Single and multi-frame auto-calibration for 3D endoscopy with differential rendering

Ryo Furukawa[1], Ryusuke Sagawa[2], Shiro Oka[3], Shinji Tanaka[3], and Hiroshi Kawasaki[4],

***Abstract*— The use of 3D measurement in endoscopic images offers practicality in cancer diagnosis, computer-assisted interventions, and making annotations for machine learning training data. An effective approach is the implementation of an active stereo system, using a micro-sized pattern projector and an endoscope camera, which has been intensively developed. One open problem for such a system is the necessity of strict and complex calibration of the projector-camera system to precisely recover the shapes. Moreover, since the head of an endoscope should have enough elasticity to avoid harming target objects, the positions of the pattern projector cannot be tightly fixed to the head, resulting in limited accuracy. A straightforward approach to the problem is applying auto-calibration. However, it requires special markers in the pattern or a highly accurate initial position for stable calibration, which is impractical for real operation. In the paper, we propose a novel auto-calibration method based on differential rendering techniques, which are recently proposed and drawing wide attention. To apply the method to an endoscopic system, where a diffractive optical element (DOE) is used, we propose a technique to simultaneously estimate the focal length of the DOE as well as the extrinsic parameters between a projector and a camera. We also propose a multi-frame optimization algorithm to jointly optimize the intrinsic and extrinsic parameters, relative pose between frames, and the entire shape.**

***Clinical relevance*— One-shot endoscopic measurement of depth information is a practical solution for cancer diagnosis, computer-assisted interventions, and making annotations for machine learning training data.**

I. Introduction

The 3D measurement of tumors in human organs during endoscopic operation has become important, as tumor size is considered one of the standard criteria for cancer diagnosis and tumor resection. An active stereo system, which consists of an ultra-small pattern projector with an endoscope camera, is a practical solution, and actual systems have been intensively researched and developed. One open problem for such a system is the necessity of strict and complex calibration of the projector-camera system to precisely recover the shapes. Moreover, since the head of an endoscope should have enough elasticity to avoid harming target objects, the positions of the pattern projector cannot be tightly fixed to the head, resulting in limited accuracy. A straightforward approach to the problem is applying auto-calibration. However, it requires special markers in the pattern or a highly accurate initial position for stable calibration, which are impractical for real operation. In the paper, we propose a novel auto-calibration method based on the differential rendering techniques, which are recently proposed and drawing wide attention. To apply the method to endoscopic systems, where a diffractive optical element (DOE) is used, we proposed a technique to simultaneously estimate the distortion parameter of the DOE as well as the extrinsic parameters between a projector and a camera.

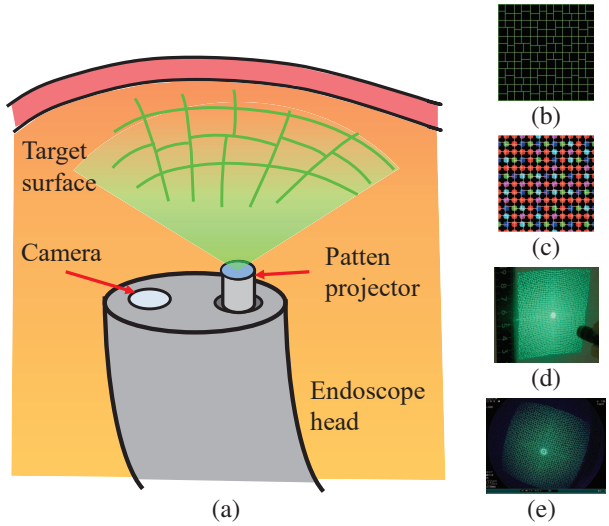

Fig. 1. The model active-stereo system. (a) System configuration. (b) The projected grid pattern. (c) Code information embedded into the pattern. (d) The pattern illuminated onto a plane. (e) Pattern-illuminated surface captured by the endoscopic camera.

