Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

Calibration-free structured-light-based 3D scanning system in laparoscope for robotic surgery

Ryo Furukawa[1] | **Elvis Chen**[2] | **Ryusuke Sagawa**[3] | **Shiro Oka**[4] | **Hiroshi Kawasaki**[5]

[1] Department of Informatics, Kindai University, Higashihiroshima, Hiroshima, Japan

[2] Robarts Research Institute, Western University, London, ON, Canada

[3] Advanced Industrial Science and Technology, Tsukuba, Ibaraki, Japan

[4] Graduate School of Biomedical and Health Sciences, Hiroshima University, Hiroshima, Hiroshima, Japan

[5] Department of Advanced Information Technology, Kyushu University, Fukuoka, Fukuoka, Japan

Abstract

Accurate 3D shape measurement is crucial for surgical support and alignment in robotic surgery systems. Stereo cameras in laparoscopes offer a potential solution; however, their accuracy in stereo image matching diminishes when the target image has few textures. Although stereo matching with deep learning has gained significant attention, supervised learning requires a large dataset of images with depth annotations, which are scarce for laparoscopes. Thus, there is a strong demand to explore alternative methods for depth reconstruction or annotation for laparoscopes. Active stereo techniques are a promising approach for achieving 3D reconstruction without textures. In this paper, we propose a 3D shape reconstruction method using an ultra-small patterned projector attached to a laparoscopic arm to address these issues. The pattern projector emits a structured light with a grid-like pattern that features node-wise modulation for positional encoding. To scan the target object, multiple images are taken while the projector is in motion, and the relative poses of the projector and a camera are auto-calibrated using a differential rendering technique. In the experiment, the proposed method is evaluated by performing 3D reconstruction using images obtained from a surgical robot and comparing the results with a ground-truth shape obtained from X-ray CT.

KEYWORDS

Auto-calibration, Structured light, surgical robots, laparoscopy.

1 | INTRODUCTION

Accurate 3D shape measurement is crucial for surgical support and alignment in robotic surgery systems and laparoscopic surgical procedures[1,2,3,4]. While stereo cameras in laparoscopes offer a potential solution, their accuracy in stereo image matching diminishes when the target image lacks texture[5]. Stereo matching with deep learning has gained significant attention in recent years[6,7,8,9,10]. However, the primary learning method, supervised learning, requires huge datasets of images with depth annotations, which are currently scarce for laparoscopy. Due to the recent significant increases in remote surgery, there is a pressing demand to explore alternative methods for depth reconstruction and annotation for laparoscopes.

Among the vast array of 3D scanning techniques, active stereo is a promising approach for achieving 3D reconstruction for laparoscopes. Lin *et al.*[11] proposed a method where colored random dots are projected from an active light source to obtain 3D shapes. Furukawa *et al.*[12] used two-dimensional patterns formed with a single-wavelength laser to obtain shape information. These methods utilize patterns projected onto the target surface to obtain correspondence information between the camera image and the pattern, even if the object has few textures. Such a system requires calibration of the projector and camera positions to accurately recover the shape. However, this process is sometimes difficult, particularly when the projector is not fixed to the camera. One potential solution is to attach a marker to the projector and track its movements[13,14]; however, this approach has severe limitations, as the markers must be placed within the field of view, which is not possible for laparoscopes in many cases.

In this paper, we propose a 3D shape reconstruction method using an ultra-small patterned projector attached to a laparoscope to address the aforementioned issues. Similar to the method of Furukawa *et al.*[12], we utilize a laser pattern projector that forms grid-like patterns to find global correspondences without the need for special patterns. By utilizing these global correspondences,

Abbreviations: ANA, anti-nuclear antibodies; APC, antigen-presenting cells; IRF, interferon regulatory factor.

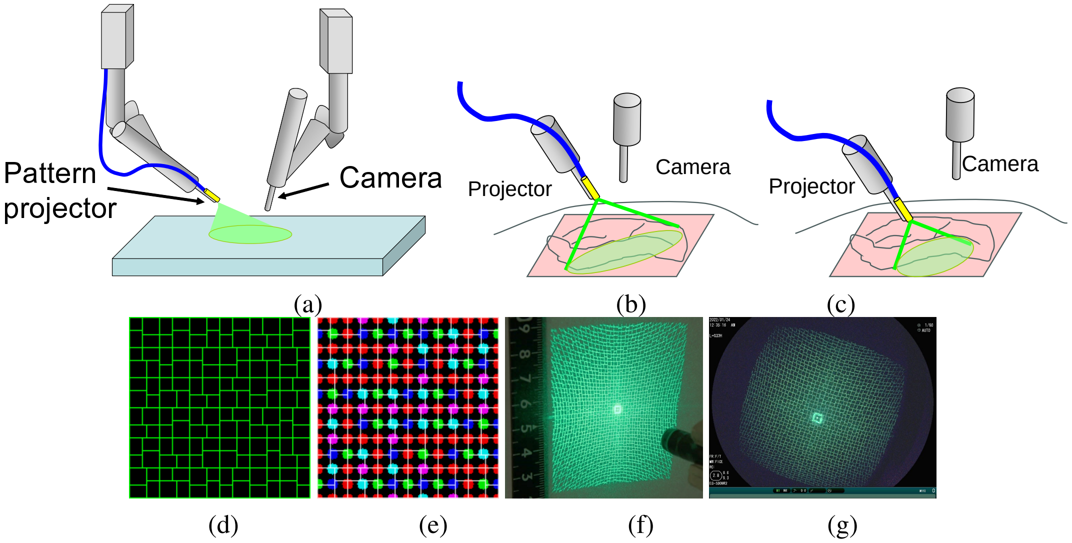


FIGURE 1 (a) Active-stereo 3D scanning system in laparoscopy. (b),(c) Scanning 3D shape by moving the projector and the camera. (d) The projected grid pattern. (e) Code information embedded into the pattern. (f) The pattern illuminated onto a plane. (g) Pattern-illuminated surface captured by the laparoscopic camera.

we can perform auto-calibration for each frame while the projector is in motion. This enables not only 3D scanning for each frame but also consistent shape integration of multiple frames.

