Self-supervised Dense Depth Prediction in Monocular Endoscope Video for 3D Liver Surface Reconstruction

Zhengtao Cao¹, Wenfeng Huang², Xiangyun Liao*, Xuesong Deng*, Qiong Wang²

¹ School of Electrical Engineering University of South China Hengyang, Hengyang, Hunan, 421000, China
² Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Guangdong, 518055, China
³ Department of Hepatobiliary Surgery, The First Affiliated Hospital of Shenzhen University, Health Science Center; Shenzhen Second People’s Hospital, Shenzhen, Guangdong, China

Corresponding author’s e-mail: xy.liao@siat.ac.cn and dengxuesong@21cn.com

Abstract. In this paper, we propose a self-supervised monocular depth prediction method which combines traditional multi-view stereo method and fully convolutional network to predict the depth map in monocular endoscopic video and achieve 3D dense reconstruction of liver surface. We adopt the sparse data generated by COLMAP supervision signal as the training data, and integrate the attention model in the fully convolutional network to effectively extract the channel features to improve the accuracy of depth prediction. Taking into account the problem of insufficient supervision ability of sparse data, the projection transformation of two images within a certain range is carried out to make up for the missing supervision points. Experimental results show that this method has achieved good results in the depth prediction of monocular endoscopic video and has good applicability to the whole liver.

1. Introduction

The key of endoscopic surgery navigation is to track the position of the endoscope and obtain the depth information of the endoscopic images. At present, the depth prediction methods in the endoscopic scene reconstruction can be divided into the traditional monocular vision algorithms and deep learning-based methods. The monocular stereo vision algorithms, such as Simultaneous Localization And Mapping (SLAM)[1] and Structure from Motion(SfM)[2], have been widely concerned and studied as they do not require additional adaptation of other hardware devices and are low in cost[3]. Moreover, tracking and 3D reconstruction have been realized in various endoscopic scenes such as gastroscopy and laparoscopy[4][5]. However, the lack of features in endoscopic images may lead to sparsely and unevenly distributed reconstructions.

In recent years, as the Convolutional Neural Networks (CNNs) can automatically and efficiently extract images features, researchers adopted the CNNs in depth prediction. Fully-supervised learning methods have achieved good results in general scene depth prediction, but it is very difficult to use CNNs in a fully-supervised form in endoscopic videos for the reason that the true depth map labels are difficult to obtained in the human body due to the size of the depth acquisition device. As the supervised learning methods are limited by depth labels, some researchers have adopted the...
unsupervised learning methods to predict depth map [6]. However, their experiments have found that these methods are not suitable for endoscopy images, as the cold light source and the camera are constantly moving, which cannot satisfy the constancy of the light source. Therefore, for the same human body organ, the prediction results may be different due to the different lighting and camera postures. To solve this problem, Liu et al. proposed a self-supervised learning method to predict the depth in monocular endoscopy [7]. On one hand, traditional multi-view stereo methods, such as SLAM and SFM, have illumination invariant characteristics and are better optimized for global depth scale prediction. On the other hand, self-supervised methods can solve the problem of lacking true depth labels. Therefore, Liu et al. used the SFM method to extract sparse data as supervisory signals to train the convolutional neural network, and achieved good results in the nasal video experiment. However, it is difficult for Liu's method to acquire accurate features due to the high reflectivity and sparse texture of liver surface, which results in inaccurate depth prediction.

To achieve accurate depth prediction for liver monocular endoscope video, in this paper we propose a new self-supervised dense depth prediction method and reconstruct 3D dense liver surface. The main contributions of this paper are as follows:

1) The COLMAP algorithm is used to obtain sparse reconstruction data, including reconstructing point cloud and camera pose, and performing projection transformation on the reconstructed point cloud to generate a sparse depth map.

2) We propose the SE-FCDenseNet, which integrates Squeeze-and-Excitation Block (SE Block)[8] into the FCDenstNet network[9]. SE Block is an attention model, which can improve the ability of extracting features of the smooth and sparse liver texture.

3) Since the sparse depth map can only provide limited supervision signals, this paper proposes a coordinate projection transformation method based on the geometric constraint principle of two images with sufficient overlap in a certain range, and mutual conversion of depth maps in different coordinate systems to compensate for supervision. The problem of insufficient data.

