised monocular depth estimation and pose estimation have more research value (Luo

*Correspondence: tanwenjun@cse.neu.edu.cn; wangzhe@dlu.edu.cn; wangruoyu@dlu.edu.cn
[2] College of Computer Science and Engineering, Northeastern University, Shenyang 110189, China
[3] Oncology Department, Affiliated Zhongshan Hospital of Dalian University, Dalian 116001, China
Full list of author information is available at the end of the article

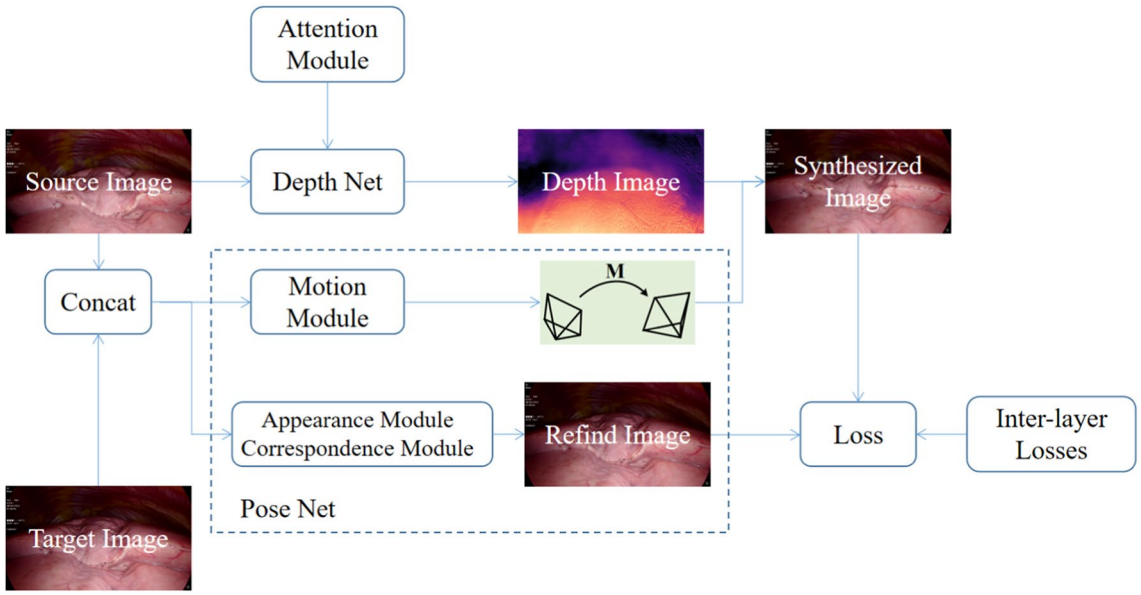

Fig. 1 The architecture of the proposed self-supervised 3D-dimensional reconstruction system

et al.[11], Ranjan et al.[12], Casser et al.[13]). Self-supervised methods simultaneously estimate scene depth and camera pose and use the obtained results to synthesize frames based on distortion. Finally, the difference between the target frame and the synthesized frame is calculated as the training supervision signal. However, network structures used in general scenes are not suitable for endoscopy scenes because the frame-to-frame photometric consistency assumption does not always hold in endoscopy videos.

Several methods have been developed to address the issue of inconsistent lighting. Liu et al.[14] used a multi-view synthesis method to generate sparse depth and camera pose first and then combine them to supervise the depth network. Spancer et al.[15] used learned dense visual representations to enhance the supervision signal when the photometric consistency condition fails. Yang et al.[16] and Ozyoruk et al.[17] used bio-inspired transformers to map the target frame to the same brightness space as the synthesized frame. Recasens et al.[18] combined traditional methods with deep learning methods, using photometric inconsistencies to track camera poses. However, these methods have drawbacks, such as high computational complexity, heavy reliance on visual representations, and the inability to handle extreme lighting changes, among other issues.

In this work, we have established a new monocular thoracoscopic three-dimensional reconstruction framework. Firstly, we address the issue of inconsistent lighting by using optical flow between adjacent frames. Optical flow introduces a generalized dynamic image constraint (GDIC), which includes both geometric and radiometric transformations. These two transformations help increase inter-frame information and compensate for differences in inter-frame brightness. Secondly, to tackle the issue of changes in the appearance of lung tissue during thoracoscopic movement, we have added an attention module to the depth estimation, allowing the network to focus more on regions with relatively rich texture information. Finally, we introduce inter-layer losses between different network layers to prevent gradient vanishing caused by convolutional layers. By supervising intermediate layers, we adequately train shallow convolutional layers and reduce underfitting in low-texture regions. We have used clinical data collected in collaboration with the hospital to validate the model's accuracy, demonstrating that the model can provide more accurate patient tissue location information to doctors during surgery (Fig. 1).

Related work

Fully supervised depth estimation

Deep convolutional networks were first proposed for depth estimation in [19]. Early depth estimation models were trained in a supervised manner using depth sensors. Eigen et al. [7] proposed using a multi-scale network and scale-invariant loss to regress depth from a single static image. Laina et al. [20] introduced a residual fully convolutional network (FCN) architecture for monocular depth estimation, which had a deeper architecture and eliminated post-processing steps. Cao et al. [9] treated depth estimation as a pixel-level classification problem and trained a residual network to predict the category

corresponding to each pixel after discretizing depth values. Xu et al. [23] used a Conditional Random Field (CRF) as a depth post-processing module. Fu et al. [10] treated depth estimation as a classification problem and introduced more robustness losses.

However, the endoscopy environment differs from the external environment, and it is challenging to train with abundant accurate RGB-D datasets under full supervision. Using computer-synthesized data has become one approach to address this. Visentini-Scarzanella et al. [21] used CT data and background-free simulated endoscopy videos to train fully supervised deep learning networks. Chen et al. [22] used color images and rendered depth maps to train a fully supervised depth network. Yang et al. [23] simulated the endoscopy imaging process using 3D modeling and rendering tools to achieve full supervision of endoscopy. However, closing the gap between the real domain and the synthetic domain is difficult by simply mimicking appearances, which may result in a performance drop.

