2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

Dense Depth Completion Based on Multi-Scale Confidence and Self-Attention Mechanism for Intestinal Endoscopy

Ruyu Liu[1,2], Zhengzhe Liu[1], Haoyu Zhang[1], Guodao Zhang[3], Zhigui Zuo[4], Weiguo Sheng[1,✉],
School of Information Science and Technology, Hangzhou Normal University, Hangzhou, 311121, China[1]
Haixi Institutes, Chinese Academy of Sciences Quanzhou Institute of Equipment Manufacturing, Quanzhou, 362000, China[2]
Department of Digital Media Technology, Hangzhou Dianzi University, Hangzhou, 310018, China[3]
Department of Colorectal Surgery, the First Affiliated Hospital of Wenzhou Medical University, Wenzhou, 325035, China[4]

***Abstract*—Doctors perform limited one-way intestine endoscopy, in which advanced surgical robots with depth sensors, such as stereo and ToF endoscopes, can only provide sparse and incomplete depth information. However, dense, accurate and instant depth estimation during endoscopy is vital for doctors to judge the 3D location and shape of intestinal tissues, which affects the human-robot interaction between doctors and surgical robots, such as the operation on the subsequent moving of the probe. In this paper, we present a deep learning-based dense depth completion method for intestine endoscopy. We utilize the scattered depth information from depth sensors to make up for the deficiency of features in the intestine and design a multi-scale confidence prediction network to extract dense geometric depth features. Then, we introduce the structure awareness module based on the self-attention mechanism in the depth completion network to enhance the geometry and texture features of the intestine. We also present a virtual multi-modal RGBD intestine dataset and conduct comprehensive experiments on a total of three intestine datasets. The experimental results clearly demonstrate that our method achieves better results in all metrics in all intestinal environments compared to state-of-the-art methods.**

***Index Terms*—Endoscopy, depth completion, self-attention mechanism, human-robot interaction**

I. Introduction

According to global cancer statistics 2020 [1], there are 1,880,725 new cases of colon and rectum cancer and 915,880 deaths in the world in 2020. Endoscopy, as the gold standard to diagnose intestinal diseases, is an important medical interactive way. However, the general operation of endoscopy is limited in one-way. The received observation often stays in the 2D image level and can't display and locate 3D information such as lesions, blood vessels, and adjacent tissues. Therefore, instantly obtaining accurate depth information corresponding to intestinal images is the key to improving endoscopy and diagnosis. With the development of endoscopic equipments and 3D vision algorithms, the depth estimation of the intestinal endoscopes mainly relies on deep learning (DL)-based depth estimation technology [2]–[4], SFM (Structure From Motion)/SLAM (Simultaneous Localization And Mapping) technology [5]–[7] and depth sensors [8].

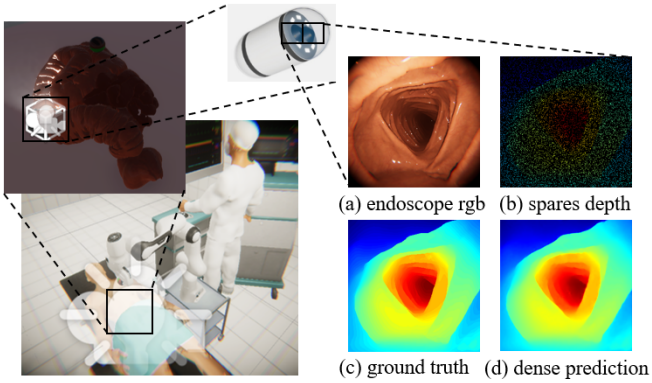

Fig. 1. Dense depth completion method for the intestine endoscope. (a) Endoscope RGB image (b) Sparse depth map (c) Ground-truth depth map (d) Dense depth map predicted by the proposed method.

The DL-based depth estimation method mainly designs an end-to-end neural network model, to predict dense depth maps. The DL models can be basically divided into two types. The first is depth regression which obtains dense depth maps directly through supervised learning of single image features [9]–[16]. Depth regression methods have recently been applied to the in-body cavities environment to get dense depth predictions [14], [15]. However, the resulting depth maps are heavily data-driven since drastic intensity variation and weak texture characteristics exist in lumen environments. The generated depth maps suffer from scale ambiguity due to a lack the utilization of environmental geometric features. The second is the depth completion methods [17]–[26], which try to reconstruct the depth map by combining images

Weiguo Sheng is the corresponding author, w.sheng@ieee.org. This work is supported in part by the National Natural Science Foundation of China under Grant 62202137; and in part by the Natural Science Foundation of Zhejiang Province under Grant LQ22F030004.

