Self-Supervised Lightweight Depth Estimation in Endoscopy Combining CNN and Transformer

Zhuoyue Yang, Junjun Pan, Ju Dai, Zhen Sun, and Yi Xiao

***Abstract*— In recent years, an increasing number of medical engineering tasks, such as surgical navigation, pre-operative registration, and surgical robotics, rely on 3D reconstruction techniques. Self-supervised depth estimation has attracted interest in endoscopic scenarios because it does not require ground truth. Most existing methods depend on expanding the size of parameters to improve their performance. There, designing a lightweight self-supervised model that can obtain competitive results is a hot topic. We propose a lightweight network with a tight coupling of convolutional neural network (CNN) and Transformer for depth estimation. Unlike other methods that use CNN and Transformer to extract features separately and then fuse them on the deepest layer, we utilize the modules of CNN and Transformer to extract features at different scales in the encoder. This hierarchical structure leverages the advantages of CNN in texture perception and Transformer in shape extraction. In the same scale of feature extraction, the CNN is used to acquire local features while the Transformer encodes global information. Finally, we add multi-head attention modules to the pose network to improve the accuracy of predicted poses. Experiments demonstrate that our approach obtains comparable results while effectively compressing the model parameters on two datasets.**

***Index Terms*— Depth and ego-motion estimation, endoscopy, lightweight architecture, self-supervised learning, transformer and CNN.**

I. Introduction

ENDOSCOPIC minimally invasive surgery is widely used because of less bleeding and shorter recovery time compared with open surgery in recent years. However, due to the narrow field of view and lack of depth perception, endoscopic surgeries place stringent demands on the experience and skills of the surgeon. Nowadays, with the rapid development of VR/AR technology, an increasing number of researchers are choosing AR-based surgical navigation to address these difficulties [1], [2], [3]. These AR systems superimpose preoperative data with intraoperative endoscopic data through registration techniques [4], [5]. The accuracy of video-CT registration algorithms primarily relies on the quality of intraoperative reconstructions from endoscopic videos [6]. In addition, there are many tasks, such as surgical robots [7], medical image segmentation [8], surgery planning assistance [9], and surgical instrument recognition [10], that can benefit from the results of depth estimation.

Previous methods for depth estimation from image sequences are based on multi-view geometry principles, such as structure from motion (SfM) [11] and simultaneous localization and mapping (SLAM) [12]. Although depth estimation tasks have been developed in natural scenes for many years, this problem is more difficult in endoscopic scenes due to inconsistent lighting, sparse texture features, and soft tissues with non-Lambertian reflection characteristics. Geometry-based methods [13] rely heavily on feature extraction and matching. The smooth and repetitive soft tissue texture usually results in sparse features and wrong feature matching. Thus, traditional methods still fall short of desirable performance.

Deep learning-based methods in harsh natural environments for depth estimation [14], segmentation [8], and detection [15] have rapidly developed due to the publication of large datasets. However, it is very difficult to obtain large amounts of data with ground truth in endoscopic scenes. Unsupervised learning methods that only use visual images have gained increasing attention in recent years. Researchers have tried to relieve these limitations for endoscopy images by utilizing self-supervised training strategy [6], [16], [17], [18]. Although many self-supervised methods have emerged, the depth networks for most of the work are similar and based on convolution layers. Some works design more complex and heavy networks to achieve better results.

For navigation applications, depth estimation networks not only need to ensure accuracy but also integrate with other modules such as registration. An effective and lightweight network structure is an important topic. Currently, there are many advanced works analyzing existing network architectures and making interesting discoveries. For example, the receptive field

Manuscript received 1 August 2023; revised 6 November 2023; accepted 4 January 2024. Date of publication 10 January 2024; date of current version 2 May 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2022ZD0115902; and in part by the National Natural Science Foundation of China under Grant U20A20195, Grant 62272017, Grant 62172437, and Grant 62102208. *(Corresponding authors: Junjun Pan; Ju Dai.)*

Zhuoyue Yang and Junjun Pan are with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Haidian, Beijing 100191, China, and also with the Peng Cheng Laboratory, Nanshan, Shenzhen 518000, China (e-mail: yangzhuoyue@buaa.edu.cn; pan_junjun@buaa.edu.cn).

Ju Dai is with the Peng Cheng Laboratory, Nanshan, Shenzhen 518000, China (e-mail: daij@pcl.ac.cn).

Zhen Sun and Yi Xiao are with the Division of Colorectal Surgery, the Department of General Surgery, the Chinese Academy of Medical Sciences, and the Peking Union Medical College, Peking Union Medical College Hospital, Dongcheng, Beijing 100730, China (e-mail: sunzhen0906@126.com; xiaoy@pumch.cn).

Digital Object Identifier 10.1109/TMI.2024.3352390

of convolution operation is limited, while Transformer [19] can model global information. The latest work has found that the most effective part of Transformer is the entire framework rather than multi-head attention (MHA) operations [20]. CNN exhibits strong texture bias, while Transformers exhibit strong shape bias [21]. Based on the above findings, we propose a lightweight self-supervised depth estimation network for endoscopic images, which combines the advantages of CNN and Transformer at a fine-grained level.

Our contributions are summarized as follows:

- For the first time, we apply the lightweight network to endoscopic scenes. We present a novel hybrid architecture with an efficient combination of CNN and Transformer at different scales. In order to extract global and shape-aware features, we insert Transformer layers into CNN layers which are sensitive to local textures.
- We propose a pose network with several multi-head attention modules. Attention modules are added at different locations in order to find a solution with better generalization. We perform experiments on several long sequences to verify the performance improvement of the methods.
- Extensive experiments have demonstrated the effectiveness of our proposed method, which compresses the number of model parameters without a significant loss of accuracy. Qualitative experiments demonstrate that our method achieves comparable results with current state-of-the-art methods on the SCARED and clinical datasets.

II. Related Work

In this section, we review the unsupervised depth estimation methods applied in endoscopic scenes, as well as the state-of-the-art (SOTA) network framework combining CNN (convolutional neural network) and Transformer applied in natural scenes.

A. Self-Supervised Learning

Depth estimation methods in natural scenes have been studied for several years and typically leverage real depth values as supervised signals to model the problem as a regression or classification problem. However, true depth values are difficult to obtain in an endoscopic environment. It is not until after unsupervised methods are widely used [22], that those deep learning methods are formally applied to endoscopic depth estimation tasks. Zhou et al. [17] propose an unsupervised training method using only monocular video sequences. The method uses the computed depth and poses as mediators and warps nearby views to the target view as supervised information. Godard et al. [14] leverage binocular videos instead of depth truth to train the fully convolutional network.

The first article [16] applies unsupervised depth estimation to endoscopy. The authors use a fully convolutional depth estimation approach with a similar structure to the method in [17]. Godard et al. [23] propose the Monodepth2 on the basis of [14]'s network framework. The predictor behind the decoder in the depth estimation network and the decoder in the pose estimation network is deleted. Most researchers find that the structure in Monodepth2 [23] including a depth network and a separate pose estimation network could achieve better performance. Following [16] and [23], this structure became the baseline for subsequent methods and the unsupervised depth estimation is regarded as an image reconstruction problem at present. To deal with edge conditions, such as object motion and occlusion, predictive interpretable masks are used. Liu et al. [24] propose a self-monitoring method to train convolutional neural networks for intensive depth estimation from monocular endoscopic data. Supervised signals are derived from the positional and sparse point clouds of the motion recovery structure. Recasens et al. [25] leverage monodepth2 [23] in this work to train an endoscopic depth estimation network to obtain the depth corresponding to each image. Ozyoruk et al. [18] put forward EndoSfMLearner, which is an unsupervised monocular depth and pose estimation method. This method combines residual networks and a spatial attention module to focus on highly textured tissue areas. Li et al. [26] add the LSTM module in the pose estimation network to model time information, thus improving the accuracy of pose estimation. Shao et al. [6] joint use optical flow appearance flow to deal with the brightness inconsistency problem. Zhang et al. [27] propose a network that shares an encoder and contains two branches in the decoder. The two branches estimate the depth information and normal information respectively. Currently, most of the self-supervised deep networks applied to endoscopic images are convolutional neural networks. Most researchers [28], [29] focus on increasing model complexity and parameters to improve the performance of the network.

