Received: 1 November 2024 | Accepted: 11 November 2024

Healthcare Technology Letters

DOI: 10.1049/htl2.70001

The Institution of Engineering and Technology WILEY

LETTER

Incremental shape integration with inter-frame shape consistency using neural SDF for a 3D endoscopic system

Ryo Furukawa[1] | **Hiroshi Kawasaki**[2] | **Ryusuke Sagawa**[3]

[1]Department of Informatics/Graduate School of System Engineering, Kindai University, Higashihiroshima, Japan

[2]Faculty of Information Science and Electrical Engineering Department of Advanced Information Technology, Kyushu University, Fukuoka, Japan

[3]Artificial Intelligence Research Center, The National Institute of Advanced Science and Technology, Tsukuba, Japan

Correspondence
Ryo Furukawa, Department of Informatics/Graduate School of System Engineering, Kindai University, 1 Umenobe, Takaya, Higashihiroshima, Higashihiroshima 739-2116, Japan.
Email: furukawa@hiro.kindai.ac.jp

Funding information
JST Startup, Grant/Award Number: JPMJSF23DR; Japan Society for the Promotion of Science, Grant/Award Numbers: JP18H04119, JP20H00611; New Energy and Industrial Technology Development Organization, Grant/Award Number: JPNP20006

Abstract

3D measurement for endoscopic systems has been largely demanded. One promising approach is to utilize active-stereo systems using a micro-sized pattern-projector attached to the head of an endoscope. Furthermore, a multi-frame integration is also desired to enlarge the reconstructed area. This paper proposes an incremental optimization technique of both the shape-field parameters and the positional parameters of the cameras and projectors. The method assumes that the input data is temporarily sequential images, that is, endoscopic videos, and the relative positions between the camera and the projector may vary continuously. As solution, a differential volume rendering algorithm in conjunction with neural signed distance field (NeuralSDF) representation is proposed to simultaneously optimize the 3D scene and the camera/projector poses. Also, an incremental optimization strategy where the optimized frames are gradually increased is proposed. In the experiment, the proposed method is evaluated by performing 3D reconstruction using both synthetic and real images, proving the effectiveness of our method.

1 | INTRODUCTION

3D measurement for endoscopic systems has large potential for various purposes, such as cancer diagnosis, robotic-surgery systems, and making annotations for endoscopic image databases for deep learning.

Furukawa et al. have proposed 3D endoscopic systems based on active stereo [1–3]. The system consists of a pattern projector inserted through the instrumental channel of a standard endoscope as shown in Figure 1a. It allows for 3D shape reconstruction from a single frame by projecting the pattern onto the target surfaces and capturing the images with the endoscopic camera.

One open problem of these systems is the pose calibration of the relative poses of the projector with respect to the camera, which inevitably occurs since the projector cannot be tightly fixed to the endoscope head. Another problem is the difficulty in capturing a sufficient area of internal organs due to the limited field of view of the camera and the restricted projection area of the micro-sized projector. One simple solution for both problems is multi-frame integration to generate consistent 3D shape by optimizing the camera/projector poses.

The straightforward approach for multi-frame integration is a step-by-step approach. For instance, 'frame-wise' auto-calibration can be applied to reconstruct the 3D shape of each frame followed by an iterative closest point (ICP) algorithm [4] to estimate ego-motion between frames. In the end, all the shapes are integrated using for example truncated signed distance field (TSDF) based algorithm [5]. However, frame-wise auto-calibration introduces frame-wise calibration errors, which causes shape inconsistency between frames.

The inter-frame shape inconsistency is a large problem for real applications. For example, if a depth-annotated endoscopic