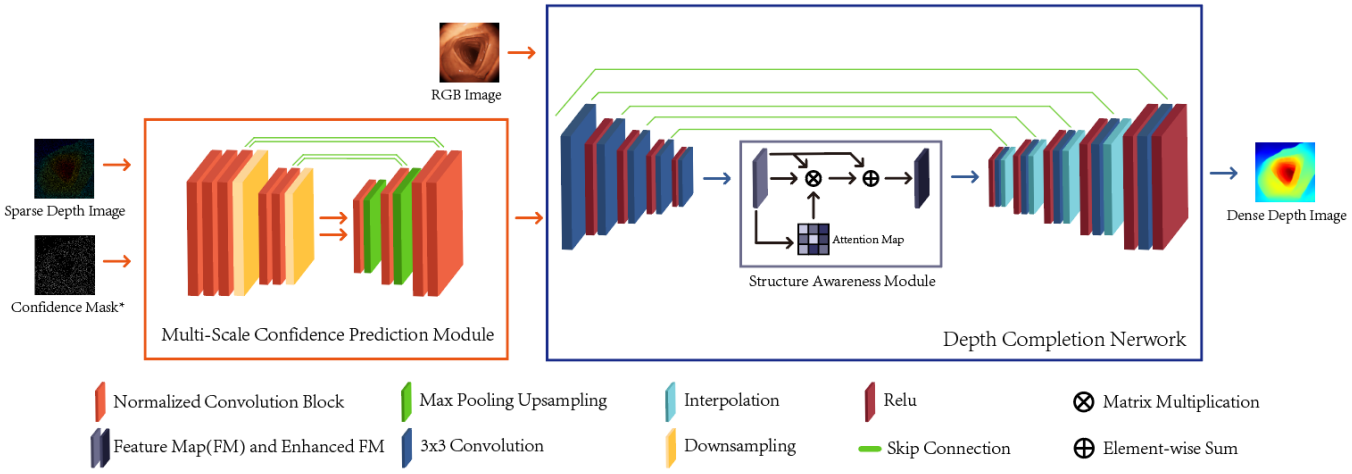

Fig. 2. **Overview of proposed network architecture.** The RGB image and the sparse depth map are simultaneously input into the network in one-to-one correspondence. The first MSC network, composed of normalized convolution, outputs the geometric confidence features. The obtained depth confidence features are contacted with RGB-SD and then input into the SA-based depth completion network to get the final dense depth map. *In the input confidence mask, the region with depth value is set to 1, and the region without depth value is set to 0.

and sparse depth information from SFM/SLAM algorithms or depth sensors. However, most methods simply concatenate or add the images and sparse depth and do not explore the promotion of deep fusion of cross-modal features for depth prediction tasks. In addition, most current methods are towards indoor and outdoor human activity scenes. To the best of our knowledge, there is no depth completion method designed for the intestine, even other in-body cavity environments.

There exists sparse and incomplete depth information in intestine endoscope from SFM/SLAM algorithms [27]–[29] or 3D endoscope cameras [8], [30]. SFM/SLAM mainly relies on the feature points detection, matching and triangulation of multi-view images for depth estimation based on monocular or stereo endoscopy. After a series of strict algorithms, the constructed depth maps are highly sparse. Besides, most SFM/SLAM methods are designed for large environments, like indoor or outdoor scenes, which is significantly different from the lumen environments with fewer textures and non-Lambertian reflection. With the development of sensors, some advanced endoscopes have probes with depth sensors to perceive distance information. However, the light source on endoscopes produces highlighted and non-Lambertian surfaces, which makes the generated depth map incompleted and with holes [31].

In response to the above challenge, we present a DL framework for dense depth completion of the intestinal environment based on RGB endoscope images and sparse depth maps (RGB-SD). First, the original SD map with scattered depth values passes through the multi-scale confidence prediction module to obtain dense confidence maps. Then, the depth map and its confidence map are taken as the input to perform the depth completion under the guidance of the corresponding RGB images. After encoding, the coarse feature map passes through the structure awareness module to generate the enhanced feature map. Finally, the enhanced feature maps are entered into the decoding part to output the final dense depth map with rich and precise details.

The main contributions of this paper is as follows:

1) We propose a dense depth completion network for the intestinal system based on multi-scale confidence and self-attention mechanism. The accuracy of the proposed method outperforms the State-of-The-Art (SoTA) methods on the various intestine datasets. To the best of our knowledge, this is the first work that combines RGB images with sparse depth information to reconstruct a dense depth map for the intestine.
2) In response to the lack of texture in the intestinal environment, the proposed multi-scale confidence module diffuses geometry depth using confidence masks to extract dense geometry features effectively. The structure awareness module further enhances geometry and texture features by combining region-specific responses from different regions. Experiments demonstrate that the designed modules can effectively reduce depth estimation error.
3) We collect and propose a near-realistic RGBD intestine dataset using a virtual reality (VR) surgery simulator to verify the depth estimation performance of different methods. We believe the RGBD intestine dataset can also benefit the surgical robot and VR surgery community.

II. The Proposed MSCSA-Net

In this section, we first introduce the architecture of the proposed network. Then, we describe the multi-scale confidence (MSC) prediction module and structure awareness (SA) module in detail.

A. The Architecture of Network

MSCSA-Net is a depth completion network designed for the intestine endoscopes based on multi-scale confidence and self-attention mechanism. Aiming at the problem of the incompleteness of the depth measurements from the intestinal endoscopes, we adopt a multi-scale confidence method to obtain the dense depth features. Then, to enhance the extracted geometry and texture features, we introduce the SA module. The SA module can further improve the accuracy of depth completion.