Another problem of 3D scanning with an endoscope is that it is difficult to obtain a sufficient area of internal organs due to the limited field of view of the camera and the limited projected area by the DOE. The straight forward approach is to solve each problem one by one. For example, first, auto-calibration techniques are applied to recover the 3D shape of each frame. Then, an iterative closest point (ICP) algorithm [1] is applied to estimate ego-motion between frames, and finally, all the shapes are merged using Truncated Signed Distance Field (TSDF) [2]. However, each step includes certain errors, and they accumulate to become a huge bias in the final results. For example, there is a scale mismatch between frames, which inevitably occurs during frame-wise auto-calibration. Thus, it is necessary to minimize the shape mismatch in different frames to achieve global optimization by using the information of multiple frames. In this paper, we propose an extension to the auto-calibration technique using a new loss function that directly models pattern projection in active stereo using the projection

*This work was supported by JSPS KAKENHI Grant Numbers JP20H00611, JP18H04119, JP21H01457 and NEDO(JPNP20006) in Japan.

[1]Kindai University, Japan, furukawa@hiro.kindai.ac.jp
[2]AIST, Japan
[3]Hiroshima University, Japan
[4]Kyushu University, Japan

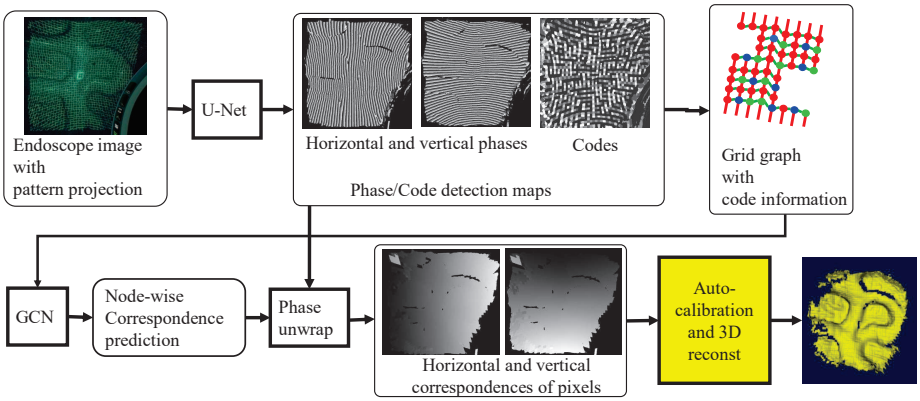

Fig. 2. Overview of the reconstruction process. Auto-calibration is simultaneously conducted with 3D reconstruction indicated by yellow box.

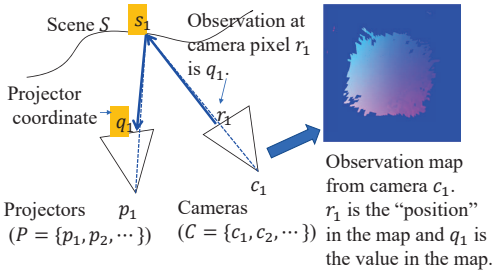

Fig. 3. Observation model of active stereo systems

mapping of projector coordinates. It is based on a multi-frame global optimization for an active stereo system where the relative position between the projector and camera is not assumed to be fixed.

In summary, our contributions are: (1) proposing an auto-calibration method for the intrinsic parameters of severe distortion of DOE projector, as well as the extrinsic parameters between a projector and a camera, by using differentiable rendering techniques, (2) proposing simultaneous optimization of relative poses of multiple frames, as well as intrinsic and extrinsic parameters for each frame of an endoscopic system based on new loss functions, and (3) conducting experimental results using both simulations and real objects are conducted to confirm the effectiveness of our method.

II. Related Works

For estimating the parameters of the projector and the camera, a typical approach is to project patterns onto a calibration object [3], [4]. These methods are used for pre-calibration, where the projector and the camera can be fixed and calibrated before measurement. 3D reconstruction based on passive cameras, such as SLAM or SfM, for endoscopic images has been researched in medical image analysis, as seen in works by Mahmoud *et al.* [5], Chen *et al.* [6], and Leonard *et al.* [7]. Recently, non-rigid SLAM has been proposed, such as Song *et al.* [8], Lamarca *et al.* [9], and Zhou *et al.* [10]. These methods need 3D feature points and thus needs textures.