To address the constraints regarding the positions of the cameras and the projectors, we employ optimization techniques using differential rendering. In this method, we render a mapping from the camera pixels to the projector coordinates and estimate the 3D shape and projector position simultaneously by minimizing the differences between the rendered images and the observations obtained from the captured images.

The contributions of this paper are summarized as follows:

1. We propose a method to automatically calibrate the extrinsic parameters between the projector and the camera using a differential renderer. We optimize the parameters by minimizing the difference between the rendered pattern and the observation obtained from the captured images, eliminating the need to detect markers or feature points.
2. To evaluate the proposed method, we conducted 3D reconstruction and compared the results with a ground-truth shape obtained from X-ray CT. We used images obtained from a surgical robot, where the projector and the camera were attached to different robotic arms, where the relative positions between the camera and the projector were dynamically changed during the operation.
3. Since the proposed method does not require any calibration, multiple images captured by freely moving the projector and camera during the operation are optimized to recover a consistent shape of a wide area, which is important in clinical settings.

2 | RELATED WORKS

For estimating the projector and the camera parameters, a typical approach is projecting patterns onto a calibration object[15,16]. These methods are for pre-calibration, where the projector and the camera can be fixed and calibrated before measurement. 3D reconstruction based on passive camera, *e.g.*, SLAM or SfM, for endoscopic images have been researched in medical image analysis, such as Mahmoud *et al.*[17], Chen *et al.*[18], and Leonard *et al.*[19]. Recently, non-rigid SLAM have been proposed, such as Song *et al.*[2], Lamarca *et al.*[20], and Zhou *et al.*[21]. These methods need 3D feature points, thus needs textures.

For 3D registration for medical purposes, ICP algorithm has been used[22,23]. In this paper, our target is not only registration of multiple 3D scenes, but simultaneously correcting inter-frame inconsistencies by taking the observation model into account. For such a purpose, Furukawa *et al.*[24] proposed a modification of bundle adjustment for passive stereo. Their method does not directly

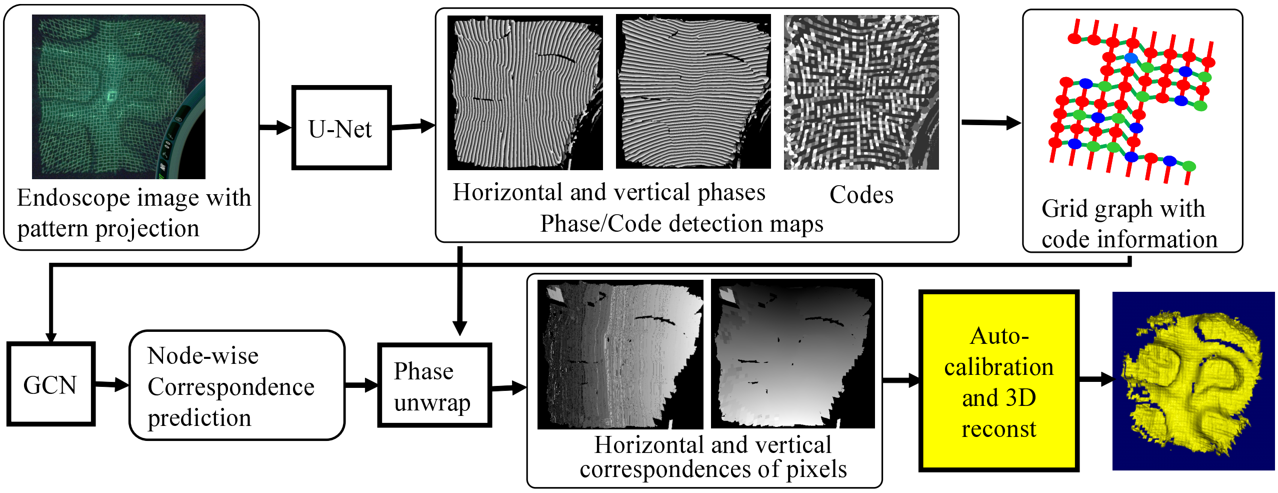

Endoscope image with pattern projection
U-Net
Horizontal and vertical phases
Codes
Phase/Code detection maps
Grid graph with code information
GCN
Node-wise Correspondence prediction
Phase unwrap
Horizontal and vertical correspondences of pixels
Auto-calibration and 3D reconst

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

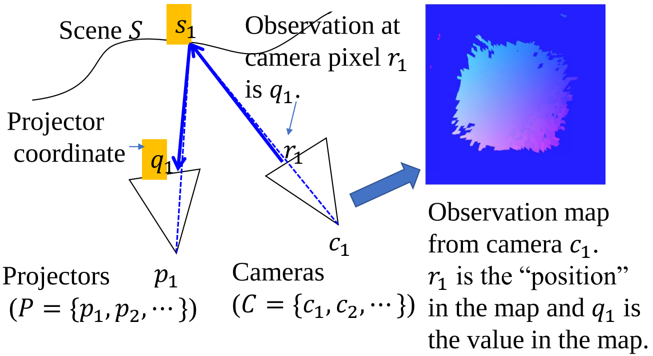

Scene S
s_1
Observation at camera pixel r_1 is q_1.
Projector coordinate
q_1
r_1
c_1
Projectors p_1
$(P = \{p_1, p_2, \cdots\})$
Cameras
$(C = \{c_1, c_2, \cdots\})$
Observation map from camera c_1.
r_1 is the "position" in the map and q_1 is the value in the map.

FIGURE 3 Observation model of active stereo systems

model dynamics of active stereo observations and has often problems in convergence. In contract, our technique is faster and stable. Note that since active stereo techniques for laparoscopes and endoscopes have been widely studied[25,11,24], auto-calibration for both single and multi-frame method proposed in this paper can be applied to any type of active stereo systems and useful.

Differentiable renderer renders CG images using a scene or camera parameters where gradient-based optimization w.r.t. the parameters can be processed. Mesh-based differentiable renderers[26,27,28,29] have been proposed to optimize various scene parameters such as geometry, illumination, textures, or materials.