Self-supervised depth and ego-motion estimation

Self-supervised networks indirectly train the network through differences in images, such as pixel, predicted depth, and appearance differences between image sequences, thus avoiding the use of depth maps. Initially, self-supervised networks were based on multi-view images. Xie et al. [24] introduced a model with discrete depth for synthetic views. Garg et al. [25] further investigated methods for predicting continuous disparity values, and Godard et al. [26] improved supervised results by adding left-right depth consistency. Various improvements based on multi-view methods include semi-supervised data [27, 28], generative adversarial networks [29, 30], additional consistency [31], temporal information [32–34], and real-time usage [35].

Various improvements have been made in the field of self-supervised estimation based on monocular images by researchers who enhanced network structures, loss functions, and more. In addition to predicting depth, self-supervised monocular training also requires the network to estimate endoscope poses between frames, which can be challenging in cases involving object motion. Zhou et al. [36] developed a self-supervised framework that views the depth estimation problem as a warping-based view synthesis task. However, self-supervised frameworks designed for general environments struggle to address issues like inter-frame brightness inconsistencies when applied to endoscopy environments.

Turan et al. [37] introduced research on self-supervised depth and ego-motion estimation in endoscopy scenes. Liu et al. [14] used sparse depth and camera poses generated by a traditional SfM pipeline as supervision, with SfM running as a preprocessing step. Li et al. [38] used Peak Signal-to-Noise Ratio (PSNR) as an additional optimization objective during training. Ozyoruk et al. [17] employed bio-inspired brightness transformers to enhance photometric robustness.

Compared to previous methods, we use optical flow to constrain image photometry, employ attention modules and inter-layer losses to handle non-Lambertian reflection and inter-reflection caused by changes in illumination inside the lung. Based on these improvements, we have established a comprehensive self-supervised framework. Our method is direct and does not require additional auxiliary information, such as CT images or depth maps generated by structured light, and it also does not necessitate multi-view images.

Methodology

In this section, we first introduce the prior knowledge of monocular 3D reconstruction. Then the proposed optical flow-based 3D reconstruction framework is elaborated. The framework consists of three parts: A. Depth estimation network with added attention mechanism, B. Motion estimation network based on optical flow, and C. Loss function. The overall framework is using a self-supervised approach to train the network, which can perform accurate 3D reconstruction of endoscopic scenes.

Self-supervised 3D reconstruction

Self-supervised 3D reconstruction involves two sub-networks: the depth estimation network and the pose estimation network. Unlike fully supervised methods that use real depth and pose as supervision signals, the supervision signal in self-supervised methods comes from view synthesis based on distortions. First, the depth estimation network estimates the pixel depth values of the current frame. Then, using the endoscope's intrinsic parameters, the pixel points on the 2D plane are projected back into the 3D camera space. The pose estimation network is then used to project the 3D point cloud onto adjacent frames. There are two frames, $I^t(p)$ and $I^s(p)$, and the frame transformation relationship is:

$$h(p^{s\to t}) = [K|0]M^{s\to t}\begin{bmatrix} D^t K^{-1} h(p^t) \\ 1 \end{bmatrix} \tag{1}$$

where $h(p^{s\to t})$ and $h(p^t)$ are the corresponding pixel coordinates on the source frame s and the target frame t, respectively, K represents the camera intrinsic parameters, $M^{s\to t}$ represents the motion from the source frame

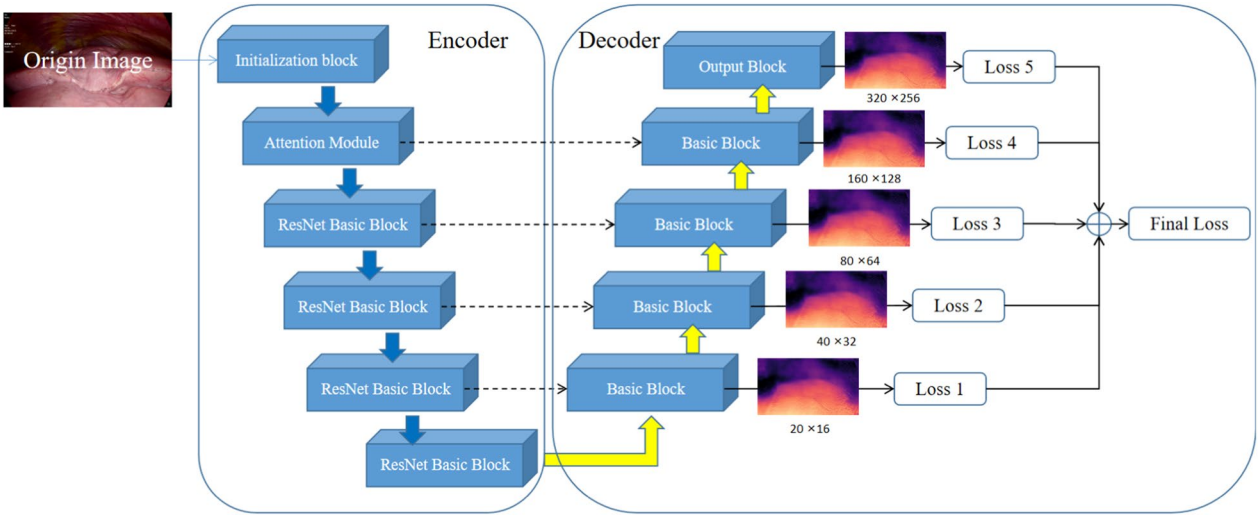

Fig. 2 Structure of the depth estimation network

to the target frame, and D^t represents the depth map of the target frame. With the above equation, the transformation relation equation between the source frame and the target frame can be obtained as follows:

$$F_{\delta}^{t\to s}(p) = p^{s\to t} - p^t \tag{2}$$

Depth estimation network

The depth estimation network(DepthNet) consists of an encoder and a decoder that takes the original frame I_s as the input and the corresponding disparity map D_s as the output. The network takes as input a 3-channel RGB image with a resolution of 320×256 and produces an output with the same resolution as the input. The overall architecture of the network is illustrated in Fig. 2.

