IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 15, 2021, accepted December 30, 2021, date of publication January 4, 2022, date of current version January 13, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3140334

3D Reconstruction of Laparoscope Images With Contrastive Learning Methods

NANNAN CHONG
Department of Information and Intelligence Engineering, Tianjin Renai College, Tianjin 301636, China
School of Electronic and Information Engineering, Hebei University of Technology, Tianjin 300401, China
e-mail: chongnannan@163.com

This work was supported by the Tianjin Municipal Education Commission Scientific Research Planned Project (Natural Science) of China under Grant No. 2018KJ268.

ABSTRACT The use of laparoscopic images and videos to reconstruct abdominal tissue structure overcomes the visual limitations of human eyes and provides great convenience for the detection and diagnosis of medical diseases. The method described in this paper is based on contrast learning and ORB-SLAM design. The contributions of this study are as follows. (i) A data preprocessing thread is introduced, which includes data augmentation and input frame evaluation. (ii) Encoding global information to avoid semantic information loss and eliminate mismatch by designing an improved U-NET network. (iii) The improved U-NET network was used to introduce a dual-branched Siamese network and GPU-accelerated intensive reconstruction thread. The dual-branched Siamese network structure was used to compute the depth information of the feature points and optical flow information in parallel, and the dense depth estimation was obtained without interrupting the original sparse reconstruction. (iv) An experimental system was established to conduct three-dimensional reconstruction tests on laparoscopic/endoscope /UBE videos of 186 patients. To effectively provide more accurate feature detection and matching support for the application of combat scenes, the influence of lens distortion was considered. Compared with the current mainstream three-dimension reconstruction and deep learning algorithms, the practicability and superiority of laparoscopic three-dimension reconstruction based on contrastive learning are demonstrated in clinical scenarios such as surgical navigation, auxiliary diagnosis, and surgical simulation.

INDEX TERMS Contrastive learning, laparoscope, endoscope, medical images, three-dimensional reconstruction.

I. INTRODUCTION

The emergence and development of laparoscopy breaks through the visual limitations of the human eyes and provides a lot of convenience for the detection and diagnosis of medical diseases. In Minimally Invasive Surgery (MIS) systems, estimating depth information from laparoscopic images, restoring dense 3D information and accurately calculating laparoscopic position is the basis for realizing computer-aided guidance, which is also a research hotspot in the medical field.

Because the traditional methods are mostly based on original pixels or traditional eigenvectors, the sparse texture areas are easily to be adversely affected, resulting in incorrect calculation and matching, which cannot solve the problems well such as lighting and ghost exposure. Therefore, we introduce a new method based on contrastive self-supervised learning to extract dense features, calculate depth information with stronger robustness to sparse textures, lighting, etc., and perform dense three-dimensional (3D) reconstruction.

As digestive endoscopy does not provide stereoscopic images, it is difficult to transfer these techniques from surgery to digestive endoscopy. In addition, because the gastrointestinal tract is located deep inside the body, which is thin and constantly moving, it is difficult to obtain the precise space needed for a 3D image. In this work, the researchers used VR in flexible digestive endoscopy by implementing different image sources in the virtual laparoscopy operation room. Digital magnification of the digestive endoscopy image relative to the examiner's position is the only one possibility

The associate editor coordinating the review of this manuscript and approving it for publication was Kumaradevan Punithakumar.

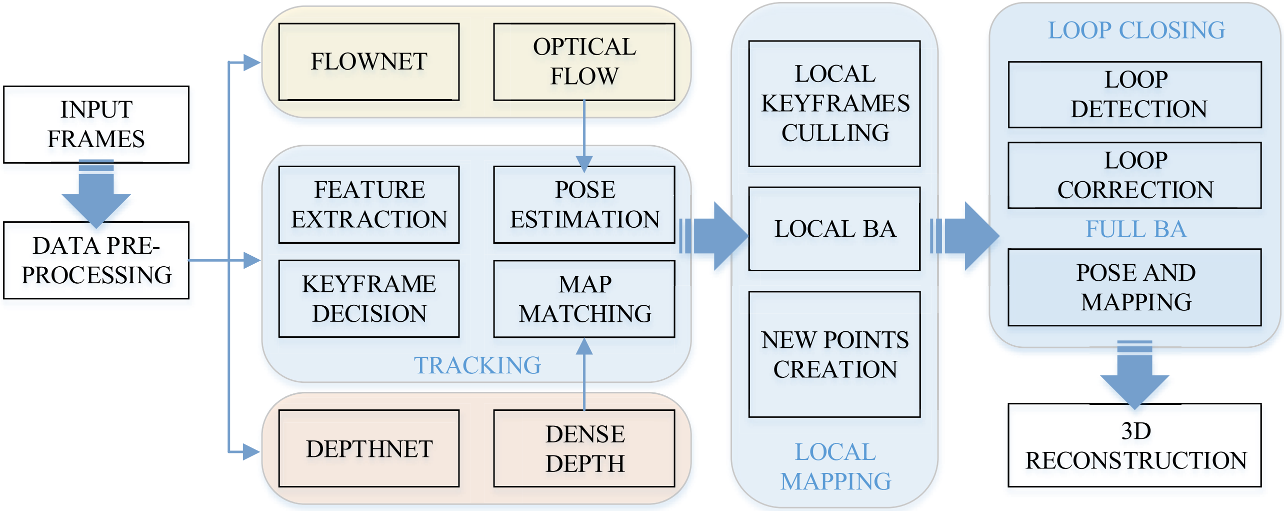


FIGURE 1. Network architecture.

for integrating these new tracking devices, which can also be integrated with electronic endoscopy. In addition, conventional endoscopic images alone can transform light-intensity information into 3D images. Other image sources that can be used in the virtual endoscopic operation room include previous CT scans and vital signs of patients during sedation [1], etc.

The main contributions of this paper are as follows.

- A dual-branch Siamese network on the basis of the traditional framework of ORB-SLAM, the addition of dense depth estimation network DepthNet and the use of optical flow to estimate camera motion FlowNet, to overcome the inconsistency of color luminance between different frames at the same position in medical images, occlusion and reflection areas in images.
- We propose an improved U-NET to implement DepthNet and FlowNet functionality.
- Qualitative and quantitative analyses of medical images from 186 patients were performed using modern deep learning methods as well as modern 3d reconstruction of medical images. Setup comparison study, cross-patient studies, respectively, to prove the superiority of our method.

II. RELATED WORK

When Several methods have been explored for depth estimation and 3D reconstruction of laparoscopy. These can be divided into traditional multiview stereo algorithms and methods based on full supervised learning.

The work of Gary [2] in 2016 has set a precedent in this respect. The network was divided into three parts. The first part is an encoder based on a traditional convolutional neural network (the author uses a full convolutional network with a skip connection structure). The input of the full convolutional network is the left image, and the corresponding prediction depth map is obtained through the network. In the second part, the depth map is combined with the image on the right to obtain an inverted image based on the corresponding distance at which the image moves along the scan line. In the third part, the optical flow error of the original left image and the deformation image is taken as the objective function to match the reconstructed image.

Liu *et al.* [3] proposed a monocular endoscope video dense 3D reconstruction based on self-supervised learning and a traditional multiview stereo algorithm, which does not require prior modeling of anatomy or shadows, and does not require human interaction and patient CT during training and application. In the cross-case study using CT scan as ground truth, submillimeter mean residual error was obtained.The system has four modules in training and application, namely depth estimation, depth map fusion, fault detection and pose map optimization. In the application phase, SfM is first run on video to estimate the camera attitude and sparse point cloud, and ends when 3D reconstruction is successfully generated. The dotted arrows indicate the execution order of the threads. Then Liu proposed another method based on deep learning which is proposed to reconstruct dense and accurate 3D surface model structures of sinuses from direct motion recovery from endoscope video, and relevant parameters were measured in clinical experiments. The results of the 3D reconstruction are in good agreement with those of CT.

Some scholars have studied the methods of 3D reconstruction based on SLAM. Engel *et al.* [4] proposed a monocular SLAM algorithm based on the direct method, which can build a large-scale and globally consistent environment map compared to the current direct method. The method can not only obtain highly accurate attitude estimation based on direct image registration, but can also reconstruct 3D environment maps into attitude maps of key frames and corresponding semi-dense depth maps in real time. Mur-Artal *et al.* [5] proposed that ORB-SLAM, which uses unified ORB features, can help the SLAM algorithm achieve endogenous consistency in feature extraction and tracking, key-frame selection, 3D reconstruction, closed-loop detection and other steps. Mahmoud *et al.* [6] proposed a novel dense SLAM method for reconstructing a dense stomach surface for handheld monocular endoscopy. He combined Zero Mean Normalized Cross Correlation (ZNCC) and a gradient Huber norm regularizer to match dense maps between cluster frames and compute them in parallel. It can outperform pure stereo reconstructions because the frame cluster can provide larger parallax from the motion of the endoscope.

III. PROPOSED APPROACH

This paper proposes a kind of three-dimensional (3D) reconstruction method for abdominal tissue, which is based on the contrast from the supervised learning by laparoscopically obtaining two-dimensional image reconstruction. The proposed model can reconstruct the 3D geometric structure of the viscera from monocular image sequences of laparoscopy, as is shown in Fig.1, from which we can optimize the operation to improve the success rate of laparoscopic surgery. The proposed network can be used in clinical surgical navigation systems and in the preoperative application of virtual simulation training. We describe a two-branch Siamese network method based on U-Net [7] and traditional ORB-SLAM threads that train the input frames using monocular laparoscopic /endoscopy. The features extracted by the differences between the two branches which could increase the density of the feature points compared with the traditional method. Based on this, the depth acquisition network DepthNet and optical flow estimation network FlowNet were introduced to increase the corresponding feature points. The obtained sparse signals predicted the dense depth information by self-monitoring method and carried out the dense 3D reconstruction of the organization. At the same time, the alignment between sparse and dense images, local and global images, and scale consistency were ensured.

