Unsupervised Learning-Based Continuous Depth and Motion Estimation with Monocular Endoscopy for Virtual Reality Minimally Invasive Surgery

Ling Li, Student Member, IEEE, Xiaojian Li*, Member, IEEE, Shulin Yang, Shuai Ding*, Member, IEEE, Alireza Jolfaei, Senior Member, IEEE, and Xi Zheng, Member, IEEE

Abstract—Three-dimensional display and virtual reality technology have been applied in minimally invasive surgery to provide doctors with a more immersive surgical experience. One of the most popular systems based on this technology is the Da Vinci surgical robot system. The key to build the virtual reality system is to estimate the depth of endoscopic images and continuous 3D spatial information has not been measured. On the one hand, the stereo endoscope is necessary, which is large and expensive. On the other hand, the view of a doctor is limited by the endoscope pose and cannot move the endoscope freely [7]. In order to provide doctors with a more realistic VR surgical experience, it is necessary to estimate the depth of endoscopic images and continuous 3D camera pose (a.k.a. ego-motion), so that the 3D model of the internal environment can be built [8].

At present, 3D reconstruction methods for in vivo measurement can be divided according to the endoscope type into three following types: monocular endoscope [9], stereo endoscope [10], and structured light endoscope [11]. Although stereo endoscopes and structured light endoscopes can obtain more spatial information, the volume, complexity, and cost of endoscopes are higher compared to the monocular endoscope. Therefore, the monocular endoscope is still the most commonly used endoscopes in clinical practice. The 3D estimation based on a monocular endoscope can be used to perform VR surgery without any change in the hardware of common endoscopic systems [9].

The 3D reconstruction of monocular endoscopes is commonly performed using the following techniques: Shape-from-Motion (SfM) [12], Simultaneous Localization and Mapping (SLAM) [13], and Shape-from-Shading (SfS) [14]. The SfM and SLAM use the images recorded during the endoscope movement at different times and directions as the input. Then, the image depth and camera motion are calculated based on the principles of feature matching and triangulation by the SfM and SLAM. The SfS is based on the optical model and uses the light-dark relationship between the observed pixel intensities to estimate pixel depth of an image. However, in feature matching, there is a large number of mismatches due to low texture features of the in vivo environment, and the accuracy of SfM and SLAM is severely affected. Reflective points, smoke, and chromatic aberration on the tissue surface affect the SfS accuracy, and the SfS cannot predict the camera motion [15].

To enhance the robustness of the existing algorithms in the in vivo environment, deep learning algorithms have been
recently used in the depth estimation of the monocular endoscope [8], [11], [16]. The deep learning-based depth estimation methods can well integrate the principles of the SfS, SLAM, and SFM, and construct a neural network model for in vivo depth estimation by adopting the end-to-end training. However, the existing neural network-based depth estimation algorithms consider only a single frame, while spatiotemporal continuity of endoscopic video is not taken into account [8]. Therefore, the depth estimation accuracy of the existing deep learning-based algorithms is unstable. Besides, although the deep learning methods have achieved good results in motion and depth estimation in the computer vision field [17], [18]. They are difficult to apply to the endoscopic image analysis due to the special in vivo environment. In a few studies, it has been shown that it is difficult to perform well on both depth and motion estimation simultaneously [19].

In this paper, an unsupervised learning-based method for depth and ego-motion estimation using the continuous monocular endoscope video is proposed, and two networks named EndoMotionNet and EndoDepthNet are developed to predict the ego-motion of an endoscopic camera and the depth of endoscopic image, respectively. The estimated depth of the previous frame is used to estimate the depth and motion of the next frame by the method based on a multi-mode fusion mechanism. By using the input consisted of continuous frames, the estimation accuracy of the proposed method is increased. Experiments with the public datasets verify that the proposed unsupervised learning-based continuous depth and motion estimation method can effectively improve the accuracy of depth and ego-motion estimation compared to existing methods, especially after processing continuous frames. The code is available online at https://github.com/cynerlee/tensorflow-endoscopy.