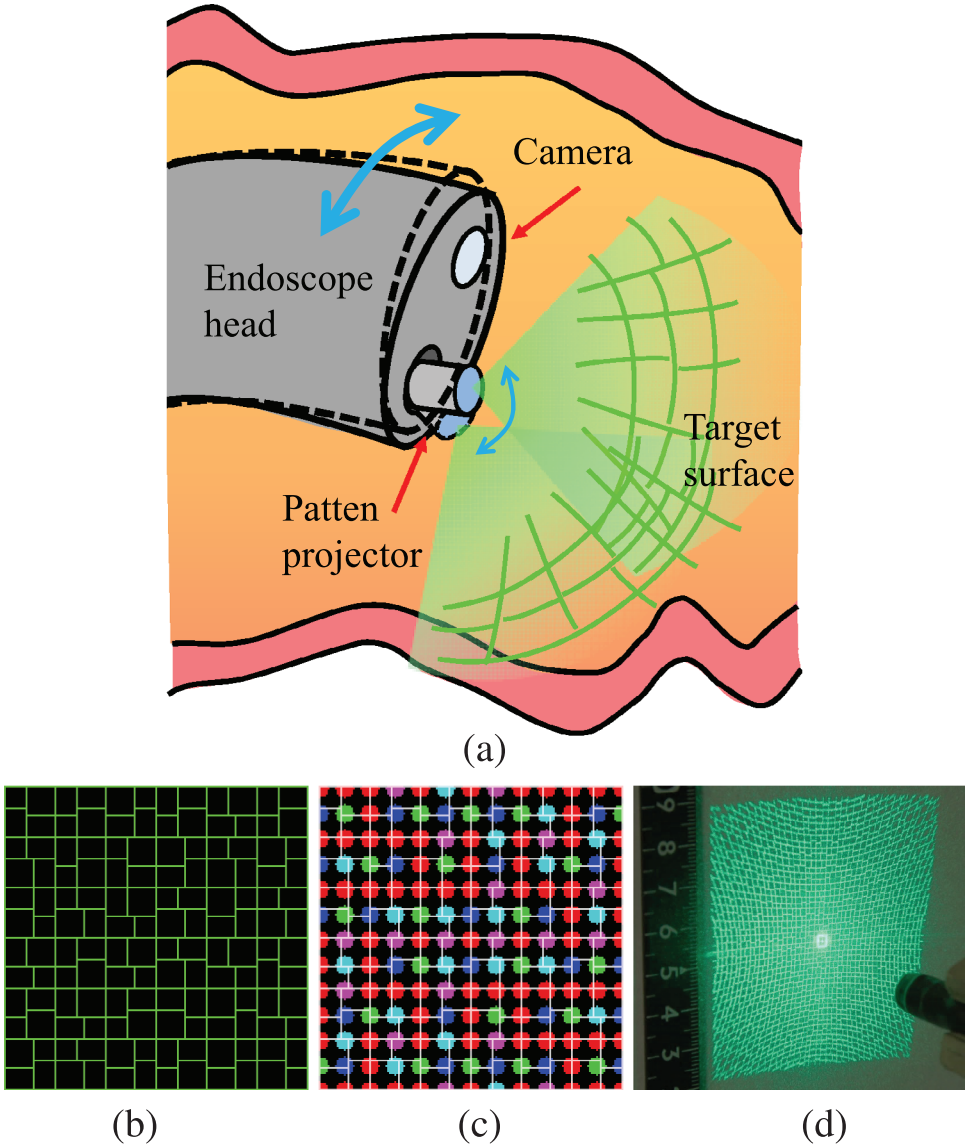

FIGURE 1 (a) Active-stereo 3D endoscopic system. (b) The projected grid pattern. (c) Code information embedded into the pattern. (d) The pattern illuminated onto a plane.

image database includes such inconsistency, it may prevent proper learning of single-frame depth estimation models.

To suppress such inter-frame shape inconsistency, all the projector and camera poses should be optimized globally, utilizing information of multiple frames. Furukawa et al. proposed a multi-frame optimization technique to solve the issue by employing a loss function that directly models active-stereo pattern projection [6]. They implemented the optimization by using triangle-mesh-based differential renderer. One problem with the approach is that it requires an initial 3D triangle mesh with sufficient accuracy to make the optimization process stable. However, it is not an easy task for real endoscopic systems, which reduces the practicality of the systems.

This paper proposes a technique to scan a target object from multiple images taken while a camera and a projector are both in motion. The relative poses are auto-calibrated by using our neural signed-distance-field (NeuralSDF or N-SDF) which is optimized by novel volumetric differential rendering technique.

In our method, similar to the method of Furukawa et al. [1], a laser pattern projector which forms grid-like patterns to find unique correspondences only from a single image was used. By utilizing the correspondences, frame-wise auto-calibration is performed to estimate initial pose parameters between the projector and the camera, which are both in motion. Since the auto-calibrated parameters include inevitable errors as well as scaling ambiguity, we employ optimization techniques using the volumetric differential rendering technique to optimize the pose parameters of the camera and the projector for all the frames. In the process, we render a mapping from the camera pixels to the projector coordinates onto the surface derived from NeuralSDF with a differential renderer. By using loss functions that estimate differences between the rendered and observed pattern and

projector coordinates, the NeuralSDF, camera poses, and projector poses can be simultaneously optimized.

In the experiment, the proposed method is evaluated by performing 3D reconstruction using both synthetic and real images obtained by a pattern projector and an endoscopic camera.

The contributions of this paper are summarized as follows:

1. We propose a method to reconstruct 3D shapes using observation from multiple view points using structured-light projection. It utilizes a differential renderer specialized for structured-light systems.
2. We utilize neural shape representation inspired by NeuS [7], although colour representation is totally modified to fit to structured-light systems. Accordingly, new loss functions utilizing both projector-camera correspondences and pattern appearances are proposed.
3. We propose an incremental optimization strategy for sequentially captured images. In the proposed method, only the first few frames are optimized initially. Then, the subsequent frames are incrementally added to the optimization system, as the shape and pose optimization proceeds. Using the method, image sequences with a large number of frames can be processed stably.

2 | RELATED WORKS

3D reconstruction algorithm using a series of images captured by monocular camera have been researched intensively in computer vision community, known as SLAM or SfM, and also applied to endoscopic systems, such as Mahmoud et al. [8], Chen et al. [9], and Leonard et al. [10]. Although common techniques assume rigid object, they are recently extended to handle non-rigid object specialized for internal organs and/or soft tissues, such as Song et al. [11], Lamarca et al. [12], and Zhou et al. [13]. One drawback of these methods is that they need 3D feature points, meaning rich textures, which is not the case for endoscopic images.

To overcome the problem of monocular camera systems, active stereo has been widely used. To calibrate the relative pose between the projector and the camera, special patterns are projected onto the calibration object to retrieve correspondences [14, 15]. Since it is usually assumed that a projector and a camera are fixed with each other, several patterns are used for pre-calibration, which cannot be true for endoscopic system, where they are not fixed with each other as shown in Figure 1a. Thus we need a solution.

To integrate several 3D shapes, usually ICP algorithm has been used [4, 16–19]. However, in our system, since a projector and a camera are not fixed to each other, 3D shapes are not consistent, and thus, ICP cannot be used. To overcome the problem, Furukawa et al. proposed a modification of bundle adjustment for active stereo systems [2, 6]. Since their method does not directly model dynamics of active stereo observations, it is problematic in convergence, if a projector and a camera are not set in front parallel configuration.

Recently, a differentiable renderer, which can render synthetic images using a scene and camera parameters where gradient-based optimization with respect to the parameters, has been proposed. Mesh-based differentiable renderers [20–23] have been proposed to optimize various scene parameters such as geometry, illumination, textures, or materials and intensively researched, however, they cannot essentially solve the occlusion problem. On the other hand, volume-based renderers [24–26] also draw wide attention because of solving occlusion problems by using neural-network-based representations called NeRF [27]. This paper uses a volume-based differentiable renderer, because of its ability on general scene and its stability. To achieve a stable process, we also adopt signed distance field (SDF) based neural representation inspired by neural-SDF (NeuS) [7] as well as hash-grid encoding technique [28].