3 | SYSTEM CONFIGURATION AND ALGORITHM

An experimental system, which allows for 3D shape reconstruction from a single frame, consist of a pattern projector inserting 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.*[30]. The structured light illumination is created by the diffractive optical element (DOE) included in the pattern projector as shown in Fig. 1(b). We can perform scanning of the target shapes by projecting the pattern onto the surface and subsequently capturing the image. Furthermore, we have the capability to reposition the camera and projector while capturing multiple images (as illustrated in Fig.1(b) and (c)). The scanned shapes obtained from these multi-frame images can be integrated, ensuring shape consistency between different frames. We used a grid pattern consisting of vertical and horizontal edges with small gaps (Fig.1(d)). The gaps represent code symbols to identify camera-to-projector mapping as shown in Fig.1(e), where the red dots mean that the vertical and horizontal edges does not have gaps, the green dots and blue dots have gaps between horizontal edges with different gap directions (green means "the left is higher" and blue means "the right is higher"),

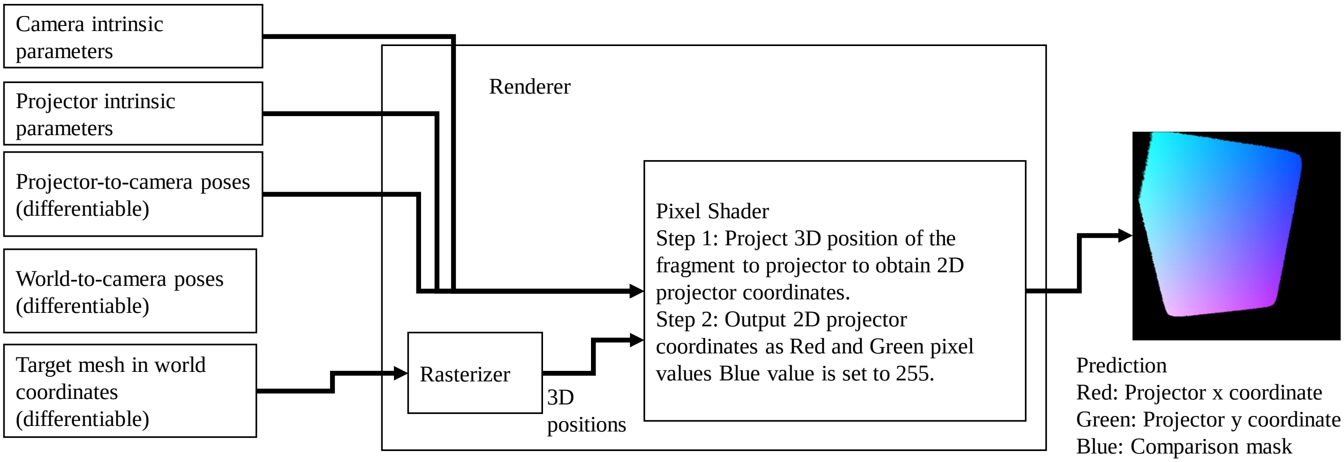

Camera intrinsic parameters
Projector intrinsic parameters
Projector-to-camera poses (differentiable)
World-to-camera poses (differentiable)
Target mesh in world coordinates (differentiable)
Renderer
Rasterizer
3D positions
Pixel Shader
Step 1: Project 3D position of the fragment to projector to obtain 2D projector coordinates.
Step 2: Output 2D projector coordinates as Red and Green pixel values Blue value is set to 255.
Prediction
Red: Projector x coordinate
Green: Projector y coordinate
Blue: Comparison mask

FIGURE 4 Differentiable renderer with differentiable variables of C, P, and S

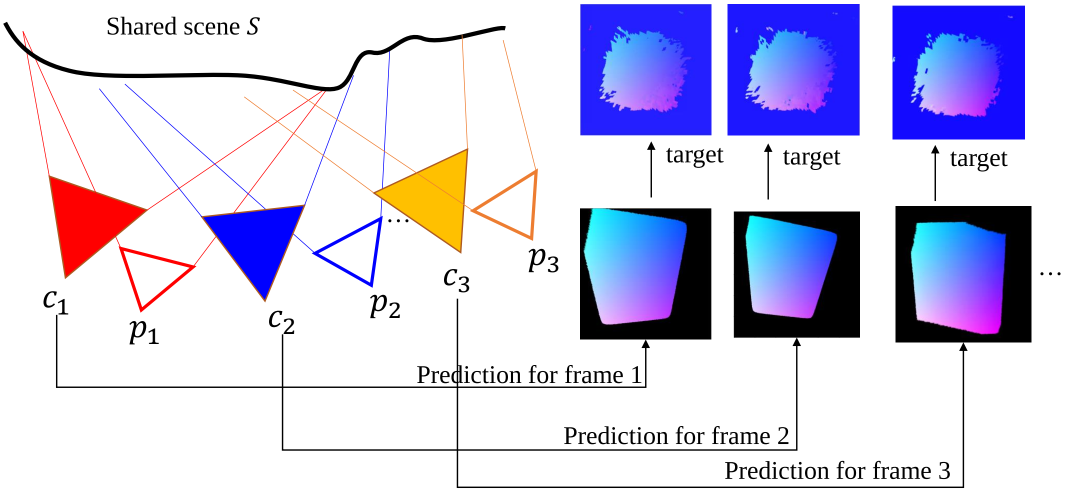


FIGURE 5 Multi-frame optimization via scene S : the prediction is projector coordinates(Red:x, Green:y) predicted for C, P, and S

and the cyan and magenta dots have gaps between vertical edges similarly. The actual patterns projected onto the object surface are shown in Fig.1(f) and (g).

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) for the horizontal direction and the vertical direction. Also, another U-Net is applined to extract code information. These U-Nets can be trained with CG images that simulates pattern projection.

Then, the grid structure and code information are converted to a grid graph (the top-left graph in Fig.2). These conversion can be done by segmentation similar to[30]. The graph is processed by a graph convolutional network (GCN), resulting in node-wise correspondences. The GCN can be trained with CG-generated graph data.