Structure of network

The initialization block of the encoder consists of 3 parts: a 3×3 convolutional layer with 64 filters (C3×3), a batch normalization layer (BN) and a rectified linear unit activation function (ReLU) with a slope of 0.01. After initialization, it passes through the spatial attention module (SAM), the details of the attention module will be introduced in the next section. Then it passes through the max-pooling layer (MP), and finally passes through four ResNet basic blocks, each of which consists of C3×3, BN, ReLU, C3×3, BN, ReLU and skip connection in turn.

The decoder consists of four basic blocks, each consisting of C3×3, exponential linear unit (ELU), C3×3, and ELU in turn.

The final output layer consists of two layers interleaved by C3×3 and ELU, and finally Sigmoid is used as the activation function. In order to establish the information flow between the encoder and decoder, a skip connection is established from layer i to layer $n-i$, where n denotes the total number of layers, $i \in \{0, 1, 2, 3\}$.

Spatial attention module

The spatial attention module guides the depth estimation network by emphasizing pixel texture details with depth differences. The spatial attention module selects a specific region of the input image and processes the features within that region. The module operates as a non-local convolution process, and for any given input $X \in R^{N\times C\times H\times W}$, the module runs with the equation:

$$Z = f(X, X^T)g(X) \tag{3}$$

where f represents the pixel-wise relationship between inputs for each pixel X. The non-local operator extracts the relative weights of all positions on the feature map.

In this module a dot product operation is used for the θ and ϕ convolution of the max-pooling, which is activated by the ReLU function:

$$P = \psi(\sigma_{relu}(\theta(X)\phi(X)^T)) \tag{4}$$

where σ_{relu} is the ReLU activation function. The dot product $\theta(X)\phi(X)^T$ gives a measure of the input covariance, which can be defined as the degree of tendency between two feature maps from different channels. We activate the ψ convolution operation in the *softmax* function to perform a matrix multiplication between g and the output of the *softmax* function. Then, we apply convolution and upsampling to the multiplication result with ϕ to extract the attention map S. Finally, an element sum

operation is performed between the attention map S and the input X to generate the output $E \in R^{N\times C\times H\times W}$.

$$S = \phi(\sigma_{softmax} Pg(X)) \tag{5}$$

$$F = S + X \tag{6}$$

where $\sigma_{softmax}$ denotes the *softmax* function. A short connection between the input X and output F finalizes the residual learned block operations.

Pose estimation network

Our pose estimation network is primarily based on the design by Shao et al. [39], and it consists of three main components: a motion module, an appearance module, and a correspondence module. The motion module serves as a 6-degree-of-freedom (6DOF) self-motion estimator, taking two consecutive frames as input and outputting a relative pose parameterized by Euler angles and a translation vector. The appearance module is used to predict appearance flow and adjusts brightness conditions through a brightness calibration process. The correspondence module handles the automatic registration step.

The encoder model of the motion module network is similar to the one described in Sect. 3.2.1. It begins with an initialization block, but it's important to note that there is no attention module at this stage. Discussion about the effects of the attention module will be covered in chapter 4. The encoder then goes through a max-pooling layer, followed by four ResNet basic blocks. The decoder consists of three basic blocks and one C3×3. Each basic block is composed of a C3×3 and ReLU in sequence.

The appearance module network has a structure similar to the depth network. During the encoding phase, a concatenated image pair passes through convolution block layers with a stride of 2, forming a five-level feature pyramid. Jump connections then propagate the pyramid's features to the decoding phase. In the decoding phase, upsampling layers, concatenated feature maps, 3×3 convolution layers with ELU activation, and the estimation layer are sequentially connected until the network's output reaches the highest resolution. Apart from the estimation layer, the correspondence module network maintains the same architecture as the appearance module network.

Loss function

The loss function consists of three parts, the residual-based smoothness loss, auxiliary loss, and smoothness loss. In order to fully utilize information across different levels, we also introduce inter-layer losses when calculating the loss. The single-layer loss function is as follows:

$$L = \lambda_1 l_{rs} + \lambda_2 l_{ax} + \lambda_3 l_{es} \tag{7}$$

Smoothness loss based on residuals

It penalizes the first-order gradients. It uses the output of the appearance module network in conjunction with the original image to calculate the result, as follows:

$$l_{rs} = \sum_p |\nabla A_\delta(p)| \tag{8}$$

where $A_\delta(p)$ represents the constraint of the appearance module on the light intensity.Additionally, the residual-based gradient is used to emphasize regions with sharp brightness changes:

$$l_{rs} = \sum_p |\nabla A_\delta(p)| \times e^{-\nabla |I^t(p) - I^{s\to t}(p)|} \tag{9}$$

Auxiliary loss

l_{ax} provides the supervisory signal for the appearance module:

$$l_{ax} = \sum_p M(p) \times \Phi(I^{s\to t}(p), I^t(p) + A_\delta(p)) \tag{10}$$

where $I^{s\to t}(p)$ is reconstructed from optical flow and spatial converters.$M(p)$ represents the mask for objects falling within the visible range.

Edge-aware smoothness loss

The smoothness property of the depth map is enforced using l_{es} with the following equation:

$$l_{es} = \sum_p |\nabla D(p)| * e^{-\nabla |I^t(p)|} \tag{11}$$

Inter-layer loss

Due to the encoding and decoding processes in the network, different levels focus on different image ranges. If only the results from the last layer of the decoder are used to compute the loss function, some local information may be lost. Therefore, we introduced inter-layer loss by adding additional branches in the decoder to compute the loss for each layer. The structure of the inter-layer loss is reflected in Fig. 2. After adding inter-layer loss, the formula for the total loss function is as follows, where k_i represents the weight parameters of the i layer and n represents the number of decoder blocks.:

$$L = \sum_{i=1}^{n} k_i(\lambda_1 l_{rs} + \lambda_2 l_{ax} + \lambda_3 l_{es}) \tag{12}$$

Experiments

To evaluate the depth estimation accuracy of the proposed framework and to investigate different design considerations, we conduct extensive experiments in this section.