In Section A, we explain the structural framework and working principle of the 3D reconstruction model of a self-supervised double-branch Siamese network system based on the U-Net. Section B introduces the pre-processing stage of the model data and the ORB-SLAM framework foundation. In Section C, we introduce a two-branch Siamese network structure based on U-Net. In addition, we provide the encoding and decoding parts of the network structure in detail in Section D. We introduce in detail the depth acquisition network DepthNet and optical flow estimation network FlowNet proposed in Sections E and F, respectively. Meanwhile, the loss function is consistent with the optical flow and depth, and the network is trained by backpropagation of useful information in the form of a gradient, so that the network can update parameters. Finally, an accurate 3D reconstruction was obtained using dense depth information. The specific methods are described as follows.

A. PROPOSED NETWORK ARCHITECTURE

This section introduces all these concepts. The SFM learner of the monocular depth estimation training framework represented cannot solve the problems of scale consistency and dynamic objects. Therefore, the system uses a sparse feature 3D reconstruction framework based on monocular ORB-SLAM to introduce an improved U-NET dual-branch Siamese network. We added the training network DepthNet for dense depth information extraction and FlowNet for optical flow estimation (extraction) in the data-processing process. The overall system architecture is illustrated in Fig. 1. The original ORB-SLAM pipeline is composed of Tracking, Local Mapping, Loop Closing and Full BA four threads, which can extract sparse depth information to form a sparse 3D reconstruction model. The reasons for the difficulty in dense stereo matching are as follows: the color and luminosity are inconsistent between different frames at the same position and there are occlusion and reflection areas in the image. Moreover, there were repeated patterns and nontextured areas in the image. The difference extraction feature of the two branches of the Siamese network based on the improved U-Net can increase the density of feature points more than that of the traditional method. Based on this, the depth acquisition network and optical flow estimation network were introduced to increase the corresponding feature points. The sparse signals obtained can predict the dense depth information using a the self-supervised method. Subsequently, the dense 3D structure is reconstructed. Simultaneously, the alignment between the sparse and dense, local, and global images is ensured. In addition, the scale consistency was taken into account.

First, DepthNet uses a sparse depth map estimated by ORB-SLAM to place pose constraints. Then it extracts dense depth information from the self-supervised dual-branch Siamese network. In addition, a patch-based method (see Section E for more details) was introduced to fill in the missing features such as reflection weak texture, forming a dense depth Map to ensure global consistency, aligning the sparse and dense depth map. Finally, the network enriches the depth-feature points in the original ORB-SLAM. Simultaneously, it enhances the map matching FlowNet to extract the camera pose through optical flow information. FlowNet is used to optimize the pose estimation and optimization in the original ORB-SLAM thread using a sparse depth map calculated by ORB-SLAM. An Optical flow composite RGB images were used as the self-supervised signals. FlowNet is also used to estimate self-motion and depth maps which are obtained by triangulation and serve as an auxiliary monitoring signal. Then, depth feature points (including corner points, corner points, SIFT edge points, etc.) are added. Through FlowNet and DepthNet, the self-supervised contrast learning method proposed by us adds depth information and 3d structure feature points, and finally forms a dense 3D reconstruction structure.

B. PRE-PROCESSING AND ORB-SLAM

In the data preprocessing stage, the input frames were randomly cropped, and the color was changed to increase the dataset. The frames are then split into two parts by odd

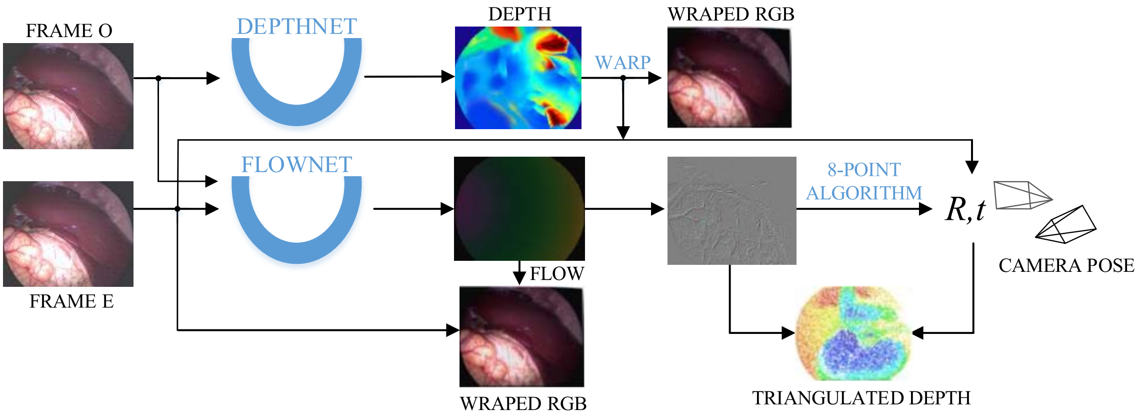


FIGURE 2. Training data generation pipeline.

and even frames. Simultaneously, the distortion coefficient estimated by the correction frame is used to correct the frame. Phase sparse tracking, sparse reconstruction and mapping are the traditional main threads of ORB-SLAM, which are used to estimate sparse reconstruction, camera pose and point visibility while ignoring the black invalid area within the frame. At the same time, we compared the internal and external neural responses of the missing regions, such as reflection without texture due to luminosity inconsistency, to ensure that the details of local small patches of texture inside and outside the region were similar. In addition, we considered outliers in sparse reconstruction and applied filters to refine the data. Without affecting the traditional sparse reconstruction ideas, we introduced contrastive self-supervised learning to extract dense ORB features in phase-dense reconstruction, align depth maps, and reconstruct the dense tissue surface photographed by monocular laparoscopy. Details of the dense reconstruction training method are described in Sections D, E, and F.

C. TWO-BRANCH SIAMESE NETWORK

We refer to Liu's network structure [8]. The difference lies in the fact that Liu *et al.* updated the network structure using SFM from a monocular endoscope, from which the self-supervised sparse signals were extracted as well as the estimated dense depth estimation. However, the training framework of monocular depth estimation represented by SFM learners cannot solve the problems of scale consistency and non-rigid surface reconstruction. Therefore, we introduced an improved U-NET based on a double-branch Siamese network, which is committed to solving the problems of scale consistency and non-rigid micro-deformation while improving the feature point extraction ability. This improves the performance of the depth feature extraction and 3D tissue reconstruction. Our two-branch Siamese network system structure is shown in Fig. 2, which is divided into DepthNet as the depth-estimation thread. DepthNet estimates the depth information of an image. FlowNet predicts the optical flow estimated of the input image, which was also used to estimate the camera pose and self-motion. The depth map was obtained using triangulation as an auxiliary monitoring signal. The proposed network automatically extracted ORB feature points, dense depth maps, and sparse optical flow using a calibration video by laparoscope. The feature extraction and matching threads were calculated in real time with GPU acceleration. The feature point lies in the center of the image patch, which is considered to be significant in multiple images of the same scene and can be detected repeatedly. Our detector recognizes the position of the feature point and codes the surrounding patches using descriptors. We used a discriminator (anti-loss) to sharpen the fill area to make the image clearer, inputting the fill area into the discriminator (spoofed discriminator).

In the process of ORB feature extraction, a two-branch Siamese network was introduced to share the weight of the preprocessed odd-even sequence. A sparse depth map, camera pose and intrinsic features were also provided. We considered the proportional consistency in geometric consistency and evaluated it using the geometric consistency loss. Considering the non-rigid structure of human tissues and the slight deformation caused by the contraction of the heartbeat and laparoscopic contact, a soft mask was introduced to correct used the photometric errors. The most commonly used networks in the field of depth estimation are VGG and ResNet. ResNet performs better than VGG, but more parameters lead to longer training and prediction times. The encoder was constructed using a new U-NET network with fewer parameters and a faster operation. The traditional network based on U-NET relies on a long connection structure to realize image recovery. The network we produced used a 12-layer network based on U-NET to complete our depth-estimation task, in which the encoding part has six layers and the decoding part also has six layers, as shown in Fig. 3. To solve the problem of information redundancy reducing network accuracy, we adopted the LFU-NET [9] structure proposed by Lei *et al.*, which can fuse the information of