The entire MSCSA-Net contains two parts, as shown in Fig. 2. The first MSC module is responsible for diffusing the valid depth in SD to generate dense geometric depth maps with confidence maps using multi-scale confidence layers. The second is the U-Net-based [32] depth completion network, which predicts the dense depth map using the depth feature maps from the MSC module together with associated RGB image input. We specially design the SA module between the encoder and decoder of the network to fuse texture and geometry features in the deep layer. The depth completion network makes the influence of sparse depth more accurate and efficient under the guidance of the RGB image features.

B. Multi-Scale Confidence Prediction Network

Image-based depth estimation often ignores the geometric information scattered in the environment. However, directly inputting RGB image with SD into the CNN network produce an inaccurate depth map with blurring edges [33]. Thereby, we first design the MSC prediction network for mining the potential geometric features and filling the holes in the SD map by diffusing the valid depth geometric value to the neighboring pixels. Finally, a dense and complete depth feature map with uniform distribution with the corresponding RGB image is obtained, which benefits subsequent RGB-SD fusion.

The MSC network adopts the encode-decode structure and replaces the standard convolution layers in the encode-decode framework with normalized convolution (NConv) layers. NConv [25] utilizes the valid pixel to represent its neighborhood pixels by the confidence-equipped signal theory, so as to achieve pixel diffusion and densification. The advantage of NConv is that it computes the applicability of the result of each level and passes it to the next layer. NConv accepts the SD map Z and its corresponding confidence map C at the same time.

The forward propagation of the depth feature and confidence feature are formula (1) and (2) respectively:

$$Z_{i,j}^{l} = \frac{\sum_{m,n} Z_{i+m,j+n}^{l-1} C_{i+m,j+n}^{l-1} \Gamma(W_{m,n}^{l})}{\sum_{m,n} C_{i+m,j+n}^{l-1} \Gamma(W_{m,n}^{l}) + \epsilon} + b^{l}, \quad (1)$$

$$C_{i,j}^{l} = \frac{\sum_{m,n} C_{i+m,j+n}^{l-1} \Gamma(W_{m,n}^{l}) + \epsilon}{\sum_{m,n} \Gamma(W_{m,n}^{l})}, \quad (2)$$

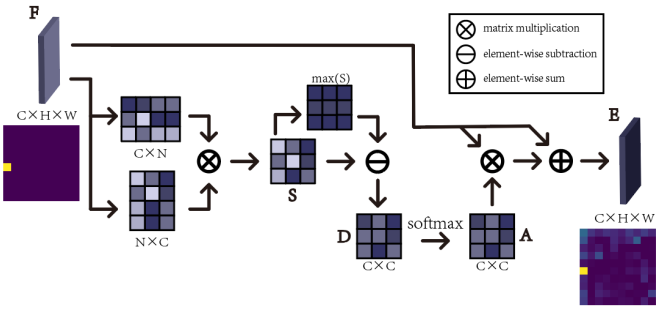

Fig. 3. The diagram of the structure awareness module. F and E are examples of channel features. N represents the size of the feature map, which value is $H \times W$

where l represents the number of the NConv layer. b^{l} is the bias. To prevent the base from being zero, a constant ϵ is added. $W_{m,n}^{l}$ is the applicability of the current result calculated by NConv. $\Gamma(\cdot) = \log(1 + \exp(\cdot))$ preserves the surface trend of W while transforming the codomain to nonnegative values.

The back-propagation of the MSC network is modified to formula (3) based on the Γ function:

$$\frac{\partial E}{\partial W_{m,n}^{l}} = \sum_{i,j} \frac{\partial E}{\partial Z_{i,j}^{l}} \cdot \frac{\partial Z_{i,j}^{l}}{\partial \Gamma(W_{m,n}^{l})} \cdot \frac{\partial \Gamma(W_{m,n}^{l})}{\partial W_{m,n}^{l}}, \quad (3)$$

where E represents the loss between the predicted depth map and ground truth.

Multi-scale module consisting of NConv produces the credibility feature corresponding to the depth map under each kind of receptive field, so as to obtain more rich geometric feature maps and benefit the depth estimation task.

C. Depth Completion Network

The features and confidence mask from the MSC network and the corresponding RGB images are fused and input into the U-Net-based depth completion network to estimate dense depth. The traditional U-Net network is a typical non-discriminatory network [32], which does not enhance any feature information. In the depth estimation, the feature map in each channel can be regarded as the response of a specific region. The responses of these specific regions are correlated with each other. Suppose each channel map captures more distinct region responses from all other channel maps. In that case, it will gain more relative depth information from farther regions and significantly enhance the perception of the scene structure. Therefore, we introduce a SA module based on the self-attention mechanism [34] in the depth completion network, which can model the interdependence between different regional responses.