Using the node-wise correspondences and the pixel-wise phase information, pixel-wise correspondences are obtained (*i.e.*, the horizontal and vertical correspondences between the camera and the projector). These pixel-wise correspondences are the inputs assumed in section 4.

4 | DIFFERENTIAL RENDERING FOR AUTO-CALIBRATION

4.1 | Cost funtion

In this paper, to achieve auto-calibration of projector-camera system from single or multi frame without special markers nor pre-calibration, a method based on differential renderer method is proposed. Note that although auto-calibration methods for active stereo for endoscopic system are already proposed, they assumed special markers[24], accurate initial parameters and cannot handle the focal length of the projector[30], which are all effectively solved by our method as follows.

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 is estimated by the process in section 3. This can be modeled as a 2D-to-2D mapping

$$\boldsymbol{H_f} \;:\; \mathbb{R}^2 \mapsto \mathbb{R}^2 \,;\, \boldsymbol{r} \mapsto \boldsymbol{q} \tag{1}$$

from a camera pixel r to a projector pixel q, where f is a frame index. For example, Fig.3 shows a mapping from r_1 to q_1. r and q are 2D vectors, thus, $(q_x, q_y) \equiv H_f(r_x, r_y)$. Map H_f can be represented as a 2D image shown in Fig.3 on the right side, where the red and blue channels of pixel (r_x, r_y) represent x and y components of $H_f(r_x, r_y)$. H_f represents all the observed information obtained from the f-th scanning.

Since the mapping H_f represents observations of a pattern projection, it should be geometrically approximated by "ray-tracing" from camera pixels r_1 to surface S, resulting in s_1, and projecting s_1 onto projector p_1, resulting in q_1 as shown in Figure 3. We represent the ray-tracing model with a mapping function

$$G_{S,c_f,p_f} : \mathbb{R}^2 \mapsto \mathbb{R}^2, \tag{2}$$

parametrized by the scene S and the parameters of camera c_f and projector p_f for frame f. The function G_{S,c_f,p_f} maps camera coordinates r to projector coordinates q using the following calculations. For example, the camera ray of q_1 is calculated using camera parameter c_f, as shown in Fig.3. Then, the intersection of the camera ray and the scene surface S is calculated as s_i. s_i is then projected onto projector p_f, resulting in 2D coordinates q_i. This process is the same calculation of “projection mapping” used in computer graphics.

We calibrate the set of cameras $C \equiv \{c_1, c_2, \cdots\}$ and the set of projectors $P \equiv \{p_1, p_2, \cdots\}$ by fitting the goemetrical model G_{S,c_f,p_f} to observation H_f. The cost function is

$$L(S, C, P) \equiv E(S, C, P) + R_e(S) + R_n(S) + R_l(S), \tag{3}$$

$$E(S, C, P) \equiv \sum_{f} \sum_{(r_x, r_y) \in D} M_f(r_x, r_y)\, \rho(\|H_f(r_x, r_y) - G_{S,c_f,p_f}(r_x, r_y)\|^2), \tag{4}$$

$$\rho(x) \equiv \log(1 + x/(a^2)), \tag{5}$$

where $E(S, C, P)$ is data term for fitting G_{S,c_f,p_f} to H_f, a is the noise level for the values of $H_f(r_x, r_y)$ defined empirically, D is the set of pixel locations of a camera image, $M_f : \mathbb{R}^2 \mapsto \{0, 1\}$ is a mask funciton where $M_f(r_x, r_y)$ is 1 if $H_f(r_x, r_y)$ is a valid observation and 0 otherwise, ρ is Cauchy-Loss funciton used for robustifying the optimization, and R_e, R_n, and R_l are mesh-regularization terms to keep S smooth. The mesh-regularization terms are edge-length loss, normal-direction loss, and laplacian loss, respectively, which are defined in Pytorch3D.

4.2 | Optimization

The loss function L is minimized by differential rendering. $\{M_f(r_x, r_y) | (r_x, r_y) \in D\}$ can be represented as an image shown in Fig.3. $\{G_{S,c_f,p_f}(r_x, r_y) | (r_x, r_y) \in D\}$ can be rendered as an image using surface mesh S, camera parameters c_f, and projector parameters p_f, as shown in Fig.4.

Note that rendering of $\{G_{S,c_f,p_f}(r_x, r_y) | (r_x, r_y) \in D\}$ can be efficiently achieved by using a *pixel shader* of GPU using differential renderer packages. For example, the rasterizers of mesh-based differential renderer packages returns the intersections of the rays corresponding to the image pixels D and surface S. The intersection points can be projected onto projectors P using, for example, pixel shaders.

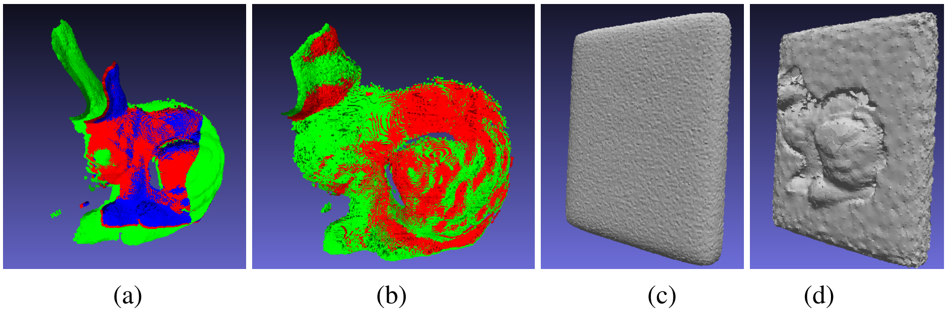


FIGURE 6 Calibration results of simulation data. (a) Shapes before and after single-frame auto-calibration. Colored point clouds are (red) the ground-truth shape , (green) before auto-calibration , and (blue) after auto-calibration. The green colored shape reconstructed by un-calibrated parameters was severely distorted, whereas the ground-truth shape and the auto-calibrated shapes were almost identical. (b) Result of 2-frame calibration, where red points are frame #1 and greens are #2. Note that multiple frames were fit to each other by disambiguation of scale inconsistency by auto-calibration. (c) Initial shape and (d) shape after optimization, showing rabbit-shape is successfully recovered.