3 | 3D RECONSTRUCTION WITH NEURAL SHAPE REPRESENTATION FROM STRUCTURED-LIGHT PROJECTION

3.1 | Overview

This paper proposes a 3D reconstruction method using pattern projection observed by multiple camera positions. The input images are multiple images captured with structured-light projection while moving the camera and the pattern projector. The overall algorithm is composed of two steps. In the first step, structured-light pattern in the captured images is analysed using deep-learning models, estimating dense projector-camera correspondences. In the second step, neural-represented shape, camera poses, and projector poses are simultaneously optimized using differential volume renderer. The rendering model is similar to NeuS [7]; however, we render structured-light features such as projector patterns and coordinates instead of scene radiance itself as described later.

3.2 | Image capturing setup

We assume a shape capturing system that consists of a camera and a pattern projector as illustrated in Figure 1a. The system is based on a similar approach proposed by Furukawa et al. [3]. The structured light illumination is generated by a diffractive optical element (DOE) incorporated into the pattern projector as shown in Figure 1b.

We use a grid pattern consisting of vertical and horizontal edges with small gaps (Figure 1b). The gaps represent five code symbols to identify camera-to-projector mapping as shown in Figure 1c, where the red dots mean that the vertical and horizontal edges do 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'), and the cyan and magenta dots have gaps between vertical edges similarly. The actual patterns projected onto the object's surface are shown in Figure 1d,e.

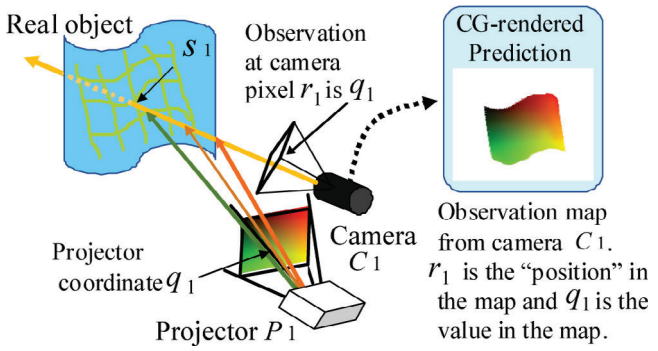

FIGURE 2 Observation model of active stereo systems.

3.3 | Obtaining camera-to-projector correspondences

By using the pattern projector and the camera, we obtain the camera-to-projector correspondences as shown in Figures 2 and 3. In the input images, the grid pattern of structured light can be seen. We call these images as 'pattern images.'

In the process, U-Nets are applied to the pattern images captured by the camera to analyse the grid-like structure of the pattern. The U-Nets outputs two kinds of images; pixel-wise phase information, which represents the grid structure of the pattern, and the code information. These U-Nets can be trained with CG and real images that simulate pattern projection.

Then, the grid structure and code information are converted to a grid graph (Figure 3). This conversion can be done by segmentation similar to [3]. 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, images representing camera-to-projector correspondence mapping are generated. For each pattern image, two images with x and y projector coordinates are obtained, where each pixel in camera-image space contains coordinates in pattern-image space (Figure 3). We call these images as (x or y) projector-coordinate images.

The estimation of the correspondence map from pattern projection in one-shot method often introduce errors. In the method in [3], the codes in the projected pattern are analysed, and global correspondences are estimated for each grid point of the grid pattern. If there is an error in the estimation of the global correspondence, the values of correspondence map in the grid region will be outliers. This causes negative impacts on shape optimization.

To deal with this problem, we implement outlier removal in the correspondence map using segmentation. As the correspondence estimation errors occur independently for each grid, the above outliers are isolated grid regions without continuity with the others. To utilize this, we segment the correspondence map into regions so that the discontinuous points of the x- or y-coordinates in the correspondence map become the boundaries of the region. Each region is either a correctly estimated correspondence of a continuous geometry, or an isolated outlier region. Outlier-isolated regions normally becomes small regions of about one or two grids, thus, regions with the number of pixels less than a threshold are removed and excluded from the correspondence map. Using the method, many of the errors in the correspondence map can be eliminated as shown in Figure 3.

3.4 | Differential volumetric rendering for structured-light projection

We reconstruct a 3D scene from multiple images captured with static structured-light pattern onto the scene. To achieve the goal, neural shape representation as same as NeuS [7]. To utilize information from active pattern projection, we render two kinds of images, which are pattern images and projector-coordinate images. These images can be rendered from the neural scene representation, camera and projector poses, and the pattern image that is projected. The rendering is done with differential volume renderer, with the same way as NeRF or NeuS.

First, we describe the method for rendering projector-coordinate images. A combination of an x-projector-coordinate image and a y-projector-coordinate image represents a 2D-to-2D mapping

$$H : \mathbb{R}^2 \mapsto \mathbb{R}^2; (r_x, r_y) \mapsto (q_x, q_y) \tag{1}$$

from camera pixels (r_x, r_y) to projector pixels (q_x, q_y). A combination of x and y projector-coordinate images (shown in Figure 3) represents a mapping H.

We use a differential volume renderer similar to NeuS to render projector-coordinate images, that is, $H(r_x, r_y)$. In original NeuS, a neural 3D field is used to represent a signed distance field, and also for colour intensities of a light field. However, in this paper, we use a neural 3D field only for a signed distance field, and not for colours. Instead, we use $\mathbf{c}$, which maps a 3D point $\mathbf{p}$ to 2D projector coordinates as shown in Figure 4. The function $\mathbf{c}$ is often used in CG rendering to achieve 'projection mapping.'

$\mathbf{c} : \mathbb{R}^3 \to \mathbb{R}^2$ is defined as

$$\mathbf{c}(\mathbf{p}) = \frac{1}{-z'} \begin{bmatrix} \alpha_x x' \\ \alpha_y x' \end{bmatrix} + \begin{bmatrix} \beta_x \\ \beta_y \end{bmatrix}, \begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \mathbf{R}_{w \to p}\mathbf{p}. \tag{2}$$

We render $\mathbf{c}$ for an SDF-represented surfaces as shown in Figure 4.