The similarity between different channel maps reflects the spatial relationship of regional responses, that is, two feature maps with high similarity have stronger responses in the same region. The SA module strengthens these corresponding responses by performing a weighted sum of each graph with the original features of other channel graphs. Finally, the

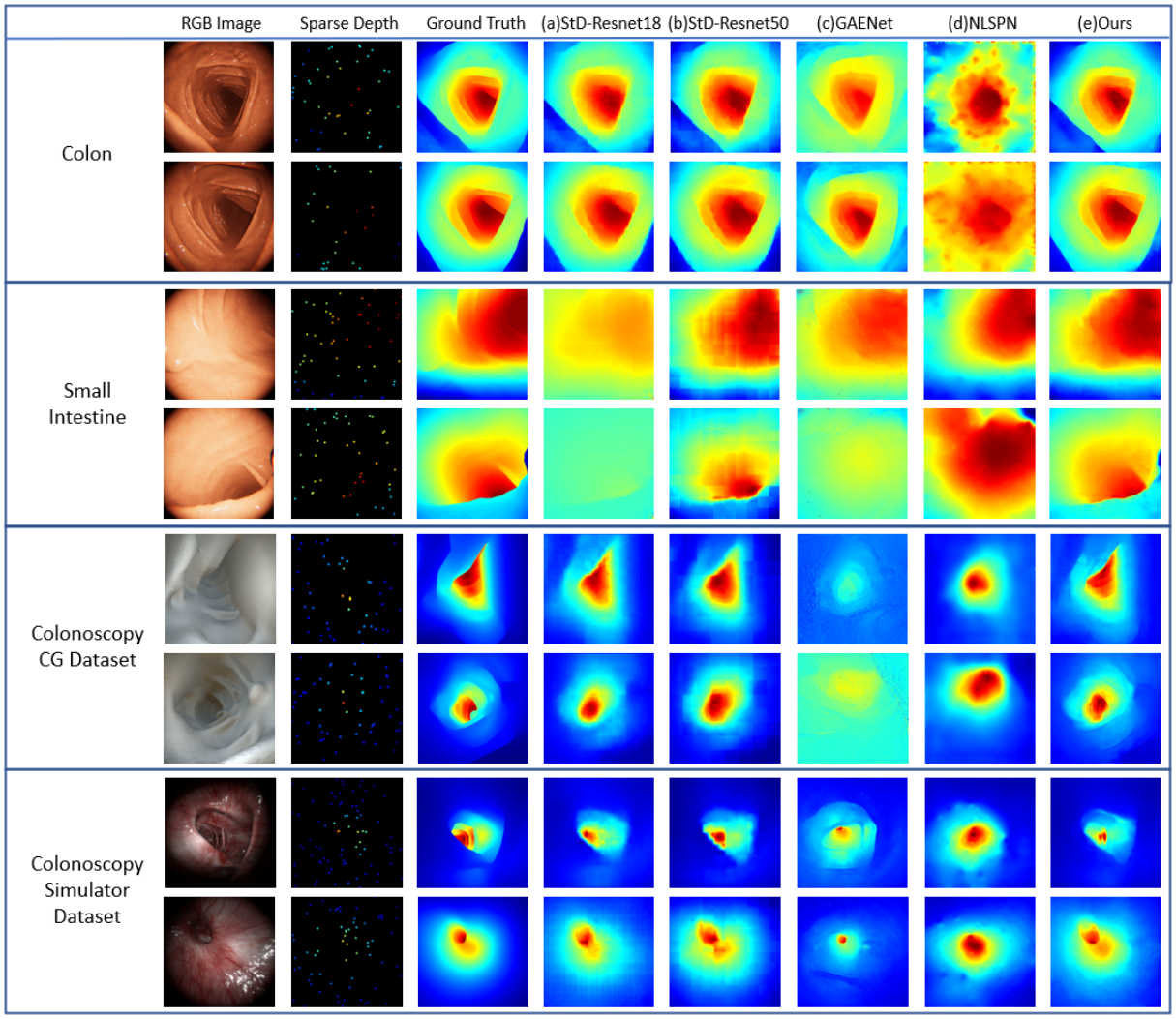

Fig. 4. Qualitative comparison between our method with SoTA depth completion methods on various RGBD intestine datasets. For visualization purposes, the sparse measurements in SD are dilated. The depth maps from our method are more explicit and close to the ground truth in all datasets.

original feature map is added to further enhance the strength of the response and strengthen the extracted features. By SA, we can get the aggregated features that encode the rich contextual information of the scene structure.

III. Experiments

To demonstrate the effectiveness of our method, we construct a new RGBD (RGB image and depth map) intestine dataset and conduct a series of experimental comparisons and ablation studies on the proposed dataset and other two public intestine datasets.

A. RGBD Intestine Dataset

We conducted SoTA comparison experiments on three RGBD intestine datasets.

1) VR-Caps Dataset: In this work, we propose an RGBD intestine dataset collected by VR-Caps [35], which is a virtual capsule endoscopy platform developed by Unity. In VR-Caps, a single-purpose virtual cap with an RGB camera and a depth camera is placed near the end of the virtual colon and then moves along the intestine from the colon to the small intestine at a constant speed of 0.0001. The parameters setting of the equipped camera are as follows: the ISO is 200, the focal length is 159.45, the FOV (Field-of-View) is 100, and the image size is 320 $\times$ 320. The generated dataset, from the colon and small intestine two environments, consists of RGB images, SD maps, and dense depth maps as ground truth (GT). In the colon scene, 15665 RGBD images are used as the training set and 1000 RGBD images as the validation set. In the small intestine scene, 7964 and 900 RGBD images are used as the training and validation sets, respectively.

