

3D Reconstruction of Whole Stomach from Endoscope Video Using Structure-from-Motion

Aji Resindra Widya¹, Yusuke Monno¹, Kosuke Imahori¹, Masatoshi Okutomi¹,
Sho Suzuki², Takuji Gotoda², and Kenji Miki³

Abstract—Gastric endoscopy is a common clinical practice that enables medical doctors to diagnose the stomach inside a body. In order to identify a gastric lesion’s location such as early gastric cancer within the stomach, this work addressed to reconstruct the 3D shape of a whole stomach with color texture information generated from a standard monocular endoscope video. Previous works have tried to reconstruct the 3D structures of various organs from endoscope images. However, they are mainly focused on a partial surface. In this work, we investigated how to enable structure-from-motion (SfM) to reconstruct the whole shape of a stomach from a standard endoscope video. We specifically investigated the combined effect of chromo-endoscopy and color channel selection on SfM. Our study found that 3D reconstruction of the whole stomach can be achieved by using red channel images captured under chromo-endoscopy by spreading indigo carmine (IC) dye on the stomach surface.

I. INTRODUCTION

Gastric endoscopy is a well-adopted procedure that enables medical doctors to diagnose the stomach inside a body. However, there still exists some challenges to doctors such as the limited point of view and the uncertainty of endoscope poses relative to a target organ. The accurate localization of a malignant lesion within the global view of the whole stomach is crucial for gastric surgeons to decide the operative procedure of the laparoscopic gastrectomy for early gastric cancer. The location of the malignant lesion is usually identified by the double contrast barium radiography [1]. However, morphological evaluations such as barium study sometimes cause the gastric surgeons difficulty in identifying flat malignant lesions. Recently, 3D computed tomography (CT) gastrography was developed for the lesion localization purpose [2]. However, 3D CT gastrography does not embed color texture information to the reconstructed 3D model. If the 3D shape of the whole stomach can be reconstructed from a standard endoscopic video, the location of the malignant lesion can be identified by the visual color information in addition to the 3D morphological information, which should be very valuable for the gastric surgeons.

This work was partly supported by JSPS KAKENHI Grant Number 17H00744.

¹A. R. Widya, Y. Monno, K. Imahori, and M. Okutomi are with the Department of Systems and Control Engineering, School of Engineering, Tokyo Institute of Technology, Meguro-ku, Tokyo 152-8550, Japan (e-mail: aresindra@ok.sc.e.titech.ac.jp; ymonno@ok.sc.e.titech.ac.jp; kima-hori@ok.sc.e.titech.ac.jp; mxo@sc.e.titech.ac.jp).

²S. Suzuki and T. Gotoda are with the Division of Gastroenterology and Hepatology, Department of Medicine, Nihon University School of Medicine, Chiyoda-ku, Tokyo 101-8309, Japan.

³K. Miki is with the Department of Internal Medicine, Tsujinaka Hospital Kashiwanoha, Kashiwa-city, Chiba 277-0871, Japan.

Previous studies have shown that 3D endoscopy systems (e.g., a stereo endoscope) have advantages over traditional 2D endoscopes in fields such as computer-aided laparoscopic surgery [3] and endoscopic surface imaging [4]. Nevertheless, those 3D systems are not widely available and the 2D counterpart is still the mainstream.

Some existing works have proposed a software solution to reconstruct the 3D structure of a target organ (e.g., colon, liver, and larynx) with the estimated endoscope poses from an endoscope video. The methods are ranging from shape-from-shading (SfS) [5], [6], visual simultaneous localization and mapping (SLAM) [7]–[9], and structure-from-motion (SfM) [10]–[15]. Even though SfS can reconstruct an organ’s surface from a single image, it requires accurate estimation of the light position, which is a difficult problem. SLAM offers a real-time solution with the reconstruction quality as a trade-off. SLAM uses a simple feature detector and descriptor and also sequential feature matching, which leads to a limited reconstruction quality. On the other hand, SfM offers an off-line solution with higher reconstruction quality. SfM uses a more accurate feature detector and descriptor to obtain higher quality features. Moreover, SfM can exhaustively use all input images to find feature correspondences and perform global reconstruction optimization applying bundle adjustment. However, since SfM relies on the detected features, it is still challenging to reconstruct texture-less surfaces, which are common in internal organs. To tackle this challenge, some systems [13], [14] exploit a projector to add a structured light pattern on the texture-less surface. Although these systems can successfully increase the number of features, they require expensive hardware modification. Enhanced imaging colonoscopy and narrow-band imaging were also applied to enhance the surface details for SfM [15]. The above-mentioned works only demonstrated the reconstruction results of a partial surface, which is not sufficient for many potential applications such as the 3D localization of a lesion within the whole shape of the organ.