The main contributions of this paper are as follows:

1) By considering continuous frames in the proposed method with the characteristics of spatiotemporal correlation and geometric correlation, the accuracy of depth and motion estimation is improved.
2) A loss function is designed as a supervision metric for depth and ego-motion prediction, and the proposed method is based on unsupervised training.
3) EndoMotionNet and EndoDepthNet are proposed to predict ego-motion and depth of endoscopic images, respectively, which are the keys to build the in vivo VR model in the MIS.

The rest of this paper is organized as follows: The related work is discussed in Section II. The framework of the proposed method is presented in Section III. The detailed network architecture of ego-motion and depth estimation is shown in Section IV. Section V and Section VI provide the experimental results and summary, respectively.

II. RELATED WORK

A. Traditional Multi-View Stereo Methods for 3D Reconstruction

With the development of multi-view stereo technology, researchers have begun to explore the reconstruction of a surgical environment using multi-view stereo algorithms, such as SFM and SLAM. Leonard et al. implemented the SFM solution on the OpenMVG library to estimate 3D structure and camera motion [12]. Mahmoud et al. added a new dense reconstruction thread to the sparse ORB SLAM pipeline to accurately compute endoscope poses and perform 3D surgical scene reconstruction [15]. Song et al. proposed a real-time MIS-SLAM system by adopting efficient large-scale stereo to estimate depth on GPU while running ORB SLAM on CPU to provide stereoscope pose [20]. Chen et al. employed the moving least-squares smoothing to gather with the Poisson surface reconstruction framework to develop dense tumor surface on the basis of ORB SLAM [13]. Turan et al. proposed a direct vision based-SLAM method to perform pose estimation and map reconstruction for endoscopic capsule robots [21].

Although the multi-view stereo algorithms can simultaneously reconstruct depth and camera motion with the support of GPU in feature-rich in vivo scenes, the adjustment process is complicated, adaptability to different surgical scenes is poor, and manual feature extraction is difficult. Therefore, the existing multi-view stereo algorithms are difficult to implement in practical applications.

B. Deep Learning-Based Methods for Motion and Depth Prediction

While the traditional multi-view stereo methods extract motion and depth information based on geometric relationships between features, deep learning-based methods try to infer them from data. According to the availability of ground truth, the existing deep learning-based methods can be divided into two categories: supervised learning-based methods and unsupervised learning-based methods. Turan et al. designed a supervised learning-based visual odometry approach for endoscopic capsule robots named the Deep EndoVO [22]. Mahmoud et al. proposed a supervised learning-based endoscopy depth estimation method using convolutional neural networks and conditional random field [16]. Visentini-Scarzanella et al. used the CT data-free, untextured endoscopic video simulation to train a fully supervised learning-based depth estimation method and convert real endoscope video frames into depth maps by utilizing another transcoder network [23]. The unsupervised learning-based method named the Endo Odometry Learner designed by Turan et al. performed well in ego-motion estimation; however, the depth estimation performance was weak [19]. Li et al. presented a self-supervised learning-based dense depth estimation method for monocular endoscopy [8].

Due to the lack of a large amount of labeled data, supervised methods are difficult to apply during the process of endoscopic image analysis. Therefore, the deep learning methods are mostly used for depth estimation of a single endoscopic image, but they rarely involve the estimation of camera motion. Moreover, in unsupervised learning-based methods, the timing information of an endoscope video is not considered.

III. PROPOSED METHOD

A. Overview

Endoscopy positioning and depth estimation are essential for constructing an internal VR model, which can provide
surgeons or surgical robots with information on the internal environment and thus ensure successful processing in the MIS. Due to the spatiotemporal correlation between continuous frames, especially geometric correlation, considering continuous frames when performing the ego-motion estimation and depth estimation can improve both stability and accuracy of the existing methods. Since there is almost no camera motion and soft tissue deformation between adjacent endoscopic frames, there will be a small change in depth between adjacent frames. The depth information of the previous frame can be used as a priori information for obtaining the depth information of the next frame.