2) Colonoscopy CG Dataset (CCD) [36]: The dataset is collected from a virtual gut environment based on real CT data. The dataset consists of three materials and two lighting conditions. We selected the intestinal environment with distinctive lighting and material conditions in CCD. The collected dataset is different from VR-Caps Dataset and

TABLE I
QUANTITATIVE COMPARISON WITH SOTA METHODS ON COLON AND SMALL INTESTINE DATASET. THE NUMBER OF SAMPLING POINTS IS 100. THE ERROR INDICATORS INCLUDE RMSE, ABS.REL AND MAE ARE ALL IN CENTIMETRES.

Dataset	Method	Error			Accuracy	
		RMSE	abs.REL	MAE	δ_1	δ_2
Colon	StD-Resnet18	1.0244	5.7732	0.7627	0.9770	0.9926
	StD-Resnet50	1.0403	5.7023	0.7878	0.9789	0.9942
	GAENet	**0.9605**	**4.8107**	0.7091	**0.9883**	**0.9980**
	NLSPN	1.1395	5.5340	**0.6582**	0.9603	0.9897
	Ours	**0.6158**	**3.0916**	**0.4097**	**0.9904**	**0.9982**
Small Intestine	StD-Resnet18	1.9058	5.2290	1.2930	0.9845	0.9987
	StD-Resnet50	1.8712	5.7232	1.4253	0.9887	**0.9992**
	GAENet	2.0940	5.8193	1.4615	0.9830	0.9981
	NLSPN	**1.2679**	**3.2163**	**0.7796**	**0.9904**	0.9975
	Ours	**0.8121**	**1.7241**	**0.4244**	**0.9954**	**0.9995**
Colonoscopy CG Dataset (CCD)	StD-Resnet18	**1.9789**	**7.9302**	**1.0927**	**0.9530**	**0.9928**
	StD-Resnet50	2.0251	7.9456	1.1552	0.9417	0.9922
	GAENet	5.9393	38.0345	3.9816	0.5762	0.8082
	NLSPN	2.7441	9.7676	1.4269	0.9111	0.9758
	Ours	**1.3501**	**5.1058**	**0.7650**	**0.9803**	**0.9971**
Colonoscopy Simulator Dataset (CSD)	StD-Resnet18	**32.7593**	**6.7433**	**18.0611**	**0.9641**	**0.9956**
	StD-Resnet50	38.4818	7.1197	18.8038	0.9504	0.9926
	GAENet	41.7737	11.0334	25.8246	0.8828	0.9845
	NLSPN	42.4963	7.2895	19.4174	0.9355	0.9834
	Ours	**30.3093**	**4.9252**	**13.5245**	**0.9736**	**0.9961**

Colonoscopy Simulator Dataset. In our work, 1820 images are used as the training set and 364 images are used as the validation set.

3) Colonoscopy Simulator Dataset (CSD) [29]: [29] provides a colonoscopy simulator that can generate RGB and depth images by colonoscopy. We use this simulator to collect 1754 RGBD data, of which 1579 are used as the training set and 175 as the validation set. The focal length is 2.162mm, and the image resolution is 320 $\times$ 320. We use this dataset to test the performance of different methods under conditions with poor lighting conditions and rich vascular texture features.

B. Experimental Setup

All experiments are performed on a workstation with Intel Core i9-10900X with 10 CPU cores, 64 GB of RAM, and an NVIDIA Quadro RTX 4000 GPU with 8 GB of memory. We use ubuntu 20.04.4 LTS during the experiment and CUDA Version is 11.6. All guided networks are trained until convergence on the full training set with a batch size of 10. The number of epochs for training is 10. We use the ADAM optimizer with an initial learning rate of 10^{-3} and a decaying factor of 10. L1 loss is adopted in our experiments. Our code and dataset will be made available at *https://github.com/liubiye/MSCSANet*.

C. Qualitative and Quantitative Comparison with the SoTA Methods

To evaluate the performance of MSCSA-Net, we conduct comparison experiments on all three datasets. We select some representative SoTA depth completions, such as StD-resnet18 [33], StD-resnet50 [33], NLSPN [19] and GAENet [22], and use their open-source codes to train the models with the same environmental configurations. The qualitative and quantitative comparisons are shown in Fig. 4 and Table I, respectively. In the experiment, the error metrics (RMSE, abs.REL, and MAE [37], [38]) and accuracy metrics (δ_1, δ_2) are used to evaluate the performance of depth estimation.

From the qualitative results in Fig. 4, StD-restnet18, GAENet, and NLSPN produce abnormal results in the small intestine environments since these scenes lack obvious texture features. The depth maps from StD have more blur edges than our method. Because StD simply treats the region with missing values in the depth map to zero values and then roughly concatenates the RGB image and the sparse depth map. GAENet relies on the geometric perception of the scene for depth estimation. Thus it performs second only to ours in the colon environment with a relatively obvious structure. However, the generated depth maps have lighter overall color in other environments with fewer geometric features, which means the predicted results have large errors. The depth maps from NLSPN only predict the approximate and rough depth change direction, and the details and edge information are incorrect. In contrast, our method produces more accurate depth maps close to GT with clear edges and details on all datasets. In the CSD, the depth estimated by all methods has different degrees of scale shift due to poor illumination.