For 3D registration for medical purposes, ICP algorithms have been used [11], [12]. In this paper, our aim is not only to register multiple 3D scenes but also to correct inter-frame inconsistencies by taking the observation model into account. For this purpose, Furukawa *et al.* [13] proposed a modification of bundle adjustment for passive stereo. However, their method does not directly model the dynamics of active stereo observations and often has problems with convergence. In contrast, our technique is faster and more stable. Note that since active stereo techniques for laparoscopes and endoscopes have been widely studied [14], [15], [13], the auto-calibration method proposed in this paper can be applied to any type of active stereo system and is useful for both single and multi-frame approaches.

A differentiable renderer renders Computer Graphics (CG) images using a scene or camera parameters, where gradient-based optimization with respect to the parameters can be processed. Mesh-based differentiable renderers [16], [17], [18], [19] have been proposed to optimize various scene parameters such as geometry, illumination, textures, or materials.

III. System Configuration and Algorithm

An experimental system has been developed which allows for 3D shape reconstruction from a single frame and consists of a pattern projector inserted through the instrumental channel of a standard endoscope, as shown in Fig.1(a). The system is based on a similar approach proposed by Furukawa *et al.* [20]. The structured light illumination is created by the Diffractive Optical Element (DOE) included in the pattern projector, as shown in Fig. 1(b). We used a grid pattern consisting of vertical and horizontal edges with small gaps. The gaps represent code symbols to identify the camera-to-projector mapping, as shown in Fig.1(c) and actual patterns projected onto the object surface are shown in Fig.1(d) and (e).

The flow of the 3D reconstruction algorithm is shown in Fig.2. In this process, the grid pattern is projected onto the target surface and captured by the endoscopic camera. Then, U-Nets are applied to captured frames to predict pixel-wise phase information (*i.e.*, repetition structure of the grid) as well as the code symbols. The grid structure and codes are represented by a graph, and graph convolutional network (GCN) is applied to predict node-wise correspondences to unwrap the phase. Finally, pixel-wise 3D reconstruction is achieved by triangulation using the unwrapped phase (*i.e.*,

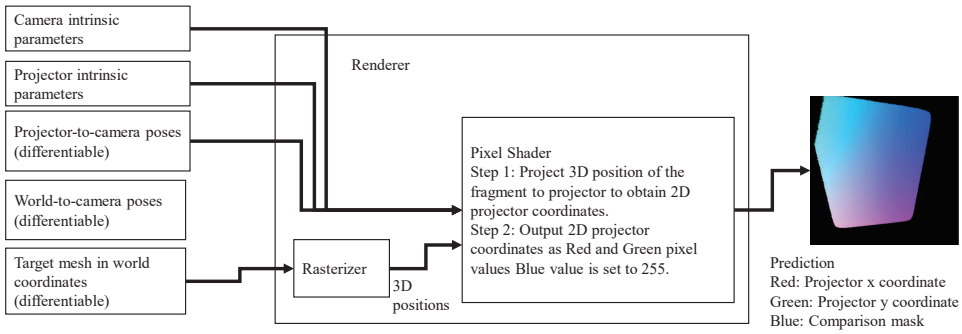

Fig. 4. differentiable renderer with differentiable variables of C, P, and S

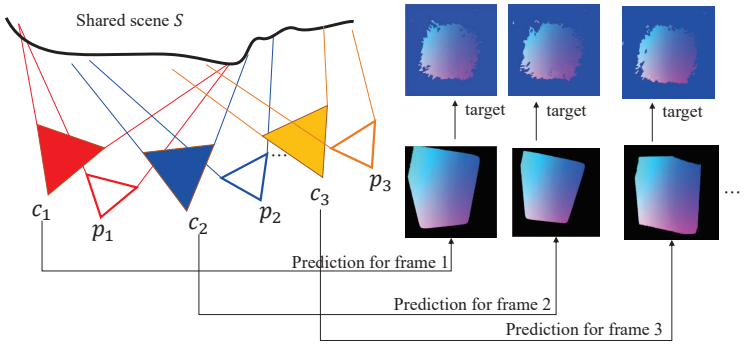

Fig. 5. Multi-frame optimization via scene S : the prediction is projector coordinates(Red:x, Green:y) predicted for C, P, and S

2D correspondences between the camera and the projector).