The surface $\mathcal{S}$ is represented as a zero level set of SDF in the same way as NeuS.

$$\mathcal{S} = \{\mathbf{x} \in \mathbb{R}^3 | f(\mathbf{x}) = 0\} \tag{3}$$

We render 'projector coordinates' by using the following rendering method. A ray from a camera can be represented by

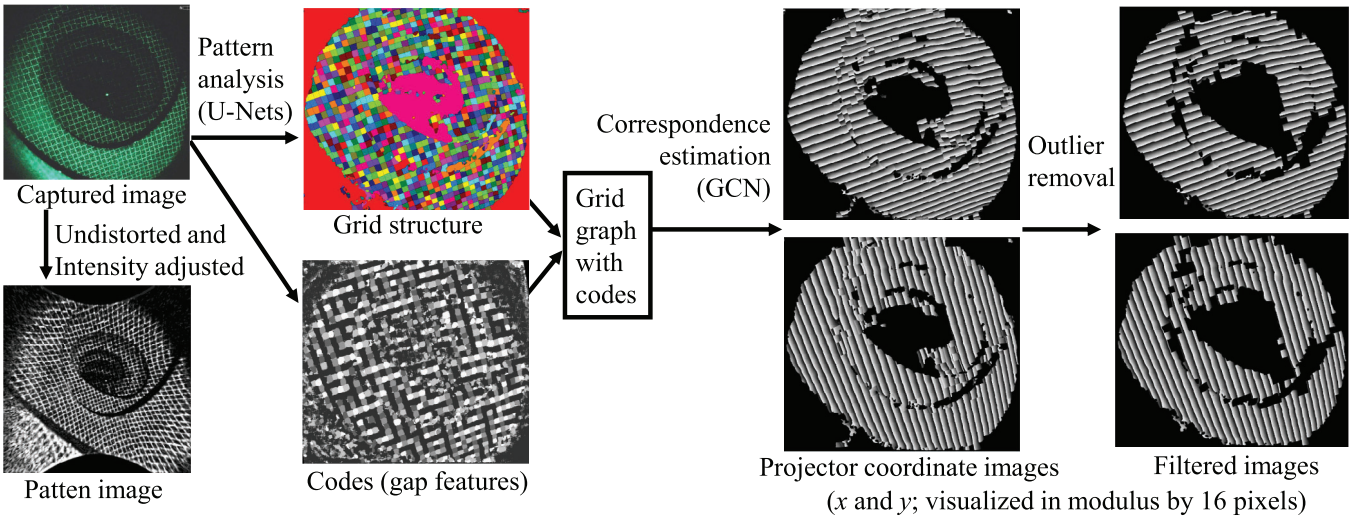

FIGURE 3 Overview of the reconstruction process. Auto-calibration is simultaneously conducted with 3D reconstruction indicated by yellow box.

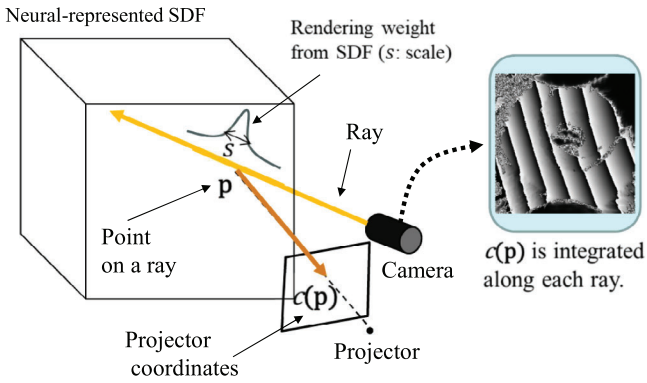

FIGURE 4 Rendering projector-coordinate images.

camera optical center $\mathbf{o}$ and ray direction $\mathbf{v}$, as $\{\mathbf{p}(t) = \mathbf{o} + t\mathbf{v} | t \geq 0\}$. The rendered value is as follows:

$$\mathbf{C}(\mathbf{o}, \mathbf{v}) = \int_0^{+\infty} w(t)\mathbf{c}(\mathbf{p}(t))dt, \tag{4}$$

where $w(t)$ is a weight function that can be calculated from SDF f in the same way as NeuS. We use bold $\mathbf{C}$ because it is a 2D vector of x and y coordinates. Since this equation renders projector coordinates $\mathbf{c}(\mathbf{p}(t))$ instead of radiance-field values, we can render projector coordinates as shown in Figure 4.

By discretizing Equation (4), we use

$$\hat{\mathbf{C}} = \sum_{i=1}^{n} \{\Pi_{j=1}^{i-1}(1 - \alpha_i)\}\alpha_i \mathbf{c}_i, \tag{5}$$

$$\alpha_i = \max\left(\frac{\Phi(s, f(\mathbf{p}(t_i))) - \Phi(s, f(\mathbf{p}(t_{i+1})))}{\Phi(s, f(\mathbf{p}(t_i)))}, 0\right), \tag{6}$$

where f is an SDF value from Equation (3), $\Phi(s, x) = (1 + e^{-sx})^{-1}$ is Sigmoid function with scale parameter s, and $\mathbf{c}_i$ is sampled values of $\mathbf{c}(\mathbf{p}(t))$, s represents 'thickness' of the surface that controls the range of integration around the zero-crossing surfaces. Controlling of s is discussed later.