From the quantitative numerical results, all experimental methods performed better in the more-textured colon than in other datasets, which is consistent with the qualitative results. In the CCD, the accuracy of the GAENet method is only 0.5762 while our accuracy can still be maintained at a high level with 0.9803. The value of error indicators in CSD is large since the GT value is large. Our method still achieves a high accuracy of 0.9736 in the CSD. Compared with StD-resnet50 in the colon scene, the RMSE, abs. REL and MAE of our method are reduced by 40.8%, 45.8%, and 49.6%. In the small intestine, the reduction in error is more significant. RMSE, abs. REL and MAE of our method are reduced by 56.6%, 69.9%, and 70.2%. In comparison, the value of our δ_1 is improved by at least 0.03 over the other methods. Even compared to the best method in all datasets, there is an average 0.01 δ_1 improvement in all scenes, which indicates that our method outperforms other SoTA methods.

D. Effectiveness of the proposed MSCSA-Net

To demonstrate the effectiveness of MSCSA-Net, we design a series of experiments to explore the impact of varying degrees of sparsity and designed modules in the framework.

1) Different Sampling Number: We explore the influence of different sampling sparsity of depth points on depth completion. The experimental results from Fig. 5 illustrate that the performance of depth completion increases with the sampling sparsity increasing. We conduct depth completion experiments on colon environments in the VR-Caps dataset using two representative SD maps, one is uniform sampling-based SD maps (US-SD), and the other is features sampling-based SD maps (FS-SD). This change is steeper in US-SD while smoother in FS-SD. In detail, when the number of

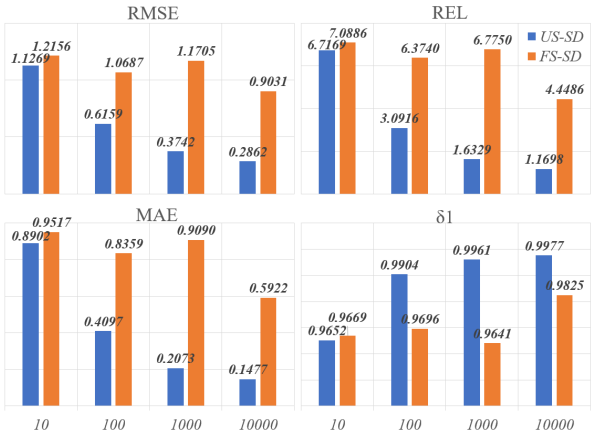

Fig. 5. Impact of varying sampling numbers on depth estimation. The input SD includes US-SD and FS-SD. Error metrics, including RMSE, MAE and abs. REL is lower the better, and the accuracy metric δ_1 is higher the better.

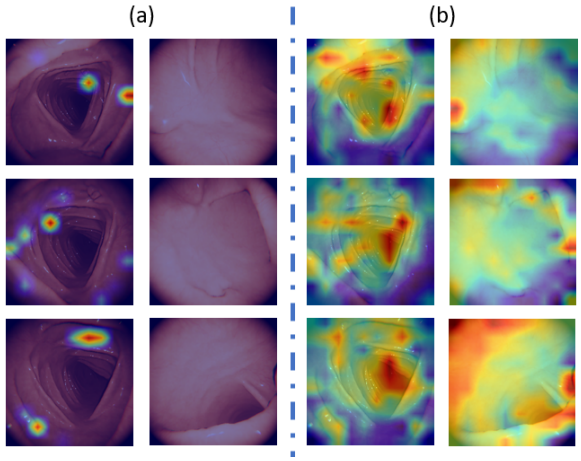

Fig. 6. The visualization of the SA module: (a) feature maps before SA enhancement (b) feature maps after SA enhancement.

TABLE II
THE ABLATION STUDY OF DESIGNED MODULES ON COLON AND SMALL INTESTINE DATASET. THE NUMBER OF SPARSE DEPTH POINTS IS 100. THE BASELINE METHOD IS BASED ON U-NET WITHOUT MSC AND SA MODULES. THE ERROR INDICATORS (RMSE, ABS.REL AND MAE) ARE ALL IN CENTIMETRES.

Scenes	Method	Error			Accuracy	
		RMSE	abs.REL	MAE	δ_1	δ_2
Colon	baseline	1.0160	6.1511	0.7977	0.9703	0.9966
	baseline + SA	0.9914	5.4719	0.7666	**0.9846**	**0.9974**
	baseline + MSC	**0.7099**	**4.2276**	**0.4892**	0.9762	0.9945
	full	**0.6158**	**3.0916**	**0.4097**	**0.9904**	**0.9982**
Small Intestine	baseline	1.9113	5.9114	1.4357	0.9844	0.9994
	baseline + SA	2.0225	6.2830	1.5497	0.9869	0.9994
	baseline + MSC	**0.8716**	**1.9867**	**0.4851**	**0.9954**	**0.9995**
	full	**0.8121**	**1.7241**	**0.4244**	**0.9954**	**0.9995**

sampling points is small (such as 10-1000), the accuracy improvement of FS-SD is not obvious compared with US-SD. However, when the number of sampling points is large enough (such as 10000), the prediction accuracy of FS-SD is still significantly improved. The reason is that there are relatively more depth pixels near the texture edge in FS-SD when sample points are fewer. The depth points located in the texture area can promote depth completion. With the increase of depth points, the increased depth values more focus on the region of texture edge in FS-SD, and the model cannot extra features from other regions, resulting in the prediction accuracy from the FS-SD being inferior to those from US-SD.