In this paper, an unsupervised learning-based continuous depth and motion estimation method with monocular endoscopy is proposed, and it is shown in Fig. 1. In order to estimate camera ego-motion and depth, two networks are developed, EndoMotionNet and EndoDepthNet. The networks are jointly trained using an unlabeled video sequence consisting of three frames. The training process consists of the following steps:

1) For each frame in the video sequence, highlighted points are identified based on HIS (Hue, Saturation, Intensity) color space, and masks are created according to the value of saturation and intensity [24].
2) The ego-motion of the camera is obtained by the EndoMotionNet, taking the video sequence and the corresponding masks as network input.
3) The depth image of the frame $i$ is obtained by EndoDepthNet, using the depth image of the frame $(i-1)$, along with the corresponding mask of the frame $(i-1)$ as input. If $i = 1$, frame $i$ is used as the EndoDepthNet input.
4) The unsupervised dual networks are trained by the backpropagation algorithm using the custom loss function.

The supervised learning-based methods require the ground truth of ego-motion and depth. In contrast, the proposed method adopts view synthesis [25] for the supervision, and considers the time-series information from continuous frames as well as the endoscopic camera internal parameters. Although the EndoMotionNet and EndoDepthNet are jointly trained, they are independently used during the test.

B. Highlighted Region Detection

In the endoscopy inspection process, light causes reflections on the skin surface due to the existence of interstitial fluid and water. The highlighted points provide key information about the image and affect the stereoscopic tasks; therefore, the examination of the highlighted points is fundamental for constructing an endoscopy VR model.

Since highlighted regions on a reflective surface usually have low saturation but high intensity, the endoscopic images are transformed from the RGB color space into the HIS color space firstly to obtain saturation and intensity of each pixel in the endoscopic images. An image pixel in the HIS color space is defined as a highlighted pixel of the reflective surface if its saturation value is less than the threshold $\alpha_1$, and its intensity value is larger than the threshold $\alpha_2$. The highlighted region of the reflective surface is obtained by extending each highlighted pixel to its five surrounding pixels. Finally, the masks are created using the highlighted region of the reflective surface, as shown in Fig. 1. The white region with 255 as the image pixel denotes the highlighted region of the reflective surface, and the black region with 0 as the image pixel represents non-reflective area.

IV. Network Architecture

A. EndoMotionNet

The EndoMotionNet network is designed in order to obtain ego-motion information of video sequences. This network is based on improved SfmLearner network architecture [18], which has been proved to be effective in endoscopy visual odometry prediction [19]. However, the SfmLearner ignores the relevance between continuous video frames and does not consider the influence of highlighted regions of reflective surfaces on the position estimation result. In order to overcome interference from highlighted regions of reflective surfaces, the input to EndoMotionNet is processed as:

$$p_i = p_i(1 - \frac{m_i}{255}),$$ (1)

where $p_i$ represents the pixel value of point $i$ in the endoscopic image, and $m_i$ represents the pixel value of point $i$ in the corresponding mask image of the endoscopic image. In addition, the long short-term memory (LSTM) layer is adopted to learn the spatial and timing characteristics of endoscopic video more completely, which can improve the performance of ego-motion estimation.

The architecture of EndoMotionNet is shown in Fig. 2, where it can be seen that it consists of convolutional layers, flattened layers, an LSTM layer, and fully connected layers. As shown in Fig. 2, seven convolutional layers sequences are connected in order to extract features of an endoscopic image. The convolutional sequence weights are shared between the three groups of convolutional layer sequences and used to extract features of each endoscopic frame. The RELU function is used as the activation function of all convolutional layers and the first fully connected layer. The activation function of the last layer is expressed as $y = \rho \times (\text{sigmoid}(x) - 0.5)$, where $\rho$ is used to limit the output value range. The first two convolution kernel sizes are $7 \times 7$ and $5 \times 5$, and the other convolution kernels have a size of $3 \times 3$. Once the features of endoscopic images are completely extracted, they are processed as a one-dimensional vector by the flattened layer. The time-series features of endoscopic video sequences are extracted by adopting the LSTM layer. Finally, the two fully connected layers are connected to obtain the ego-motion vectors of a camera.