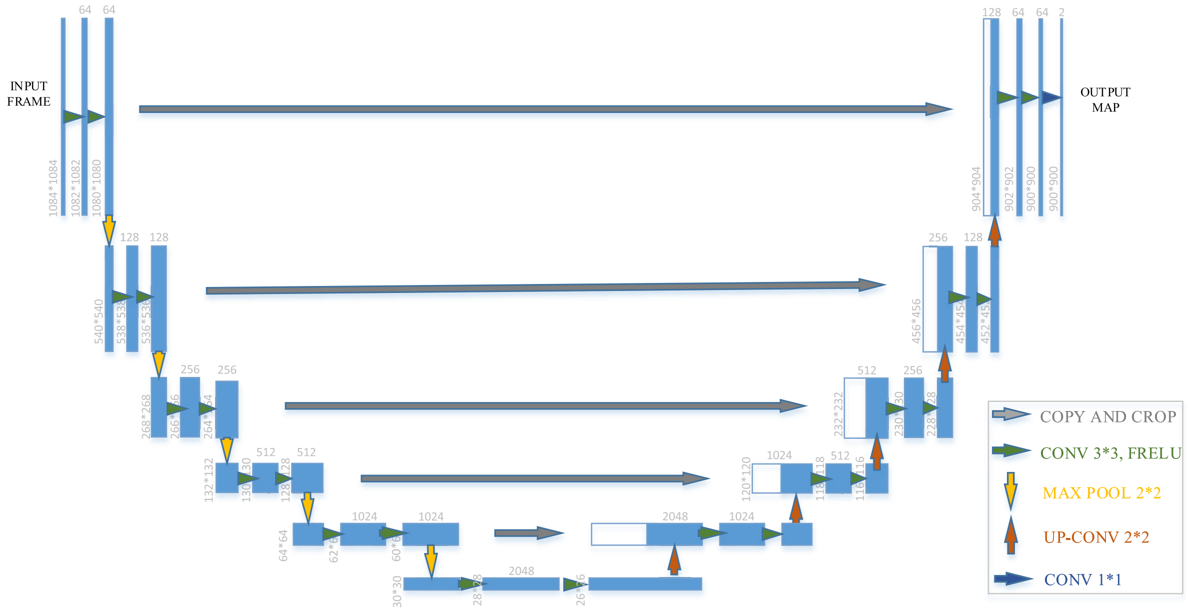


FIGURE 3. New U-Net architecture.

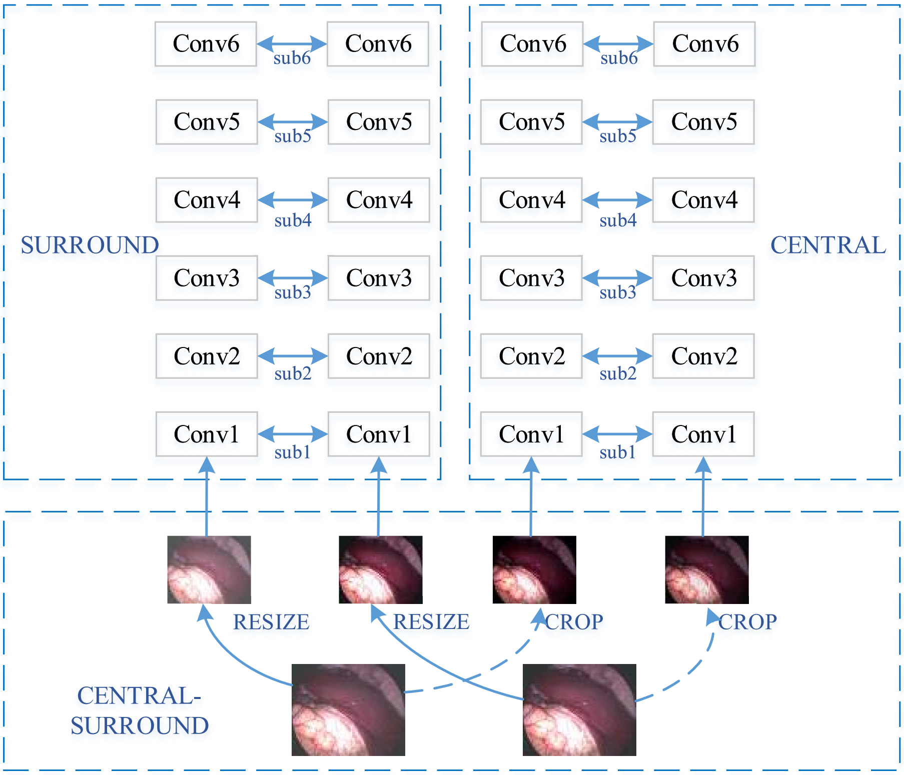
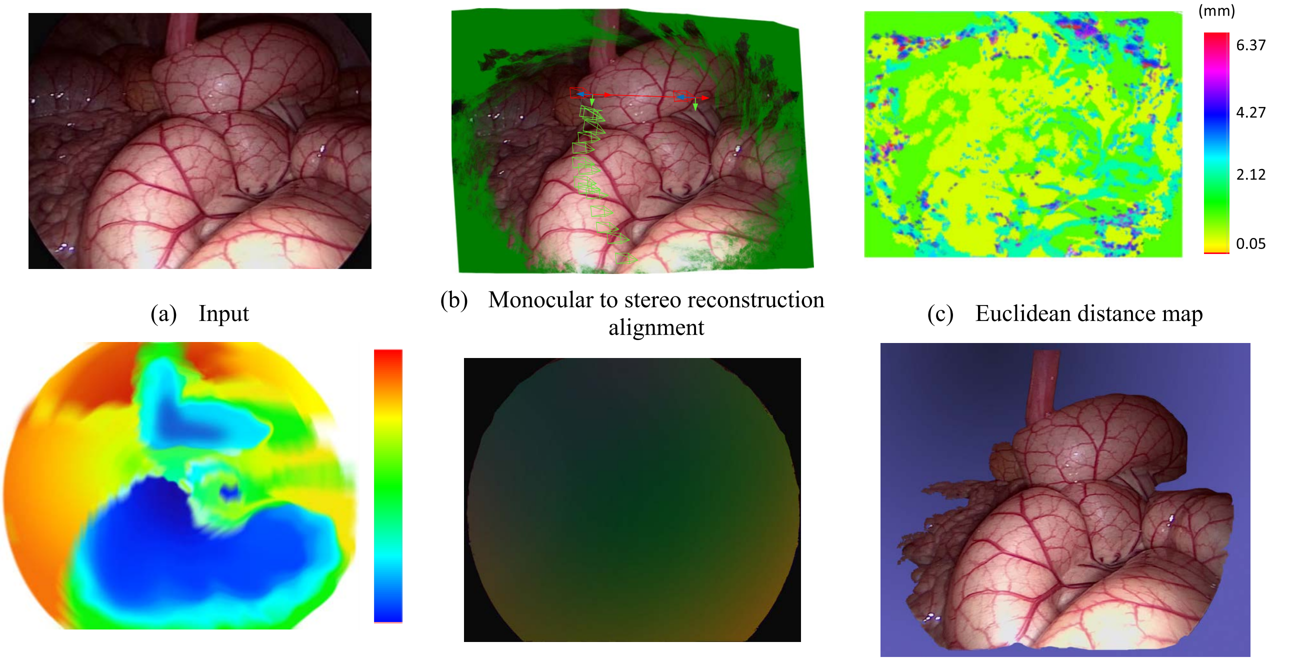


FIGURE 4. Encoding structure.

different levels extracted by the decoding layer and control the number of network parameters, which leads to a better effects that can be learned with fewer parameters.

D. ENCODER AND DECODER

The connection between the Encoder and Decoder was used by Decoder to repair details. It was found that increasing the score can improve the output resolution, so we integrated the idea of residual network and added a branch structure. We added 3, 4, 4, 2, and 2 convolution layers to the U-Net encoder part conv0-conv4 respectively. The decoder parts decode1-decode5 were successively added with 2, 3, 3, 2, and 2 convolution layers, and the Resize_Biliner layer was introduced in decode6. The latest stereo matching methods



(a) Input

(b) Monocular to stereo reconstruction alignment

(c) Euclidean distance map

(d) Depth Map

(e) Flow Map

(f) Reconstruction

FIGURE 5. Results of different steps of the proposed method.

used a deep training model to improve the reconstruction accuracy, but the optimization results increased the computational cost, even though a medium GPU network may not be suitable. When a model must be deployed on a device with limited resources, it can break down or even fail. Therefore, we propose a network training model that can reduce complexity without sacrificing accuracy. The codec model of the combined 2D convolution and 3D convolution was designed according to the cost volume.

The coding part consists of the Central region and Surround region, as shown in the Fig. 4. The downside box is a schematic diagram of processing the images of the two periods before and after the input. Conv is the convolution operation and sub is the difference image. When images from the Central region and Surround region are input to the coding end of the improved U-NET two-branch Siamese network, the depth information can be extracted with the help of image differences.

$$S_{subi}, C_{subi} = \left|f_{1,i} - f_{2,i}\right|, \quad i = 1, 2, 3, 4, 5, 6 \tag{1}$$

where S_{subi} and C_{subi} are the different images of the convolution layer of layer i surrounding the central region of the region respectively. f_i is the feature image obtained from the images of the two periods in the convolution layer of layer i. Then the band dimension feature fusion is performed for the difference images between the central region and the surrounding region of the universal convolution layer. The fused image was considered as the feature image extracted from the coding end.

We use 2D MobileNet blocks to extend it to 3D stereo-vision applications, and propose new costs to improve the 3D model accuracy to make its performance close to the real value. It not only reduces the network parameters, but also improves the network learning ability. Our network consisted of DepthNet and FlowNet. DepthNet can estimate the depth information whereas FlowNet estimates the optical flow. The deep network consisted of a debugging-link codec network with a general U-NET architecture. We used ResNet18 as the encoding network. The deep decoder was similar to using sigmoid to output different scale (1/8, 1/4, 1/2, 1/1) depth maps [10], using FReLU instead of depth decoding where other places are exponential linear units (ELU) nonlinear. Adding FReLU to the encoder can improve the detail of the depth map, increase the attention to high-dimensional features, and reduce noise in depth estimation. For the optical flow network, we used lucas-Kanade (LK) [11] optical flow method to represent the predicted optical flow. It then represents the pose and rotation of the camera through the axis angle, similar to [12]. During monocular sequence training, the network inputs two sequences and outputs a single 6-dof relative pose using ResNet18.