IV. Differential Rendering for Auto-Calibration

In this paper, a differential-renderer-based method is proposed to achieve auto-calibration of a projector-camera system from a single frame without special markers or pre-calibration. Note that although auto-calibration methods for active stereo for endoscopic systems have already been proposed, they assumed special markers [13] or accurate initial parameters, and cannot handle estimation of the focal length of the projector [20]. These issues are effectively solved by our method.

In our auto-calibration process, correspondences between the camera and the projector pixels are represented as a mapping from camera pixels to projector coordinates, as shown in Fig.3. For each camera pixel r_i, the corresponding projector coordinates q_i are estimated. This can be modeled as "ray-tracing" from camera pixels r_1 to surface S, resulting in s_1, and projection of s_1 to projector p_1, resulting in q_1. The correspondence map represents all the information obtained from a single scan of our "one-shot" active-stereo system. By using this information, a CG scene can be rendered as a projection mapping. Note that this can be efficiently achieved using a *pixel shader* of GPU, where the *pixel shader* receives a 3D point on surface S and projects it onto a projector, as shown in Fig.4. This image represents a mapping from the camera 2D coordinates to the 2D coordinates of the projected pattern and is a function of surface S, camera parameters C, and projector parameters P. Let the images obtained by the projection mapping be denoted as $R(S, C, P)$. In active stereo, $R(S, C, P)$ is the same as the 2D correspondence map from the camera to the projector, which is obtained by analyzing a pattern-projected camera image. Thus, for an optimized geometrical solution, $R(S, C, P)$ should mimic the 2D correspondence map.

TABLE I

Comparison with per-frame auto-calibration approach, where (A) estimating independent projector-to-camera poses and (B) estimating a single projector-to-camera pose for all frames.

Methods	Proposed	Per-frame (A)	Per-frame (B)
RMSE(mm)	1.41	2.59	2.15

Auto-calibration can be implemented by minimizing the loss function $L(S, C, P)$, which represents the deference between $R(S, C, P)$ and the observed correspondence mapping, with respect to S,C, and P, which can be single-frame or multi-frame. We used Cauchy Loss for $L(S, C, p)$. Although Fig.5 shows the multi-frame case, single-frame optimization is also possible. Note that the rendered and optimized images are not the projected images themselves, but the projector coordinates. Images of the projector coordinates have much better properties for optimization, since projector coordinates change monotonously in smooth surfaces, whereas the projected images tend to have repetitive grid structure, which causes local minimums for image similarities.

The optimization of $L(S, C, P)$ is done using Adaptive Moment Estimation (Adam). The inter-frame positions are also optimized with this method, as all the views share a single scene model S.

V. Experiments

A. Validation by simulation data

To validate the auto-calibration with the proposed method, we synthesized simulation data with a rabbit-shaped mesh

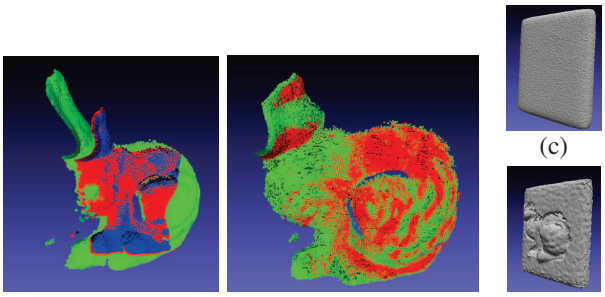

Fig. 6. Calibration results of simulation data. (a) Shapes before and after single-frame auto-calibration. Colored point clouds are GT shape (red), before auto-calibration (green), and after auto-calibration (blue). The shape reconstructed by un-calibrated parameters (green) was severely distorted, whereas GT shape (red) and auto-calibrated shapes coincided. (b) Result of 2-frame calibration, where red points are frame 1 and green points are frame 2. Note that multiple frames were fit to each other without scale inconsistency. (c) Initial shape S. (d) Shape S after optimization.