B. EndoDepthNet

Due to the continuity of video sequences and small camera motion between neighboring frames, the depth information between neighboring frames has high relevance. Due to the existence of a highlighted region on a reflective surface, the depth information of any previous frame with a non-highlighted region in the endoscopic image has quite high
reliability, while the current frame with a highlighted region has low reliability. Assume that for frame \( j \), the predicted depth image is \( \hat{D}_j \). Based on a difference in reliabilities, pixel \( \hat{d}_i \) in the predicted depth image \( \hat{D}_i \) can be expressed as:

\[
\hat{d}_i = \hat{d}_i(1 - \frac{m_j}{255}),
\]

where \( \hat{d}_i \) is the predicted depth value of point \( i \) in the endoscopic image, and the processed depth image can be denoted as \( \hat{D}_j \). The processed image depth of the previous image frame is \( \hat{D}'_{j-1} \) and current frame \( j \) are input into the EndoDepthNet. In the EndoDepthNet, the DispNet codec network architecture is adapted to output the depth image of the current frame. The processed predicted depth image \( \hat{D}'_{j-1} \) and current frame \( j \) are input into the EndoDepthNet simultaneously via different channels, and the dual-channel weights are shared. For the first frame of each video sequence, the single-channel mode is used, and only the current frame is input into the EndoDepthNet. All convolutional layers have ReLU activation function except for the prediction layer, whose activation function is given by:

\[
y = \begin{cases} 
1/ (\alpha \times \text{sigmoid}(x) + \beta) & j = 1 \\
0.5/ (\alpha \times \text{sigmoid}(x) + \beta) + 0.5/ \hat{D}'_{j-1} & j \neq 1
\end{cases},
\]

where \( \alpha = 1/d_{min} - 1/d_{max} \), and \( \beta = 1/d_{max} \); \( d_{min} \) and \( d_{max} \) denote the shortest and longest depth relative to the endoscopic camera. The activation function of the prediction layer constrains the predicted depth to a positive value within a reasonable range. Further, \( 1/y \) denotes the predicted depth value of frame \( i \).

C. Loss Function

Using the known 3D geometrical relationship of different views, a new view can be generated from the neighboring view. Therefore, view synthesis is used as a supervision metric. The camera model is given by:

\[
Z \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \triangleq KP,
\]

where \( P = [X \ Y \ Z]^T \) denotes the coordinate set of a point in the camera coordinate system, \([u \ v]^T\) is the coordinate set of the point projected onto the image coordinate system; \( f_x, f_y, c_x, c_y \) are the internal camera parameters, and \( K \) is the camera intrinsic matrix. Suppose a point \( P \) in the world coordinate system converted to the target view’s camera coordinate system is \( P_t \), and converted to the original view’s camera coordinate system is \( P_s \), then \( P_s = T_{t \rightarrow s}P_t \). Assuming that \( P_t \) is the projection of point \( P \) onto the target view and \( P_s \) is the projection of point \( P \) onto the original view, then according to the camera model, it holds that:

\[
d_t \begin{bmatrix} p_t \\ 1 \end{bmatrix} = KP_t, \quad \text{and} \quad d_s \begin{bmatrix} p_s \\ 1 \end{bmatrix} = KP_s = KT_{t \rightarrow s}P_t. \quad \text{Thus,} \quad \begin{bmatrix} p_t \\ 1 \end{bmatrix} = KT_{t \rightarrow s}d_s^{-1}K^{-1} \begin{bmatrix} p_t \\ 1 \end{bmatrix}.
\]

A video sequence composed of three frames can be expressed as \(< F_{i-1}, F_i, F_{i+1} >\). Taking \( F_i \) as a target
The smoothness loss function is defined as:

\[ L_1 = \sum_{p_i} \left( |F_i(p_i) - \hat{F}_{i-1}(p_i)| + \sum_{p_q \in \text{neighbors of } p_i} |F_i(p_i) - \hat{F}_{i+1}(p_i)| \right) \]  

where \( F_i(p_i) \) denotes the pixel value of point \( p_i \) on \( F_i \), and \( \hat{F}_{i-1} \) and \( \hat{F}_{i+1} \) respectively denote \( F_{i-1} \) and \( F_{i+1} \) converted to the target frame using the estimated depth image \( D_t \) and the camera ego-motion transformation matrix \( T_{t \rightarrow s} \); \( \hat{F}_{i-1} \) and \( \hat{F}_{i+1} \) are computed by a bilinear interpolation method as follows:

\[ \hat{F}_{i-1}(p_i) = F_{i-1}(p_s) = \sum_{p_q \in \text{neighbors of } p_s} w F_{i-1}(p_q), \]  

\[ \hat{F}_{i+1}(p_i) = F_{i+1}(p_s) = \sum_{p_q \in \text{neighbors of } p_s} w F_{i+1}(p_q), \]

and \( w = (1 - |u - u_s|)(1 - |v - v_s|) \), and \( p_s = [u, v, v] \) denotes the four-pixel neighbors (top-left, top-right, bottom-left, and bottom-right) of \( p_s \).