B. Network Architectures

With the development of the technology, Transformer shows great potential for depth estimation tasks in natural scenes. Varma et al. [30] first evaluate the impact of transformer on self-supervised monocular depth estimation. DPT [31] directly uses the Transformer as the encoder, and then fuses the results of each layer of the Transformer separately to generate depth estimation results. AdaBins [32] uses ViT after general encoders and decoders, and then adaptively divides depth values based on the dynamic changes of the scene. TransDepth [33] also adds Transformer blocks to the ResNet [34] results to obtain long-distance information, then uses a decoder based on attention and Gate to fuse features, and finally performs depth estimation through prediction head. Vision Adaptor [35] designs an adapter that runs in parallel with ViT, incorporating prior knowledge of images into the ViT backbone to provide reconstructed multi-scale features for dense depth estimation problems, preserving the flexibility of ViT and improving performance. DeepFormer [36] performs ViT and convolution operations separately in the encoder stage and designs a layered aggregation and interaction module to combine the two parts. To summarize, some researchers build independent Transformer-based encoders to obtain feature maps or add several modules to fuse features from CNN.

MonoViT [37] is the current state-of-the-art work in natural scene depth estimation tasks. The encoder of MonoViT [37] is constructed by stacking several MPViT [38], and the decoder

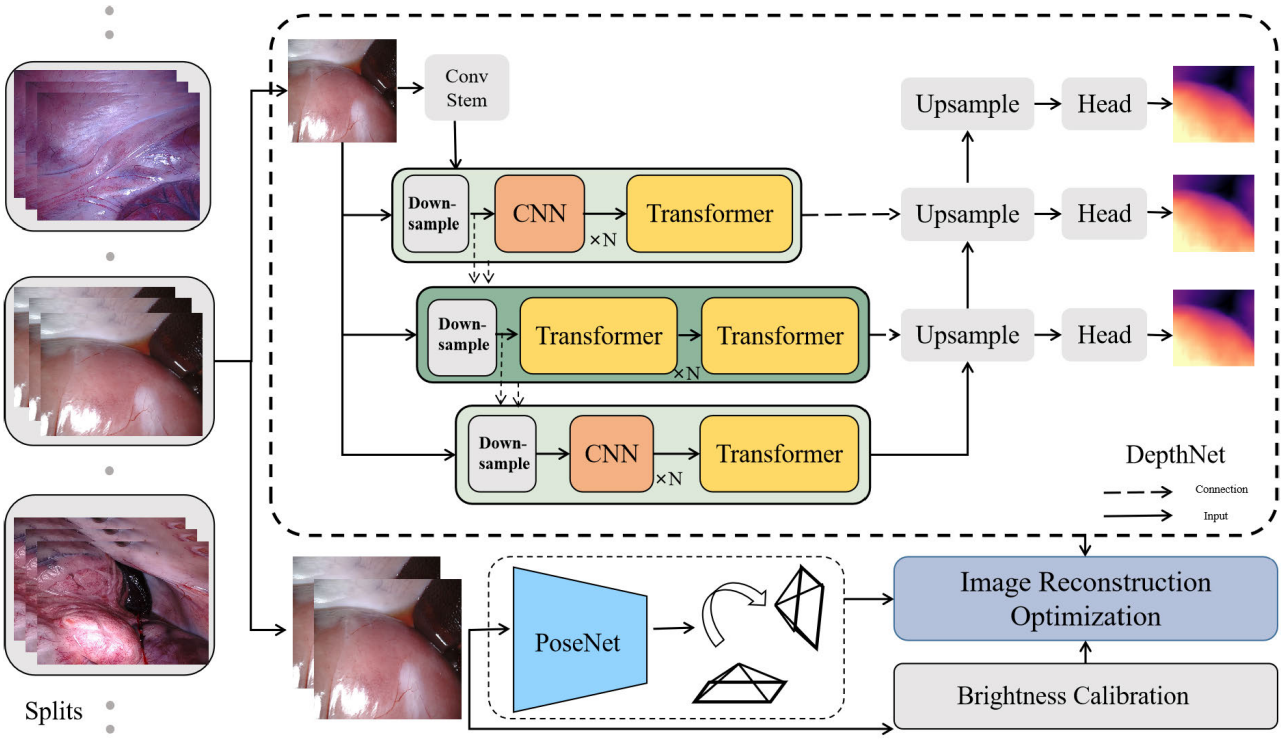

Fig. 1. **Overview of the proposed method.** Our method includes a depth network (DepthNet) and a pose network (PoseNet). Our DepthNet consists of an encoder with a combination of CNN and Transformer and a decoder. Our PoseNet is enhanced by the multi-head attention modules.

is from HR-Depth [39]. Each layer of MPViT has three transformer heads and a convolution head. MonoFormer [21] still relies on ViT [40], mainly by proposing the attention connection module and feature fusion decoder. Zhang et al. [41] propose a dilated convolutional module to extract rich multi-scale local features and a self-attention-based feature interaction module to encode remote global information into features. Yu et al. [20] prove that the general architecture of the Transformers, instead of the specific token mixer module, is more essential to the model's performance. CMT [42] inserts Transformer structures between different convolutional layers of CNN. The ablation experiments have shown that the widely used phased design in CNN is a better choice for promoting Transformer-based architectures. In summary, the integration of CNN and Transformer in architecture has evolved from coarse-grained stacking to fine-grained information exchange. The difference between our method and the above methods is that we stack the CNN layers and the Transformer layers alternately. We utilize this hybrid structure to obtain local and global features while also using textures and contours.

III. Method

A. Overall Architecture

The framework includes a depth estimation network (DepthNet), a pose estimation network (PoseNet), and a brightness calibration network, as shown in Fig. 1. Endoscopic images are segmented into groups of three in chronological order. The DepthNet estimates the multi-scale depth map of a single endoscopy image, while the PoseNet estimates the camera motion between adjacent images. We combine convolutional layers with transformer structures to build a hybrid DepthNet. We use the brightness calibration module proposed in [6] to compensate for lighting changes caused by endoscope movement. Then, according to the predicted camera poses and camera internal parameters, the estimated depth is re-projected back to the two-dimensional plane, and the model is supervised and optimized by calculating the loss between the reconstructed image and the target image. The details of DepthNet and PoseNet are described below. The utilized loss functions are listed.

B. DepthNet

Following [14] and [23], we design our method as an encoder-decoder architecture. CNN has better performance in extracting local textures and Transformers are sensitive to global information and contours [21]. We present a novel hybrid encoder that is able to focus on both texture and contour features. The first and third layers are stacked with multiple layers of CNN modules, and then several Transformer blocks are placed in sequence in the middle layer. Multi-scale features from the encoder are connected into a concise decoder.

1) Depth Encoder: The input image is first passed through a convolution stem, containing three 3×3 convolutions. The first convolution with a stride of 2 and the next two with a stride of 1. The output channel is C_1, and the size of the output feature map is $H/2 \times W/2$. From the following stages, the CNN-based layer and Transformer-based layer are alternately stacked. Firstly, several symbols are defined to describe the input and output of each stage. We use $\mathbf{F}_i$ to represent the feature map output from the i-th layer. The image that has been pooled in the i-th layer is labeled as $\mathbf{I}_i$. The feature obtained through downsampling modules for each layer is $\mathbf{D}_i$.

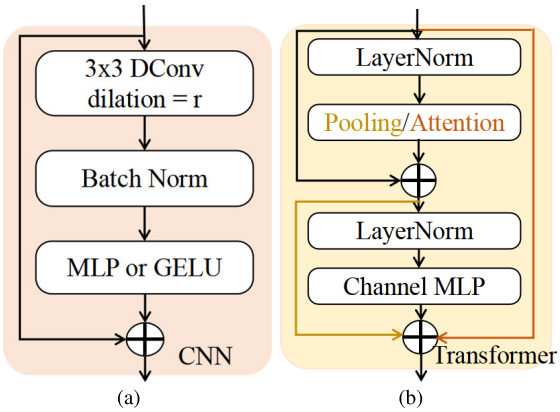

Fig. 2. **CNN and Transformer blocks that are adopted in the depth encoder of DepthNet.** (a) is the structure of the CNN block. (b) shows the architecture of the Transformer blocks. To distinguish between two different Transformer blocks, we name them based on the different operations used in the framework.