Using the rendered images G_{S,c_f,p_f}, $L(S,C,P)$ can be calculated. By minimizing $L(S,C,P)$ with respect to S,C and P using differential renderer's capability, the rendered images $\{G_{S,c_f,p_f}(r_x,r_y)\,|\,(r_x,r_y) \in D\}$ approaches to the observed mapping $\{M_f(r_x,r_y)\,|\,(r_x,r_y) \in D\}$ as shown in Fig.5. Then, we can estimate C, P and S.

In the experiments shown in this paper, c_f includes camera pose parameters, p_f includes projector pose parameters and projector focal length. S is a polygon mesh. The optimization of $L(S,C,P)$ is done by Adam. The inter-frame positions for the cameras C and the projectors P are also optimized with this method, because all the views share a single scene model S.

Although Fig.5 shows the multi-frame case, single-frame optimization is also possible. It is important to note that the rendered and optimized images are not the projected raw patterns, but rather the projector coordinates. The image of the projector coordinates has much better properties for optimization. The projector coordinates change monotonously on smooth surfaces, whereas the projected images tend to exhibit a repetitive grid structure, which can lead to local minima in terms of image similarities.

4.3 | Remeshing

In the process of optimizing the data term $E(S,C,P)$ in the cost function, a large deformation of the 3D mesh surface S may be required. However, since the mesh regularization terms R_e, R_n, and R_l aim to suppress large deformations, they may compete with each other. To alleviate this problem, we have implemented a remeshing technique for S, which can be executed periodically during the optimization steps.

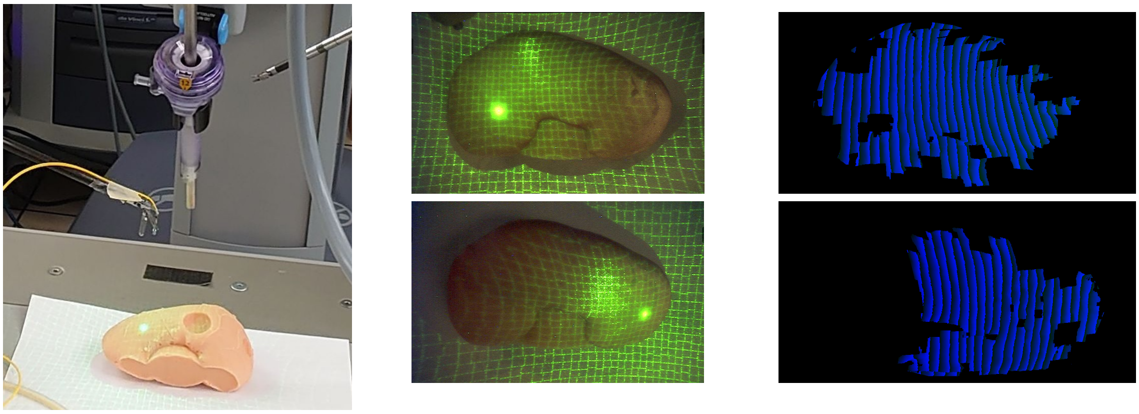
The remeshing process involves “collapsing short edges” and “splitting long edges”. In the process of “collapsing short edges”, edges less than half of the average edge length are collapsed. In the process of “splitting long edges” edges longer than twice the average edge length are split. These operations do not alter the mesh topology but help to correct the non-uniformity of polygon sizes caused by large deformations. It's important to note that the non-uniformity of polygon sizes is also detrimental to estimating the regularization terms R_e, R_n, and R_l appropriately.

5 | EXPERIMENTS

5.1 | Validation by simulation data

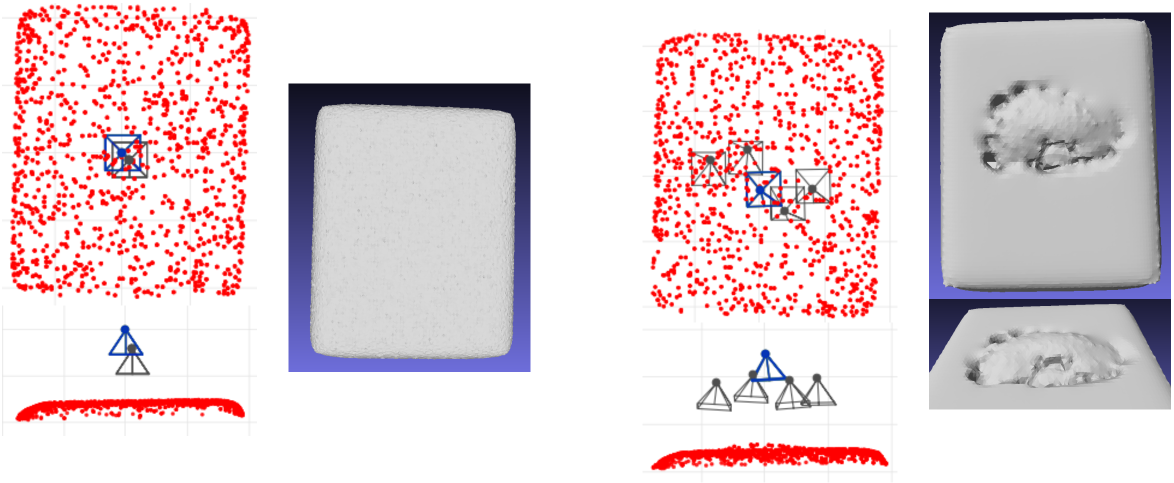
We implemented the proposed method with Pytorch(1.7.0) and Pytorch3D(0.7.2). on a system with 16GB of GPU memory. A typical execution time was about 20 min for 1000 iterations of optimization step.