Dataset

- SCARED[40].The SCARED dataset was collected from fresh pig cadaver abdominal dissections and contains 35 endoscopic videos as well as realistic depth and pose information.
- Clinical Data Set. In collaboration with the hospital, we used Olympus' endoscope, which comes with a video recording function, when performing lung surgery, and asked the surgeon to film the patient's thoracoscopy from as many angles as possible before the surgery. Included is endoscopic video of 11 complete surgeries.

For depth estimation and pose estimation, we conducted extensive experiments on clinical datasets. Faced with the challenge of not having access to ground truth depth and endoscope motion paths in clinical datasets, we utilized normalized local cross-correlation as a quantitative evaluation metric, which will be detailed in Section 4.2. Then, to demonstrate the generalization capability of this model, we will use the model trained on clinical datasets, without any adjustments, directly for experiments on the SCARED dataset.

Training parameter

Training: Our framework is implemented in the Pytorch library and trained on a single NVIDIA RTX 3060. We use the Adam optimizer, where $\beta_1 = 0.9, \beta_2 = 0.99$, the batch size is 4,$\alpha = 0.85, \lambda_1 = 0.1, \lambda_2 = 0.01, \lambda_3 = 0.001$. For inter-layer losses, the number of decoders $n = 5$, and the loss factor $k_i = 0.2$ for each layer. We employed a pre-trained ResNet-18 encoder on ImageNet. The input image resolution for the entire network is 320×256. In each epoch, we divided the training into two stages. First, we trained the correspondence module network using edge-aware smoothness loss. After backpropagation and parameter updates, we proceeded to train the depth network, motion module network, and appearance module network. A total of 20 epochs were trained. In these two stages, the initial learning rate was set to 1e-4 and was scaled by a factor of 0.1 after 10 epochs.

Performance metrics: In response to the challenge of not having access to real depth and motion trajectories in clinical data, we use the cross-correlation coefficient [41, 42] to quantitatively evaluate the model's performance.

Table 1 The error and accuracy metrics for depth evaluation

Metric	Definition
Abs Rel	$\frac{1}{D}\sum_{d\in D} \lvert d^* - d\rvert / d^*$
Sq Rel	$\frac{1}{D}\sum_{d\in D} \lvert d^* - d\rvert^2 / d^*$
RMSE	$\sqrt{\frac{1}{D}\sum_{d\in D} \lvert d^* - d\rvert^2}$
RMSE log	$\sqrt{\frac{1}{D}\sum_{d\in D} \lvert log d^* - log d\rvert^2}$
δ	$\frac{1}{D}\left\{d \in D \lvert max(\frac{d^*}{d}, \frac{d}{d^*} < 1.25)\right\} \times 100\%$

This metric has been employed in medical image registration research to measure the similarity between images before and after registration. We adapt this metric for monocular endoscope 3D reconstruction. After inputting the original image s into the model, we obtain estimated depth and pose. Then, using depth and pose, we warp s to the target image t, resulting in a synthesized frame $\hat{t}$. We then calculate the cross-correlation coefficient between t and $\hat{t}$, and after normalization, a coefficient closer to 1 indicates greater similarity between the synthesized frame and the target frame, reflecting better model performance. The formula for the cross-correlation coefficient is as follows:

$$CC(s,t) = \sum_{p\in\Omega} \frac{\left(\sum_{p_i}(s(p_i) - \hat{s}(p))(t(p_i) - \hat{t}(p))\right)^2}{\left(\sum_{p_i}(s(p_i) - \hat{s}(p))\right)\left(\sum_{p_i}(t(p_i) - \hat{t}(p))\right)} \tag{13}$$

where p denotes the pixel point on the image, and $\hat{s}$ and $\hat{t}$ are the synthetic frames obtained from the estimated depth and bit pose warping of the original image s and the target image t, respectively.

Additionally, to validate the model's generalization, experiments were conducted on the SCARED dataset using other evaluation metrics, as specified in Table 1. In this table, d and d^* represent the predicted depth values and the corresponding ground truth values, and D represents a set of predicted depth values. During validation, we use the median scaling method to scale the predicted depth values, and the formula for this scaling is as follows:

$$D_{scaled} = D_{pred} * (median(D_g t)/median(D_{pred})) \tag{14}$$

On the SCARED dataset, the depth maps are scaled proportionally with an upper limit of 150 mms. We have chosen 150 mms as the scaling limit.

For pose estimation, we evaluate using the Absolute Trajectory Error (ATE) [43], as well as the mean and standard deviation of angle errors.

Table 2 Quantitative comparisons of correlation coefficient

Methods	1	2	3	4	5	6
EndoSLAM	0.5991	0.6732	0.8173	0.7470	0.7134	0.6214
Endo D &M	0.6006	0.6712	0.8221	0.7485	0.7001	0.6635
AF-SFM	0.5991	0.6694	0.8165	0.7746	0.7638	0.6618
Ours	0.6007	0.6739	0.8415	0.8205	0.7435	0.7204

Quantitative evaluation of the cross-correlation coefficient

We evaluated the depth estimation accuracy of our framework against several typical self-supervised methods used for endoscopy, including EndoSLAM[17], Endo-Depth-and-Motion[18], and AF-SFM[39]. We sliced the clinical data collected from hospital surgeries and organized 6370 endoscopic RGB images from 8 surgeries. We selected six of these surgeries containing 4410 images for training. Images from the remaining two surgeries were used to evaluate the training effect. Additionally, to demonstrate the generality of our model, we selected data from 4 scenes in the SCARED dataset and conducted the same tests. The experimental results on both datasets are shown in the following Table 2:

In the evaluation based on the cross-correlation coefficient, values closer to 1 indicate a higher degree of similarity between the synthesized frames computed from estimated depth and pose and the target frames. This suggests better accuracy of the method. In the table above, entries 1 and 2 represent results from clinical data, while entries 3 to 6 represent results from the SCARED dataset. Comparing the results among different methods, except for the experiment at entry 5 where our method slightly underperformed compared to AF-SFM, our framework showed a significant advantage in the other test groups. This demonstrates that our framework can more accurately simulate camera motion within the thoracic cavity, enabling more precise 3D reconstruction.