Following [21] and [41], $\mathbf{F}_{i-1}$, $\mathbf{D}_{i-1}$ and $\mathbf{I}_i$ are concatenated together and fed into the i-th layer. Inspired by [41], in the CNN-based layer, local and long-range features are extracted by stacking several dilated convolution blocks and a Transformer block. In the second layer, Transformer-based architecture is adopted to enhance shape information, resulting in the depth feature with size $H/8 \times W/8 \times C_2$. Then, the aggregating features are fed into dilated convolution modules, and depth maps of $H/16 \times W/16 \times C_3$ are generated by the Transformer block.

The encoder consists of three layers, each of which is composed of multiple stacked blocks. We first introduce the basic blocks used in the depth encoder, and then explain the structures of the CNN-based layer and Transformer-based layer. The dilated convolutions and Transformer [20], [41] blocks are shown in Fig. 2. We first define the symbols used in this paper for convenience. $\mathbf{X}$ denotes the input features. And $\hat{\mathbf{X}}$ represent the output of the dilated convolution module. BN is a batch normalization and LN is a layer normalization. MLP is the abbreviation of a multi-layer perceptron. As shown in Fig.2(a), $\hat{\mathbf{X}}$ is defined as follows:

$$\hat{\mathbf{X}} = \mathbf{X} + \mathrm{MLP}(\mathrm{BN}(\mathrm{DConv}(\mathbf{X}))), \tag{1}$$

where DConv is the depth-wise dilated convolution operation with the dilatation rate. We replace MLP with activation function (GELU) [43] in some blocks to reduce model parameters. There are two types of Transformer blocks, as shown in Fig.2(b). $\hat{\mathbf{Y}}$ is the output of the Transformer module and can be computed as follows:

$$\begin{aligned} \tilde{\mathbf{X}} &= \mathrm{MixToken}(\mathrm{LN}(\mathbf{X})) + \mathbf{X}, \\ \hat{\mathbf{Y}} &= \tilde{\mathbf{X}} + \mathrm{MLP}(\mathrm{LN}(\tilde{\mathbf{X}})). \end{aligned} \tag{2}$$

Pooling [20] and cross-covariance attention [44] operations are utilized as MixToken. As shown in Fig. 2(b), for the output of cross-covariance attention block, $\hat{\mathbf{Y}} = \mathbf{X} + \mathrm{MLP}(\mathrm{LN}(\tilde{\mathbf{X}}))$.

The Transformer-based layer is illustrated in Fig. 3. The input of the Transformer layer is the concatenation of $\mathbf{F}_{i-1}$, $\mathbf{D}_{i-1}$, and $\mathbf{I}_i$. We first perform a convolution to reduce the dimensionality of the input. Then, two Transformer blocks

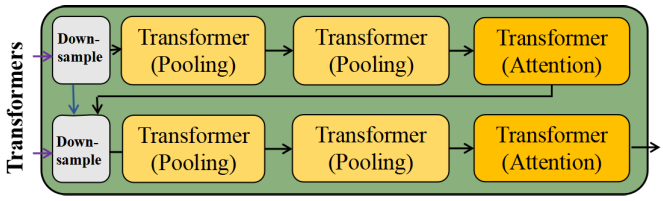

Fig. 3. **Transformer-based layers that are adopted in the depth encoder of DepthNet.** Transformer blocks using different operations are distinguished by different shades of yellow. The specific structure of each type of block is shown in Fig. 2.

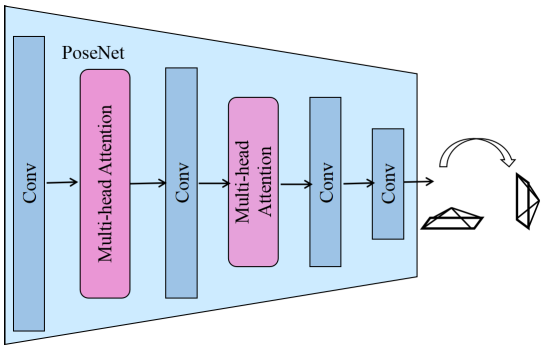

Fig. 4. **PoseNet.** Multi-head attention.

with a pooling operation and one Transformer block with an attention block are used to extract shape-aware and long-range features ($\mathbf{F}$). Subsequently, we concatenate $\mathbf{F}$, $\mathbf{D}_i$, and $\mathbf{I}_{i+1}$ together, and fed them into the stacked three Transformer blocks again. The CNN-based layer consists of several convolution blocks and one Transformer block, as shown in Fig. 1. The number of CNN blocks in the third layer is twice the number of CNN blocks in the first layer.

2) Depth Decoder: Our decoder adopts the concise and effective U-Net [45] structure, in [23]. Convolution layers and skip connections are employed in the decoder to receive multi-scale features from the encoder. Then, cross-layer connections and upsamples are used to increase the resolution. Finally, three prediction heads output inverse depth maps at different resolutions, according to the aggregated features. Each prediction head consists of a convolution layer, a bilinear upsample, and a sigmoid layer. All predicted multi-scale depth maps participate in self-supervised learning optimization.

C. PoseNet

Most of the networks [46] utilize a pose estimation network similar to monodepth2 [23], which takes two adjacent color pictures as input and outputs the 6-DoF relative pose between the pictures. PoseNet uses the pre-trained ResNet [34], i.e. a structure with four superimposed convolutional layers, as an encoder. Considering the influence of light in the medical scene, we add multi-head attention modules [19] into the above architecture to improve the performance of the pose estimation network, as shown in Fig. 4.

Two adjacent images ($H \times W \times 3$) are first fed into a convolution stem to obtain a feature map $\mathbf{F}$ of size $H/2 \times W/2$. After passing through the maximum pooling layer, the output ($\widetilde{\mathbf{F}}$) of multi-head attention can be defined as:

$$\widetilde{\mathbf{F}} = \mathrm{MultiHeadAtten}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) + \mathbf{F}, \tag{3}$$

where MultiHeadAtten($\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$) is the concatenated output of k self-attention operations, which is applied as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}})\mathbf{V}, \tag{4}$$

where $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$ are projected from $\widetilde{\mathbf{F}}$ and d is the dimension of the input. Then, the feature extraction was performed through two superimposed ResNet [34] blocks to obtain feature maps with scales of $H/4 \times W/4$ and $H/8 \times W/8$, respectively. In addition, the extracted feature map passes through the multi-head attention layer again. Finally, the last two feature maps are obtained through two basic blocks. The feature maps are converted into pose matrices through convolutions.

D. Self-Supervised Learning

Like other unsupervised learning methods, we also transform the task as 2D image reconstruction and supervise the consistency and accuracy of depth estimation by minimizing the similarity between the re-projected image and the target image. The image reconstruction loss consists of the photometric loss ($\mathcal{L}_p$) and edge-aware loss ($\mathcal{L}_e$). We define the source image as $\mathbf{I}^\dagger$. Utilizing the pose estimation $\mathbf{T}$ and intrinsic parameters of the camera $\mathbf{P}$, the reconstructed image ($\tilde{\mathbf{I}}$) can be re-projected (π) from the depth estimation $\mathbf{D}$ and $\mathbf{I}^\dagger$. The reconstructed image ($\tilde{\mathbf{I}}$) is defined as follows:

$$\tilde{\mathbf{I}} = \pi(\mathbf{I}^\dagger, \mathbf{T}, \mathbf{D}, \mathbf{P}). \tag{5}$$

Due to inconsistent lighting in the endoscopic environment, the photometric loss is inaccurate. We apply a pre-trained optical flow network to calibrate the rotation and translation changes between two input images and use a pre-trained appearance flow network that results in $\mathbf{C}$ to supplement the illumination. The modified image ($\hat{\mathbf{I}}$) resulting from the target image $\mathbf{I}$ is as follows:

$$\hat{\mathbf{I}} = \mathbf{I} + \mathbf{C}. \tag{6}$$

The image similarity ($\mathcal{F}$) between the modified image ($\hat{\mathbf{I}}$) and the reconstructed image ($\tilde{\mathbf{I}}$) is defined as follows:

$$\mathcal{F} = \alpha \cdot \frac{1 - SSIM(\hat{\mathbf{I}}, \tilde{\mathbf{I}})}{2} + (1-\alpha) \cdot \left|\hat{\mathbf{I}} - \tilde{\mathbf{I}}\right|, \tag{7}$$

where $SSIM$ is the structural similarity index [47] and $\alpha = 0.85$. The photometric loss $\mathcal{L}_p$ is the minimum value of $\mathcal{F}$ among two adjacent images with the visibility mak [6], [23]. In order to maintain the edges, edge-aware loss is also used. As in previous work [17] and [23], the edge-aware loss is defined as:

$$\mathcal{L}_e = |\partial_x d| e^{-|\partial_x \mathbf{I}|} + |\partial_y d| e^{-|\partial_y \mathbf{I}|}, \tag{8}$$

where d represents the mean-normalized inverse depth of $\mathbf{I}$.