In this work, we aimed at reconstructing the 3D model of a whole stomach with color texture information from a standard endoscopic video using SfM. We specifically investigated the combined effect of chromo-endoscopy and color channel selection on SfM to achieve better reconstruction quality. Our study found that 3D reconstruction of the whole stomach can be achieved by using red channel images captured under chromo-endoscopy by spreading indigo carmine (IC) dye onto the stomach surface. To the best of our knowledge, this is the first paper to report a successful

3D reconstruction of a whole stomach and visualize the color details of the mucosal surface of it by texture mapping generated from the standard monocular endoscopy video.

We also demonstrate our custom viewer that can visualize a particular image frame's location in the 3D model.

II. MATERIALS AND METHODS

In this section, we briefly describe the data collection and the 3D reconstruction method. We first explain our endoscopy hardware setup and the captured video sequences information (Section II-A). Then, we explain each component of our method starting from the input images extraction for SfM (Section II-B), the SfM pipeline (Section II-C), and the mesh and texture generation (Section II-D).

A. Data collection

This study was conducted in accordance with the Declaration of Helsinki. The Institutional Review Board at Nihon University Hospital approved the study protocol on March 8, 2018, before patient recruitment. Informed consent was obtained from all patients before they were enrolled. This study was registered with the University Hospital Medical Information Network (UMIN) Clinical Trials Registry (identification No.: UMIN000031776) on March 17, 2018. This study was also approved by the research ethics committee of Tokyo Institute of Technology, where 3D reconstruction experiments were conducted.

We captured the endoscopy video using a standard endoscopy system. We used an Olympus IMH-20 image management hub coupled with a GIF-H290 scope. To prevent any compression and unwanted artifacts such as image interlacing, we used an Ehipan video grabber to capture unprocessed data from the image management hub. The video data was saved as an AVI format in 30 frames per second with full HD resolution.

The videos used for 3D reconstruction were captured on three different subjects undergoing general gastrointestinal endoscopy. As shown in Figure 1 (a) and (b), each video contains two image sequences captured without and with spraying the IC blue color dye onto the stomach surface as chromo-endoscopy, which is widely applied in endoscopy to enhance the surface visualization. For the dye, we used $C_{16}H_8N_2Na_2O_8S_2$ manufactured by Daiichi Sankyo Company, Limited, Tokyo, Japan. Additionally, we captured images of a planar checkerboard pattern from multiple orientations for the camera calibration purpose.

B. Pre-processing of the collected data

The pre-processing of the collected data is performed to estimate intrinsic camera parameters and to extract input images for SfM. This process includes camera calibration, frame extraction, and color channel separation as follows.

An endoscopy camera generally uses an ultra-wide lens to provide a large angle of view inside the stomach. As a trade-off, the ultra wide lens introduces a strong visual distortion and produces images with a convex non-rectilinear