The dense depth-estimation network mainly consists of a detailed encoder and decoder network architecture. The encoder-decoder learns to map data points from the input domain to the output domain through a two-stage mechanism in the network. In the first stage, the encoder function $f = f(x)$ represents the compression of the input into a latent-space representation. The decoder function $y = g(f)$ predictes the output in the second stage. In the encoder, MobileNet is used to decompose the CNN layer into depth and point direction layers based on the depth decomposition process. Each depth layer uses a filter function to extract low-resolution or weak-texture features from an input image. The extracted features are fed into a decoder that extracts, merges, and upsamples them to produce a dense output depth map. In the second stage of the network, the decoder consists of six upsamples and a singleton layer. Each upsampling layer performs a 3 × 3 CNN to reduce the number of channels in a 2:1 ratio. Three jump connections are applied to reconstruct a more detailed secret-level information for the final depth map. To minimize reconstruction errors, the mixed loss function measures the differences between the ground truth and predicted depth.

E. DEPTHNET

We constructed the proposed self-supervised deep completion network based on a previous self-supervised deep-learning framework. The main difference is the use of two different depth-dependent monitoring signals. We relied on the successful framework of ORB-SLAM [13] as the basic framework. The entire network consist of DepthNet as a dense depth network. The dense depth estimation network, called DepthNet, is used to constrain the camera pose and ensure the alignment of sparse and dense depth maps to maintain local and global consistency. An optical flow estimation network was used to estimate the optical flow and self-motion of the camera. We focused on the corner and edge points of the monocular laparoscopic image sequence. In the process of endoscopy moving in the body, according to the characteristics of the motion speed of pixels at different depths in the monocular image, pixels far from the camera move slowly, while pixels close to the camera move fastly. Aiming at a multi-scale feature detection scheme, we localized the activation of each channel in the maximum stable extremum region by detecting local maxima in the activation graph of multiple activation layers. We matched the detection of the feature points in the training image with the double-branched Siamese network. Lens distortion was considered at the same time. Using a group of laparoscopes as an example, we calculated the depth map of the input frame. The sequence of the input frame is shown in Fig. 5 (a). The calculated Euclidean distance map and depth map are shown in Figs. 5 (c) and (d), respectively. The obtained depth map of the comparison study is shown in Figs. 7 (b) and (d). Fig.7 (b) shows the depth score calculated by different methods for pre-selected feature points and a comparison of the real depth. The results showed that the comparison method proposed in [3] was high, but the selection line was poor. Method [17] scored similarly to ours, but it was easy to fall into a local optimum. Method [18] also tended to fall into a local optimum with a low score. Our method scored higher in terms of true depth and was more selective. Fig. 7 (d) shows the average and median depth values calculated using the endoscope keyframe sequence. When the number of pixels of the screened corresponding feature points decreases, the consistency also decreases.

First, we used short-range depth (SFM/ORB-SLAM/ multiview geometry) as the monitoring signal. ORB-SLAM tracked the motion of the camera in the last frame. Then, it internally extracts the key frames that are passed on to the local and global mapping threads. Specifically, for pixels where a depth measurement exists, we calculated the difference between DepthNet's estimated depth of odd and even frames. Let p be a pixel, $D(p)$ denote an even frame depth map, and $\hat{D}(p)$ represent an odd frame depth map. The loss function is defined as follows:

$$L_{dd} = \frac{1}{|M_d|} \sum_{p \in M_d} \left| D(p) - \hat{D}(p) \right| \tag{2}$$

Furthermore, 3D points corresponding to the optical flow recovery were used as monitoring signals. Similar to the camera attitude estimation, k pixels are used to correspond and dual-view triangulation [14] is applied to obtain the depth of these pixels. Because the camera's attitude is relative in proportion, the depth obtained by triangulation is also relative. To align the depth ratio between the depth D_t obtained by triangulation and the depth estimated by DepthNet D, a single-scale factor s is introduced. The depth was considered to be $D = sD_t$. Using s, which minimizes the error, we calculate the loss as follows:

$$L_{dt} = \frac{1}{|M_t|} \sum_{p \in M_t} \left| D(p) - \hat{D}_t(p) \right| \tag{3}$$

To solve the problem of low texture and lack of information in a uniform region, we adopted the depth-consistent loss algorithm. The depth inconsistency of each pixel can be described as

$$D_{diff}\ (p) = \frac{\left| D_b^a\ (p) - D_b'\ (p) \right|}{D_b^a\ (p) + D_b'\ (p)} \tag{4}$$

However, the depth scale predicted by the network was consistent for both frames and sequences. Therefore, through the estimated pose transformation, pixel-level differences can be directly compared in the same coordinate system, thereby quantifying the scale consistency between and. To strengthen the geometric constraint between the two independently predicted depth maps, a scale-consistent loss is introduced, which is expressed as

$$L_{ds} = \frac{1}{|V|} \sum_{p \in V} D_{diff}\ (p) \tag{5}$$

Above all, the loss function for DepthNet is defined as follows:

$$L_{depth} = \omega_1^d L_{dd} + \omega_2^d L_{dt} + \omega_3^d L_{ds} \tag{6}$$

In the training, the five weights are respectively set as $(\omega_1^d, \omega_2^d, \omega_3^d) = (0.5, 0.0001, 0.2)$.

F. FLOWNET

Optical flow allows the calculation of the correct amount for each pixel to produce the relative motion between two images. Ego-motion estimation is an important part of the self-supervised learning in this system. The optical flow network is trained in an unsupervised manner. Therefore, the entire process can be regarded as unsupervised learning. Lv *et al.* [13] proposed a CNN that uses optical flow images as input and self-motion as output to improve the depth estimation performance in an indoor environment. Zhou *et al.* [12] showed that it is more effective to estimate self-motion by calculating the basic matrix based on the correspondence obtained from optical flow than by inputting optical flow images into a CNN. In this study, we used a corresponding-based self-motion estimation method. In this study, we used a corresponding-based self-motion estimation method. The obtained optical flow diagram is shown in Fig. 5 (e).

The core idea is to obtain a reliable correspondence from the optical flow. FlowNet is used to estimate the forward and backward optical flows of the input image. The occlusion mask [14] and the flow consistency score [10] were calculated. We then randomly sampled k correspondences in the top 20% of the unoccluded area consistency score. Once the correspondence is found, the basic matrix is calculated using the 8-point algorithm [15] and RANSAC [16]. The appropriate camera pose is estimated using geometric verification.

Non-rigid bodies, occlusion, and inaccurate depth prediction of complex regions all lead to a greater loss of geometric consistency. To solve the problem of inconsistent dynamic object mapping, a soft mask and optical flow loss are introduced, so that a soft mask with the same proportion of sparse and dense mapping can explain the difference in error distribution of each point in the spherical SLAM results, alleviating the influence of sparse 3D reconstruction errors.

$$M = \begin{cases} 1 - \alpha \cdot e^{-\sum_i b_n^i/\sigma} - \beta \cdot D_{diff}, & if\ b_n^i = 1 \\ 0, & if\ b_n^i = 0 \end{cases} \tag{7}$$

Mask can be understood to mean that the current pixel-predicted depth is not affected by the above three reasons, which can reduce the adverse effects of dynamic objects and occlusion on training. In network training, the ratio of losses in this region to the weight of the total losses is reduced, thus weakening the influence of back propagation on parameter updating. It is also a detection map for dynamic object and scene occlusion.

The sparse flow graph of two-dimensional position in frame K is organized into a regular two-dimensional grid, which can be expressed as

$$F_k = \left(\frac{U_k - U}{W}, \frac{V_k - V}{H} \right) \tag{8}$$

As a regular grid, U consists of H rows and V consists of W columns.

To generate the correct dense depth map consistent with the sparse reconstruction of ORB-SLAM, a sparse optical flow loss training network is introduced to minimize the difference between dense optical flow maps and sparse optical flow maps, so as to solve the problem of data imbalance caused by inconsistent scales.

$$L_{df} = \frac{1}{\sum M_j} \sum \left(M_j \left| F_j^s - F_k \right| \right) + \frac{1}{\sum M_k} \sum \left(M_k \left| F_k^s - F_j \right| \right) \tag{9}$$

The overall loss function for the network training from the pre-processing data is

$$L_{reconst} = \lambda_1 L_{depth} + \lambda_2 L_{flow} \tag{10}$$

G. LIVE ALIGNMENT OF KEYFRAME DEPTHMAPS AND BUNDLE ADJUSTMENT

We combined the dense depth maps from the SLAM map into a single coordinate system to obtain a globally consistent reconstruction. Odd frames were used as anchors to keep the depth map aligned with the sparse SLAM map. Therefore, any update of the depth distortion layer and compact depth map may result in a rearrangement of the compact depth map. Various alignment methods have been studied extensively. The RANSAC algorithm is widely used to align two images and determine the necessary matrices. The geometric features of each 3D point are encoded using feature vectors. The FPFH can fully express the 3D features of points without occupying excessive computing resources. Feature points that do not meet the geometric constraints can be disqualified. All matching points were recorded and returned. This monocular scale recovery was computed only once for each sequence. It was then used to scale all keyframe depth maps from the same sequence. The Euclidean distances between all pixels from the two reconstructions are shown in Fig. 5(c).