IV. Dataset and Results

A. Dataset

1) SCARED Dataset: We utilize SCARED [48] dataset to evaluate our methods' performance. The SCARED dataset is published on the endoscopic sub-challenge organized by MICCAI2019, containing 9 different sub-datasets collected from porcine cadavers. Each sub-dataset contains an endoscope video, the ground truth of the pose recorded by the surgical robot, and the ground truth of depth collected by structured light equipment. Therefore, we can evaluate the performance of pose estimation and depth estimation methods using this dataset. Following [6], we also refer to the Eigen-Zhou [17], [49] evaluation protocol to separate the training, validation, and test datasets, respectively.

2) Clinical Dataset: In order to verify the generalization performance of the method, we also collect videos during right hemicolectomy surgery with the assistance of surgeons. Four representative video clips are selected for quantitative experiments. Each video contains 150-200 images. The contents of the images include live, colon, small intestine, fat, etc. in the abdominal cavity. These four sequences are representative image sequences during the surgical navigation phase. This dataset is not utilized in the training process.

B. Implementation Details

Our method is implemented by PyTorch. In our experiments, we utilize a single NVIDIA V100 and the batch size is 12. The following training augmentations are performed, with 50% chance: random brightness, contrast, saturation, and hue jitter with respective ranges of ± 0.2, ± 0.2, ± 0.2, and ± 0.1. Our depth estimation network and pose estimation network use two AdamW [50] optimizers respectively. The initial values of learning rates are 1e-4. Drop-path is used to mitigate overfitting and the training epoch is set to 50. The specific values of C_1, C_2 and C_3 are 48, 80 and 128.

Following [6], [17], and [23], we compute the 5 standard metrics (Abs Rel, Sq Rel, RMSE, RMSE log, $\delta < 1.25$) proposed in [49] for evaluation. These metrics are defined as follows:

$$Abs\ Rel = \frac{1}{|\mathbf{D}|} \sum_{d \in \mathbf{D}} |d^* - d| / d^* \tag{9}$$

$$Sq\ Rel = \frac{1}{|\mathbf{D}|} \sum_{d \in \mathbf{D}} |d^* - d|^2 / d^* \tag{10}$$

$$RMSE\ log = \sqrt{\frac{1}{|\mathbf{D}|} \sum_{d \in \mathbf{D}} |\log d^* - \log d|^2} \tag{11}$$

$$RMSE = \sqrt{\frac{1}{|\mathbf{D}|} \sum_{d \in \mathbf{D}} |d^* - d|^2}, \tag{12}$$

$$\delta = \frac{1}{|\mathbf{D}|} \left| \left\{ d \in \mathbf{D} | max(\frac{d^*}{d}, \frac{d}{d^*} < 1.25) \right\} \right| \times 100\% \tag{13}$$

where $\mathbf{D}$ is the set of the predicted depth. d and d^* denote the predicted depth and the ground truth, respectively. We perform a 5-frame pose evaluation following [17] and adopt the metric of absolute trajectory error (ATE) [51].

C. Depth Estimation

1) Performance Comparision: We run experiments on the SCARED dataset to evaluate the depth error and accuracy of

TABLE I

DEPTHNET PERFORMANCE. 'ENCODER', 'DECODER', AND 'OVERALL' REPRESENTS THE NUMBER OF PARAMETERS UTILIZED IN DEPTHNET. AUXILIARY REPRESENTS THE NUMBER OF AUXILIARY MODELS' PARAMETERS. †MEANS THE BEST RESULT WE REPRODUCE ON OUR MACHINE

Method	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE log↓	δ↑	Encoder (M)	Decoder (M)	Overall (M)	Auxiliary (M)
Fang et.al [52]	0.078	0.794	6.794	0.109	0.946	131.9	4.9	136.8	—
DeFeat-Net [53]	0.077	0.792	6.688	0.108	0.941	11.2	3.1	14.3	3.2
SC-SfMLearner [54]	0.068	0.645	5.988	0.097	0.957	11.2	3.1	14.3	—
Monodepth2 [23]	0.071	0.590	5.606	0.094	0.953	11.2	3.1	14.3	—
Endo-SfM [18]	0.062	0.606	5.726	0.093	0.957	11.2	3.1	14.3	—
AF-SfM [6]	0.059 0.062†	0.435 0.489†	4.925 5.142†	0.082 0.087†	0.974 0.959†	11.2	3.1	14.3	28.6
Ours	**0.062**	**0.558**	**5.585**	**0.090**	**0.962**	**1.73**	**0.22**	**1.95**	**28.6**

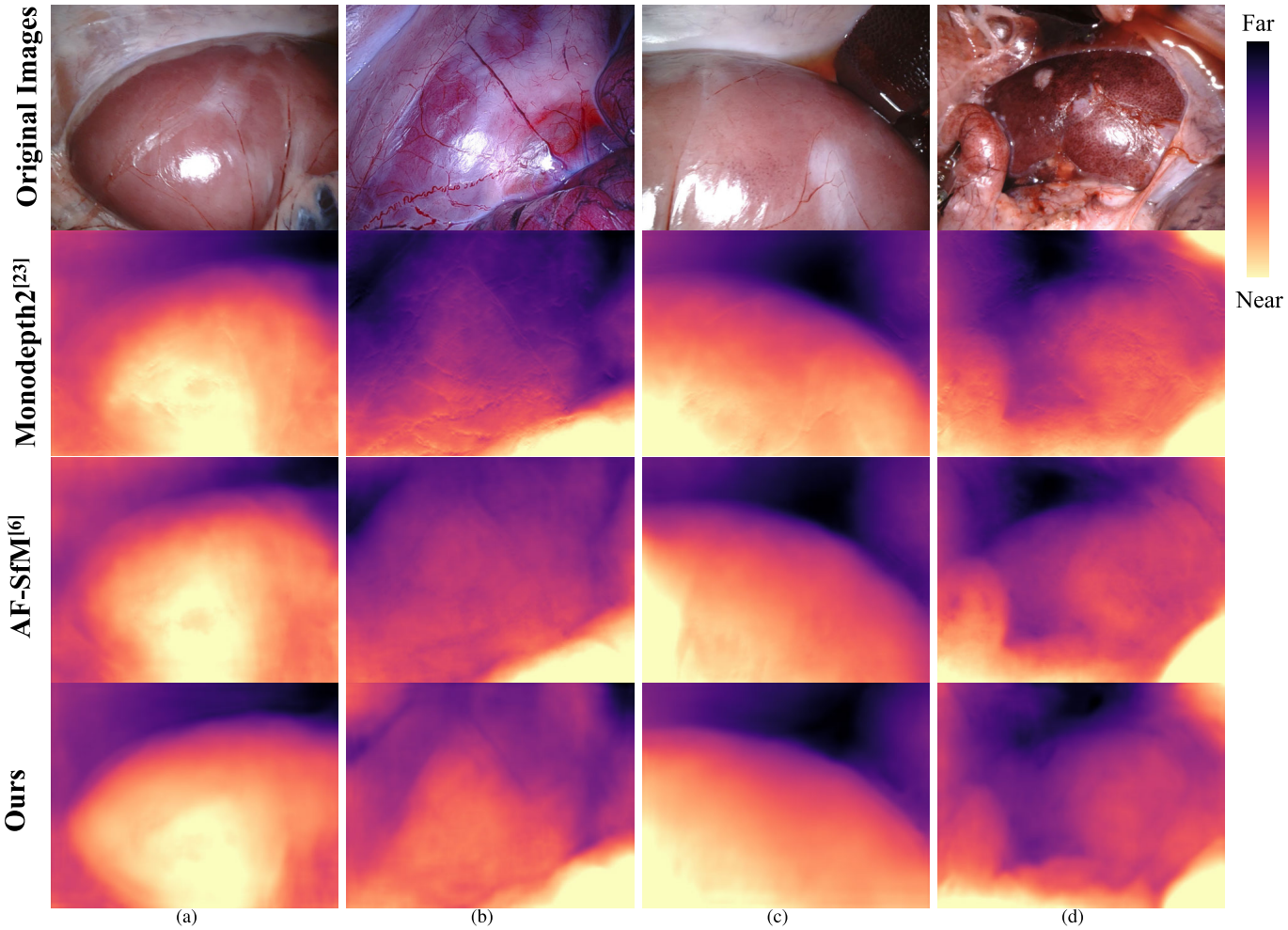

Fig. 5. **Qualitative depth comparison.** There are four examples from the SCARED dataset. The first row is original images and the others are depth maps. The second and third rows are results from [6] and [23]. The last row shows our results.