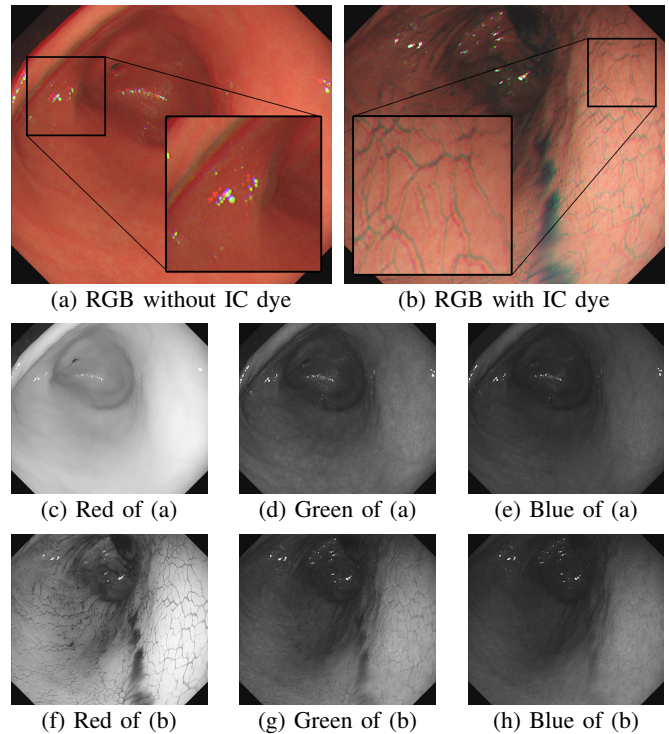


Fig. 1: Examples of endoscope images captured without IC dye (a) and with IC dye (b). The color channel misalignment problem is observed in (a) and (b). The images (c)-(h) are six channel images extracted from (a) and (b). We can observe that the IC dye adds textures on the stomach surface, especially in the red channel (f).

appearance, which leads to incorrectly estimated 3D structure. Therefore, camera calibration is needed to obtain the intrinsic camera parameters such as focal length, projection center, and distortion parameters. We used the previously captured planar checkerboard pattern images and a fish-eye camera model [16] for the camera calibration. The acquired intrinsic camera parameters were used to optimize the 3D reconstruction process in SfM and to correct the image's distortion.

In the input images extraction process, we first extracted all RGB frames from each video. Then, we extracted two image sequences from each video, where the first one consists of the images captured without IC dye (see Fig. 1(a)), while the second one consists of the images captured with IC dye (see Fig. 1(b)). After an in-depth inspection, we found that there are many color artifacts in the RGB images caused by color channel misalignment as shown in Fig. 1(a) and (b). To minimize the effect of the artifacts, we decided to use single channel images as SfM inputs. We also removed any duplicate frames that have almost no movement between successive frames. We used six channel images, as shown in Fig. 1(c)-(h), as SfM inputs and investigated the combined effect of chromo-endoscopy and color channel selection.

C. Stomach 3D reconstruction

The stomach 3D reconstruction follows the general flow of an SfM pipeline, assuming that the stomach has minimum