Depth quantitative evaluation of conventional indicators

In addition to comparing using the cross-correlation coefficient, we also conducted a quantitative evaluation on the SCARED dataset using conventional metrics. We directly validated the model trained on clinical data on the SCARED dataset without any fine-tuning. The experimental results are as shown in the following Table 3:

From the table above, it is evident that our method achieved better results in various parameters, demonstrating its strong performance across different patients and endoscopes.

Figure 3 provides a qualitative comparison. It can be observed that the three methods do not differ significantly on the SCARED dataset, with our method and AF-SFM showing slightly better results. However, on the clinical dataset, the Endo D &M method is no longer capable of accurate depth estimation, while our method can effectively capture the depth values of two protruding regions.

Table 3 Quantitative comparison of depth on the SCARED dataset

Methods	Abs Rel	Sq Rel	RMSE	RMSE log	δ
EndoSLAM	0.062	0.606	5.726	0.093	0.957
Endo D &M	0.070	0.761	5.221	0.084	0.970
AF-SFM	0.074	0.807	6.442	0.097	0.922
Ours	0.051	0.337	4.797	0.072	0.986

Pose evaluation on the SCARED dataset

It can be seen that, except for the standard deviation, our framework slightly underperforms EndoSLAM. In other metrics, our model outperforms the others. This may be because the EndoSLAM model uses an attention module in its pose estimation network, which allows the pose estimation network to handle regions with missing textures more effectively. We also attempted to incorporate this into our framework, but experimental results showed that the performance gain was not significant. This could be due to the complexity of adding an attention module in the pose network in the challenging clinical environment, which might decrease accuracy rather than improve it (Table 4).

Figure 4 shows qualitative experiments for pose estimation. Figure (a) displays the predicted trajectory for the AF-SFM method, where the blue line represents the ground truth trajectory and the green line represents the estimated trajectory. From the image, it is evident that our predictions are closer to the actual results.

Ablation experiments

We divided our framework into four models: the baseline model (ID1) without an attention mechanism and without inter-layer loss, directly computing the loss only on the last layer; the attention model (ID2) with only the attention mechanism and no inter-layer loss; the

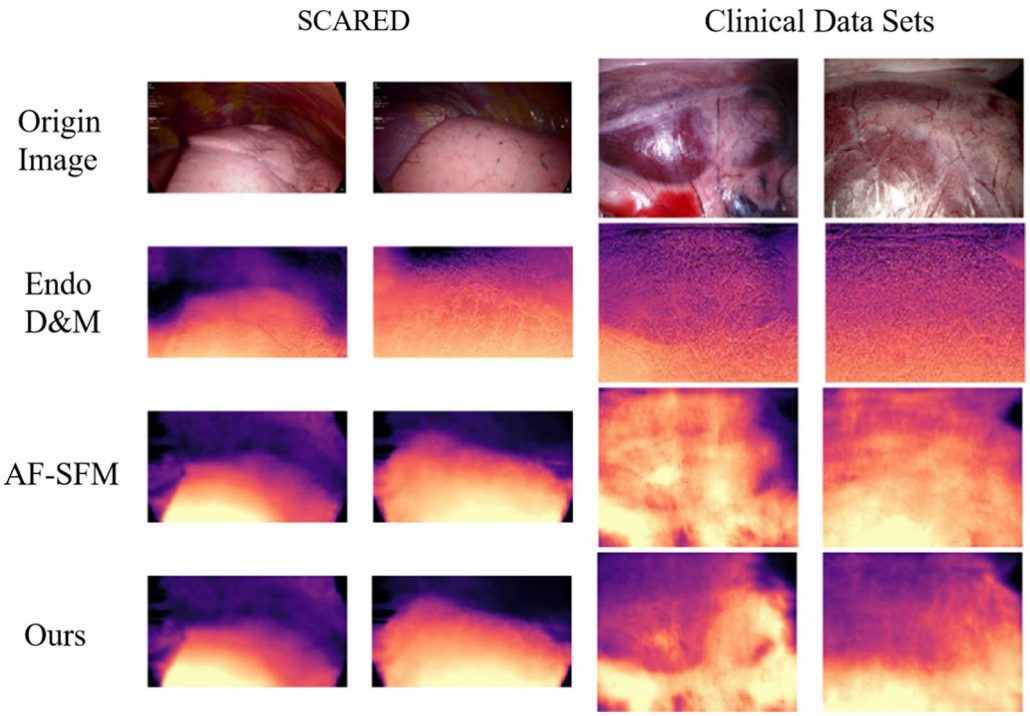

Fig. 3 Qualitative comparison results of different methods

Table 4 Quantitative comparison of motion on the SCARED dataset

Methods	Trajectory		Rotation	
	Error	Std	Error	Std
EndoSLAM	0.0376	0.0117	0.0027	0.0014
Endo D &M	0.0743	0.0729	0.0030	0.0024
AF-SFM	0.1075	0.0936	0.0036	0.0024
Ours	0.0272	0.0220	0.0017	0.0015

inter-layer model (ID3) without an attention mechanism, using only inter-layer loss; and our full model (ID4). We compared these models and conducted experiments on both the SCARED dataset and clinical dataset. The Table 5 shows the quantitative depth evaluation results on the SCARED dataset:

The table clearly demonstrates that the proposed improvements indeed enhance the accuracy of depth prediction. After incorporating the attention module and inter-layer loss, all the experimental metrics improve, with δ values approaching 1. This indicates that more predicted values fall within the range of 75% to 125% of the true values. By comparing the differences between ID2, ID3, and ID1, it is apparent that inter-layer loss has a more significant impact on the entire framework. This may be because inter-layer loss leverages features at different resolutions, providing the framework with a better understanding of both image details and the overall structure.