Additionally, due to the existence of highlighted regions on a reflective surface, there is efficacy loss in the view synthesis in that region. Assuming \( M_i \) is the mask of \( F_i \), in order to overcome the influence by highlighted regions on reflective surfaces, \( L_1 \) is redefined as a new loss function \( L_v \):

\[ L_v = \sum_{p_i} \left( (1 - M_{i-1}(p_i)) |F_i(p_i) - \hat{F}_{i-1}(p_i)| + (1 - M_{i+1}(p_i)) |F_i(p_i) - \hat{F}_{i+1}(p_i)| \right), \]

where \( M_{i-1}(p_i) \) and \( M_{i+1}(p_i) \) denote the results of \( M_{i-1} \) and \( M_{i+1} \) converted to \( M_i \) corresponding to the original views \( F_{i-1} \) and \( F_{i+1} \), respectively, using the method given by (5).

The Peak-Signal-to-Noise Ratio (PSNR) is a measure that can be used to evaluate estimation accuracy based on the similarities between synthesized images and target images and is particularly effective for endoscopic images [26]. Inspired by the study of Godard et al. [27], PSNR is considered as an optimization objective during training. The larger the value of PSNR is, the higher the image similarity is. Therefore, a new similarity loss function \( L_p \) is defined as:

\[ L_p = 1 - \frac{10}{\lambda} \log_{10} \frac{\sum_{p_i} |F_i(p_i) - \hat{F}_{i-1}(p_i)|^2}{\varepsilon^2}, \]

where \( \lambda \) denotes the set threshold, and \( w \) and \( h \) are the width and height of the endoscopic image, respectively.

Since \( \hat{F}_{i-1} \) and other parameters are obtained by solving the pixel values of neighbors of the point \( p_s \) in the original view, the depth images will lack smoothness. Therefore, a smoothness loss function is required, and in this paper, L1 regularization is adopted to smooth the depth information.

The smoothness loss function is defined as:

\[ L_s = \sum_{p_i} \left( |\partial_x D(p_i)| e^{-\|\partial_x F_i(p_i)\|} + |\partial_y D(p_i)| e^{-\|\partial_y F_i(p_i)\|} \right). \]

In summary, the final loss function \( L \) of the complete process is expressed as:

\[ L = k_1 L_v + k_2 L_p + k_3 L_s, \]

where \( k_1 \), \( k_2 \), and \( k_3 \) represent the weights of the three sub-loss functions.

V. EXPERIMENTS

A. Datasets and Processing

The experiments were performed using the Stereo Correspondence and Reconstruction of Endoscopic Data available on the following link: https://endoivissub2019-scared.grand-challenge.org/. The datasets contain images of fresh porcine cadaver abdominal anatomy collected by a Da Vinci Xi endoscopy and a projector. Three sub-datasets from different pigs were used to train and verify the performance of the proposed method. Each sub-dataset consists of five keyframes, and each keyframe contains camera calibration, a sequence of video frames, a camera transformation matrix relative to the initial camera position, and point clouds as seen by the left and right cameras. Although the datasets are binocular, they were regarded and used as two monocular datasets and adopted to train and test the depth estimation and visual odometer based on monocular vision.

Binocular video frames containing \( 2048 \times 1280 \) pixels were classified into a left monocular video (with \( 1024 \times 1280 \) pixels) and a right monocular video (with \( 1024 \times 1280 \) pixels). OpenCV was used to read and combine the video frames for each monocular video. A picture was formed by combining three image frames containing \( 1024 \times 3840 \) pixels, where the second image was regarded as a target image, and the first and third images were regarded as source images. Next, the highlighted regions of the image were extracted to create the masks. The point clouds corresponding to each frame image were saved in the tiff format so that the z-axis coordinates could be extracted and reshaped into a matrix of the same dimensions as the frame. The depth value was expressed by pixels to obtain the depth image. An original left endoscopic frame, the corresponding mask map, and the corresponding depth image are displayed in Fig. 3. In the depth image, the black areas indicated areas with unknown depth information, while white pixels in the mask map represented specular highlighted regions.