our model. The proposed method is compared with several SOTA self-supervised methods, including AF-SfM [6], Endo-SfM [18], Monodepth2 [23], Fang et al. [52], DeFeat-Net [53] and SC-SfMLearner [54]. To make up the monocular scale ambiguity, following the same strategies indicated in [6] and [23], the estimated depth is scaled by the per-image median ground truth. Table I collects the quantitative results of our model against other typical self-supervised methods. The encoder, decoder, and overall columns in Table I report the size of parameters in the DepthNet. Our method achieves comparable performance to the state-of-the-art methods with the smallest parameters in the inference phase. We achieve the second-highest ranking result in accuracy. In Table I, the auxiliary parameters refer to the network parameters proposed in AF-SfM [6] for correcting illumination. With these two auxiliary networks only utilizing the training phase, both our model and AF-SfM [6] can achieve better performance. The performance of compared methods on depth estimation is from [6]. According to Table I, our method achieves a lower result on RMSE. Fig. 5 shows that our model obtains satisfactory results compared with other methods. We can observe that our method provides a more accurate depth

TABLE II
ABLATION STUDY ON THE NUMBER OF TRANSFORMER BLOCKS IN ONE LAYER

Structure	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE log↓	δ ↑
Baseline	0.069	0.649	6.032	0.099	0.949
Pool×2	0.068	0.690	6.070	0.101	0.949
Pool×3	**0.064**	**0.577**	**5.589**	**0.091**	**0.961**
Pool×4	0.068	0.664	5.989	0.097	0.955

estimation of the edge of organs while maintaining the global smoothness of soft tissues. These quantitative and qualitative results demonstrate the superiority of our method.

2) Ablation Study on DepthNet Architecture: To further demonstrate the validity of the proposed model, an ablation study is conducted to assess the importance of different designs in the architecture. We conduct experiments on the number of Transformer blocks and the structure of Transformers layers. The number of modules in a single layer determines the foundation of the framework.

a) Ablation study on the number of transformer blocks: Table II collects the results with different numbers of transformer blocks with pooling operation in the middle Transformer layer. The transformer block with attention remains unchanged in each experiment. The baseline is a simplified model in the second layer that does not replace CNN layers with Transformer layers. We test the depth estimation results of 2, 3, and 4 Transformer blocks. Based on the results of the second and third rows, we find that adding a block with pooing can improve the performance of the model. However, based on the results of the third and fourth rows, we find that consistently stacking pooling Transformer blocks result in a decrease in performance. Therefore, we use the structure in Fig. 3 to achieve stable performance improvement while increasing pool formers through cascading and convolution operations. Based on the results in the last row of Table I, our current structure can strike a balance between the depth estimation accuracy and the model size.

b) Ablation study on the architecture of transformer layers: The influence of different architecture on accuracy has been studied. We compare the following three frameworks, as shown in Fig. 6. These three subgraphs show the basic hybrid structure, each consisting of 3, 4, and 5 layers. In both Fig. 6(a) and Fig. 6(c), CNN-based layers are used as the first and last layers. In both (b) and (c), there are two Transformer layers in the architecture.

Table III shows the different results obtained by these three structures. Both (a) and (c) achieve good performance, which is comparable to the most advanced methods. However, the model parameters of (a) are the smallest. So in the performance analysis experiment, we report the results of (a). However, the structure in (c) can achieve smaller errors on Sq Rel, RMSE, and RMSE log metrics.

D. Pose Estimation

We select two sequences with longer trajectories [6] in the SCARED dataset and label them as Sequence-1 (Seq.1) and Sequence-2 (Seq.2) respectively. Table IV shows the

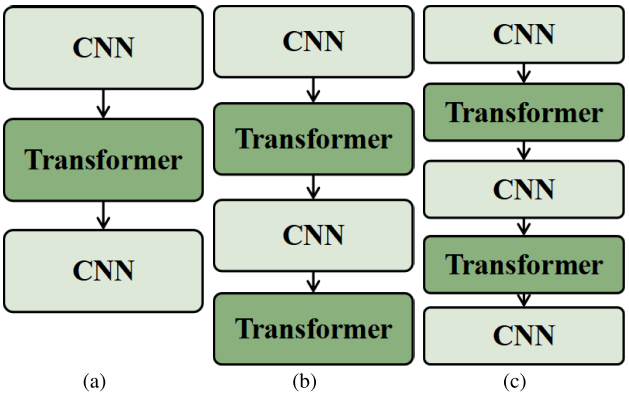

Fig. 6. **CNN and Transformer architectures that can be adopted in the depth encoder of DepthNet.** (a), (b), and (c) are the hybrid architecture with 3, 4 and 5 layers.

TABLE III
ABLATION STUDY ON TRANSFORMER AND CNN ARCHITECTURES

Architecture	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE log↓	δ ↑
Fig.6(a)	**0.062**	**0.558**	**5.585**	**0.090**	**0.962**
Fig.6(b)	0.068	0.634	6.001	0.097	0.954
Fig.6(c)	**0.062**	**0.523**	**5.402**	**0.087**	**0.969**

TABLE IV
POSE PERFORMANCE

Method	Backbone	ATE↓ (Seq.1)	ATE↓ (Seq.2)
Defeat-Net [53]	ResNet-18	0.1765	0.0995
SC-SfMLearner [54]	ResNet-18	0.0767	0.0509
Monodepth2 [23]	ResNet-18	0.0769	0.0554
Endo-SfM [18]	ResNet-18	0.0759	0.0500
AF-SfM [6]	ResNet-18	0.0742	0.0478
Ours	**ResNet-18+MHA2**	**0.0723**	**0.0474**

comparison of the proposed method with the other five methods. The performance of compared methods is from AF-SfM [6]. Our method achieves the lowest error on the ATE. Most of the work use the same pose estimation network. We concatenate two input images and then estimate the 6DoF between the two images using features extracted by ResNet [34]. Feature-dependent approaches have higher immunity against light variations. We add attention mechanisms to enhance features, emphasize differences, and thus improve performance.

To further analyze the effect of the multi-head attention mechanism on the pose estimation network, we conduct ablation experiments. Table V collects the results of adding multiple attention mechanisms at different locations. These insertion locations include the first layer of convolution, the second layer of convolution, the third layer of convolution, and various combinations of these locations. For the scheme, after MHA is added to the first layer of convolution, we note that while it achieves better results than Monodepth2 [23], it is not as good as the combined use of appearance flow. Interestingly for Sequence-1, we find that adding multi-head attention in both the second and third layers achieves lower errors. However, for Sequence-2, only the MHA in the second layer yields a performance gain. Therefore, we use the addition of the MHA mechanism in the middle layer to obtain better