To validate the auto-calibration with the proposed method, we synthesized a simulation data with a rabbit-shaped mesh model using a projection mapping technique. The size of the rabbit shape was about 1.0 in vertical length and the distance from the camera was about 2.0. The distance between the camera and the projector, *i.e.*, the baseline length, was 0.1. Multi-frame data



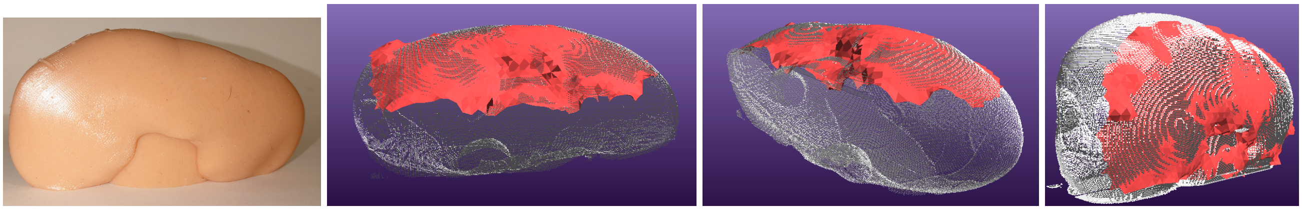
(a) Laparoscope (b) Capture image (c) X coordinates of pattern

FIGURE 7 (a) A head of surgical robotic system, (b) captured images with structured light and (c) X coordinate of pattern projector.



(a) Before optimization (b) After optimization

FIGURE 8 Camera and projector positions and reconstructed shape. Blue lines represents cameras, gray lines represents projectors, and red points are represents shape. (a) Before optimization. (b) After optimization.



(a) Phantom (b) Reconstructed shape with the ground-truth shape

FIGURE 9 (a) Phantom of kidney, (b) 3D reconstructed shapes (red mesh of the three right images) of a 'front' shape of a kidney-shaped phantom (the left-most image). The shape is fit to the ground-truth point cloud obtained from X-ray CT scan (white points in the three right images).

with two frames was generated by moving the object with horizontal motions by length of 0.2. All the lengths had no units since the data was generated by CG.

Then, we intentionally added Gaussian noise to projector and camera poses. The translation elements in the x, y, and z directions were perturbed with a standard deviation of 0.05. The quaternion rotation elements of x, y and z components were purturbed with a standard deviation of 0.05, where the quaternion elements consists of x, y, z and w components. After the addition of perturbations, the perturbed quaternions were normalized to maintain a length of 1 unit. For projector focal length, noise level was $\sigma = 5.0$ for the actual focal legth of 500.0 pixels.

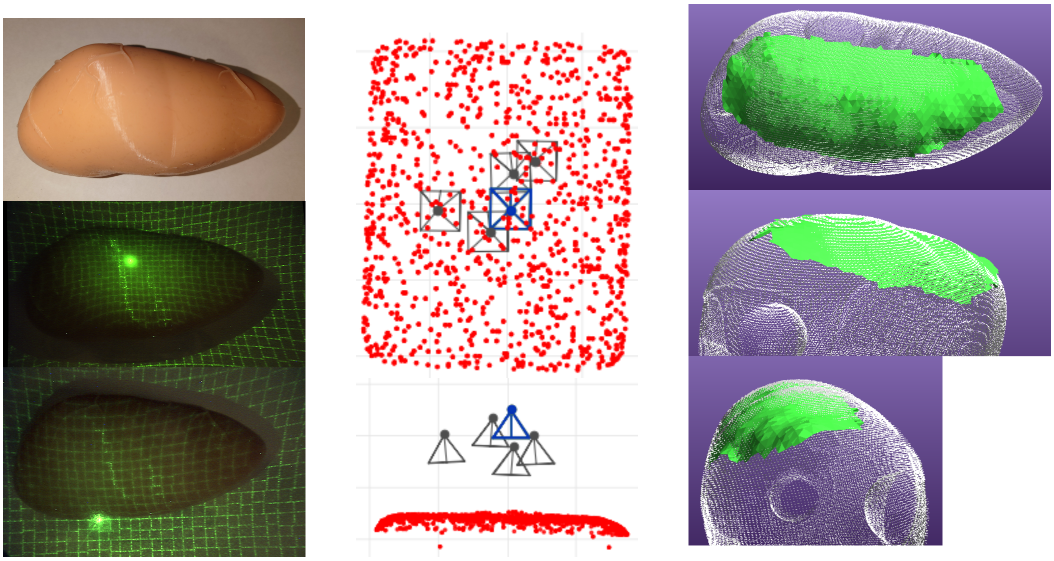


FIGURE 10 Scanning and calibration results of the 'top' region of the kidney-shaped phantom. The left column:he appearance of the surface and captured images. The middle column: the positions of the projector and camera after optimization. The initial position was the same as Fig.8(a). The right column: reconstructed shape fit to the ground-truth shape.

The reconstructed shape using the perturbed parameters is shown as green shape of Fig.6(a). By comparing it to the ground-truth shape (red in (a)), it is confirmed that the shape which is not calibrated is largely distorted and apart from the ground truth. Then, the proposed method was applied to auto-calibrate the camera and projector poses as well as estimate the target shapes. Using the optimized C and P, we obtained blue shape in (a), where we can confirm that the reconstructed shape is almost identical to the ground-truth shape. Fig.6(c) and (d) are the shapes before and after optimization, where we can confirm that original shape is successfully recovered by our method. Also, 2-frame calibration was processed by optimizing multiple frames, where baselines were set largely different. From Fig.6(b), it is confirmed that ambiguity of scaling between frames were successfully solved to make consistent shapes.

For the single-frame auto-calibration experiment, we assessed the errors in relative pose parameters between the camera and the projector before and after optimization. We also examined shape errors before and after optimization. In the experiment, we generated point clouds using the correspondence map. To evaluate the disparities between shapes, we employed the Iterative Closest Point (ICP) algorithm, aligning shapes by minimizing the Root of Mean Squared Errors (RMSE) between them. The residual alignment cost represented as RMSE was utilized as a measure of shape difference. To avoid genuine scale ambiguity, projector-camera baselines after optimization were normalized to be actual length of 0.1. The errors in relative pose between the projector and the camera, the focal length error, and shape errors are presented in Tab.1.