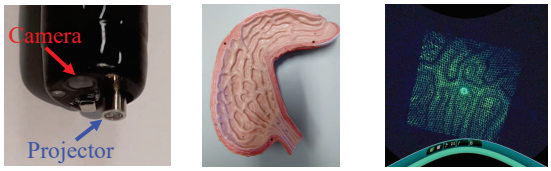

Fig. 7. (a) A head of endoscope, (b) a stomach phantom and (c) captured image with structured light.

model and projection mapping. Then, we intentionally added errors to the projector-to-camera pose parameters and the focal length of the projector. The reconstructed shape with erroneous parameters is shown as the green shape in Fig.6(a). By comparing it to the GT shape (red in (a)), we confirmed that the shape that is not calibrated is largely distorted. Then, we applied the proposed method to estimate camera and projector poses, using an initial shape of a plane , and S, C and P were optimized (Fig.6(c) and (d)). Using the optimized C and P, we can obtained the blue shape in (a), which coincided with the GT shape. Also, a 2-frame calibration was processed. By optimizing multiple frames, we obtained frame-wise shapes with consistent scale and positions, as shown in (b).

B. Single and multi-frame calibration of real data

Next, we captured sequential images using a medical endoscopic system, where its head consists of a camera and a projector, as shown in Fig.7(a). Note that the projector is not attached anywhere, and thus it may move around inside the instrument channel during a measurement. The target shape is a medical shape model (phantom) of a stomach, as shown in Fig.7(b). One example of a captured image is shown in Fig.7(c). By using the proposed method, we can obtain single-frame and multi-frame shapes without pre-calibration. Fig.8 shows two examples of measuring colon and stomach phantom models using single-frame calibrations. For both samples, the projector-to-camera poses, and focal length were automatically estimated, and 3D shapes were obtained.

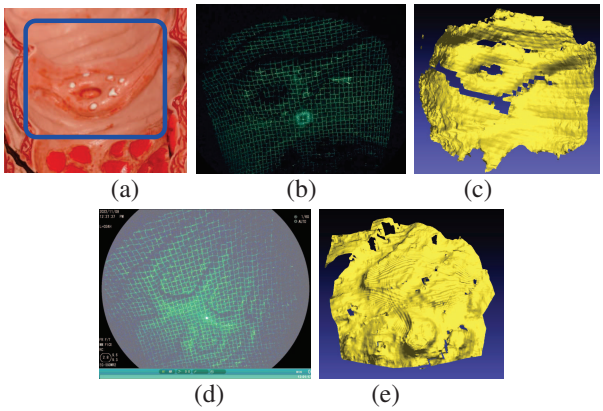

Fig. 8. Measurement of colon and stomach phantoms. (a) Appearance (colon). (b)(d) Captured images. (c)(e) Shapes obtained with optimized camera and projector parameters.

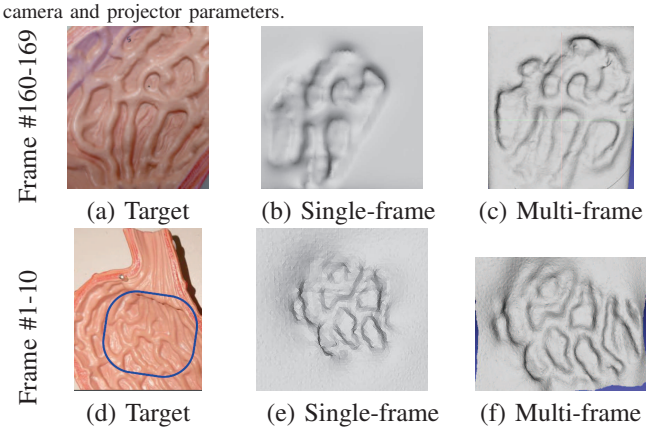

Fig. 9. Reconstruction results of a stomach phantom using different frames. (a)(d) Appearance and captured region. (b)(e) Shape S estimated from one frame. (c)(f) S estimated using all the frames.

Fig.9 shows examples of both single and multi-frame measurements, where single-frame calibration is done with the proposed method, and then multiple frames are added and simultaneously optimized. Note that all of the results shown in Fig.9 were obtained from a fixed initial projector-to-camera pose parameters, where the projector is located at the front-parallel position with the camera (*i.e.*, rigid transformation from the projector coordinates to the camera coordinates is a translation in the xy plane with no rotation).