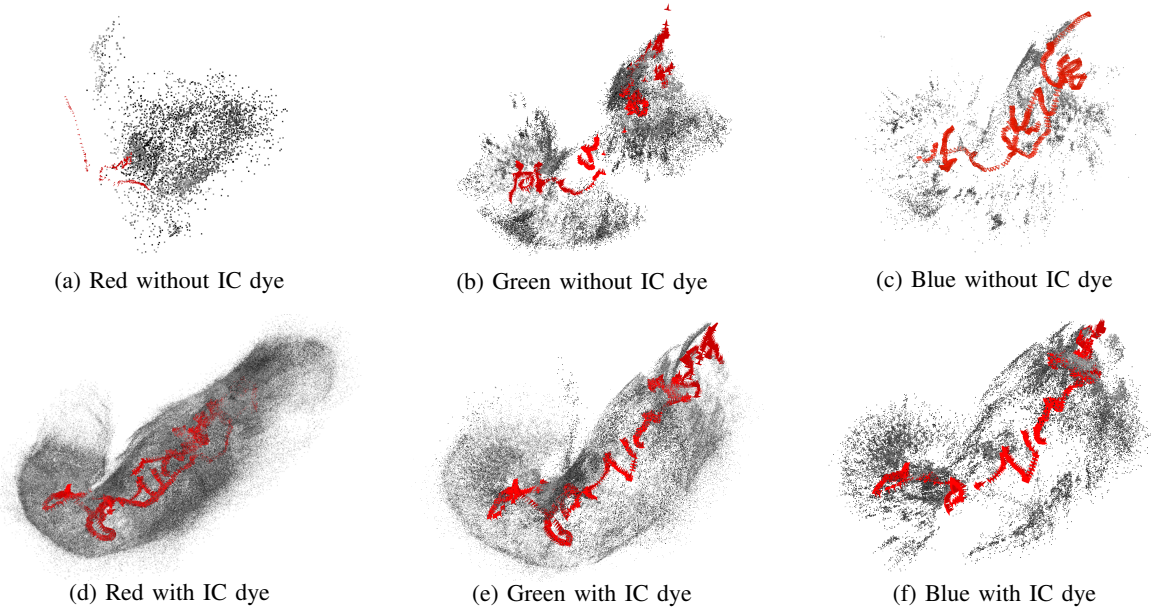


Fig. 2: The 3D point cloud results on Subject A. The gray dots represent the reconstructed 3D points and the red pyramids represent the estimated endoscope poses. There is a significant difference between the cases with and without IC dye. Only a sparse and small part of the stomach can be reconstructed in the case of without IC dye. On the other hand, the whole stomach can be reconstructed using the red channel with IC dye.

TABLE I: The objective evaluation of the 3D point cloud results using each color channel without and with IC dye.

		Subject A			Subject B			Subject C		
		Red	Green	Blue	Red	Green	Blue	Red	Green	Blue
Without IC dye	Input images	2980	2975	3000	734	729	731	2959	2790	2977
	Reconstructed images	207 (6.9%)	687 (23.1%)	497(16.6%)	177(24.1%)	226 (31.0%)	138 (18.9%)	1064 (36.0%)	1142 (40.9%)	946 (31.8%)
	3D points	5740	70610	22764	8252	15319	3117	47960	79733	40073
	Average observation	173	670	282	385	509	158	329	467	283
With IC dye	Input images	1483	1489	1476	2329	2331	2319	2327	2304	2323
	Reconstructed images	1481 (99.8%)	1249 (83.9%)	567 (38.4%)	2246 (96.4%)	1488 (63.8%)	335 (15.3%)	2297 (98.7%)	891 (38.7%)	361(15.5%)
	3D points	731070	359418	49982	515762	100114	12035	727954	152223	14022
	Average observation	4188	1988	487	1971	503	207	2656	1195	221

movements. The algorithm starts with extracting features from the input images, matching the extracted features, and followed by the endoscope poses estimation and the feature points triangulation in parallel. This step generates a sparse point cloud of the stomach based on the endoscope motion and estimates each frame's pose with respect to each other.

We used SIFT [17] for feature extraction and exhaustively search to find the feature correspondences among all image pairs. We also applied bundle adjustment [18] to optimize the 3D points and the endoscope poses.

D. Mesh and texture generation

Mesh and texture representation enables better visualization of the reconstructed 3D model. Our mesh generation starts by downsampling the original point cloud from the SfM result to a n number of 3D points and removes outlier

points using statistical outlier removal to generate a smooth mesh. The outlier removal starts by calculating the each 3D point's mean distance (\bar{x}_p) with its $n \times 0.1$ closest neighboring points. Assuming the distance distribution is Gaussian, the global distance mean (\bar{x}_g) and the standard deviation (σ_g) are then computed. Any 3D points whose mean distance \bar{x}_p is over a threshold $\bar{x}_g + 2\sigma$ are removed as outliers, leaving k numbers of inlier 3D points. Then, the normal of each inlier 3D point is estimated based on its $k \times 0.1$ closest neighboring points. Each of the estimated normal is further refined using the related endoscope camera poses information to prevent it pointing outward. Finally, the triangle mesh is generated based on the outlier-removed 3D point cloud and its per-point estimated normal by using screened Poisson surface reconstruction [19]. To add more visual detail and functionality, we then applied a texture

to the generated mesh based on the registered endoscope cameras in the SfM step. For each triangle mesh, we searched the best registered image for texturing based on the triangle-to-camera angle and distance.

III. RESULTS AND DISCUSSION

We performed the endoscope camera calibration using the OpenCV camera calibration library. The SfM pipeline was implemented on Colmap [20]. For filtering the point cloud, we set as $n = 10000$ to generate a smooth triangle mesh. We applied screened Poisson reconstruction [19] for triangle mesh generation. For the texturing purpose, we applied the parameterization and texturing function from Meshlab [21].

Figure 2 shows the 3D point cloud results on subject A, which are reconstructed using different color channels of the cases without and with IC dye. In general, the channels with the IC dye (Fig. 2(d)-(f)) give a more complete reconstruction result compared to the channels without the IC dye (Fig. 2(a)-(c)). In the case without the IC dye, the green channel gives the best result even though the model is full of holes. In the case with the IC dye, the results of all the three channels have the shape of the stomach. Among the RGB channels, the red channel gives the most complete and densest result. Some holes still exist in the result using the green channel, while the blue channel is only able to reconstruct around 3/4 of the whole stomach.