Table 6 represents six experiments conducted on the clinical dataset, using the cross-correlation coefficient as the evaluation standard.It's evident that on the clinical dataset, there is a similar trend in the data among different models (ID1-4), leading to the same conclusions as observed on the SCARED dataset.

We performed a quantitative evaluation of the pose under the same conditions, and Table 7 shows the results of the quantitative evaluation of the pose:

In terms of bit pose, there is not much difference between ID1 and ID2, indicating that the attention module in the deep network cannot have a positive effect on bit pose prediction. In contrast, the results of ID3 and ID4 are better than those of ID1 and ID2, with the best result for ID4, which also shows the superiority of our complete model on the bit-pose prediction.

Point clouds for SCARED and clinical datasets

After obtaining depth estimation results and endoscope ego-motion results, along with the camera intrinsic parameters, you can derive a 3D point cloud representation of the endoscopic scene. We use the Truncated Signed Distance Function (TSDF) proposed by Recasens

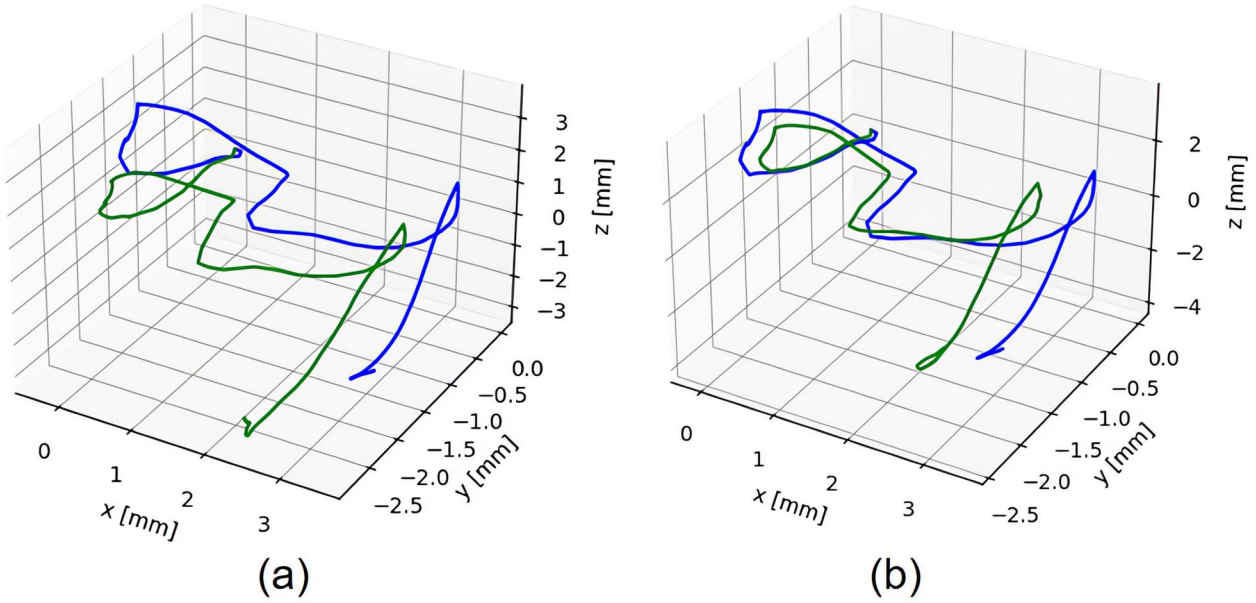

Fig. 4 Qualitative comparison of motion on the SCARED dataset

Table 5 Experimental study on the ablation of attention modules and inter-layer losses

ID	Attention Module	Inter-layer Loss	Abs Rel	Sq Rel	RMSE	RMSE log	δ
1			0.109	0.683	9.428	0.146	0.841
2	✓		0.109	1.373	8.772	0.138	0.867
3		✓	0.087	0.956	8.507	0.127	0.928
4	✓	✓	0.051	0.337	4.797	0.072	0.986

Table 6 Validation of ablation experiments on a clinical data set

ID	1	2	3	4	5	6
1	0.4808	0.5799	0.6569	0.6786	0.5934	0.5478
2	0.5189	0.6115	0.7035	0.6826	0.6100	0.5505
3	0.5749	0.6344	0.7557	0.7419	0.6813	0.5811
4	0.6007	0.6694	0.8165	0.7746	0.7638	0.6618

Table 7 Validation of pose estimation by ablation experiments on the SCARED dataset

ID	Trajectory		Rotation	
	Error	Std	Error	Std
1	0.0375	0.0270	0.0137	0.0088
2	0.0329	0.0319	0.0122	0.0088
3	0.0316	0.0230	0.0095	0.0053
4	0.0272	0.0220	0.0017	0.0015

et al. [18] to represent and fuse the depth predictions into a high-quality surface reconstruction. Quantitative evaluation of point cloud reconstruction results may be challenging, but qualitatively, you can observe the reconstruction results in Fig. 5. Figures (a) and (c) are results from the SCARED dataset, while (b) and (d) are from the clinical dataset. It is apparent that our framework can faithfully and accurately reconstruct the 3D structure of the endoscopic scenes in both environments.