For further validation, we compared our calibration method with a baseline method of per-frame auto-calibration proposed in Furukawa *et al.* [21]. In the baseline method, we reconstructed 3D shapes by applying per-frame auto-calibration for each frame, and registered the results using ICP [1]. Then, finally shapes are integrated by TSDF [2]. For per-frame auto-calibration, we applied two configurations: configuration (A) is estimating projector-to-camera poses independently for each frame, and configuration (B) is estimating a single projector-to-camera pose, assuming that projector poses with respect to the camera are constant for all the frames. Both our results and the per-frame auto-calibration results were compared to the Ground-truth shape. The Roots of Mean Squared Errors (RMSEs) from the ground-truth shape were shown in Tab.I, where it is confirmed that the accuracy of the proposed method was

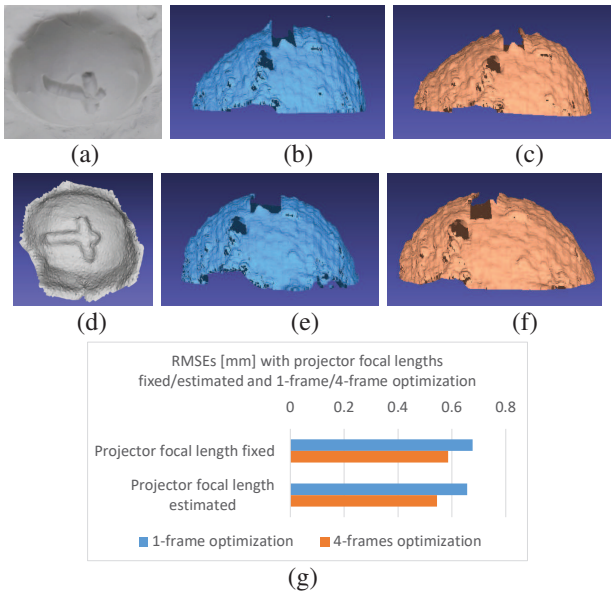

Fig. 10. Measurement of concave sphere made of clay. (a)Object appearances. (b) Result with f_p fixed, optimized with 1 frame. (c) Result with f_p fixed, 4-frames optimization. (d) Estimated mesh optimized with the projector focal length (f_p) estimated and optimized with 4 frames. (e) Result with f_p estimated, 1-frame optimization. (f) Result with f_p estimated, 4-frames optimization. (g) RMSEs for (b)(c)(e)(f).

better than ICP-registered per-frame calibration results. We can also confirm that configuration B is better than A. This might be because a projector head does not move drastically within a short period of time.

C. Projector-focal-length estimation for real images

To show the effectiveness of projector-focal-length estimation and multi-frame optimization in contrast to multi-frame optimization, we measured a concave sphere-shaped object made of clay. We intentionally started from an erroneous focal length f_p of the projector. The results are shown in Fig.10. We can see that the focal-length estimation effectively corrected distortions caused by the erroneous f_p, and also that results of 4-frame optimization were better than single-frame optimization.

VI. Conclusion

In this paper, a novel single and multi-frame auto-calibration method for active-stereo scanning is proposed. Our approach used optimization process with a differentiable renderer to estimate the projector and camera poses, along with the projector focal length. By directly modeling the active-stereo observation process as CG rendering and minimizing the active-stereo observation errors with differentiable rendering, multi-frame shape integration with shape consistencies over frames is achieved. The proposed method has been validated with scanned data using a 3D endoscopic system, and in the future, the method could be extended to integrate more complex shapes.

References

[1] P. J. Besl and N. D. McKay, "Method for registration of 3-d shapes," in *Robotics-DL tentative*. International Society for Optics and Photonics, 1992, pp. 586–606.

[2] B. Curless and M. Levoy, "A volumetric method for building complex models from range images," in *SIGGRAPH 96. ACM*, 1996, pp. 303–312.

[3] J. Liao and L. Cai, "A calibration method for uncoupling projector and camera of a structured light system," in *2008 IEEE/ASME International Conference on Advanced Intelligent Mechatronics*. IEEE, 2008, pp. 770–774.