We used [20] to fuse depth data, which is achieved by minimizing the energy function, including the depth information error and optical flow information error. The depth was the most accurate when the error function was minimized. Abdominal densification was calculated by iteration and minimization of the total loss. After each BA of SLAM, the attitudes of the sparse points and keyframes are refined. This refinement may result in a misalignment of the density depth map with respect to the SLAM map. Such improvements may involve not only rotation and translation, but also scale changes. Therefore, we aligned each depth map by using the method proposed in [9]. In addition, in order to improve the alignment efficiency, we propose a mixed attention mechanism to optimize our network with the help of [9], and fine-tune the loss function of the unsupervised 3D reconstruction algorithm. The intrinsic parameters of the camera were extracted via camera calibration, and the external parameters were calculated using ORB-SLAM. These parameters are input into two multiview stereo vision methods, and the results of the two methods are compared to obtain better reconstruction results than unstructured datasets. The monocular-to-stereo reconstruction alignment results are shown in Fig. 5 (b). The final 3D reconstruction results are shown in Fig. 5 (f).

IV. EXPERIMENTS AND EVALUATION

A. DATA COLLECTION AND PRE-PROCESSING

1) DATASETS

Endoscope images were obtained using a discontinuous optical amplification endoscope series (GIF-HQ 290 model) and a CV-290 video system (Olympus Medical System). In addition to the open dataset from the Hamlyn Center Laparoscopic/Endoscopic dataset [15], the dataset was used for training and testing the model proposed in this paper. In addition, medical image video data approved by anonymous patients were obtained from the Peking University BinHai Hospital, including 73 cases of liver laparoscopic video (46 cases with CT images of corresponding positions) and 52 cases of gastroscopy video (without CT). 12 patients underwent sinus endoscopic video (without CT) and 49 patients underwent Unilateral Biportal endoscopy (UBE) video (all patients had CT [16] or MR [17] images at the corresponding position). We mainly used open datasets and liver laparoscopic video sequences from hospital patients to train, evaluate and test the model proposed in this paper. We used gastroscopy, sinus scopy, and UBE video as a comparative study of the reconstruction effect of the model.

The system was implemented using C++ and OpenCV. The workstation uses two Nvidia Tesla M60 GPU, each with 16GB of memory, and an Intel(R) Core i7 CPU with 3.4GHz. This method uses PyTorch. In order to facilitate calculation comparison, we sampled all frame resolutions to 1084 × 1084 in training and testing, which were recorded the average time consumpting of each step in Table 1. Sparse Tracking and Reconstruction, KeyFrame Selection, Pose Estimation, Bundle Adjustment, Computation of loss function by GPU and CPU, Depth Map, Flow Map, Alignment of KeyFrame time (in seconds) Time comparison was made with the lower sampling resolution of 640 × 640 and 320 × 320 respectively. The higher the resolution, the greater is the computation and time consumption. Meanwhile, the reconstruction effect deteriorates with the feature loss. Among them, the calculation amount of the dense depth map mapping step was the largest, and the time-consuming was the longest. The calculation of the dense depth map of the input frame with 1084 × 1084 resolution reaches 425s, while when the input frame was sampled to 320 × 320, the time consumption was reduced to 93s. The fastest step was the sparse tracking portion, which could reach as fast as 0.025s, owing to the mature traditional ORB-SLAM threads and the fast and efficient ORB feature points.

Simultaneously, we reconstructed the key steps in the results listed in Fig. 5. Fig. 5. (a) denotes the input sequence of frames, and (b) aligned for the monocular stereo reconstruction process, the green marked as camera path and posture, (c) shows the Euclidean distance map, (d) was the depth map, blue to red indicate the distance from the camera, (e) is the light flow diagram, and (f) is the reconstruction result. The effective key frames for 3D reconstruction for each patient ranged from 1705 to 4322 frames, except for frames that were blurred or too bright/weak to extract feature points.

2) PRE-PROCESSING

None of the training data is marked. We corrected the frames by using the distortion coefficients estimated from the corresponding calibration video. Then randomly crop frames, change colors, and sharpen to add datasets and enhance the features. The new extended frames are then divided into odd and even frames. They were then trained using the self-supervised double-branch connected network described in the data training section. Simultaneously, the distortion coefficient estimated by the correction frame is used to correct the frame. The frames contain similar luminosity and texture

TABLE 1. Average processing time (in seconds).

	SPARSE	KEYFRAME	POSE	BA	LOSS FUNCTION		DEPTH	FLOW	ALIGNMENT
	TRACKING	SELECTION	ESTIMATION		GPU	CPU	MAP	MAP	
1084*1084	0.042	0.15	1.2	0.097	10.43	876	425	261	312
640*640	0.031	0.11	0.9	0.072	4.82	701	276	97	127
320*320	0.025	0.09	0.7	0.060	3.40	584	93	34	61

TABLE 2. Reconstruction error of different methods.

	MAE	MRE	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
[17]	0.2851	**0.3092**	0.2913	2.1003	0.1017	0.8235	0.9926	0.9971
[18]	0.4264	0.4953	0.3984	2.8192	0.1215	0.7021	0.9893	0.9915
[3]	0.5131	0.6271	0.6089	3.0519	0.1928	0.6017	0.9389	0.9960
[19]	3.7851	1.8126	1.5723	6.3751	0.2284	0.5033	0.9217	0.9770
[20]	1.1923	1.0322	1.0037	5.1129	0.3022	0.6950	0.9534	0.9872
[21]	1.8237	1.9723	1.0980	5.8516	0.2917	0.6532	0.9511	0.9713
Ours	**0.2645**	0.3132	**0.2789**	**1.9826**	**0.0910**	**0.8861**	**0.9962**	**0.9978**

TABLE 3. Reconstruction error in proposed framework of different surface.

	MAE	MRE	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Laparoscopic	**0.2645**	**0.3132**	0.2789	1.9826	**0.0910**	**0.8861**	**0.9962**	**0.9978**
Gastroscope	1.3236	1.5328	**0.1654**	2.2543	0.2321	0.8573	0.9297	0.9968
Sinus	0.6112	1.2336	0.4552	**1.5565**	0.3165	0.8562	0.9315	0.9937
UBE	5.2518	2.0743	1.0679	4.1216	0.3517	0.8037	0.9154	0.9871

information. Private and invariant feature extractors are used to absorb orthogonality and similarity, reduce the negative effects of interference items (lighting, etc.), and helping to obtain a better depth map. At the same time, odd-even information and depth maps were efficiently estimated by minimizing the photometric loss.

B. EXPERIMENTAL SETUP

A retention training model was used, in which data from one patient were used for testing and data from other patients for training. The experimental group was also evaluated by three primary laparoscopists, who had performed more than 2,000 endoscopies. The original figure and the reconstruction results are presented in Fig. 6. The top line represents the original frame image input. [3], [17], [18], respectively, the current 3d reconstruction method of endoscopic images; the fifth act is the reconstruction result of our method; [19]–[21] are the reconstruction results of the computational network in this model replaced by other deep learning methods. Since the model proposed by [3], [17] and [18] was designed and trained for endoscopic images, the reconstruction results were better than those of [19], [20] and [21], but some key feature points were still missing. Our reconstruction results could better extract most of the feature points, and the reconstruction effect was good. In our method, endoscope has the best effect, because the model is trained according to the data of endoscope, and our method has the worst effect of arthroscopy reconstruction. Owing to the weak texture of the arthroscope, the imaging quality is reduced because of the operation in normal saline, and some feature points cannot be effectively extracted.

1) FREE PARAMETERS

In the model training, FlowNet and DepthNet are trained with a batch size of 64 using the Adam Optimizer (beta1 = 0.9, beta2 = 0.999). The learning rate was set to 0.001, reduced to 20 epochs, and then to 0.0007, 0.0002 and 0.0001. We studied the effects of different epoch settings on MSE, as shown in Fig. 7(a). The Epoch significantly improved the training accuracy between 20 and 40, but when the epoch was set to 80 or more, the accuracy improvement was not obvious.

In order to verify the actual performance of the algorithm, except for special instructions, the parallax graph used in this study was the initial parallax graph without any parallax refinement and post-processing. The error limit of the comparison experiment was 3 pixels, that is, when the difference between the matching result and the real parallax map was more than 3 pixels, the point was considered to be mismatched. The main sensitive parameter is the overlap between cluster frames controlled by and. For experimental indicators, we used the following indicators to evaluate our depth information and 3D reconstruction methods: sqRel, Mae, RMSE and log RMSE(the smaller the better). We compute the accuracy of $\delta < 1.25$, $\delta < 1.25^2$, and $\delta < 1.25^3$ respectively. We quantitatively compared the reconstruction results of the different methods, and the endoscopic reconstruction of different sites is shown in Tables 2 and 3. The detailed procedures are presented in Section C.

C. COMPARISON STUDY

We conducted a comparative study with the 3D reconstruction method proposed in [19]–[21] by other scholars recently, the deep learning method and the application of our method in medical endoscopy videos of different parts, respectively, to evaluate the performance of the model quantitatively and qualitatively. For quantitative evaluation, MAE, MRE, MSE and RMSE were used to test the effect of the reconstruction. The reconstruction results of the different methods are shown in Fig. 6. We also used CT images of the same patients for the registration observation of the feature point clouds, and the qualitative comparison results are shown in Fig. 8.

1) 3D RECONSTRUCTION METHODS

We compared the 3D reconstruction results of our method with those proposed by other scholars recently introduced in [3], [17] and [18]. Fig. 6 shows the qualitative comparison results of the liver greater omental membrane obtained by laparoscopy, gastric fundus taken by gastroscope, sinus and UBE respectively. The average render parallax is 12.4 degrees. The proposed system and [18] produce similar accuracy, both presenting higher monocular parallax than the stereoscopic method, but the proposed system is several orders of magnitude faster. [18] took 4.5 min for intensive reconstruction and 1.5 min for subsequent filtering.