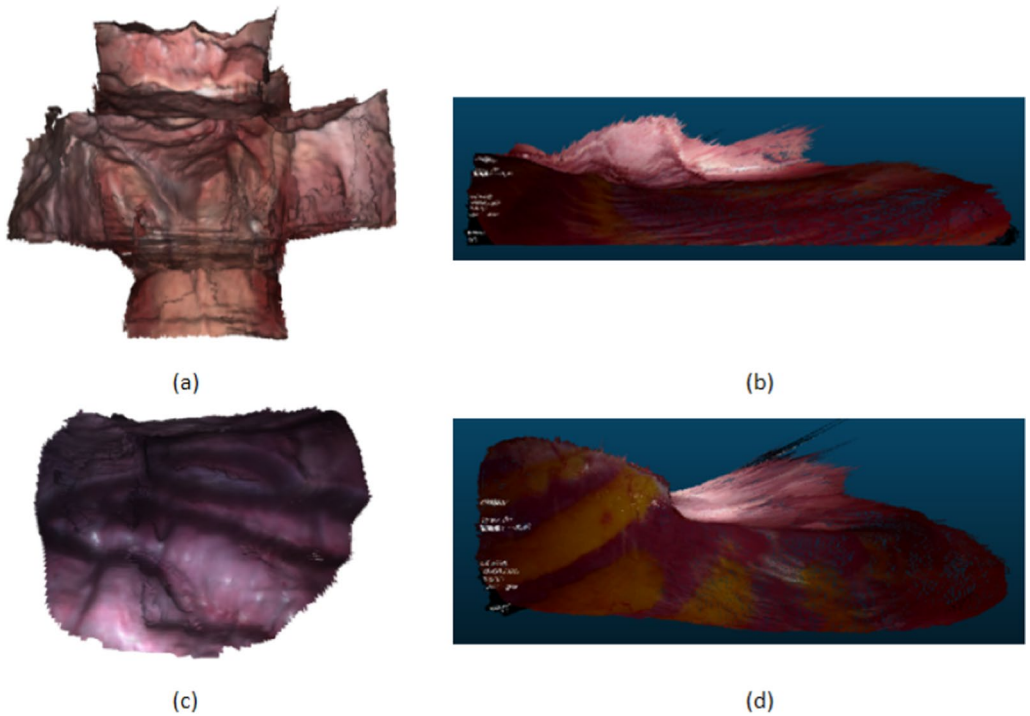

Fig. 5 Results of dense reconstruction of SCARED dataset and clinical dataset

Conclusion

In this research, a novel self-supervised framework was designed. Leveraging an optical flow network as a base, we incorporated an attention module and introduced inter-layer loss into the network to address the challenges presented by endoscopic clinical datasets, such as severe inter-frame brightness fluctuations and significant scene variations. When faced with the difficulty of obtaining real depth and camera motion in clinical datasets, we used the cross-correlation coefficient as a quantitative evaluation metric. After assessing the performance using the cross-correlation coefficient, our framework exhibited superior mapping relationships between frames, which was attributed to the accuracy of depth estimation and endoscope ego-motion estimation. Finally, we conducted generalization experiments on the SCARED dataset, which also demonstrated the accuracy and generalization capabilities of our network.

Limitations and future work

In current research, researchers have primarily focused on static environments. For instance, datasets like SCARED and SERV-CT are collected on deceased pigs. However, in actual surgeries, the intraoperative environment is subject to real-time changes. Especially in the lung region, the lungs expand and contract periodically with the patient's breathing. Our study, compared to previous research, has increased scene complexity but hasn't addressed the dynamic aspects. Establishing dynamic lung models is currently a relatively underdeveloped aspect of endoscopic 3D reconstruction research. While some progress has been made in typical environments, incorporating these achievements into the field of endoscopic 3D reconstruction is a direction for our future research.

Acknowledgements
This work is supported by the National Natural Science Foundation of China(61971118), Fundamental Research Funds for the Central Universities(N2216014),Science and Technology Plan of Liaoning Province(2021JH1/10400051)

Declarations

Conflict of interest
All authors declare that there is no conflict of interest.

Author details
[1]Key Laboratory of Intelligent Computing in Medical Image, Ministry of Education, Northeastern University, Shenyang 110189, China. [2]College of Computer Science and Engineering, Northeastern University, Shenyang 110189, China. [3]Oncology Department, Affiliated Zhongshan Hospital of Dalian University, Dalian 116001, China.

Received: 1 June 2023 Accepted: 8 November 2023
Published online: 11 December 2023