TABLE 1 Errors before and after single-frame auto-calibration

	Initial values (Perturbed by Gaussian noise)	Optimized values
Pose translation errors (RMSEs of x, y and z)	8.5×10^{-2}	4.2×10^{-4}
Pose rotation errors (RMSEs of x, y and z rotations in radian)	5.0×10^{-2}	2.7×10^{-3}
Projector focal-length errors from the ground truth	4.2	0.71
Shape errors from the ground truth (RMSEs of ICP errors)	1.8×10^{-2}	2.1×10^{-3}

For the multi-frame calibratin experiment aimed at demonstrating scale-consistency optimization, we assessed scale-consistency through the comparison of baseline length ratios. In the experiment, the ground-truth baselines are the same for the two frames. Thus, the ground-truth value of the baseline ratios are 1.0. The baseline ratios alongside the shape errors, before and after optimization, are provided in Tab.2.

TABLE 2 Errors before and after single-frame auto-calibration. The ground-truth value of ratio of baseline lengths is 1.0.

	Initial values (Scaled by baseline lengths)	Optimized values
Ratio of baseline lengths of two frames $\left(\frac{\max(baseline1,baseline2)}{\min(baseline1,baseline2)}\right)$	1.2	1.000996
Shape errors (RMSEs of ICP errors)	1.3×10^{-2}	2.0×10^{-3}

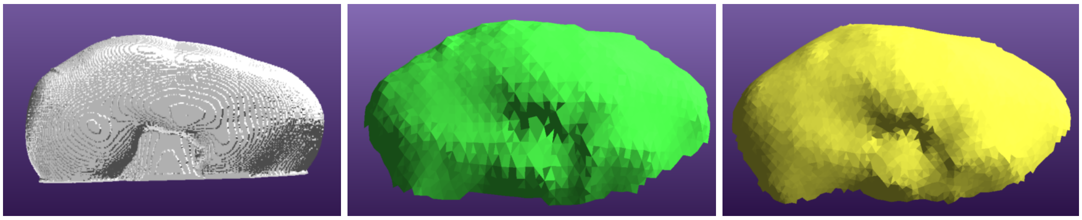

(a) (b) (c)

FIGURE 11 Optimization results with and without periodical remeshing: (a) The ground-truth shape. (b) The result without periodical remeshing. (c) The result with periodical remeshing.

5.2 | Single and multi-frame auto-calibration of real data

Next, we captured sequential images using a medical surgery system, where its head consists of a camera and a projector as shown in Fig.7(a). Note that the projector is attached to the robotics arm which is different from the one with camera, and thus, it will be freely moved relative to the camera during the operation. The target shape is a medial shape model (phantom) of a kidney as shown in Fig.9(a). Examples of captured images are shown in Fig.7(b).

In this example, we used four-frame sequence. Fig.8 shows the position of the camera and the projector for each frame and reconstructed shapes (a) before and (b) after optimization. Note that all the results shown in Fig.8(b) were obtained from a fixed initial projector-to-camera pose parameters as shown in Fig.8(a), 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 *xy* plane with no rotation).

In this example, we used the optimized polygon mesh for evaluation. As shown in Fig.8 (b), the mesh had large portion of unoptimized regions. Thus, we manually removed these unoptimized regions. Then, the output point clould wass generated by extracting vertices from the mesh. Also, the ground-truth point could was generated from X-ray CT by extracting borders of frame-wise segmented regions.

Then, ICP registration was performed between the ground-turth point cloud and the output point cloud. Fig.9(b) shows the reconstructed shapes with the ground truth captured by X-ray CT scan, confirming the reconstructed shape is aligned closely to the ground truth. Fig.10 shows an experiment on another surface of the shape. The RMSEs calculated as ICP errors between reconstructed shapes and the ground-truth shape (Fig.9 and Fig.10) were 0.482 mm and 0.742 mm, where the longest-axis length of the phantom was about 120mm. Thus, we can say that our proposed method can achieve high-precision 3D reconstruction without any pre-calibration.

5.3 | Effect of remeshing in optimization process

To validate the effect of the remeshing capability, we compared the optimization results with and without periodic remeshing. In the case of periodic remeshing, the surface *S* was remeshed every 80 iterations of optimization. During the remeshing process, edges shorter than half of the mean edge length were collapsed, and edges longer than a threshold were split. These processes were implemented using the PyMesh library.

Fig.11 shows the results. The result with periodic remeshing maintains finer shape features compared to the case without remeshing because mesh triangles were stretched during the optimization process when remeshing was not applied. The ICP errors from the ground-truth shape were 1.93mm for the result shown in Fig.11(b) and 1.79mm for the result shown in Fig.11(c).

5.4 | Limitations

The limitation of the proposed method is the condition where the target shape is too simple, such as near plane. For such situation, auto-calibration using projector-camera correspondences becomes unstable.

The proposed method largely depends on projector-camera correspondence prediction shown in Fig.2 Although the cost function in section 4.1 include Cauchy-Loss for robustifying the optimization, noise in projector-camera correspondence prediction may cause unstable reconstruction. Thus, improving the accuracy of the projector-camera correspondence prediction and implementing outlier removal for optimization would be needed for future study.

Another limitation is that the polygon mesh does not chenge in topology, and discontinuities in target shapes are not processed properly. For example, in Fig.8, boundary between the phantom and the background was not reconstructed. Reconstruction that can handle such situation is future study.

6 | CONCLUSION

In the paper, a novel multi-frame auto-calibration for active-stereo scanning with laparoscope is proposed, where an optimization with a differentiable renderer estimates the projector and camera poses, even if initial positions are largely apart from the ground truth. In the method, active-stereo observation process is directly modeled by using a CG renderer with a customized pixel shader, and minimizing the observation errors by differentiable rendering technique. The proposed method was confirmed to work with both single-frame and multi-frame conditions in the simulation data, and also for some scanned data using a remote surgery system.

We argue that the proposed method holds the potential to advance research in the ‘marker-less structured-light approach’ for laparoscopic and endoscopic 3D scanning, which has not been sufficiently explored. However, in the current state, the clinical practicality of the proposed method is not sufficiently validated.