The method introduced in [17] designed a patch embedding network based on Multiview Stereo(MVS) algorithms, projected each candidate image onto other images in the sequence, used a patch embedding network to map each image patch into a compact embedding, and measured the matching score. Finally, the candidate with the highest score is selected for each pixel. The reconstruction result is valid and complete, but there's a disadvantage of which the speed is too slow. The model training time was 21 minutes 38 seconds, and the model could not be used in real-time. [18] The method is based on a U-NET convolutional neural network structure to extract carotid artery features and segment them, which is sensitive to contour structures. However, it is insensitive to the non-contour parts of the viscera with weak texture, resulting in sparse model results. [3] Methods Based on patient-specific learning, sinus surface anatomy was reconstructed directly and only from endoscopic videos. Inheriting the characteristics of SFM, this method is not ideal for fuzzy images or tissue deformation.

2) COMPARISON WITH EXISTING DEEP LEARNING METHODS

We used the contrastive learning method proposed in this paper to conduct a comparative study on the reconstruction results of other recent deep learning methods. We used the network structure proposed by [19], [20] and [21] respectively to replace the double-branch network proposed in this study to extract the features of the frame sequence and perform 3D reconstruction calculations. The results are shown in Fig. 6. As the comparison method [19]–[21] is not a model training method for medical images, although it has high computational accuracy in their respective scenes, it is not suitable for medical endoscopic images. Therefore, the reconstruction results are not good.

In the method in [19], a pre-trained encoder is introduced on the basis of U-NET, and the decoder part is redesigned. The transpose convolution operation is replaced by an upsampling operation based on Nearest Neighbor (NN) interpolation. However, this method exhibits insufficient performance in identifying scenes and feature point types. The 3D reconstruction of human endoscopy is still sparse. This is not robust to artifacts. [20] Methods VGG-19 was replaced with a deeper residual neural network (RESNET-50) as the feature extraction tool. A more robust target representation feature was obtained by the fusion of special additional layer structures and convolution layer features in RESNET-50. Subsequently, the feature result and filter are correlated to the filter. The target location is determined based on the maximum response value. However, the effect of 3D reconstruction of the laparoscopic endoscopy application scene needs to be improved owing to increased significance monitoring and increased computation. The feature extraction method in [20] is based on the multitasking network and real-time monitoring of the depth of the feature points learning (visual SLAM system, with the simplified design of CNN network detection feature point and the descriptor instead of the traditional feature extraction. Its test based on the indoor scene, has the obvious characteristic, in view of the weak texture, the micro-deformation characteristics of the laparoscopic scene effect remain to be improved, and there is still a gap in real time.

In view of the above comparative study, we performed the reconstruction of the proximal jejunum of liver in Laparoscopic by methods introduced in [3], [17]–[21], calculated MAE, MRE, Sq Rel, RMSE, RMSE-log, and the accuracy of $\delta < 1.25$, $\delta < 1.25^2$, $\delta < 1.25^3$ respectively, and recorded the results in Table 2. In addition, we performed 3D reconstruction of jejunum in the proximal liver of Laparoscopic, and calculated the scores of different methods with an average depth of 0.85mm to 1.15mm, that is, the accurate probability of depth characteristic information, as shown in Fig. 7(b). The results show that the depth score of our method is more selective in the real value when extracting a certain feature point, whereas other methods have no obvious advantage in the nearby range.

Fig. 7(c) shows that the accuracy of different methods decreases with an increase of keyframe determination books in the sparse tracking thread, but the accuracy of our method is more advantageous than other methods. By extracting the depth of all feature points, the average and median values were calculated respectively as shown in Fig. 7(d), which shows that consistency was negatively correlated with the number of feature points. Comparative data show that the proposed method can obtain better results on an existing basis.

3) COMPARISON OF 3D RECONSTRUCTION OF DIFFERENT BODY PARTS

We applied the proposed training model to 3D reconstruction of gastroscopy, sinus and UBE. Then a comparative study is conducted on the above reconstruction results. The results showed that the model trained in this study had a good reconstruction effect on different parts of the human body. As shown in Fig. 4. The reconstruction effect is affected by frame sequence resolution, lighting, texture, contrast, and sharpness, etc. At the same time, MAE, MRE, Sq Rel, RMSE, RMSE-log, and we also calculated the accuracy of $\delta < 1.25$, $\delta < 1.25^2$, $\delta < 1.25^3$. The quantitative comparison results are presented in Table 3. Since the proximal jejunum of the liver in laparoscopic surgery was used for training in our model, the reconstruction effect was better in laparoscopic, which is shown in Fig. 6. The reconstruction result also depend on the quality of the video frames. Sinus image quality was better in the sinus mirror, depth data quality was better, and therefore showed better reconstruction.

4) QUALITATIVE EVALUATION COMBINED WITH CT

CT images of the same patient were used to register observation of feature point clouds. CT selected a window of 77 and a window width of 2202. A fault image at −130mm to −158mm in height, because this position was consistent with the position of the proximal jejunum that we selected for 3D reconstruction. Methods [3], [17] and [18] were respectively used to conduct qualitative comparisons with the reconstruction results of the method proposed in this paper. It was found that the feature points and depth information extracted by the proposed method can well fit the patient CT model well. In contrast, [3] extracted the fewest feature points, but it can still better fit the real organ of the patient CT. [3] extracted more feature points, but there will still be a large number of feature point interferences outside the target organ, which will further interfere with the final reconstruction effect. Through qualitative comparison, it was proved that the model presented in this paper has high accuracy and anti-interference performance. The fitting results are shown in Fig. 6.

D. CROSS-PATIENT STUDY

In addition to the open dataset from the Hamlyn Center Laparoscopic dataset [15], In addition, medical image video data approved by anonymous patients were obtained from Peking University BinHai Hospital, including 73 cases of liver laparoscopic video (among which 46 patients had CT images of the corresponding positions) and 52 cases of gastroscopy video. There were 12 sinus endoscopic videos and 49 UBE videos (all patients underwent CT or MR images of the corresponding position). In this case crossover study, images of 46 laparoscopic patients with CT scan sequence data from 49 UBE patients were qualitatively evaluated. To demonstrate the generality of our method, we performed a

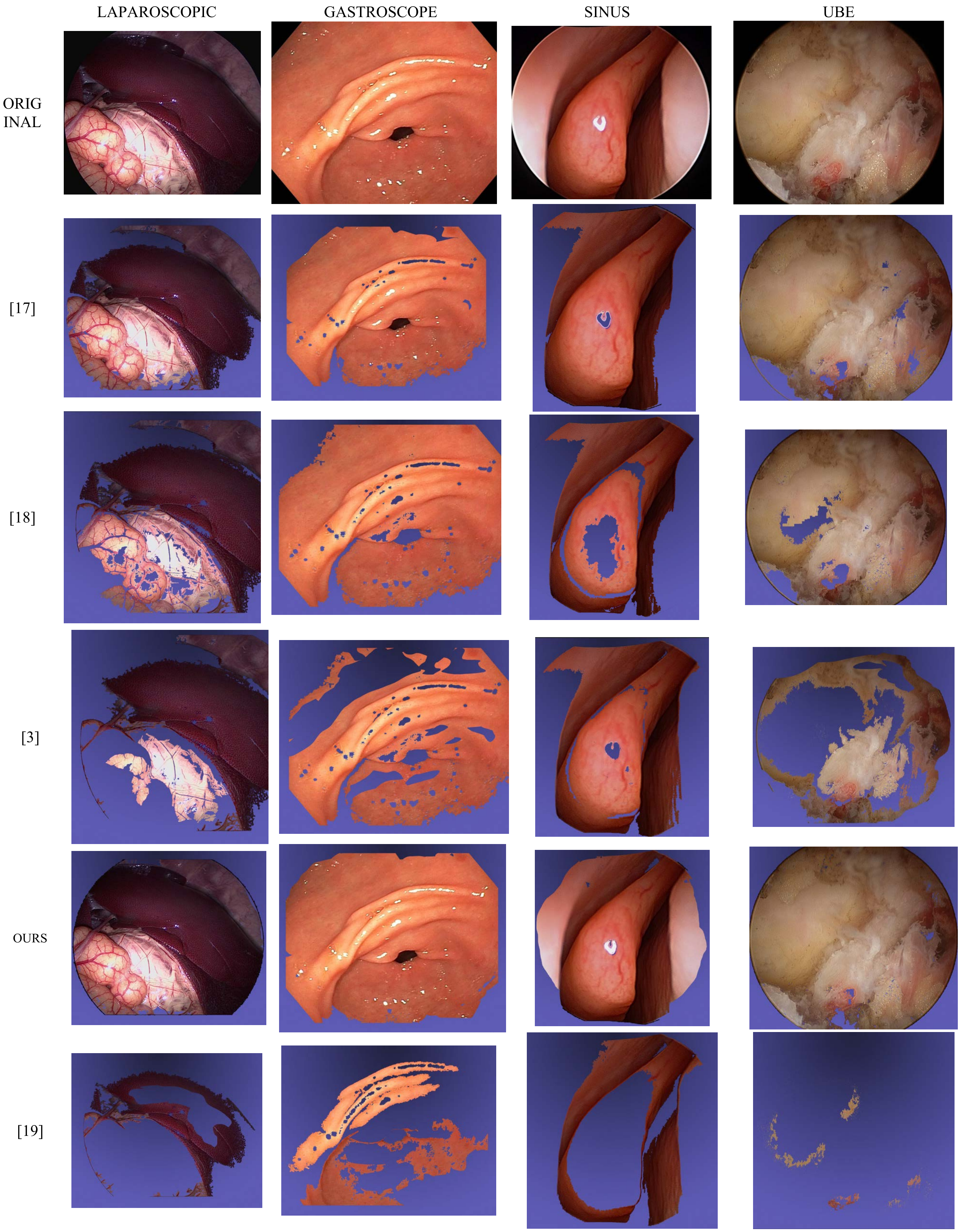


FIGURE 6. Comparition study of different methods.