With the aim to improve the robustness of the proposed unsupervised learning-based continuous depth and motion estimation method and avoid overfitting when training, a data augmentation strategy was adopted to expand the data volume and increase the diversity of the training data [28]. The camera parameters were also modified during data expansion using random scaling and cutting.

B. Performance Evaluation

The performance of the proposed unsupervised learning-based continuous depth and motion estimation method was evaluated using multiple indicators to assess the tasks of ego-motion and depth estimation. Absolute trajectory error was an indicator adopted to evaluate the accuracy of camera trajectory prediction, including the absolute translational error (ATE) and the absolute rotational error (ARE). Other evaluation indicators included the absolute relative error (AbsRel), the square relative error (SqRel), the root mean squared error (RMSE),
the root mean square logarithmic error \((\log RMSE)\), and accuracy, and they were adopted to evaluate the accuracy of the depth estimation, using the following formulas:

\[
ATE = \frac{1}{l-1} \sum_{j=1}^{l-1} \sqrt{(\hat{x}_j - x_j)^2 + (\hat{y}_j - y_j)^2 + (\hat{z}_j - z_j)^2},
\]

\[
ARE = \frac{1}{l-1} \sum_{j=1}^{l-1} \sqrt{(\hat{\varphi}_j - \varphi_j)^2 + (\hat{\psi}_j - \psi_j)^2},
\]

\[
AbsRel = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{\hat{d}_i - d_i}{d_i} \right|,
\]

\[
SqRel = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{\hat{d}_i - d_i}{d_i} \right|^2,
\]

\[
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} |\hat{d}_i - d_i|^2},
\]

\[
\log RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \log |\hat{d}_i - d_i|^2},
\]

\[
Accuracy = \frac{N(\max\{\frac{\hat{d}_i}{d_i}, \frac{d_i}{\hat{d}_i}\} < \delta)}{n},
\]

where \(l\) denotes the frames of an endoscopic video, \(j\) denotes the sequence number consisting of two frames of endoscopic video; \(\hat{x}_j, \hat{y}_j, \) and \(\hat{z}_j\) denote the positions of the endoscope camera in the second frame relative to the first frame predicted by the proposed method; \(\hat{\varphi}_j, \hat{\psi}_j\) denote the directions of the endoscope camera in the second frame relative to the first frame predicted by the proposed method; \(x_j, y_j, z_j, \varphi_j, \) and \(\psi_j\) represent the ground truth of the corresponding variables; \(n\) denotes the number of pixels, \(\hat{d}_i\) represents the predicted depth of pixel \(i\), \(d_i\) is the corresponding real depth; \(\delta\) denotes a threshold and three different thresholds \((1.25, 1.25^2, 1.25^3)\) are used in this work \([8], [18], [29]\); lastly, \(N(\max\{\frac{\hat{d}_i}{d_i}, \frac{d_i}{\hat{d}_i}\} < \delta)\) denotes the number of pixels where \(\max\{\frac{\hat{d}_i}{d_i}, \frac{d_i}{\hat{d}_i}\} < \delta\).

### C. Experimental Setup

The EndoMotionNet and EndoDepthNet were trained in an end-to-end manner using Ubuntu 16.04 LTS operating system, and the training workstation is with an Intel Xeon Gold 5218 16 Core 2.3 GHz 32 Threads CPU, 10 NVIDIA RTX 2080TI GPU, and 256 GB of RAM. After data preprocessing, the EndoMotionNet and EndoDepthNet were trained, compiled, and run using a Python TensorFlow framework.