2. Methods

In this section, we first introduce the data pre-processing process based on COLMAP and the supervised data obtained after pre-processing. The supervised data are sparse depth maps and camera poses. The pre-processing data involved in the training are endoscopic images, camera intrinsics, sparse depth maps and camera poses. Then, we introduce the overall architecture of our method, including the SE-FCDenseNet, scaled depth layer, Coordinate projection transformation layer. Finally, we present the proposed loss functions.

2.1. Training Data

COLMAP is an automatic reconstruction tool based on the SFM method. It only needs to input the images and camera intrinsics, select the camera model and matching method to automatically generate a sparse reconstruction map. The data pre-processing pipeline is shown in figure 1. First, the endoscopic video frames are sent to COLMAP for feature extraction and matching. Then perform sparse reconstruction of the known camera intrinsics and matching feature points, and obtain the camera pose and reconstruct the point cloud. After that, perform projection mapping on the reconstructed 3D points to obtain a sparse depth maps.

Figure 1. Training data generation pipeline.
Sparse Depth Map. The sparse depth map is obtained by 2D projection of the sparse point clouds. We use $K$ to denote the camera intrinsics matrix, $j$ to denote the frame index of the same video sequence, and $n$ to denote the index of the 3D point in the sparse reconstruction. The camera pose of frame $j$ with respect to the world coordinate is $T_j^w$, where $w$ stands for world coordinate system. The homogeneous coordinate of $n^{th}$ 3D point of the sparse reconstruction in the world coordinate is $P^w_n$.

The coordinate of $n^{th}$ 3D point relative to the frame $j$, $P_j^j$, is

$$P_j^j = KT_j^w P^w_n$$

(1)

The depth of $n^{th}$ 3D point relative to the frame $j$, $D_j^j$, is the z-axis component of $P_j^j$. The 2D projection location of $n^{th}$ 3D point relative to the frame $j$, $(u_n^j, v_n^j)$, is

$$(u_n^j, v_n^j) = \frac{P_n^j}{D_n^j}$$

(2)

Because COLMAP does not use all video frames when triangulating one particular 3D point, we only project the 3D point onto the relevant image planes. $b_{n,j} = 1$ indicates that frame $j$ is used to reconstruct the $n^{th}$ 3D point, and $b_{n,j} = 0$ indicates that the reconstruction of the $n^{th}$ 3D point has nothing to do with frame $j$. The sparse depth map of frame $j$ is represented by $Y^*_j$, as shown in Equation (3). Since the reconstruction is sparse, there are invalid depth values in $Y^*_j$. For invalid depth values, we set it to 0.

$$Y^*_j = \begin{cases} D_j^j W_j & \text{if } b_{n,j} = 1 \\ 0 & \text{if } b_{n,j} = 0 \end{cases}$$

(3)

$D_j^j$ is the depth value of the 2D projection position of the $n^{th}$ 3D point in frame $j$. $W_j$ is a weight related to the number of frames used to reconstruct 3D point $n$ and the accumulated parallax of the projected 2D locations of this point in these frames. $D_j^j$ can be expressed as:

$$W_j = 1 - e^{-\sum b_{n,j}^w}$$

(4)

$\alpha$ represents the average number of frames used to reconstruct the 3D point, $w$ represents the number of frames used to reconstruct the nth 3D point.

2.2. Network Architecture

The overall network architecture is shown in the figure 2, it is a self-supervised dual-branch twin network during the training phase. Input data includes video sequence frames, camera poses, camera intrinsics, sparse depth map. First, two frames with sufficient overlap are selected from the same video sequence to enter two identical branch networks for training, which we denote as frame $j$ and frame $k$.