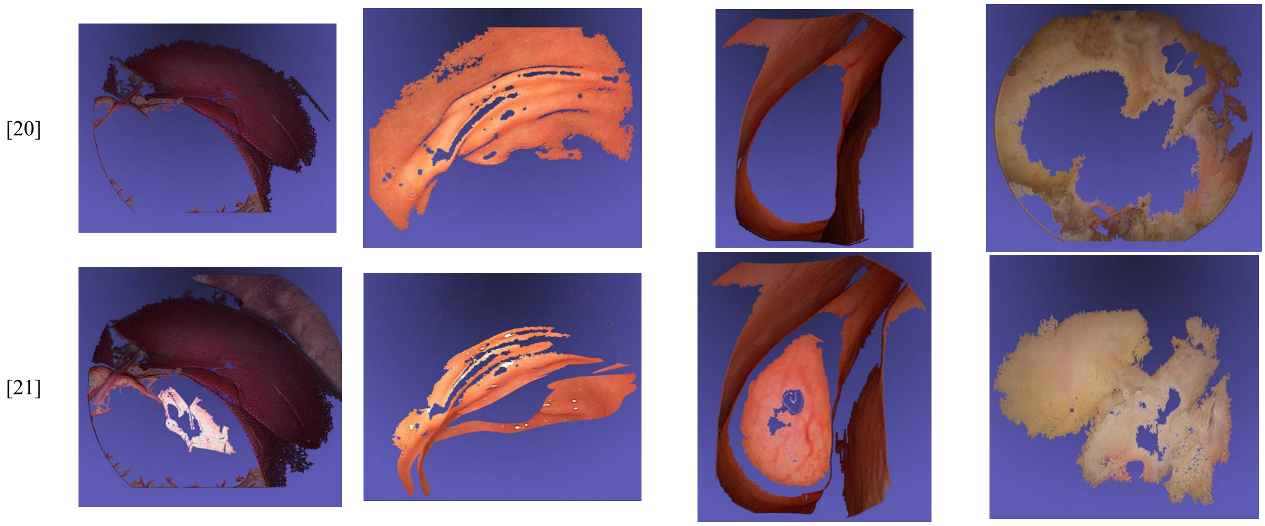
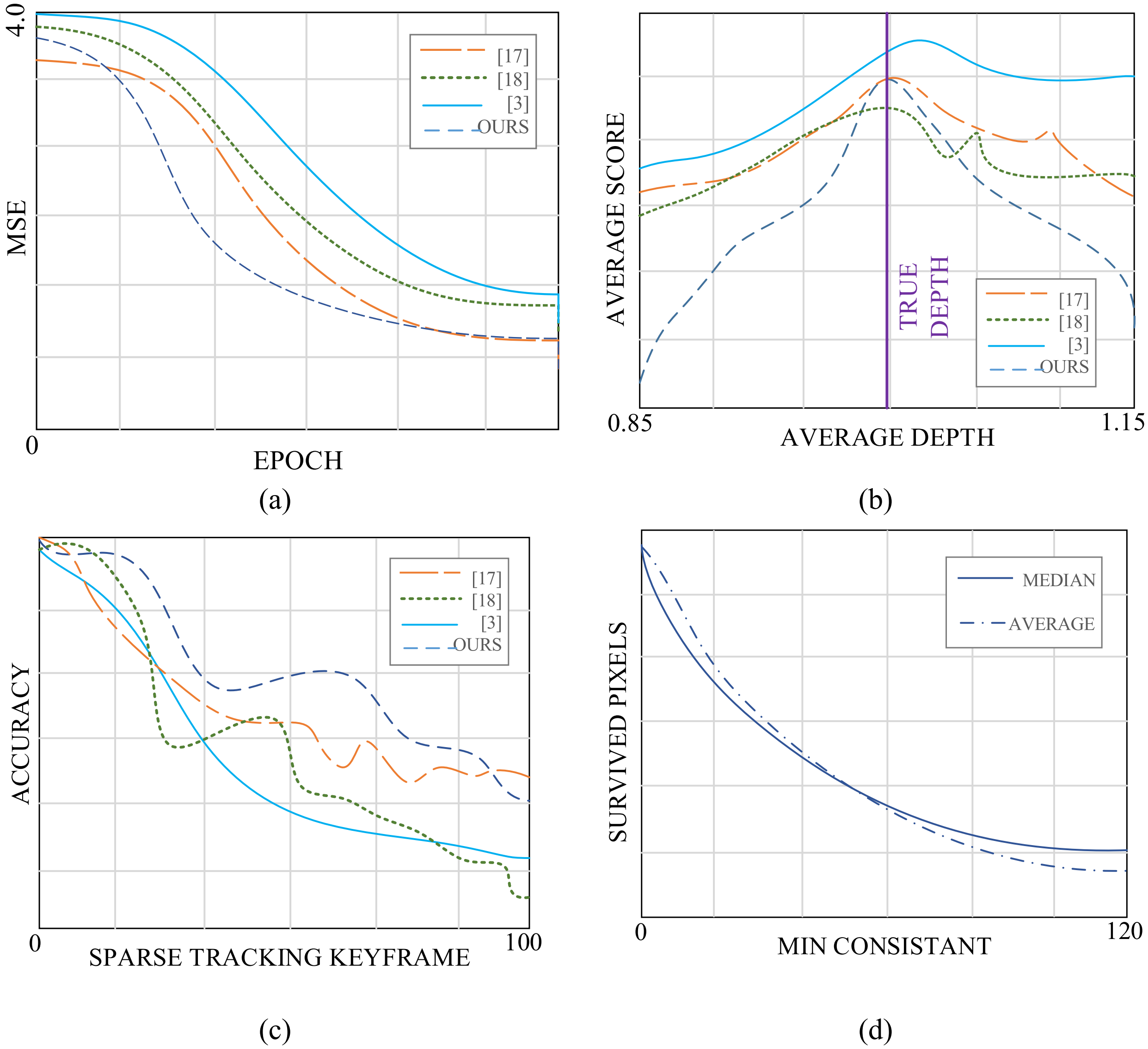


FIGURE 6. ***(Continued.)*** **Comparition study of different methods.**



(a)

(b)

(c)

(d)

FIGURE 7. **Quantitative comparison curve.**

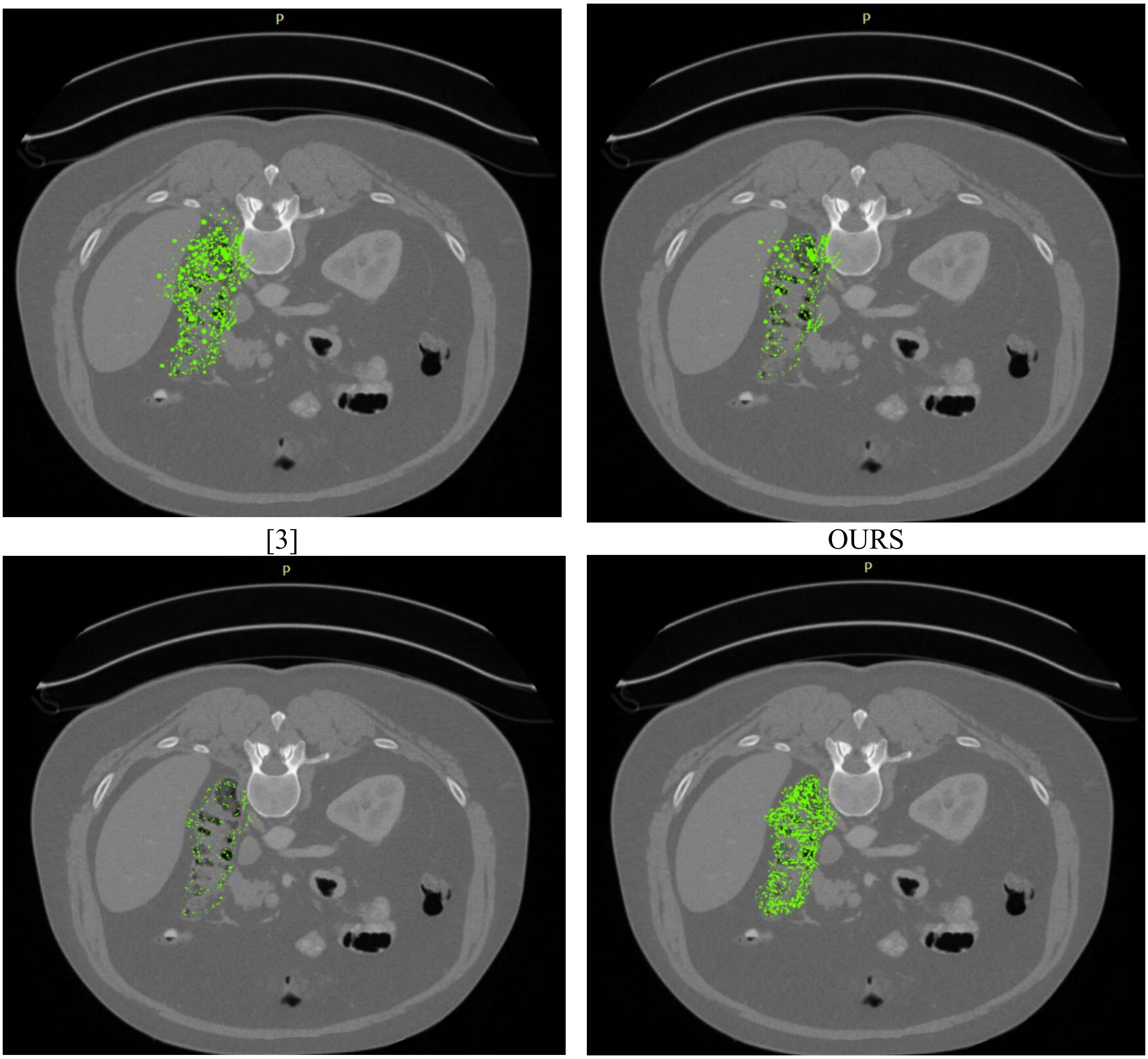


FIGURE 8. Qualitative comparison with fitting CT.