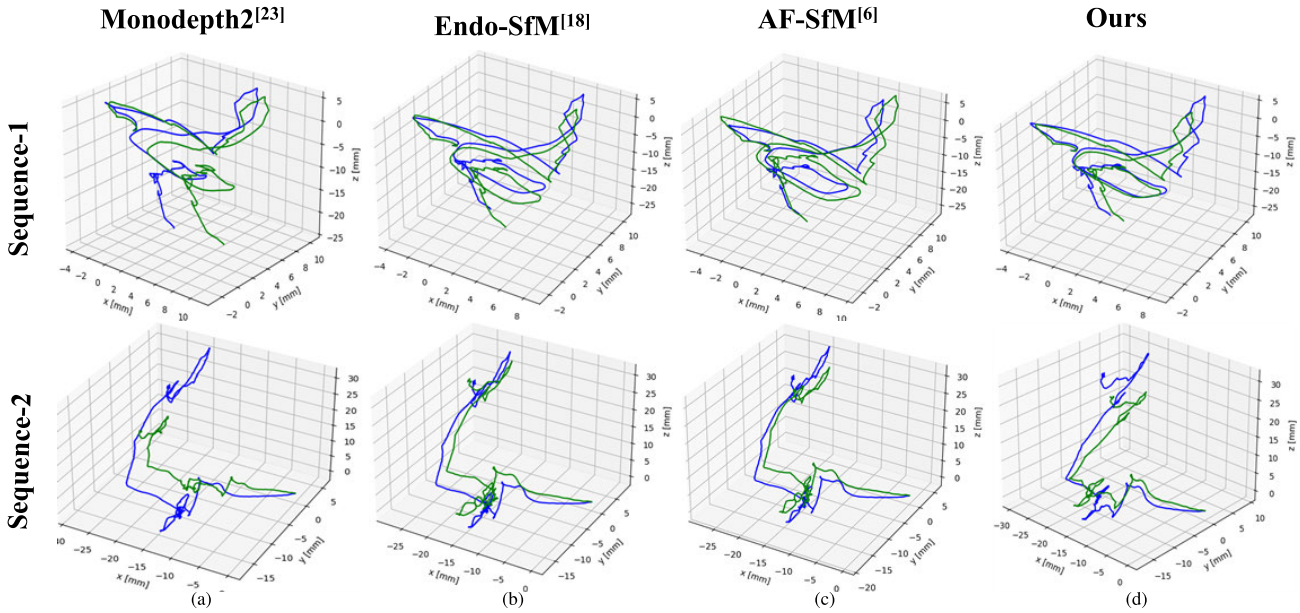

Fig. 7. **Qualitative pose comparison.** The first three columns are the trajectory results by using comparative methods([6], [18], [23]). The results in the last column are our trajectory results.

TABLE V
ABLATION STUDY ON POSENET

Method	Backbone	ATE↓ (Seq.1)	ATE↓ (Seq.2)
Baseline (monodepth2)	ResNet-18	0.0769	0.0554
SOTA (afsfm)	ResNet-18	0.0742	0.0478
MHA×1	ResNet-18+MHA	0.0754	0.0548
MHA×2	**ResNet-18+MHA×2**	**0.0723**	**0.0474**
MHA×3	ResNet-18+MHA×3	0.0732	0.0514

generalization. Fig. 7 reports qualitative examples from these two trajectories. The performance of our model is superior to other competitors in the middle of trajectories.

E. Surface Reconstruction

We can recover point clouds from camera intrinsics and depth estimates, as shown in Fig. 8. The point cloud shown in Fig. 8 does not have any added colors, in order to display the geometric structure. We use truncated signed distance function (TSDF) [55] to fuse multiple point clouds in order to extend the 3D model of the tissue surface. The implementation is developed by Open3d [56]. Readers can reference [25] to get the procedure of expanding multiple point clouds based on pose estimates. We further utilize laparoscopic images obtained from surgery for visual performance analysis.

Fig. 9 shows the surface reconstructed from the SCARED dataset. Subfigures in Fig. 10 are the surfaces recovered from the clinical dataset. The images in the first row demonstrate the texture and the second row shows the mesh. Through mesh, we can more clearly see the structure of soft tissues in different scenarios. By adding textures, the entire scene can be visually reflected. Fig. 9 shows that our method preserves distinct tissue structures and keeps local soft tissues smooth and continuous. TableVI reflects our scenes containing a large number of verts.

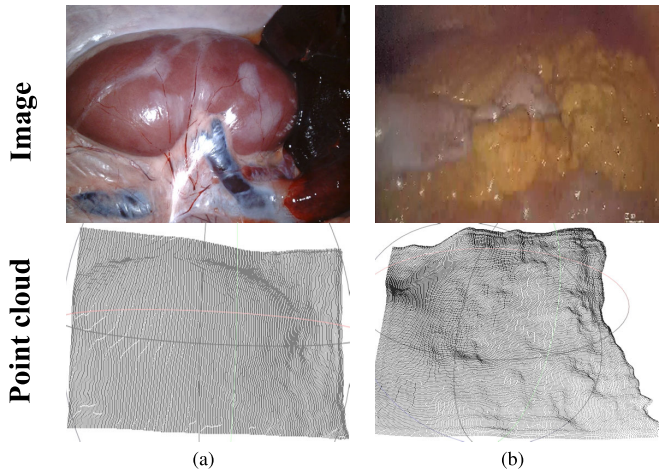

Fig. 8. **Point clouds on the SCARED and clinical datasets.** (a) (b) show two examples. Images in the first row are original images and figures in the second row are reconstructed point clouds.

TABLE VI
SURFACE RECONSTRUCTION

Scene	Mesh (million)	Time (per image)	Scene	Mesh (million)	Time (per image)
Fig.9(a)	1.5	0.22s	Fig.10(a)	1.4	0.21s
Fig.9(b)	1.3	0.24s	Fig.10(b)	1.8	0.22s
Fig.9(c)	1.4	0.23s	Fig.10(c)	1.2	0.21s
Fig.9(d)	3.0	0.22s	Fig.10(d)	1.6	0.23s

The average number of points for surface models in the scared dataset and the real dataset is 1.8 and 1.5 million, respectively. The average processing time for each image is 0.2 seconds. We do not include network inference time here. The inference time of our method and other methods are shown in TableVII. Our method also reduces inference time.

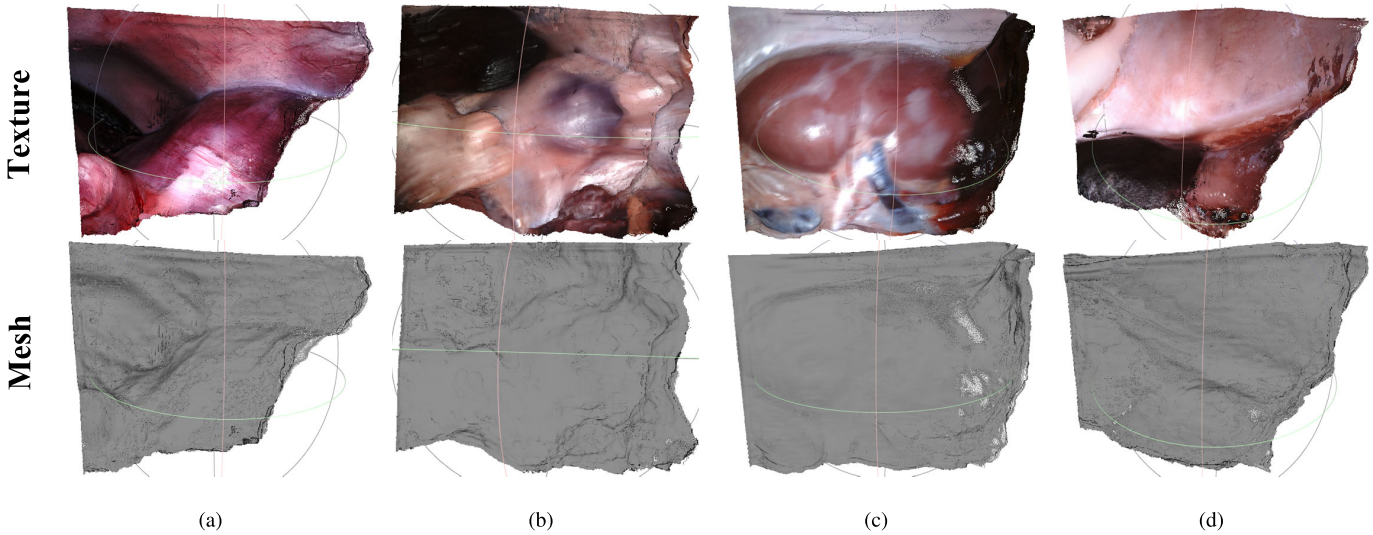

Fig. 9. **Our surface reconstructions on the SCARED dataset.** (a), (b), (c) and (d) are 3D surfaces recovered from four images captured from porcine cadavers.