For rendering 'pattern images', we use $t(\mathbf{c}(\mathbf{p}(t))$ instead of $\mathbf{c}(\mathbf{p}(t))$, where t is a bilinear texture access of the projected pattern. By replacing $\mathbf{c}$ with $\mathbf{d}$, the above rendering pipeline renders 'pattern images.' Thus, pattern images can be rendered by

$$P(\mathbf{o}, \mathbf{v}) = \int_0^{+\infty} w(t)t(\mathbf{c}(\mathbf{p}(t)))dt. \tag{7}$$

We use non-bold t and P because it is a 1D intensity of texture image. The discretized form is

$$\hat{P} = \sum_{i=1}^{n} \left\{\Pi_{j=1}^{i-1}(1 - \alpha_i)\right\}\alpha_i t(\mathbf{c}_i). \tag{8}$$

3.5 | Optimization strategy

We optimize a neural surface representation and camera/projector poses by minimizing the discrepancy between the rendered pattern images and the projector-coordinate images towards respective target images. For loss functions for the optimization, we utilize L1 losses for the projector-coordinate images and cosine losses for the pattern images. This choice stems from the necessity to evaluate direct value differences for projector-coordinate images and to match the overall brightness distributions for pattern images. The preference for using L1 loss over L2 loss is for robustness.

Let the set of camera poses be $L \equiv \{l_1, l_2, ...\}$, the set of projectors be $M \equiv \{m_1, m_2, ...\}$, and the neural SDF be $f(x, y, z)$. The cost function is

$$L(f, L, M) \equiv w_{\text{coord}} E_{\text{coord}}(f, L, M) + w_{\text{pattern}} E_{\text{pattern}}(f, L, M),$$

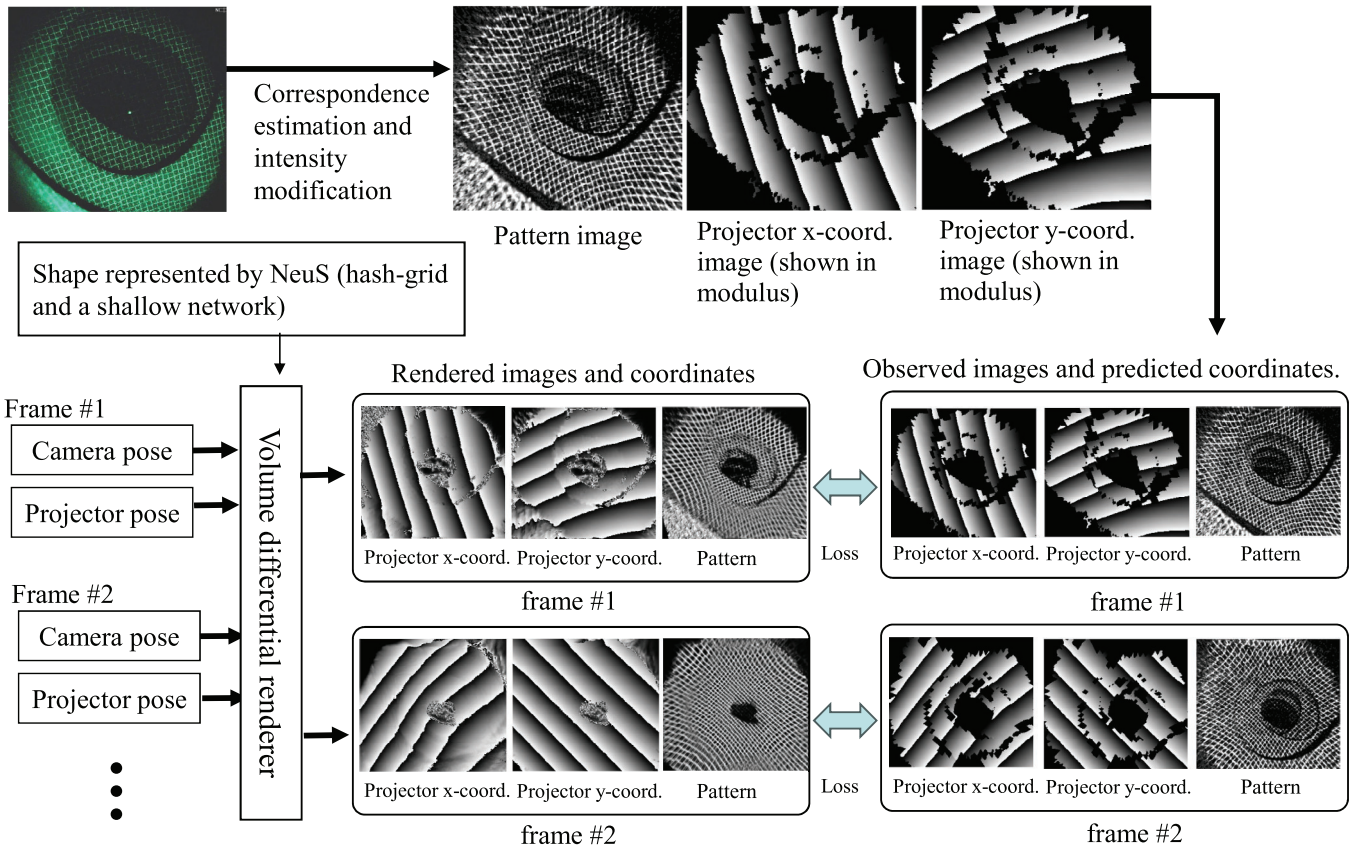

FIGURE 5 Multi-frame optimization using volumetric differential rendering.