The convolutional neural network involved in training is the SE-FCDenseNet network. There is a weight sharing mechanism between the two branches of the network, which can not only reduce the amount of network parameters, but also strengthen the correlation between the two frames of images. Then, we use the sparse depth map as an anchor to scale the prediction depth map to obtain the scaled depth map. The sparse depth map and the scaled depth map are used to construct an effective depth loss to guide the generation of an accurate scaled depth map. Since there are some invalid depth regions in the sparse depth map, considering two sufficiently overlapping frames $j$ and frame $k$, one of which is obtained by moving the camera of the other frame, the depth map predicted by the two frames of image separately has relevant. Perform coordinate transformation between the scaled depth map of frame $j$ and the scaled depth map of frame $k$ to generate the depth map $Y^*_{j,k}$ under the frame $k$. 
coordinate system and the depth map $Y_{k,j}$ under the frame $j$ coordinate system. Finally, we use the scaled depth map and the depth map under coordinate changes. Construct a depth difference loss function to optimize the scaled depth map. Finally, we use the scaled depth map and the depth map under coordinate changes to construct the depth difference loss function to optimize the scaled depth map. The predicted depth map of image frame $j$ is denoted as $Y_j^\wedge$, and the scaled depth map is denoted as $Y_j$.

Figure 2. Network architecture of the proposed method.

**Depth Convolutional Network.** The FCDenseNet network uses the output of all previous layers as the input of the current layer, which realizes the reuse of features, reduces the amount of network parameters, and also alleviates the problem of gradient disappearance. SE Block is to focus on the relationship between feature channels and learn the importance of different channel features. We combine the advantages of the two to propose the SE-FCDenseNet network, which can not only reuse the feature maps of the multilayer network, but also assign weights to the channels according to the information of the channel features, and realize the dynamic use of images features. As shown in the figure 3, the left half of the figure is the overall convolutional network structure, and the right half of the figure is the SE Block. We used the FC-DenseNet57-layer architecture and added SE Blocks in the TD and TU phases. In the feature dimension squeeze stage, we set the feature dimension to be compressed to $1/4$ of the original. In order to make the network output suitable for the depth prediction task, in the up-sampling stage, the number of channels in the last convolutional layer is changed to 1, and the activation function is changed to a linear activation function.
Scaling Depth Layer. The depth map predicted by the depth convolutional network and the sparse depth map are not in the same depth scale. Therefore, we use the sparse depth map as the standard to perform the depth scaling transformation to obtain the scaled depth map, which is also the final dense depth map that we want to get. The scaled depth map prepares for the subsequent calculation of effective depth loss and coordinate transformation.

Coordinate Transformation Layer. The sparse depth map can only provide sparse supervision signals, and the depth information cannot be accurately predicted for invalid regions that are not supervised. Two sufficiently overlapped frames $j$ and $k$, one of which is obtained by moving the camera in the other frame, so the depth maps predicted by the two frames are related. The method of Zhang et al. [7] is to generate camera poses through a camera pose estimation CNN, while this article uses COLMAP reconstruction to generate camera poses. Through the depth map of the image frame $j$, the camera internal parameters and the camera pose, the 3D point cloud of the image frame $j$ can be obtained, and then the 3D point cloud projection is mapped to the image frame $k$, and then bilinear interpolation is performed to obtain the corresponding position Pixel value, a depth map $^Y_{j,k}$ transformed to the frame $k$ coordinate system can be obtained. The pixel coordinates of frame $j$ converted to frame $k$ coordinate system are $(u_n',v_n') \sim KT_{j,k}Y_jK^{-1}(u_n',v_n')$. $T_{j,k}$ is the camera pose of frame $j$ relative to frame $k$, and $Y_j$ is the scaled depth map. Using the bilinear difference method, the dense depth map transformed from frame $j$ to frame $k$ can be obtained as $^Y_{j,k}$.

2.3. Loss Function
We propose two loss functions. The effective depth loss is to directly monitor the correlation between the sparse depth and the target depth, and the depth difference loss is to compensate for the two dense the relative difference between the depth maps.

Effective Depth Loss. Effective depth loss can effectively update network parameters and make the loss function converge faster. The effective depth loss is defined as
\[
L_{ed,(j,k)} = \frac{1}{\sum_{n} M_j} \sum_{n} M_j \left( \log Y_j - \log Y^*_j \right)^2 + \sum_{n} M_k \sum_{n} M_k \left( \log Y_k - \log Y^*_k \right)^2
\]

where, \( M_j \) is a sparse mask, used to ignore areas without depth in the training data. \( M_j \) is defined as
\[
M_j = \begin{cases} 
W_j & \text{if } b_{n,j} = 1 \\
0 & \text{if } b_{n,j} = 0 
\end{cases}
\]

**Depth Difference Loss.** This loss function increases the geometric constraints between two frames in the same video sequence, and compensates for the relative difference between the two depth maps. The depth difference loss also has a better prediction effect on relatively flat organ tissues. The depth difference loss is defined as
\[
L_{dcl,(j,k)} = \frac{1}{N} \sum_{n} |Y^*_{j,k} - Y_j| + \frac{1}{N} \sum_{n} |Y^*_{k,j} - Y_k|
\]

Where \( Y_k \) is the scaled depth map, and \( Y^*_{k,j} \) is the depth map transformed from the image frame \( k \) to the image frame \( j \) coordinate system.