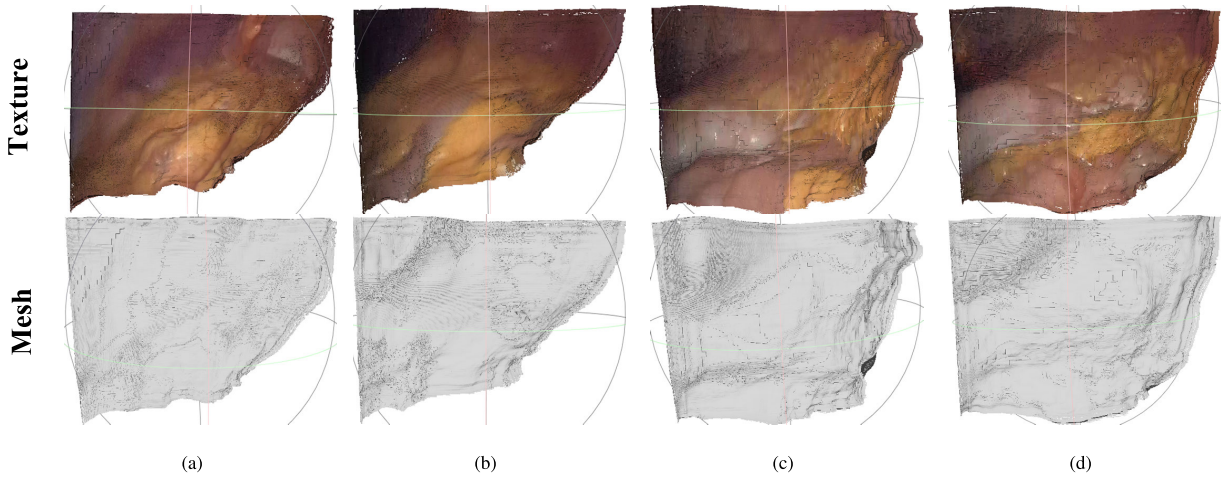

Fig. 10. **Recoverd surfaces on the clinical dataset.**(a), (b), (c), and (d) are 3D surfaces recovered from four representative laparoscopic images obtained during surgery, mainly including fat, intestines, and liver.

TABLE VII
DEPTHNET INFERENCE SPEED

Method	Speed
EndoSfM [18]	4.2ms
Monodepth2 [23]	3.8ms
Ours	**3.6ms**

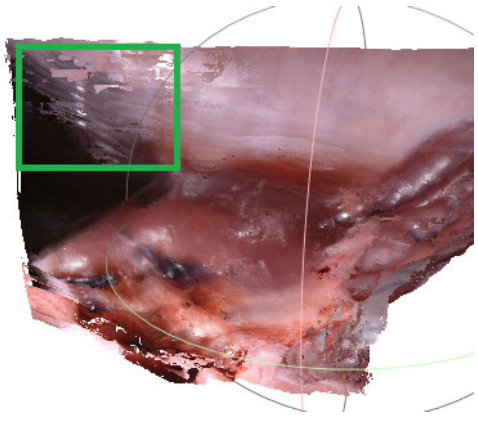

Fig. 11. **The example of unsatisfactory reconstruction result.** The green box region shows the overlap.

F. Limitations

Although our method is mainly trained and tested on laparoscopic images, we have also tested it in clinical experiments. However, there are still some disadvantages to the depth fusion, such as the presence of discrete points on the edge of the fourth image in Fig. 9. In addition, for dynamic scenarios, such as device movement and interaction between devices and soft tissues, current fusion methods may have a significant overlap (Fig. 11). The reason for this phenomenon may be due to inconsistent depth estimates across multiple images. The problem of inconsistent depth between different laparoscopic images still exists due to the similar texture.

V. Conclusion

A lightweight depth estimation network is first applied for endoscopy images in this paper. We propose a self-supervised depth estimation network with a combination of CNN and Transformer for endoscopy images. CNN-based layers mixed with transformer-based layers are utilized as the encoder to aggregate local texture information and global contour features. Our method achieves competitive results while also reducing the number of parameters. The proposed pose network obtains the minimum error on the SCARED dataset compared to the previous approaches. Detailed quantitative and qualitative experiments demonstrate the effectiveness of our method.

However, there are still some issues that need to be improved in the depth fusion task. The newest implicit scene representation methods, such as NeRF [57], can be used to solve the above challenge. In the future, we attempt to improve the performance of networks in dynamic object scenarios, such as surgical instruments and deformable tissues. Further validation is needed to apply our method in actual surgical scenarios. Animal studies with pigs will be done in the future. Pigs' gut environment and structure resemble those of humans. We attempt to extend the method in this study for use in human research after conducting animal tests.

References

[1] T. Collins et al., "Augmented reality guided laparoscopic surgery of the uterus," *IEEE Trans. Med. Imag.*, vol. 40, no. 1, pp. 371–380, Jan. 2021, doi: 10.1109/TMI.2020.3027442.

[2] P. Zhang et al., "Real-time navigation for laparoscopic hepatectomy using image fusion of preoperative 3D surgical plan and intraoperative indocyanine green fluorescence imaging," *Surgical Endoscopy*, vol. 34, no. 8, pp. 3449–3459, Aug. 2020, doi: 10.1007/s00464-019-07121-1.

[3] R. Hussain, A. Lalande, R. Marroquin, K. B. Girum, C. Guigou, and A. B. Grayeli, "Real-time augmented reality for ear surgery," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, Sep. 2018, pp. 324–331, doi: 10.1007/978-3-030-00937-3_38.

[4] H. Luo et al., "Augmented reality navigation for liver resection with a stereoscopic laparoscope," *Comput. Methods Programs Biomed.*, vol. 187, Apr. 2020, Art. no. 105099, doi: 10.1016/j.cmpb.2019.105099.

[5] R. Wei et al., "Stereo dense scene reconstruction and accurate localization for learning-based navigation of laparoscope in minimally invasive surgery," *IEEE Trans. Biomed. Eng.*, vol. 70, no. 2, pp. 488–500, Feb. 2023, doi: 10.1109/TBME.2022.3195027.

[6] S. Shao et al., "Self-supervised monocular depth and ego-motion estimation in endoscopy: Appearance flow to the rescue," *Med. Image Anal.*, vol. 77, Apr. 2022, Art. no. 102338, doi: 10.1016/j.media.2021.102338.

[7] Y. Li et al., "SuPer: A surgical perception framework for endoscopic tissue manipulation with surgical robotics," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 2294–2301, Apr. 2020, doi: 10.1109/LRA.2020.2970659.

[8] H. Itoh et al., "Binary polyp-size classification based on deep-learned spatial information," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 16, no. 10, pp. 1817–1828, Oct. 2021, doi: 10.1007/s11548-021-02477-z.

[9] R. Tang et al., "Augmented reality technology for preoperative planning and intraoperative navigation during hepatobiliary surgery: A review of current methods," *Hepatobiliary Pancreatic Diseases Int.*, vol. 17, no. 2, pp. 101–112, Apr. 2018, doi: 10.1016/j.hbpd.2018.02.002.

[10] D. Psychogyios, E. Mazomenos, F. Vasconcelos, and D. Stoyanov, "MSDESIS: Multitask stereo disparity estimation and surgical instrument segmentation," *IEEE Trans. Med. Imag.*, vol. 41, no. 11, pp. 3218–3230, Nov. 2022, doi: 10.1109/TMI.2022.3181229.

[11] S. Rattanalappaiboon, T. Bhongmakapat, and P. Ritthipravat, "Fuzzy zoning for feature matching technique in 3D reconstruction of nasal endoscopic images," *Comput. Biol. Med.*, vol. 67, pp. 83–94, Dec. 2015, doi: 10.1016/j.compbiomed.2015.09.021.

[12] Ó. G. Grasa, E. Bernal, S. Casado, I. Gil, and J. M. M. Montiel, "Visual SLAM for handheld monocular endoscope," *IEEE Trans. Med. Imag.*, vol. 33, no. 1, pp. 135–146, Jan. 2014, doi: 10.1109/TMI.2013.2282997.

[13] M. Ye, S. Giannarou, A. Meining, and G.-Z. Yang, "Online tracking and retargeting with applications to optical biopsy in gastrointestinal endoscopic examinations," *Med. Image Anal.*, vol. 30, pp. 144–157, May 2016, doi: 10.1016/j.media.2015.10.003.