$$
\begin{aligned}
E_{\text{coord}}(f, L, M) &\equiv \|\hat{\mathbf{C}} - \tilde{\mathbf{C}}\|_1 \\
E_{\text{pattern}}(f, L, M) &\equiv \left(1 - \frac{\hat{\mathbf{P}} \cdot \tilde{\mathbf{P}}}{\sqrt{\hat{\mathbf{P}} \cdot \hat{\mathbf{P}}}\sqrt{\tilde{\mathbf{P}} \cdot \tilde{\mathbf{P}}}}\right) \\
\hat{\mathbf{C}} &\equiv (\hat{\mathbf{C}}_1, \hat{\mathbf{C}}_2, \hat{\mathbf{C}}_3, \ldots) \\
\tilde{\mathbf{C}} &\equiv (\tilde{\mathbf{C}}_1, \tilde{\mathbf{C}}_2, \tilde{\mathbf{C}}_3, \ldots) \\
\hat{\mathbf{P}} &\equiv (\hat{P}_1, \hat{P}_2, \hat{P}_3, \ldots) \\
\tilde{\mathbf{P}} &\equiv (\tilde{P}_1, \tilde{P}_2, \tilde{P}_3, \ldots)
\end{aligned} \tag{9}
$$

where $\hat{\mathbf{C}}_i$ and $\hat{P}_i$ are rendered results of pattern-coordinate images and pattern images for the i-th sample pixel, $\tilde{\mathbf{C}}_i$ and $\tilde{P}_i$ are target values for the same sample, $\|\cdot\|_1$ is L1 loss, $\hat{\mathbf{C}}$ and $\tilde{\mathbf{C}}$ are stacked vectors of $\hat{\mathbf{C}}_i$ and $\tilde{\mathbf{C}}_i$. Weights w_{coord} and w_{pattern} are manually defined.

The loss function L is minimized by differential rendering the Monte Carlo sampled pixels of projector-coordinate images $\hat{\mathbf{C}}_1, \hat{\mathbf{C}}_2, \hat{\mathbf{C}}_3, \ldots$ and pattern images $\hat{P}_1, \hat{P}_2, \hat{P}_3, \ldots$, calculating L, back propagating L, and updating f, L and M as shown in Figure 5.

When volumetric rendering from the neural-represented signed distance function, the scale parameter s of the sigmoid function determines 'thickness' of rendered surfaces. A large value of s results in integrating values for long segments of 3D points along each ray, leading to a 'blurry' image in the rendering. Therefore, s is desirable to be small in the last stages of the optimization. However, if this value is too small from the initial stages, only a small fraction of volume information for each ray is utilized, leading to unstable optimization. Thus, we start with a large value for s and gradually decrease it during optimization. This process also works as a coarse-to-fine strategy, where first blurry but global shape information is optimized, and shaper but local shape information is optimized later.

The effect of value s is more distinct for pattern images, compared to projector-coordinate images as shown in Figure 6. For rendering projector-coordinate images, projector coordinates are integrated, which locally vary monotonously along each ray in the 3D space. As a result, even if the integration range expands around the surface due to large s, the output projector coordinates do not change much (similar to applying moving average to monotonically increasing/decreasing functions). In the case of the pattern image, the values being integrated are the intensities of the pattern image. In this case, the pattern becomes blurry for large s.

By gradually decreasing s in the optimization process pattern images are first blurry while projector-coordinate images are not much effected, but they are rendered clearly in the later stages where s is small.

Additionally, projector-coordinate images may contain missing regions due to outlier removal described in Section 3.3. For such regions, losses of projector-coordinate images are

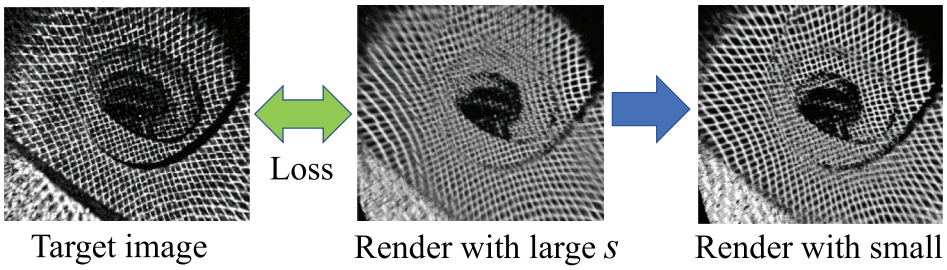

FIGURE 6 Effects of s on pattern image. Leftmost is the target image, and middle and rightmost are rendered images with $s = 1/32$, and $s = 1/200$, respectively. Pattern images are rendered with stronger blur for larger s, which results in that coordinates images are less affected by s.

ineffective. However, even for such regions, rendering and loss function of the pattern image remains effective. Therefore, we can expect the complementary effects by using both projector-coordinate images and pattern images.

4 | INCREMENTAL OPTIMIZATION OF FOR SEQUENTIAL IMAGES

By using the neural shape representation and optimization described so far in this paper, we can reconstruct shapes that are captured with the structured-light projection. In the optimization process, the camera and the projector positions can be corrected. However, in a situation where the camera and the projector positions are unknown, the initial positional parameters should be estimated. If these parameters include considerable errors, the optimization becomes unstable.

This paper proposes incremental optimization of both the shape-field parameters and the positional parameters of the cameras and projectors. The method assumes that the input data is temporarily sequential images which captures a rigid scene, and the camera and projector positions are continuous. The relative positions between the camera and the projector may vary, thus, pre-calibration of the projector-camera system cannot be applied.

The following is the steps of the proposed method.

Step 1. Calculate the projector-xy-coordinate images for all the sequence.
Step 2. Estimate the projector-to-camera pose parameters (i.e. extrinsic parameters) using the first frame porjector-xy-coordiantes from Step 1.
Step 3. Initialize the SDF neural representation and the pose parameters of the first frame.
Step 4. Optimize the SDF neural representation and the camera/projector pose parameters with high learning rates for both the SDF field and the pose parameters.
Step 5. Add a camera pose parameter, a projector pose parameter, and projector-xy-coordinate images of the next frame to the optimization system. The initial parameters can be the same as the cameras and the projectors of the last-optimized frame, since those poses are continuous.
Step 6. Optimize the SDF neural representation and the camera/projector pose parameters with a low learning rate for the SDF field, and high learning rates for the camera and the projector poses.
Step 7. Repeat steps 5 and 6 until all the frames are optimized.