Table I shows the objective evaluation of the 3D point cloud results on all three subjects. Table I shows that the number of 3D points is generally higher when the IC dye is present. We also notice that the average observation, which represents the per-image average number of the 2D feature points that can be triangulated into the 3D points, is generally increased when the IC dye exists. In addition, the percentage of reconstructed images over input images is significantly increased by using the IC dye. Among all the results, the red channel with the IC dye gives the best result, where more than 95% images are reconstructed. When the IC dye is not present, the green channels gives the best result.

The above subjective and objective evaluation consistently shows that the red channel with the IC dye gives the best result. As shown in Fig. 1(c)-(h), this is because that the red channel leverages the effect of the IC dye more than the other channels. In Fig. 1(f), many textures, from which many distinctive features can be extracted, are apparent in the red channel. When the IC dye is not used, the green channel has better contrasts compared to the others channels. The blue channel is the least preferable among those three channels for both cases without and with the IC dye.

Figure 3 shows the results of triangle mesh and texture generation using the red channel with the IC dye. We can confirm that the generated meshes resemble the whole shape of a stomach. Moreover, the textured representation makes the generated 3D model more perceptible for viewers.

Figure 4 shows our custom viewer that can project any selected reconstructed images to the generated triangle mesh based on the estimated endoscope poses in SfM. This custom viewer provides viewers with the estimated location of a

particular image frame, which can be used for the 3D localization of a malignant lesion. Our viewer should be very valuable for gastric surgeons to make a medical decision.

IV. CONCLUSION

In this paper, we have presented an offline solution to reconstruct the whole shape of a stomach from a standard monocular endoscope video. To obtain better reconstruction quality using SfM, we used a single channel images without color channel misalignment artifact. We found that the chromo-endoscopy with IC blue color dye generally gives significant improvement to the completeness of the reconstruction result. Furthermore, we found that the red channel with the IC dye provides the most complete 3D model compared to the other channels. A custom viewer that can localize a particular image frame in the reconstructed 3D model was also presented. In future work, we plan to refine the mesh generation process for more detail representation considering more effective downsampling and outlier removal approaches. To view the results in more detail, please visit our project page in the following link (<http://www.ok.sc.e.titech.ac.jp/res/Stomach3D/>).

REFERENCES

- [1] N. Yamamichi, C. Hirano, Y. Takahashi, C. Minatsuki, C. Nakayama, R. Matsuda, T. Shimamoto, C. Takeuchi, S. Kodashima, S. Ono, Y. Tsuji, M. Fujishiro, R. Wada, T. Mitsushi, and M. Koike, "Comparative analysis of upper gastrointestinal endoscopy, double-contrast upper gastrointestinal barium X-ray radiography, and the titer of serum anti-Helicobacter pylori IgG focusing on the diagnosis of atrophic gastritis," *Gastric cancer*, vol. 19, no. 2, pp. 670–675, 2016.
- [2] J. W. Kim, S. S. Shin, S. H. Heo, H. S. Lim, N. Y. Lim, Y. K. Park, Y. Y. Jeong, and H. K. Kang, "The role of three-dimensional multidetector CT gastrography in the preoperative imaging of stomach cancer: Emphasis on detection and localization of the tumor," *Korean Journal of Radiology*, vol. 16, no. 1, pp. 80–89, 2015.
- [3] L. Maier-Hein, P. Mountney, A. Bartoli, H. Elhawary, D. Elson, A. Groch, A. Kolb, M. Rodrigues, J. Sorger, S. Speidel, and D. Stoyanov, "Optical techniques for 3D surface reconstruction in computer-assisted laparoscopic surgery," *Medical Image Analysis*, vol. 17, no. 8, pp. 974–996, 2013.
- [4] J. Geng and J. Xie, "Review of 3-D endoscopic surface imaging techniques," *IEEE Sensors Journal*, vol. 14, no. 4, pp. 945–960, 2014.
- [5] T. Okatani and K. Deguchi, "Shape reconstruction from an endoscope image by shape from shading technique for a point light source at the projection center," *Computer Vision and Image Understanding*, vol. 66, no. 2, pp. 119–131, 1997.
- [6] C. H. Q. Foster and C. Tozzi, "Towards 3D reconstruction of endoscope images using shape from shading," in *Proc. of Brazilian Symposium on Computer Graphics and Image Processing*, pp. 90–96, 2000.
- [7] O. G. Grasa, E. Bernal, S. Casado, I. Gil, and J. M. M. Montiel, "Visual SLAM for handheld monocular endoscope," *IEEE Trans. on Medical Imaging*, vol. 33, no. 1, pp. 135–146, 2014.
- [8] 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 *Proc. of International Workshop on Computer-Assisted and Robotic Endoscopy (CARE)*, pp. 72–83, 2016.
- [9] N. Mahmoud, C. Toby, A. Hostettler, L. Soler, C. Doignon, and J. M. M. Montiel, "Live tracking and dense reconstruction for handheld monocular endoscopy," *IEEE Trans. on Medical Imaging*, vol. 38, no. 1, pp. 79–88, 2019.
- [10] S. Mills, L. Szymanski, and R. Johnson, "Hierarchical structure from motion from endoscopic video," in *Proc. of Int. Conf. on Image and Vision Computing New Zealand (IVCNZ)*, pp. 102–107, 2014.