In the future, more realistic setup should be tested, *e.g.*, closely resembling abdominal laparoscopic procedures. Also, comparison with previous methods, such as marker-based approach (*e.g.*, [25]) should also needed to validate clinical practicality of the proposed method. Development of AR/VR system with realtime process for actual remote surgery is also planned.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

ACKNOWLEDGMENTS

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

References

1. Song J, Wang J, Zhao L, Huang S, Dissanayake G. Dynamic reconstruction of deformable soft-tissue with stereo scope in minimal invasive surgery. *IEEE Robotics and Automation Letters.* 2017;3(1):155–162.

2. Song J, Wang J, Zhao L, Huang S, Dissanayake G. Mis-slam: Real-time large-scale dense deformable slam system in minimal invasive surgery based on heterogeneous computing. *IEEE Robotics and Automation Letters.* 2018;3(4):4068–4075.

3. Widya AR, Monno Y, Imahori K, et al. 3D reconstruction of whole stomach from endoscope video using structure-from-motion. In: IEEE. 2019:3900–3904.

4. Haouchine N, Dequidt J, Peterlik I, Kerrien E, Berger MO, Cotin S. Image-guided simulation of heterogeneous tissue deformation for augmented reality during hepatic surgery. In: IEEE. 2013:199–208.

5. Stoyanov D, Scarzanella MV, Pratt P, Yang GZ. Real-time stereo reconstruction in robotically assisted minimally invasive surgery. In: Springer. 2010:275–282.

6. Brandao P, Psychogyios D, Mazomenos E, Stoyanov D, Janatka M. HAPNet: hierarchically aggregated pyramid network for real-time stereo matching. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization.* 2021;9(3):219–224.

7. Long Y, Li Z, Yee CH, et al. E-dssr: efficient dynamic surgical scene reconstruction with transformer-based stereoscopic depth perception. In: Springer. 2021:415–425.

8. Turan M, Almalioglu Y, Araujo H, Konukoglu E, Sitti M. Deep endovo: A recurrent convolutional neural network (rcnn) based visual odometry approach for endoscopic capsule robots. *Neurocomputing.* 2018;275:1861–1870.

9. Yang G, Manela J, Happold M, Ramanan D. Hierarchical deep stereo matching on high-resolution images. In: 2019:5515–5524.

10. Ye M, Johns E, Handa A, Zhang L, Pratt P, Yang GZ. Self-supervised siamese learning on stereo image pairs for depth estimation in robotic surgery. *arXiv preprint arXiv:1705.08260.* 2017.

11. Lin J, Clancy NT, Stoyanov D, Elson DS. Tissue surface reconstruction aided by local normal information using a self-calibrated endoscopic structured light system. In: Springer. 2015:405–412.

12. Furukawa R, Mizomori M, Hiura S, Oka S, Tanaka S, Kawasaki H. Wide-area shape reconstruction by 3D endoscopic system based on CNN decoding, shape registration and fusion. In: , , Springer, 2018:139–150.

13. Globus Medical . Spine surgery surgical navigation system ExcelsiusGPS. 2020. https://www.globusmedical.com/musculoskeletal-solutions/excelsiustechnology/excelsiusgps/.

14. Stryker . Surgical navigation platform Stryker NAV3i. 2019. https://www.stryker.com/.

15. Liao J, Cai L. A calibration method for uncoupling projector and camera of a structured light system. In: IEEE. 2008:770–774.

16. Yamauchi K, Saito H, Sato Y. Calibration of a structured light system by observing planar object from unknown viewpoints. In: IEEE. 2008:1–4.

17. Mahmoud N, Collins T, Hostettler A, Soler L, Doignon C, Montiel JMM. Live tracking and dense reconstruction for handheld monocular endoscopy. *IEEE transactions on medical imaging.* 2018;38(1):79–89.

18. Chen L, Tang W, John NW, Wan TR, Zhang JJ. SLAM-based dense surface reconstruction in monocular minimally invasive surgery and its application to augmented reality. *Computer methods and programs in biomedicine.* 2018;158:135–146.

19. Leonard S, Sinha A, Reiter A, et al. Evaluation and stability analysis of video-based navigation system for functional endoscopic sinus surgery on in vivo clinical data. *IEEE transactions on medical imaging.* 2018;37(10):2185–2195.

20. Lamarca J, Parashar S, Bartoli A, Montiel J. Defslam: Tracking and mapping of deforming scenes from monocular sequences. *IEEE Transactions on robotics.* 2020;37(1):291–303.

21. Zhou H, Jayender J. EMDQ-SLAM: Real-Time High-Resolution Reconstruction of Soft Tissue Surface from Stereo Laparoscopy Videos. In: Springer. 2021:331–340.

22. Combès B, Prima S. An efficient EM-ICP algorithm for symmetric consistent non-linear registration of point sets. In: Springer. 2010:594–601.

23. Sinko M, Kamencay P, Hudec R, Benco M. 3D registration of the point cloud data using ICP algorithm in medical image analysis. In: IEEE. 2018:1–6.

24. Furukawa R, Nagamatsu G, Oka S, et al. Simultaneous shape and camera-projector parameter estimation for 3D endoscopic system using CNN-based grid-oneshot scan. *Healthcare Technology Letters.* 2019;6(6):249–254.

25. Geurten J, Xia W, Jayarathne U, Peters TM, Chen EC. Endoscopic laser surface scanner for minimally invasive abdominal surgeries. In: Springer. 2018:143–150.

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

27. Palazzi A, Bergamini L, Calderara S, Cucchiara R. End-to-end 6-dof object pose estimation through differentiable rasterization. In: 2018:0–0.

28. Kato H, Ushiku Y, Harada T. Neural 3d mesh renderer. In: 2018:3907–3916.

29. Liu HTD, Tao M, Jacobson A. Paparazzi: surface editing by way of multi-view image processing.. *ACM Trans. Graph..* 2018;37(6):221–1.

30. Furukawa R, Oka S, Kotachi T, et al. Fully Auto-calibrated Active-stereo-based 3D Endoscopic System using Correspondence Estimation with Graph Convolutional Network. In: IEEE. 2020:4357–4360.