2) Ablation Study: In order to prove the positive effect of each module on depth estimation, we disassemble each module step by step and conduct ablation experiments, as shown in Table. II. In colon and small intestine scenes, the full method (full) is significantly better than the baseline approach. The most significant improvement of abs.REL is reduced by 49.7% and 70.8%, respectively. With MSC-only (baseline+MSC), the RMSE is reduced by 30.1% and 54.4% in two scenes, respectively. Even in the small intestine, the MAE is diminished by 38.7%. With SA-only (baseline+SA), the abs.REL is decreased by 11.0% compared to the baseline method in the colon. In the small intestine scene, the features enhanced by SA are also prone to be inaccurate due to the lack of geometric features provided by MSC. Therefore, the performance of SA-only is inferior to the baseline. Compared to MSC-only, the abs.REL of the full method is decreased by 26.9% and 13.2% in the colon and small intestine scenes. Our method (full) is superior to the MSC-only approach, demonstrating that the SA module can improve the overall accuracy of depth completion. These experimental results also indicate that the combination of these two modules can further improve the performance of depth completion.

In addition, we qualitatively analyze the effectiveness of the SA module by visualizing the extraction of intestinal features by the network before and after using SA. Fig. 6 shows that our model captures richer features after using SA, especially in the small intestine environment with more scarce textures. SA adds features to the edge and near and far regions, respectively. Therefore, it is also proved that the SA module can effectively improve the feature extraction ability of the network for the intestinal environment.

IV. CONCLUSION

In this paper, the MSCSA-Net is proposed to predict the dense depth map for intestinal scenes using RGB endoscope images with sparse depth measurements. The proposed MSC and SA modules can effectively extract and enhance the geometry and texture features from RGB-SD input. Benefiting from the MSC and SA modules in our network, the depth estimation performance can be significantly improved. Experimental results on the three RGBD intestine datasets show that the proposed MSCSA-Net can achieve effective and accurate depth completion in intestinal environments and outperforms SoTA methods. In the future, we hope to extend the deep completion into the SLAM systems to provide doctors with accurate pose estimation and 3D reconstruction of the intestine.

REFERENCES

[1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[2] X. Liu, A. Sinha, M. Ishii, G. D. Hager, A. Reiter, R. H. Taylor, and M. Unberath, "Dense depth estimation in monocular endoscopy with self-supervised learning methods," *IEEE Transactions on Medical Imaging*, vol. 39, no. 5, pp. 1438–1447, 2019.

[3] Y. Ming, X. Meng, C. Fan, and H. Yu, "Deep learning for monocular depth estimation: a review," *Neurocomputing*, vol. 438, pp. 14–33, 2021.

[4] J. Hu, C. Bao, M. Ozay, C. Fan, Q. Gao, H. Liu, and T. L. Lam, "Deep depth completion: a survey," *ArXiv Preprint ArXiv:2205.05335*, 2022.

[5] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "Orb-slam: a versatile and accurate monocular slam system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[6] R. Mur-Artal and J. D. Tardós, "Orb-slam2: an open-source slam system for monocular, stereo, and rgb-d cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[7] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 3, pp. 611–625, 2017.

[8] J. Geng and J. Xie, "Review of 3-d endoscopic surface imaging techniques," *IEEE Sensors Journal*, vol. 14, no. 4, pp. 945–960, 2013.

[9] J.-W. Bian, H. Zhan, N. Wang, Z. Li, L. Zhang, C. Shen, M.-M. Cheng, and I. Reid, "Unsupervised scale-consistent depth learning from video," *International Journal of Computer Vision*, vol. 129, no. 9, pp. 2548–2564, 2021.

[10] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6612–6619, 2017.

[11] X. Liu, A. Sinha, M. Unberath, M. Ishii, G. Hager, R. H. Taylor, and A. Reiter, "Self-supervised learning for dense depth estimation in monocular endoscopy," in *Lecture Notes in Computer Science*, vol. 11041, pp. 128–138, 2018.

[12] M. Hu, S. Wang, B. Li, S. Ning, L. Fan, and X. Gong, "PENet: towards precise and efficient image guided depth completion," in *IEEE International Conference on Robotics and Automation (ICRA)*, vol. 2021, pp. 13 656–13 662, 2021.

[13] K. Zhou, K. Yang, and K. Wang, "Panoramic depth estimation via supervised and unsupervised learning in indoor scenes," *Applied Optics*, vol. 60, no. 26, pp. 8188–8197, 2021.

[14] A. R. Widya, Y. Monno, M. Okutomi, S. Suzuki, T. Gotoda, and K. Miki, "Learning-based depth and pose estimation for monocular endoscope with loss generalization," *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 2021, pp. 3547–3552, 2021.

[15] K. B. Ozyoruk, G. I. Gokceler, T. L. Bobrow, G. Coskun, K. Incetan, Y. Almalioglu, F. Mahmood, E. Curto, L. Perdigoto, M. Oliveira, H. Sahin, H. Araujo, H. Alexandrino, N. J. Durr, H. B. Gilbert, and M. Turan, "EndoSLAM dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos," *Medical Image Analysis*, vol. 71, p. 102058, 2021.