**Overall Loss.** The overall loss function of the network is shown below, which is a weighted combination of tow loss functions.
\[
L_{total,(j,k)} = \lambda_1 L_{ed,(j,k)} + \lambda_2 L_{dcl,(j,k)}
\]

### 3. Experimental results and analysis

#### 3.1. Experimental environment and equipment

The experimental software environment is the Ubuntu18.04, and the experimental hardware configuration mainly includes two NVIDIA GTX 1080 Ti and an Intel Xeon CPU E5-2637. In order to verify the feasibility and effectiveness of the algorithm in this paper, the experiment uses the endoscopic video data set of isolated pig liver for testing. Pig liver can be divided into caudate lobe, right lobe, middle-right lobe, middle-left lobe and left lobe. During the data collection, we conducted video collection on both the front and back of the pig liver. We use the same laparoscope to collect multiple sets of videos for each segment of multiple pig livers. In the data collection, the length of each group of videos is about 20s, and the number of frames per second is 60 frames. 90% of the groups are used for training and 10% are used for testing.

Figure 4 illustrates several sample images from endoscopic video. Each row of data is from the same video sequence. The first row and the second row are collected from the back of the pig liver, and the third row and fourth row are from the front of the pig liver. Due to the large liver organs, we need to collect video data of pig livers in different regions. First, we mark each area of the pig liver to facilitate subsequent experimental comparisons. Second, we take multi-angle shots of each area of the liver. The purpose is to collect images from multiple angles for sparse reconstruction to ensure that the reconstruction depth information of the region is more accurate. COLMAP performs global reconstruction, estimates the camera pose according to the geometric parallax between images and performs global reconstruction, and finally reconstructs a 3D sparse scene based on a video sequence. The endoscope is constantly moving during data collection, because the COLMAP sparse reconstruction needs to depend on the camera poses changes.
In order to test the robustness of the proposed method, a number of data enhancement methods such as Gaussian noise, Gaussian blur, random brightness, and random contrast are used in the training process. The loss function adopts the stochastic gradient descent method to converge the loss function, and dynamically set the learning rate from $1.0 \times 10^{-3}$ to $1.0 \times 10^{-4}$. The effective depth loss function weight is 5, and the depth difference loss function weight is 2. The range between two random frames from the same video sequence is 2 to 20. The batch size of training video frames is 4.

### 3.2. Experimental Results and Comparison

In order to demonstrate the feasibility of the algorithm and the reliability of the experimental results, we conduct two experiments. First, we compare the experimental depth map results produced by our method with Liu et al. [8]. Second, we use the effective depth values of the sparse depth map as the true depth map for data comparison with the prediction depth map to analyse the results quantitatively. The experimental results show that our method is superior to Liu et al. [8].

We showed two sets of video sequences in a comparative experiment. As shown in the figure 5. The first and fourth columns are the original images, the second and third columns are the results obtained by our method, and the fifth and sixth columns are the results obtained by Liu et al. Comparing the experimental results, it can be seen that the depth map obtained by Liu et al. is not accurate or even the depth information is lost for areas with small depth changes. This is because the depth consistency loss used by Liu et al. reduces the influence of noise points, but also smooths the depth information. In addition, for the reflective area, the results obtained by Liu et al.'s method have obvious prediction distortions, because the liver surface had high reflectance and sparse texture, which reduced the ability of the network to acquire features. From the results of 3D reconstruction, Liu et al.'s reconstruction results are somewhat distorted. To solve this problem, this paper adds an attention model to increase the weight of important feature channels, enhance the network's ability to acquire important features, and reduce the distortion and deformation of the prediction results.
In order to analyze the accuracy of the predicted depth map, the sparse depth map is used as the true depth map, and the evaluation measurement is performed only at the effective pixel positions. In terms of error analysis, this article uses the average relative difference as the evaluation index:

$$\text{AbsRel} = \frac{1}{N} \sum_{i=1}^{N} \frac{|y_i^s - y_i^r|}{y_i^r}$$  

(10)