The EndoMotionNet and EndoDepthNet were jointly trained, setting the batch size of each forward input image sequence to four, with a total of 12 original images and 12 mask images. The training included 2000 epochs and the learning rate was updated as the epoch number increased. For the first 200 epochs, the learning rate was 0.001; for the epochs from 200 to 1000, the learning rate was 0.0005; after the 1000th epoch, the learning rate was 0.0002. In this way, the method was prevented from falling into a local optimum. Additionally, Adam optimizer was adopted for backpropagation training, using the following parameters: \(\beta_1 = 0.9, \beta_2 = 0.999, \alpha_1 = 25, \alpha_2 = 200, \rho = 0.4, \lambda = 20, d_{max} = 150, d_{min} = 40, k_1 = 1, k_2 = 0.3, k_3 = 0.5\).

### D. Experimental Results

1) Ego-motion estimation: The performance of the ego-motion estimation was evaluated by comparing the proposed method to three mainstream algorithms: ORB SLAM2 [30], Deep EndoVO [22], and Endo Odometry Learner [19]. Two sections of video sequences were randomly selected for the test. Fig. 4 shows the camera trajectory image obtained by the above methods, where the black, red, green, blue, and yellow curves denote the ground truth, the predicted trajectory images obtained by the proposed method, ORB SLAM2, Deep EndoVO, and Endo Odometry Learner, respectively. As presented in Fig. 4, when predicting trajectory 1, the translational vector predicted by the proposed method on Z-axis is a litter higher than ground truth in some continuous frames, while the translational vector predicted by Deep EndoVO on Z-axis is a litter smaller than ground truth. Although it seems that the prediction by EndoMotionNet is above the ground truth in trajectory 1, the error is small. When predicting both trajectory 1 and trajectory 2, the trajectories obtained by all
the methods except for the ORB SLAM2 method basically followed the ground truth. The EndoMotionNet performed better than the Endo Odometry Learner in odometry estimation and the performance is second only to the supervised learning-based method Deep EndoVO.

The $ATE$ and $ARE$ of the EndoMotionNet and competing methods are given in Table I, where it can be seen that our unsupervised odometry outperformed the ORB SLAM2 and Endo Odometry Learner in both rotational and translational motion estimations. In addition, the proposed EndoMotionNet achieved the same performance as the Deep EndoVO.

2) Depth estimation: In the depth estimation of endoscopic image, the proposed method was compared to SfmLearner [18] proposed by Zhou et al. and GeoNet [29] proposed by Yin et al. The performance comparison results are provided in Table II, where it can be seen that the proposed EndoDepthNet was superior to the competing methods regarding all performances. Since the SfmLearner and GeoNet could process only a single frame during depth estimation, it has been proven that processing continuous frames could improve the performance of depth estimation.

The 3D displays of an endoscopic image obtained by different methods and the corresponding ground truth are presented in Fig. 5. In the 3D display of the ground truth, the depth of invalid pixels was replaced with the mean depth of adjacent pixels. The 3D display predicted by SfmLearner is the worst, while the 3D displays predicted by GeoNet and EndoDepthNet can show the 3D effect. The prediction result of EndoDepthNet is better than GeoNet according to the ground truth. Fig. 5 intuitively shows that the EndoDepthNet was superior to the competing methods.

VI. CONCLUSION

This work presents an unsupervised learning-based continuous depth and motion estimation method with monocular endoscopy. The EndoMotionNet and EndoDepthNet are used to predict the ego-motion of an endoscopic camera and the depth of an endoscopic image, respectively. First, to improve the accuracy of the proposed unsupervised learning-based continuous depth and motion estimation method, the specular highlight regions are detected. Next, the LSTM layers are used in the EndoMotionNet to consider the timing of an endoscopic video and enhance the connection between video frames. The estimated depth information of the previous frame is used to estimate the depth of the next frame using a multi-mode fusion mechanism. Lastly, a loss function is introduced as a supervision metric. The experiments with public datasets demonstrate that the proposed unsupervised learning-based continuous depth and motion estimation method can effectively improve the accuracy of depth and motion estimation.