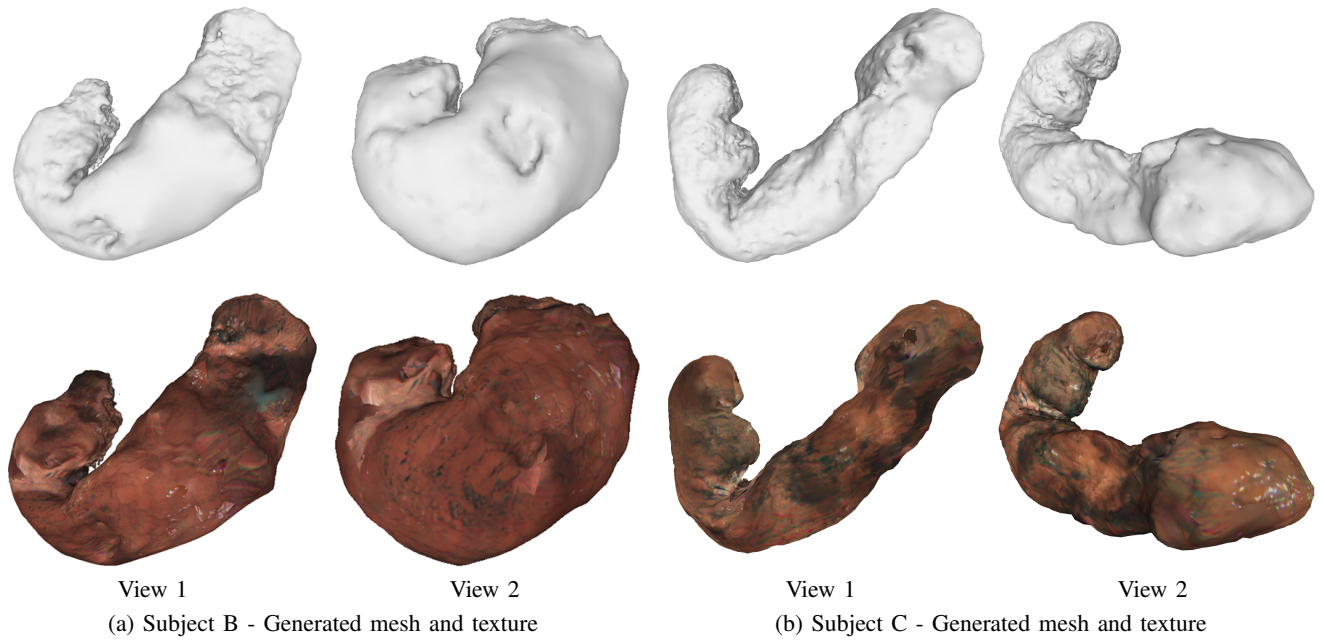


Fig. 3: The triangle mesh and texture models generated from the point clouds reconstructed using the red channel with IC dye. The shown texture is the inner texture of the stomach. The video version can be seen from the following link (<http://www.ok.sc.e.titech.ac.jp/res/Stomach3D/>).