[14] 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. (CVPR)*, Jul. 2017, pp. 6602–6611, doi: 10.1109/CVPR.2017.699.

[15] Q. Xu, Y. Li, M. Zhang, and W. Li, "COCO-Net: A dual-supervised network with unified ROI-loss for low-resolution ship detection from optical satellite image sequences," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022, Art. no. 5629115, doi: 10.1109/TGRS.2022.3201530.

[16] M. Turan et al., "Unsupervised odometry and depth learning for endoscopic capsule robots," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2018, pp. 1801–1807, doi: 10.1109/IROS.2018.8593623.

[17] 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.

[18] K. B. Ozyoruk et al., "EndoSLAM dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos," *Med. Image Anal.*, vol. 71, Jul. 2021, Art. no. 102058, doi: 10.1016/j.media.2021.102058.

[19] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Dec. 2017, vol. 30, pp. 6000–6010, doi: 10.48550/arXiv.1706.03762.

[20] W. Yu et al., "MetaFormer is actually what you need for vision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10809–10819, doi: 10.1109/CVPR52688.2022.01055.

[21] J. Bae, S. Moon, and S. Im, "Deep digging into the generalization of self-supervised monocular depth estimation," presented at the Proc. AAAI Conf. Artif. Intell., 2023, doi: 10.48550/arXiv.2205.11083.

[22] Q. Xu, Y. Li, J. Nie, Q. Liu, and M. Guo, "UPanGAN: Unsupervised pansharpening based on the spectral and spatial loss constrained generative adversarial network," *Inf. Fusion*, vol. 91, pp. 31–46, Mar. 2023, doi: 10.1016/j.inffus.2022.10.001.

[23] C. Godard, O. M. Aodha, M. Firman, and G. Brostow, "Digging into self-supervised monocular depth estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3827–3837, doi: 10.1109/ICCV.2019.00393.

[24] X. Liu et al., "Dense depth estimation in monocular endoscopy with self-supervised learning methods," *IEEE Trans. Med. Imag.*, vol. 39, no. 5, pp. 1438–1447, May 2020, doi: 10.1109/TMI.2019.2950936.

[25] D. Recasens, J. Lamarca, J. M. Fácil, J. M. M. Montiel, and J. Civera, "Endo-depth-and-motion: Reconstruction and tracking in endoscopic videos using depth networks and photometric constraints," *IEEE Robot. Autom. Lett.*, vol. 6, no. 4, pp. 7225–7232, Oct. 2021, doi: 10.1109/LRA.2021.3095528.

[26] L. Li, X. Li, S. Yang, S. Ding, A. Jolfaei, and X. Zheng, "Unsupervised-learning-based continuous depth and motion estimation with monocular endoscopy for virtual reality minimally invasive surgery," *IEEE Trans. Ind. Informat.*, vol. 17, no. 6, pp. 3920–3928, Jun. 2021, doi: 10.1109/TII.2020.3011067.

[27] Y. Zhang et al., "Colde: A depth estimation framework for colonoscopy reconstruction," Nov. 2021, *arXiv:2111.10371.*, doi: 10.48550/arXiv.2111.10371.

[28] Y. Liu and S. Zuo, "Self-supervised monocular depth estimation for gastrointestinal endoscopy," *Comput. Methods Programs Biomed.*, vol. 238, Aug. 2023, Art. no. 107619, doi: 10.1016/j.cmpb.2023.107619.

[29] Y. Yang et al., "A geometry-aware deep network for depth estimation in monocular endoscopy," *Eng. Appl. Artif. Intell.*, vol. 122, Jun. 2023, Art. no. 105989, doi: 10.1016/j.engappai.2023.105989.

[30] A. Varma, H. Chawla, B. Zonooz, and E. Arani, "Transformers in self-supervised monocular depth estimation with unknown camera intrinsics," Feb. 2022, *arXiv:2202.03131*.

[31] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12159–12168, doi: 10.1109/ICCV48922.2021.01196.

[32] S. Farooq Bhat, I. Alhashim, and P. Wonka, "AdaBins: Depth estimation using adaptive bins," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4008–4017, doi: 10.1109/CVPR46437.2021.00400.

[33] G. Yang, H. Tang, M. Ding, N. Sebe, and E. Ricci, "Transformer-based attention networks for continuous pixel-wise prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 16249–16259, doi: 10.1109/ICCV48922.2021.01596.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[35] Z. Chen et al., "Vision transformer adapter for dense predictions," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Feb. 2023, doi: 10.48550/arXiv.2205.08534.

[36] Z. Li, Z. Chen, X. Liu, and J. Jiang, "DepthFormer: Exploiting long-range correlation and local information for accurate monocular depth estimation," 2022, *arXiv:2203.14211*.

[37] C. Zhao et al., "MonoViT: Self-supervised monocular depth estimation with a vision transformer," in *Proc. Int. Conf. 3D Vis. (3DV)*, Sep. 2022, pp. 668–678, doi: 10.1109/3DV57658.2022.00077.

[38] Y. Lee, J. Kim, J. Willette, and S. J. Hwang, "MPViT: Multi-path vision transformer for dense prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 7277–7286, doi: 10.1109/CVPR52688.2022.00714.

[39] X. Lyu et al., "HR-depth: High resolution self-supervised monocular depth estimation," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 3, pp. 2294–2301, doi: 10.1609/aaai.v35i3.16329.

[40] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Jan. 2021, doi: 10.48550/arXiv.2010.11929.

[41] N. Zhang, F. Nex, G. Vosselman, and N. Kerle, "Lite-mono: A lightweight CNN and transformer architecture for self-supervised monocular depth estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 18537–18546, doi: 10.48550/arXiv.2211.13202.

[42] J. Guo et al., "CMT: Convolutional neural networks meet vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12165–12175, doi: 10.1109/CVPR52688.2022.01186.

[43] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," Jun. 2016, *arXiv:1606.08415*, doi: 10.48550/arXiv.1606.08415.

[44] A. Ali et al., "XCiT: Cross-covariance image transformers," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 34, Dec. 2021, pp. 20014–20027, doi: 10.48550/arXiv.2106.0968.

[45] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, Nov. 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.

[46] Z. Zhou, X. Fan, P. Shi, and Y. Xin, "R-MSFM: Recurrent multi-scale feature modulation for monocular depth estimating," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12757–12766, doi: 10.1109/ICCV48922.2021.01254.

[47] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004, doi: 10.1109/TIP.2003.819861.

[48] M. Allan et al., "Stereo correspondence and reconstruction of endoscopic data challenge," 2021, *arXiv:2101.01133*.

[49] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Proc. Int. Conf. Neural Inf. Process. Syst. (NIPS)*, Dec. 2014, pp. 2366–2374, doi: 10.48550/arXiv.1406.2283.

[50] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Dec. 2018, doi: 10.48550/arXiv.1711.05101.

[51] 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.

[52] Z. Fang, X. Chen, Y. Chen, and L. Van Gool, "Towards good practice for CNN-based monocular depth estimation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1080–1089, doi: 10.1109/WACV45572.2020.9093334.

[53] J. Spencer, R. Bowden, and S. Hadfield, "DeFeat-Net: General monocular depth via simultaneous unsupervised representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14390–14401, doi: 10.1109/CVPR42600.2020.01441.

[54] J. Bian et al., "Unsupervised scale-consistent depth and ego-motion learning from monocular video," in *Proc. 33rd Conf. Neural Inf. Process. Syst.*, Dec. 2019, pp. 35–45, doi: 10.48550/arXiv.1908.10553.

[55] B. Curless and M. Levoy, "A volumetric method for building complex models from range images," in *Proc. 23rd Annu. Conf. Comput. Graph. Interact. Techn.*, Aug. 1996, pp. 303–312, doi: 10.1145/237170.237269.

[56] Q.-Y. Zhou, J. Park, and V. Koltun, "Open3D: A modern library for 3D data processing," Jan. 2018, *arXiv:1801.09847*, doi: 10.48550/arXiv.1801.09847.

[57] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," *Commun. ACM*, vol. 65, no. 1, pp. 99–106, Dec. 2021, doi: 10.1145/3503250.