For step 6, we decrease the learning rate for the shape field while keeping the learning rate for the camera/projector pose parameters. It is because we want to 'incrementally' update the shape for the new frames without disturbing the shape regions that are already optimized, and want to optimize the newly-added camera/projector pose positions rapidly.

In step 3, we can use several frames in the first optimization instead of only the first frame. In step 5, several frames can be added as long as they disturb the shape field that is already optimized. For the samples described in Section 5, we first used three frames for the step 3, and three or five frames were added for the step 5.

5 | EXPERIMENTS

5.1 | Validation by simulation data

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. We synthesized 50 frames of data surrounding the rabbit shape. Although we have ground-truth pose parameters for the projector and the camera, we intentionally added Gaussian noises to the parameters to confirm effectiveness of the pose estimation. Figure 7 shows the reconstructed shapes with ground-truth camera/projector poses and disturbed camera/projector poses by random noise. Figure 7a is reconstructed with the ground-truth poses. (b, c) are reconstructed without pose-parameter estimation and randomized initial poses, where (b) is the result after 500 iterations and (c) is after 1500 iterations. (d, e) are reconstructed with pose-parameter estimation and randomized initial poses, where (d) and (e) are after 500 iterations and after 1500 iterations, respectively. The optimization was done in Monte Carlo method, where 2024 rays are sampled for a single iteration. The execution time was about ten iterations per second using Nvidia GeForce RTX 3090 with 24 GB GPU memory. Execution time

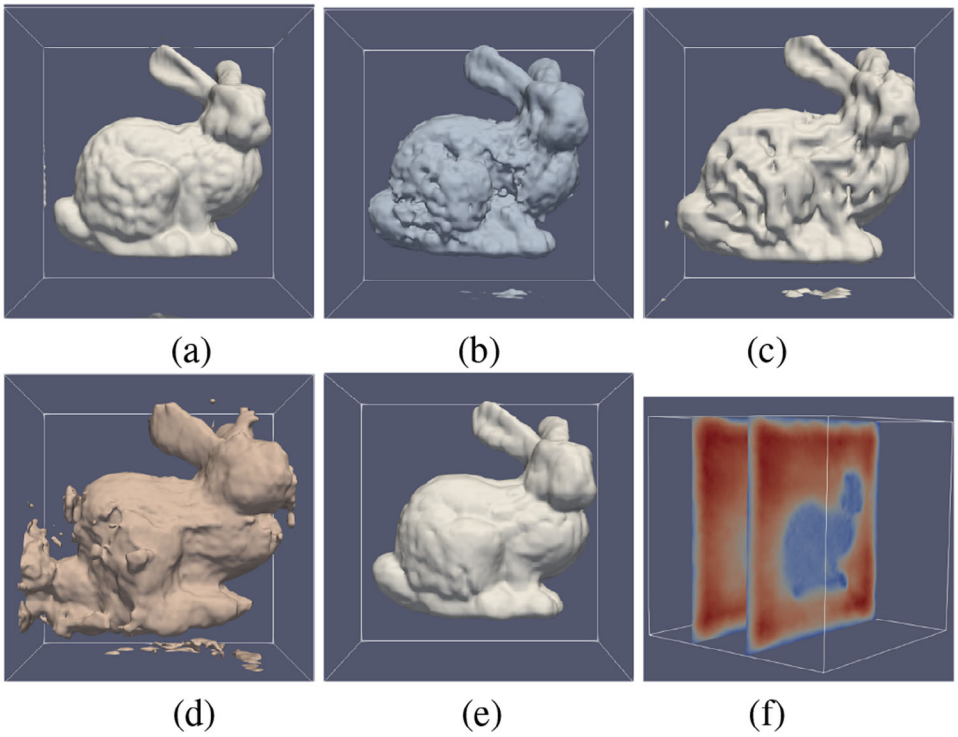

FIGURE 7 Reconstruction results of simulated rabbit shape. (a) Result with ground-truth camera/projector poses. (b,c) Results with disturbed camera/projector poses by random noise without pose optimization. (b) is after 500 iterations and (c) is after 1500 iterations. (d,e) Results with disturbed camera/projector poses by random noise with pose optimization. (d) is after 500 iterations and (e) is after 1500 iterations. (f) Volume values for (a).

was about 110 s for 1000 iterations. The results confirmed that, by applying our camera/projector pose optimization algorithm, pose parameters were refined and the shape was successfully integrated with minimized inter-frame shape inconsistencies.

In the example of Figure 7, the camera positions surround the target object. This situation is not a typical situation in endoscopic diagnosis. To test the proposed method for a more realistic situation related to endoscopy, we made a model of wrinkled cylinder as shown in Figure 8a,b, which is a course model for a colon. We generated a sequence of simulated images where 20 pattern-projected images were captured while the endoscope camera went backward within the model as shown in Figure 8c,d.

The captured images were reconstructed using frame-wise auto-calibration. Figure 8e shows depth estimated from image (c). Then, the 20-frame shapes were optimized using the proposed method. The initial camera positions of all the frames were the same. We use the incremental optimization (Section 4), where initially the first four frames were optimized, with 2000 iterations, and the rest of the frames were added one by one. For each added frame, optimization was performed by 1000 iterations. Figure 8f–h shows the merged shape, GT shape, and the merged and GT shape registered by ICP, respectively. To show effectiveness of the incremental optimization, we also conducted optimization without the incremental frame adding (i.e.global optimization for all the frames from the beginning). The result was a failure for proper optimization as shown in Figure 8i. The ICP residual RMS errors to GT shape were 0.031 for (f) and 0.057 for (i), where the cylinder radius was about 1.0 and max neighbour distance was set to 0.1.