References

1. Global Cancer Statistics Report 2020. Chin J Prev Med 2021;55(03):398–398
2. Schonberger JL, Frahm JM, Ieee. Structure-from-Motion Revisited. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016; pp. 4104–4113.
3. Mur-Artal R, Tardos JD. ORB-SLAM2: an open-source SLAM system for monocular, stereo, and RGB-D cameras. Ieee Trans Robot. 2017;33(5):1255–62.
4. Engel J, Schoeps T, Cremers D. LSD-SLAM: large-scale direct monocular SLAM. In: 13th European Conference on Computer Vision (ECCV), 2014;834–849
5. Lee SH, Civera J. Loosely-coupled semi-direct monocular SLAM. Ieee Robot Autom Lett. 2019;4(2):399–406.
6. Czarnowski J, Laidlow T, Clark R, et al. DeepFactors: real-time probabilistic dense monocular SLAM. Ieee Robot Autom Lett. 2020;5(2):721–8.
7. Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network. In: 28th Conference on Neural Information Processing Systems (NIPS), 2014.
8. Xu D, Ricci E, Ouyang W, et al. Multi-Scale Continuous CRFs as Sequential Deep Networks for Monocular Depth Estimation. In: 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017;161–169.
9. Cao Y, Wu Z, Shen C. Estimating depth from monocular images as classification using deep fully convolutional residual networks. Ieee Trans Circ Syst Video Technol. 2018;28(11):3174–82.
10. Fu H, Gong M, Wang C, et al. Deep ordinal regression network for monocular depth estimation. In: 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018;2002–2011.
11. Luo C, Yang Z, Wang P, et al. Every pixel counts plus plus : joint learning of geometry and motion with 3D holistic understanding. Ieee Trans Pattern Anal Mach Intell. 2020;42(10):2624–41.
12. Ranjan A, Jampani V, Balles L, et al. Competitive collaboration: joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In: 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019;12232–12241.
13. Casser V, Pirk S, Mahjourian R, et al. Depth Prediction without the Sensors: Leveraging Structure for Unsupervised Learning from Monocular Videos. In: 33rd AAAI Conference on Artificial Intelligence / 31st Innovative Applications of Artificial Intelligence Conference/9th AAAI Symposium on Educational Advances in Artificial Intelligence, 2019;8001–8008.
14. Liu X, Sinha A, Ishii M, et al. Dense depth estimation in monocular endoscopy with self-supervised learning methods. Ieee Trans Med Imaging. 2020;39(5):1438–47.
15. Spencer J, Bowden R, Hadfield S. Defeat-net: General monocular depth via simultaneous unsupervised representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020;14402–14413.
16. Yang N, Von Stumberg L, Wang R, et al. D3VO: Deep Depth, Deep Pose and Deep Uncertainty for Monocular Visual Odometry. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020; 1278–1289.
17. Ozyoruk KB, Gokceler GI, Bobrow TL, et al. EndoSLAM dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos. Med Image Anal. 2021;71:102058.
18. Recasens D, Lamarca J, Facil JM, et al. Endo-depth-and-motion: reconstruction and tracking in endoscopic videos using depth networks and photometric constraints. Ieee Robot Autom Lett. 2021;6(4):7225–32.
19. Eigen D, Fergus R, Ieee. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: IEEE International Conference on Computer Vision, 2015;2650–2658.
20. Laina I, Rupprecht C, Belagiannis V, et al. Deeper depth prediction with fully convolutional residual networks. In: 4th IEEE International Conference on 3D Vision (3DV), 2016;239–248.
21. Visentini-Scarzanella M, Sugiura T, Kaneko T, et al. Deep monocular 3D reconstruction for assisted navigation in bronchoscopy. Int J Comput Assist Radiol Surg. 2017;12(7):1089–99.
22. Chen RJ, Bobrow TL, Athey T, et al. Slam endoscopy enhanced by adversarial depth prediction. arXiv preprint arXiv:1907.00283, 2019.
23. Yang YM, Shao SW, Yang T, et al. A geometry-aware deep network for depth estimation in monocular endoscopy. Eng Appl Artif Intell. 2023;122:105989.
24. Xie J, Girshick R, Farhadi A. Deep3D: fully automatic 2D-to-3D video conversion with deep convolutional neural networks. In: 14th European Conference on Computer Vision (ECCV), 2016;842–857.
25. Garg R, Vijaykumar B G, Carneiro G, et al. Unsupervised CNN for single view depth estimation: geometry to the rescue. In: 14th European Conference on Computer Vision (ECCV), 2016;740–756.
26. Godard C, Mac Aodha O, Brostow GJ, et al. Unsupervised monocular depth estimation with left-right consistency. In: 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017;6602–6611.
27. Kuznietsov Y, Stuckle J, Leibe B, et al. Semi-supervised deep learning for monocular depth map prediction. In: 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017;2215–2223.
28. Luo Y, Ren J, Lin M, et al. Single view stereo matching. In: 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018;155–163.
29. Aleotti F, Tosi F, Poggi M, et al. Generative adversarial networks for unsupervised monocular depth prediction. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops. 2018;0-0.
30. Pilzer A, Xu D, Puscas MM, et al. Unsupervised adversarial depth estimation using cycled generative networks. In: 6th International Conference on 3D Vision (3DV), 2018;587–595.
31. Poggi M, Tosi F, Mattoccia S, et al. Learning monocular depth estimation with unsupervised trinocular assumptions. In: 6th International Conference on 3D Vision (3DV), 2018;324–333.
32. Li R, Wang S, Long Z, et al. UnDeepVO: monocular visual odometry through unsupervised deep learning. IEEE International Conference on Robotics and Automation (ICRA), 2018;7286–7291.
33. Zhan H, Garg R, Weerasekera C S, et al. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In: 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018;340–349.
34. Babu M V, Das K, Majumdar A, et al. UnDEMoN: Unsupervised deep network for depth and ego-motion estimation. In: 25th IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2018;1082–1088.
35. Poggi M, Aleotti F, Tosi F, et al. Towards real-time unsupervised monocular depth estimation on CPU. In: 25th IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2018;5848–5854.
36. Zhou T, Brown M, Snavely N, et al. Unsupervised learning of depth and ego-motion from video. In: 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017;1851–1858.
37. Turan M, Ornek E P, Ibrahimli N, et al. Unsupervised odometry and depth learning for endoscopic capsule robots. In: 25th IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2018: 1801–1807.
38. Li L, Li X, Yang S, et al. Unsupervised-learning-based continuous depth and motion estimation with monocular endoscopy for virtual reality minimally invasive surgery. Ieee Trans Ind Inform. 2021;17(6):3920–8.
39. Shao SW, Pei ZC, Chen WH, et al. Self-Supervised monocular depth and ego-motion estimation in endoscopy: appearance flow to the rescue. Med Image Anal. 2022;77:102338.
40. Allan M, Mcleod J, Wang C, et al. Stereo correspondence and reconstruction of endoscopic data challenge. arXiv preprint arXiv:2101.01133, 2021.
41. Fang Q, Gu X, Yan J, et al. A FCN-based Unsupervised Learning Model for Deformable Chest CT Image Registration. In: IEEE Nuclear Science Symposium / Medical Imaging Conference (NSS/MIC), 2019.
42. Fang Q, Yan J, Gu X, et al. Unsupervised learning-based deformable registration of temporal chest radiographs to detect interval change. In: Medical Imaging Conference—Image Processing, 2020.
43. Mur-Artal R, Montiel JMM, Tardos JD. ORB-SLAM: a versatile and accurate monocular SLAM system. Ieee Trans Robot. 2015;31(5):1147–63.