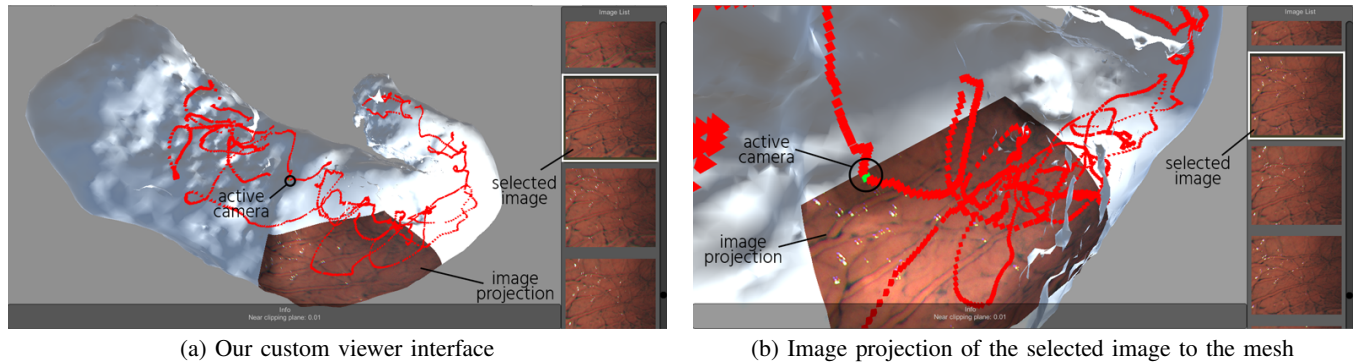


Fig. 4: Demonstration of our custom viewer. The figure (a) illustrates the viewer interface. It loads the generated mesh and the estimated endoscope camera positions. The red pyramids in (a) and (b) represent the estimated cameras and the camera trajectory. The user can select either a camera or an image to project the related image to the mesh, as shown in (b). The selected or active camera is shown as green in both (a) and (b).

- [11] D. Sun, J. Liu, C. A. Linte, H. Duan, and R. A. Robb, "Surface reconstruction from tracked endoscopic video using the structure from motion approach," in *Proc. of Augmented Reality Environments for Medical Imaging and Computer-Assisted Interventions (AE-CAI)*, pp. 127–135, 2013.
- [12] K. L. Lurie, R. Angst, D. V. Zlatev, J. C. Liao, and A. K. E. Bowden, "3D reconstruction of cystoscopy videos for comprehensive bladder records," *Biomedical Optics Express*, vol. 8, no. 4, pp. 2106–2123, 2017.
- [13] R. Furukawa, H. Morinaga, Y. Sanomura, S. Tanaka, S. Yoshida, and H. Kawasaki, "Shape acquisition and registration for 3D endoscope based on grid pattern projection," in *Proc. of European Conf. on Computer Vision (ECCV)*, pp. 399–415, 2016.
- [14] C. Schmalz, F. Forster, A. Schick, and E. Angelopoulou, "An endoscopic 3D scanner based on structured light," *Medical Image Analysis*, vol. 16, no. 5, pp. 1063–1072, 2012.
- [15] P. F. Alcantarilla, A. Bartoli, F. Chadebecq, C. Tilmant, and V. Lepilliez, "Enhanced imaging colonoscopy facilitates dense motion-based 3D reconstruction," in *Proc. of Int. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 7346–7349, 2013.
- [16] J. Kannala and S. S. Brandt, "A generic camera model and calibration method for conventional, wide-angle, and fish-eye lenses," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 8, pp. 1335–1340, 2006.
- [17] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
- [18] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, "Bundle adjustment – A modern synthesis," in *Proc. of Int. Workshop on Vision Algorithms*, pp. 298–372, 1999.
- [19] M. Kazhdan and H. Hoppe, "Screened Poisson surface reconstruction," *ACM Trans. on Graphics*, vol. 32, no. 3, p. 29, 2013.
- [20] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 4104–4113, 2016.
- [21] "Meshlab," <http://www.meshlab.net/>, (Accessed on 01/10/2019).