[4] K. Yamauchi, H. Saito, and Y. Sato, "Calibration of a structured light system by observing planar object from unknown viewpoints," in *ICPR*. IEEE, 2008, pp. 1–4.

[5] 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.

[6] L. Chen, W. Tang, N. W. John, T. R. Wan, and J. J. Zhang, "Slam-based dense surface reconstruction in monocular minimally invasive surgery and its application to augmented reality," *Computer methods and programs in biomedicine*, vol. 158, pp. 135–146, 2018.

[7] S. Leonard, A. Sinha, A. Reiter, M. Ishii, G. L. Gallia, R. H. Taylor, and G. D. Hager, "Evaluation and stability analysis of video-based navigation system for functional endoscopic sinus surgery on in vivo clinical data," *IEEE transactions on medical imaging*, vol. 37, no. 10, pp. 2185–2195, 2018.

[8] J. Song, J. Wang, L. Zhao, S. Huang, and G. Dissanayake, "Mis-slam: Real-time large-scale dense deformable slam system in minimal invasive surgery based on heterogeneous computing," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 4068–4075, 2018.

[9] J. Lamarca, S. Parashar, A. Bartoli, and J. Montiel, "Defslam: Tracking and mapping of deforming scenes from monocular sequences," *IEEE Transactions on robotics*, vol. 37, no. 1, pp. 291–303, 2020.

[10] H. Zhou and J. Jayender, "Emdq-slam: Real-time high-resolution reconstruction of soft tissue surface from stereo laparoscopy videos," in *MICCAI*. Springer, 2021, pp. 331–340.

[11] B. Combès and S. Prima, "An efficient em-icp algorithm for symmetric consistent non-linear registration of point sets," in *MICCAI*. Springer, 2010, pp. 594–601.

[12] M. Sinko, P. Kamencay, R. Hudec, and M. Benco, "3d registration of the point cloud data using icp algorithm in medical image analysis," in *2018 ELEKTRO*. IEEE, 2018, pp. 1–6.

[13] R. Furukawa, G. Nagamatsu, S. Oka, T. Kotachi, Y. Okamoto, S. Tanaka, and H. Kawasaki, "Simultaneous shape and camera-projector parameter estimation for 3d endoscopic system using cnn-based grid-oneshot scan," *Healthcare Technology Letters*, vol. 6, no. 6, pp. 249–254, 2019.

[14] J. Geurten, W. Xia, U. Jayarathne, T. M. Peters, and E. C. Chen, "Endoscopic laser surface scanner for minimally invasive abdominal surgeries," in *MICCAI*. Springer, 2018, pp. 143–150.

[15] J. Lin, N. T. Clancy, D. Stoyanov, and D. S. Elson, "Tissue surface reconstruction aided by local normal information using a self-calibrated endoscopic structured light system," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 405–412.

[16] P. Henderson and V. Ferrari, "Learning to generate and reconstruct 3d meshes with only 2d supervision," *arXiv preprint arXiv:1807.09259*, 2018.

[17] A. Palazzi, L. Bergamini, S. Calderara, and R. Cucchiara, "End-to-end 6-dof object pose estimation through differentiable rasterization," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0–0.

[18] H. Kato, Y. Ushiku, and T. Harada, "Neural 3d mesh renderer," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3907–3916.

[19] H.-T. D. Liu, M. Tao, and A. Jacobson, "Paparazzi: surface editing by way of multi-view image processing." *ACM Trans. Graph.*, vol. 37, no. 6, pp. 221–1, 2018.

[20] R. Furukawa, S. Oka, T. Kotachi, Y. Okamoto, S. Tanaka, R. Sagawa, and H. Kawasaki, "Fully auto-calibrated active-stereo-based 3d endoscopic system using correspondence estimation with graph convolutional network," in *EMBC*. IEEE, 2020, pp. 4357–4360.

[21] R. Furukawa and H. Kawasaki, "Uncalibrated multiple image stereo system with arbitrarily movable camera and projector for wide range scanning," in *Fifth International Conference on 3-D Digital Imaging and Modeling (3DIM'05)*. IEEE, 2005, pp. 302–309.