This study strengthens the existing research on VR technology applications in the MIS. The research presented in this work can be extended to many fields, such as brain, urology, and orthopedics, by integrating it into VR scenes, and thus contributing to the MIS development.
REFERENCES


**Ling Li** received her B.S. degree in Electronic Commerce from Hefei University of Technology in 2017. She is working toward the Ph.D. degree in Management Science and Engineering at the School of Management, Hefei University of Technology. Her research focuses on surgical robotics and medical imaging. She won the honor of “Person of the Year for Graduate Students of Hefei University of Technology” in 2019. She has attended in many international conferences such as the IEEE/ASME International Conference on Advanced Intelligent Mechatronics and the IEEE/RSJ International Conference on Intelligent Robots and Systems, where she gave speeches about her research fields. Moreover, she won the Award of Excellent Paper in the 20th China Annual Conference of Management Science. She showed an ardent attitude towards the Innovation and Entrepreneurship competition and led her team members to win the silver Award of the fifth “Internet +” National Competition.

**Xiaojuan Li** is currently a professor of the School of Management, Hefei University of Technology, China. He received his B.S. degree in Automation from Southeast University in 2011, and his Ph.D. degree jointly from the Department of Automation, the University of Science and Technology of China, and the Department of Mechanical and Biomedical Engineering, the City University of Hong Kong in 2017. His research interests include robot-aided *in vivo* cell manipulation, surgical robotics, and *in vivo* SLAM. He has published many papers in top journals, such as Science Robotics, Automatica, IEEE Transactions on Robotics. He has served as the chairman of many conferences, such as the 2019 IEEE/ASME International Conference on Advanced Intelligent Mechatronics, the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems, and the 2019 International Conference on Cybernetics.

**Shanlin Yang** is a member of the Chinese Academy of Engineering and the leading Professor in management science and information system at the School of Management, Hefei University of Technology. He is the director of the academic board of Hefei University of Technology, and the director of National-Local Joint Engineering Research Center of “Intelligent Decision and Information System”. He has won 2-second-class prizes for State Scientific and Technological Progress Award, and 6 first-class prizes for provincial and ministerial level science and technology award. His research interests include surgical robotics, information systems and artificial intelligence. He has published more than 300 academic papers in top journals and international conferences at home and abroad.

**Shuai Ding** is a professor of information systems at the School of Management, Hefei University of Technology, China. He has been a visiting scholar at the University of Pittsburgh. His research interests include data mining, knowledge discovery and their applications in clinical decision support. He has published papers in refereed journals and conference proceedings, such as IEEE Transactions on Knowledge and Data Engineering, Transactions on Fuzzy Systems, Decision Support Systems, and Physics of Life Reviews. He is a member of the China Medical Equipment Journal Life Support Editorial Board, a member of the China Systems Engineering Society Service System Engineering Branch and the Intelligent Manufacturing Professional Committee. He is a reviewer of international journals such as Decision Support Systems, IEEE Transactions on Cybernetics, IEEE Transactions on Industrial Informatics, IEEE Systems Journal, Future Generation Computer Systems, Applied Mathematics and Computation, Knowledge-Based Systems.

**Alireza Jafarzadeh** received the Ph.D. degree in Applied Cryptography from Griffith University, Gold Coast, Australia. He is the Program Leader of Master of IT in Cyber Security in the Department of Computing at Macquarie University, Sydney, Australia. His current research areas include Cyber and Cyber-Physical Systems Security. He has authored over 100 peer-reviewed articles on topics related to cybersecurity and received multiple awards for Academic Excellence, University Contribution, and Inclusion and Diversity Support. He received the prestigious IEEE Australian council award and a recognition diploma with a cash award from the IEEE Industrial Electronics Society. He is a founding member of the Federation University IEEE Student Branch. He served as the Chairman of the Computational Intelligence Society in the IEEE Victoria Section and also as the Chairman of Professional and Career Activities for the IEEE Queensland Section.

**Xi Zheng** holds a PhD in Software Engineering from the University of Texas at Austin, USA, and a Chief Solution Architect for Menulog Australia. Now he is the director of the Intelligent Systems research center (itseg.org), the deputy director of software engineering, global engagement, and assistant professor in Software Engineering at Macquarie University. He is specialized in Service Computing, IoT Security and Reliability Analysis. He published more than 80 high-quality publications in top journals and conferences. He has awarded multiple best papers in leading conferences and a few highly competitive funding, including ARC LP and Data61 CRP as the Lead Investigator.