The accuracy rate is the percentage that meets the following different threshold conditions:

$$\max \left( \frac{y_i^s}{y_i}, \frac{y_i^r}{y_i^s} \right) = \delta < \text{thr}$$  

(11)

where, \(\text{thr}\) is the threshold, \(y_i^s\) is the sparse depth map, \(y_i\) is the scaled depth map, and \(N\) is the sum of effective pixels.

This paper randomly selects a group of video frames to test the average relative error within a certain threshold range and the accuracy of the predicted depth map. The results are shown in Table 1. It can be seen from the table that the average relative difference of the predicted depth map is lower than the error result of Liu et al., and the accuracy rate is increased by 1.6% (\(\text{thr}<1.25\)) compared with Liu.

| method | AbsRel | \(\text{thr}<1.25\) | \(\text{thr}<1.25^2\) | \(\text{thr}<1.25^3\) |
|--------|--------|----------------|----------------|----------------|
| Liu    | 0.151  | 0.784         | 0.935         | 0.969         |
| our    | 0.143  | 0.886         | 0.949         | 0.985         |

4. Conclusion

In this work, we propose a depth estimation method for monocular endoscopy applied to the liver. The proposed method adopts the COLMAP to obtain sparse reconstruction data as the supervision signal and does not require any artificially labelled depth maps, and can accurately reconstruct the 3D surface with smooth and sparse texture. The experimental results have demonstrated the effectiveness our method, which has a great potential to be widely applied in the surgical navigation of robotic surgery.

Acknowledgment

This work is supported by multiple grants, including: Key Area Research and Development Program of Guangdong Province, China (No: 2020B010165004), National Natural Science Foundation of China
(No:61902386, 62072452), Natural Science Foundation of Guangdong (No:2018A030313100), and Shenzhen Science and Technology Program (No: JCYJ20180507182415428). Xiangyun Liao is the corresponding author of this paper, Zhengtao Cao and Wenfeng Huang contributed equally to this work.

References
[1] Grasa O G, Bernal E, Casado S, et al.(2014) Visual SLAM for Handheld Monocular Endoscope. Ieee T Med Imaging, 33(1):135-146.
[2] Leonard S, Sinha A, Reiter A, et al. (2018) Evaluation and Stability Analysis of Video-Based Navigation System for Functional Endoscopic Sinus Surgery on In Vivo Clinical Data. Ieee T Med Imaging, 37(10):2185-2195.
[3] Lin B, Sun Y, Qian X, et al. (2016) Video-based 3D reconstruction, laparoscope localization and deformation recovery for abdominal minimally invasive surgery: a survey. The International Journal of Medical Robotics and Computer Assisted Surgery, 12(2):158-178.
[4] Mahmoud N, Cirauqui I, Hostettler A, et al. (2017) ORBSLAM-based endoscope tracking and 3d reconstruction[Z]. Athens, Greece: Springer Verlag, 10170 LNCS, 72-83.
[5] Sun D, Liu J, Linte C A, et al. (2013) Surface reconstruction from tracked endoscopic video using the structure from motion approach[Z]. Nagoya, Japan: Springer Verlag, 8090 LNCS, 127-135.
[6] Zhou T, Brown M, Snavely N, et al. (2017) Unsupervised learning of depth and ego-motion from video[Z]. Honolulu, HI, United states: Institute of Electrical and Electronics Engineers Inc., 6612-6621.
[7] Liu X, Sinha A, Ishii M, et al. (2020,) Dense Depth Estimation in Monocular Endoscopy with Self-Supervised Learning Methods. Ieee T Med Imaging, 39(5):1438-1447.
[8] Hu J, Shen L, Sun G. (2018) Squeeze-and-Excitation Networks[Z]. Salt Lake City, UT, United states: IEEE Computer Society, 7132-7141.
[9] Jegou S, Drozdzal M, Vazquez D, et al. (2017) The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for Semantic Segmentation[Z]. Honolulu, HI, United states: IEEE Computer Society , 1175-1183.