3D reconstruction experiment on four images. We included 95 patients during the evaluation training. As shown in Table 3, we can see that our method has achieved submillimeter residuals for all test reconstructions. The results showed that this method is effective for different patients.

E. ABLATION STUDY

In order to evaluate the effect of each loss component, i.e., consistent scale and depth estimation, only using consistent scale and depth to train the network, we conducted computational tests on all the medical image video data obtained from Peking University BinHai Hospital with consent of anonymous patients. ORB-SLAM introduces the optical flow estimation network FlowNet and depth estimation network DepthNet. The system introduces FlowNet and DepthNet network model calculation comparisons. The average accuracy was calculated using the average MAE, MRE, SqRel, RMSE, log RMSE(the smaller the better), $\delta < 1.25$, $\delta < 1.25^2$ and $\delta < 1.25^3$ respectively. The quantitative results are presented in Table 4. The results show that the proposed method is complete and the reconstruction quality is high by introducing the optical flow network FlowNet and depth information extraction network DepthNet which are trained with uniform loss of scale and uniform loss of depth.

V. CONCLUSION

In this paper, a 3D reconstruction from laparoscope images method is presented, based on contrastive learning. It can produce a high-quality dense reconstruction in the surgical scene, establish medical school training programs based on existing models, and document postoperative case analysis. It has been validated and evaluated on abdominal tissue sequences and is robust to various illumination changes and different scene textures. The trained model was tested and validated by gastroscopy, sinus endoscopy, and UBE. Evaluations such as the stereo methods provide a similar measure of the accuracy in the laparoscope pose estimation on the surface. Compared with the same type of medical image 3d reconstruction method and the latest deep learning method, it has several advantages. The accuracy of the camera rotation and translation with respect to the estimated surface was tested mainly in this study. The model we designed model was proven to be superior to other methods in terms of reconstruction accuracy.

Although much progress has been made in depth estimation and 3D reconstruction of monocular laparoscopic images, there is still much work remains to be done in the future. Firstly, the use of monocular image depth estimation is limited, a big reason is that there is no good algorithm to deal with the scene matching problem of weak texture features. Most of these methods are limited to data from a single common dataset. In particular, for scenes with complex lighting and weak textures, the generalization ability of the algorithm must be improved simultaneously. In addition, we should consider improving the accuracy and speed of calculations. Secondly, local constraints are extracted according to the features of the pixels. Globally consistent constraints need to be optimized to obtain a ground-way fit or additional information as well as real data, resulting in increased network complexity and computation. Therefore, it is necessary to explore a more efficient deep learning neural network and a monocular endoscope suitable for clinical medicine to obtain a real-time dense depth map.

Based on our early feasibility data, 3D reconstruction of 2D images facilitates flexible examination and surgical navigation using the existing laparoscopic equipment. The proposed method can be applied to medical VR scenarios in the future to help digestive laparoscopists perform real laparoscopic interventions.

ACKNOWLEDGMENT

The author would like to thank Tianjin Wujing Hospital, Peking University BinHai Hospital, and Tianjin Hongqiao Hospital for providing medical guidance and helping them run their experiments. Finally, they sincerely thank the anonymous reviewers for their positive comments and helpful suggestions.

REFERENCES

[1] A. Hann, B. M. Walter, N. Mehlhase, and A. Meining, "Virtual reality in GI endoscopy: Intuitive zoom for improving diagnostics and training," *Gut*, vol. 68, no. 6, pp. 957–960, 2019.

[2] R. Garg, V. K. Bg, G. Carneiro, and I. Reid, "Unsupervised CNN for single view depth estimation: Geometry to the rescue," in *Computer Vision—ECCV 2016*. (Lecture Notes in Computer Science), vol. 9912. Cham, Switzerland: Springer, 2016, pp. 740–756.

[3] X. Liu, A. Sinha, M. Ishii, G. D. Hager, R. H. Taylor, and M. Unberath, "Self-supervised dense 3D reconstruction from monocular endoscopic video," 2019, *arXiv:1909.03101*.

[4] J. Engel, J. Stückler, and D. Cremers, "Large-scale direct SLAM with stereo cameras," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2015, pp. 1935–1942, doi: 10.1109/IROS.2015.7353631.

[5] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE Trans. Robot.*, vol. 31, no. 5, pp. 1147–1163, Oct. 2015, doi: 10.1109/TRO.2015.2463671.

[6] N. Mahmoud, T. Collins, A. Hostettler, L. Soler, C. Doignon, and J. M. M. Montiel, "Live tracking and dense reconstruction for handheld monocular endoscopy," *IEEE Trans. Med. Imag.*, vol. 38, no. 1, pp. 79–89, Jan. 2019, doi: 10.1109/TMI.2018.2856109.

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervent*, 2015, pp. 234–241.

[8] X. Liu, A. Sinha, M. Ishii, G. D. Hager, A. Reiter, R. H. Taylor, and M. Unberath, "Dense depth estimation in monocular endoscopy with self-supervised learning methods," *IEEE Trans. Med. Imag.*, vol. 39, no. 5, pp. 1438–1447, May 2020.

[9] Z. Lei, Y. Wang, Z. Li, and J. Yang, "Attention based multilayer feature fusion convolutional neural network for unsupervised monocular depth estimation," *Neurocomputing*, vol. 423, pp. 343–352, Jan. 2021.

[10] C. Godard, O. M. Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, Jul. 2017, pp. 270–279.

[11] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," 2015, *arXiv:1511.07289*.

[12] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6612–6619, doi: 10.1109/CVPR.2017.700.

[13] Q. Lv, H. Lin, G. Wang, H. Wei, and Y. Wang, "ORB-SLAM-based tracing and 3D reconstruction for robot using Kinect 2.0," in *Proc. 29th Chin. Control Decis. Conf. (CCDC)*, May 2017, pp. 3319–3324, doi: 10.1109/CCDC.2017.7979079.

[14] V. G. Turaev and O. Y. Viro, "State sum invariants of 3-manifolds and quantum 6j-symbols," *Topology*, vol. 31, no. 4, pp. 865–902, Oct. 1992.

[15] P. Mountney, D. Stoyanov, and G.-Z. Yang, "Three-dimensional tissue deformation recovery and tracking," *IEEE Signal Process. Mag.*, vol. 27, no. 4, pp. 14–24, Jul. 2010.

[16] D. Kim, J. Choi, D. Lee, H. Kim, J. Jung, M. Cho, and K.-Y. Lee, "Motion correction for routine X-ray lung CT imaging," *Sci. Rep.*, vol. 11, no. 1, p. 3695, Dec. 2021.

[17] K. P. Pruessmann, M. Weiger, M. B. Scheidegger, and P. Boesiger, "SENSE: Sensitivity encoding for fast MRI," *Magn. Reson. Med.*, vol. 42, no. 5, pp. 952–962, 1999.

[18] O. D. T. Catala, I. S. Igual, F. J. Perez-Benito, D. M. Escriva, V. O. Castello, R. Llobet, and J.-C. Perez-Cortes, "Bias analysis on public X-ray image datasets of pneumonia and COVID-19 patients," *IEEE Access*, vol. 9, pp. 42370–42383, 2021.

[19] S. M. Kamrul Hasan and C. A. Linte, "U-NetPlus: A modified encoder-decoder U-Net architecture for semantic and instance segmentation of surgical instruments from laparoscopic images," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 7205–7211.

[20] H. Jin and X. Y. Li, "Image saliency detection based residual network feature fusion tracking algorithm," *Laser Optoelectron. Prog.*, vol. 57, no. 18, pp. 257–262, 2020.

[21] G. Li, L. Yu, and S. Fei, "A deep-learning real-time visual SLAM system based on multi-task feature extraction network and self-supervised feature points," *Measurement*, vol. 168, Jan. 2021, Art. no. 108403.

NANNAN CHONG was born in Baoding, Hebei, China, in 1987. She received the M.S. degree in signal and information system from Nankai University, Tianjin, China, in 2013. She is currently pursuing the Ph.D. degree majored in electronic science and technology with the School of Electronic and Computer Engineering, Hebei University of Technology, Tianjin. She was a Lecturer with the School of Information and Communication Engineering, Tianjin Renai College. Her research interests include the fields of image processing, deep neural network applications, and big data analytics.