[16] Z. Yin and J. Shi, "GeoNet: unsupervised learning of dense depth, optical Flow and camera pose," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1983–1992, 2018.

[17] W. Van Gansbeke, D. Neven, B. De Brabandere, and L. Van Gool, "Sparse and noisy LiDAR completion with RGB guidance and uncertainty," in *International Conference on Machine Vision Applications*, pp. 1–6, 2019.

[18] F. Ma, G. V. Cavalheiro, and S. Karaman, "Self-supervised sparse-to-dense: self-supervised depth completion from LiDAR and monocular camera," in *IEEE International Conference on Robotics and Automation (ICRA)*, vol. 2019, pp. 3288–3295, 2019.

[19] J. Park, K. Joo, Z. Hu, C.-K. Liu, and I. S. Kweon, "Non-local spatial propagation network for depth completion," in *European Conference on Computer Vision (ECCV)*, vol. 12358, pp. 120–136, 2020.

[20] Y. Chen, B. Yang, M. Liang, and R. Urtasun, "Learning joint 2D-3D representations for depth completion," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10 022–10 031, 2019.

[21] S. S. Shivakumar, T. Nguyen, I. D. Miller, S. W. Chen, V. Kumar, and C. J. Taylor, "DFuseNet: deep fusion of RGB and sparse depth information for image guided dense depth completion," in *IEEE Intelligent Transportation Systems Conference*, pp. 13–20, 2019.

[22] H. Chen, H. Yang, Y. Zhang *et al.*, "Depth completion using geometry-aware embedding," in *2022 International Conference on Robotics and Automation (ICRA)*, pp. 8680–8686, 2022.

[23] Y. Xu, X. Zhu, J. Shi, G. Zhang, H. Bao, and H. Li, "Depth completion from sparse LiDAR data with depth-normal constraints," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2811–2820, 2019.

[24] J. Qiu, Z. Cui, Y. Zhang, X. Zhang, S. Liu, B. Zeng, and M. Pollefeys, "DeepLiDAR: deep surface normal guided depth prediction for outdoor scene from sparse LiDAR data and single color image," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3308–3317, 2019.

[25] A. Eldesokey, M. Felsberg, and F. S. Khan, "Confidence propagation through CNNs for guided sparse depth regression," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 10, pp. 2423–2436, 2020.

[26] X. Liang and C. Jung, "AGNet: attention guided sparse depth completion using convolutional neural networks," *IEEE Access*, vol. 10, pp. 10 514–10 522, 2022.

[27] N. Mahmoud, I. Cirauqui, A. Hostettler, C. Doignon, L. Soler, J. Marescaux, and J. M. M. Montiel, "Orbslam-based endoscope tracking and 3d reconstruction," in *International Workshop on Computer-Assisted and Robotic Endoscopy*, pp. 72–83, 2016.

[28] N. Mahmoud, T. Collins, A. Hostettler, L. Soler, C. Doignon, and J. M. M. Montiel, "Live tracking and dense reconstruction for handheld monocular endoscopy," *IEEE Transactions on Medical Imaging*, vol. 38, no. 1, pp. 79–89, 2018.

[29] S. Zhang, L. Zhao, S. Huang, M. Ye, and Q. Hao, "A template-based 3d reconstruction of colon structures and textures from stereo colonoscopic images," *IEEE Transactions on Medical Robotics and Bionics*, vol. 3, no. 1, pp. 85–95, 2020.

[30] T. Köhler, S. Haase, S. Bauer, J. Wasza, T. Kilgus, L. Maier-Hein, H. Feußner, and J. Hornegger, "Tof meets rgb: novel multi-sensor super-resolution for hybrid 3-d endoscopy," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 139–146, 2013.

[31] D. Stoyanov, "Stereoscopic scene flow for robotic assisted minimally invasive surgery," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 479–486, 2012.

[32] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, 2015.

[33] F. Mal and S. Karaman, "Sparse-to-dense: depth prediction from sparse depth samples and a single image," in *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4796 – 4803, 2018.

[34] J. Yan, H. Zhao, P. Bu, and Y. Jin, "Channel-wise attention-based network for self-supervised monocular depth estimation," in *International Conference on 3D Vision (3DV)*, pp. 464–473, 2021.

[35] K. Incetan, I. O. Celik, A. Obeid, G. I. Gokceler, K. B. Ozyoruk, Y. Almalioglu, R. J. Chen, F. Mahmood, H. Gilbert, N. J. Durr, and M. Turan, "VR-Caps: a virtual environment for capsule endoscopy," *Medical Image Analysis*, vol. 70, p. 101990, 2021.

[36] A. Rau, P. E. Edwards, O. F. Ahmad, P. Riordan, M. Janatka, L. B. Lovat, and D. Stoyanov, "Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy," *International journal of computer assisted radiology and surgery*, pp. 1–10.

[37] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *Advances in Neural Information Processing Systems*, vol. 27, pp. 2366–2374, 2014.

[38] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, and A. Geiger, "Sparsity invariant cnns," in *International Conference on 3D Vision (3DV)*, pp. 11–20, 2017.