5.2 | Validation by real data (colon phantom model)

We applied our technique to real object using real endoscopic systems. We used a Fujifilm EG-590WR endoscope and a pattern projector with a diffractive optical element (DOE) to generate structured-light illumination as shown in Figure 1b. The test object was a training phantom model for colonoscopy. Figure 9a shows a phantom used for the experiment, Figure 9b shows an example of internal appearance with normal illumination. We captured a sequence of images while the endoscope was moved in the backward direction. Ten frames were captured. For this sample, we first aligned the ten frames with manual ICP. Because of the scale and shape inconsistency, shapes from different frames were not aligned as shown in Figure 9f,h. Using the aligned frames as the initial state, we optimize the multi-frame shape and camera/projector positions. After optimization, the inter-frame inconsistencies were largely reduced as shown in Figure 9g,i. The optimized shape is shown in Figure 9j–l. The ICP residual RMS errors between the ten frames were 0.186 cm before optimization (i.e. Figure 9f and 0.114 cm after optimization (i.e. Figure 9i), where the cylinder

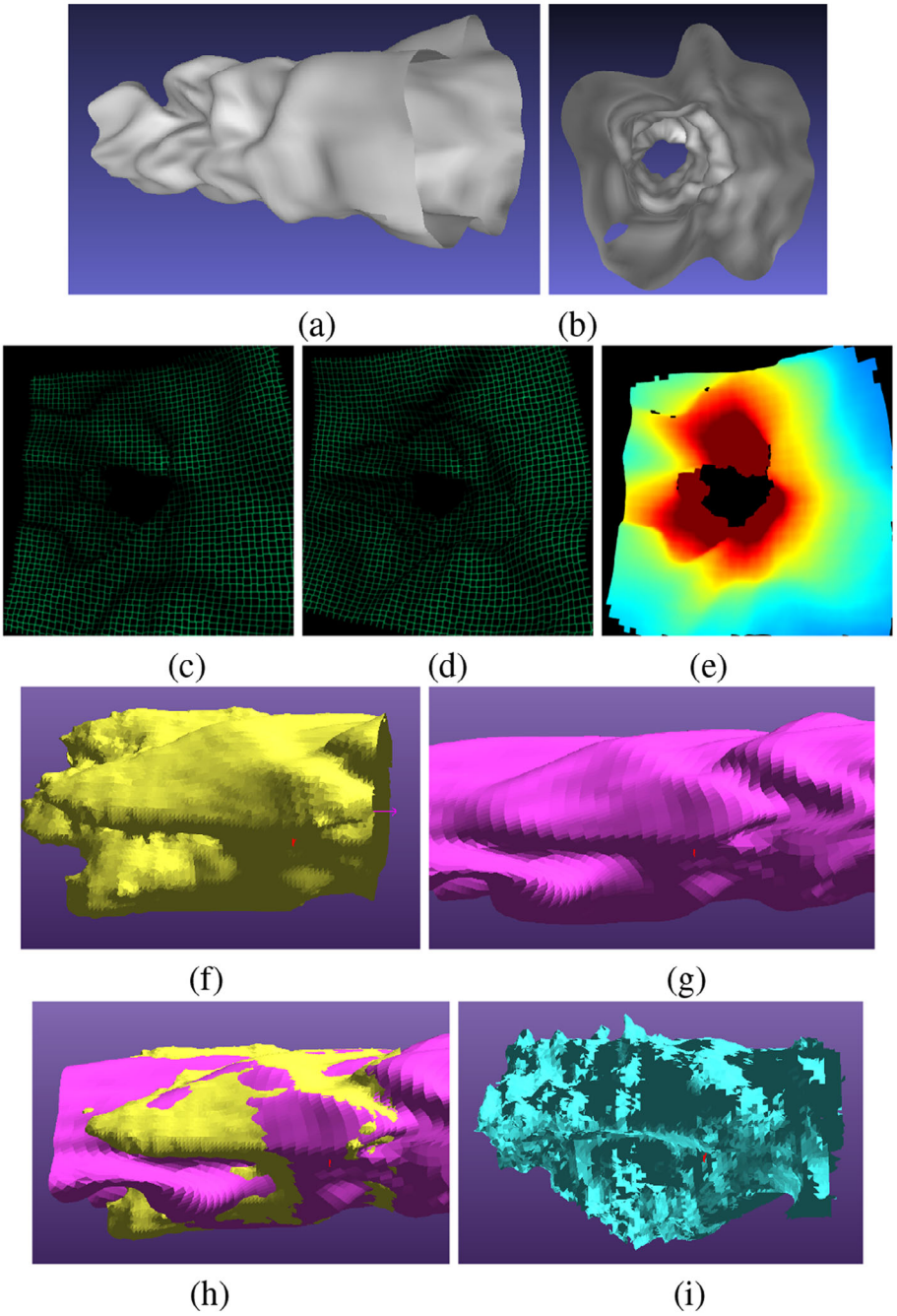

FIGURE 8 Reconstruction results of simulated wrinkled-cylinder shape. (a,b) Appearances of the target object. (c,d) Pattern-projected images (frames 10 and 16). (e) Estimated frame-wise depth for image (c). (f) Result of the proposed method (20 frames were merged). (g) GT shape of (f). (h) GT shape and merged shape (f) registered by ICP. (i) Result of multi-frame optimization without using incremental optimization.

radius was about 2.5 cm, and max neighbour distance was set to 0.1 cm.

Next, we captured a longer image sequence while moving the endoscope within the colon phantom. Then we processed 70 continuous frames with the proposed method. Figure 10a–c are samples of the projector xy-coordinate image estimation ((b) and (c) are projector xy-coordinate images of (a)).

Using the method described in Section 4, we processed the sequential optimization for the sequence. The first three images are initially optimized, and then, subsequent images were added, by five frames-batch. Each time the batches were added, the optimization was processed by 1000 iterations.

After processing all the frames, the neural shape was extracted as zero-level surfaces. Figure 10f–h shows the extracted shape.

Figure 10i–w shows images captured by the endoscope and reconstructed depth images. The left column shows the captured images (fish-eye lens distortions are removed). The middle column shows frame-wise depth calculation results where the

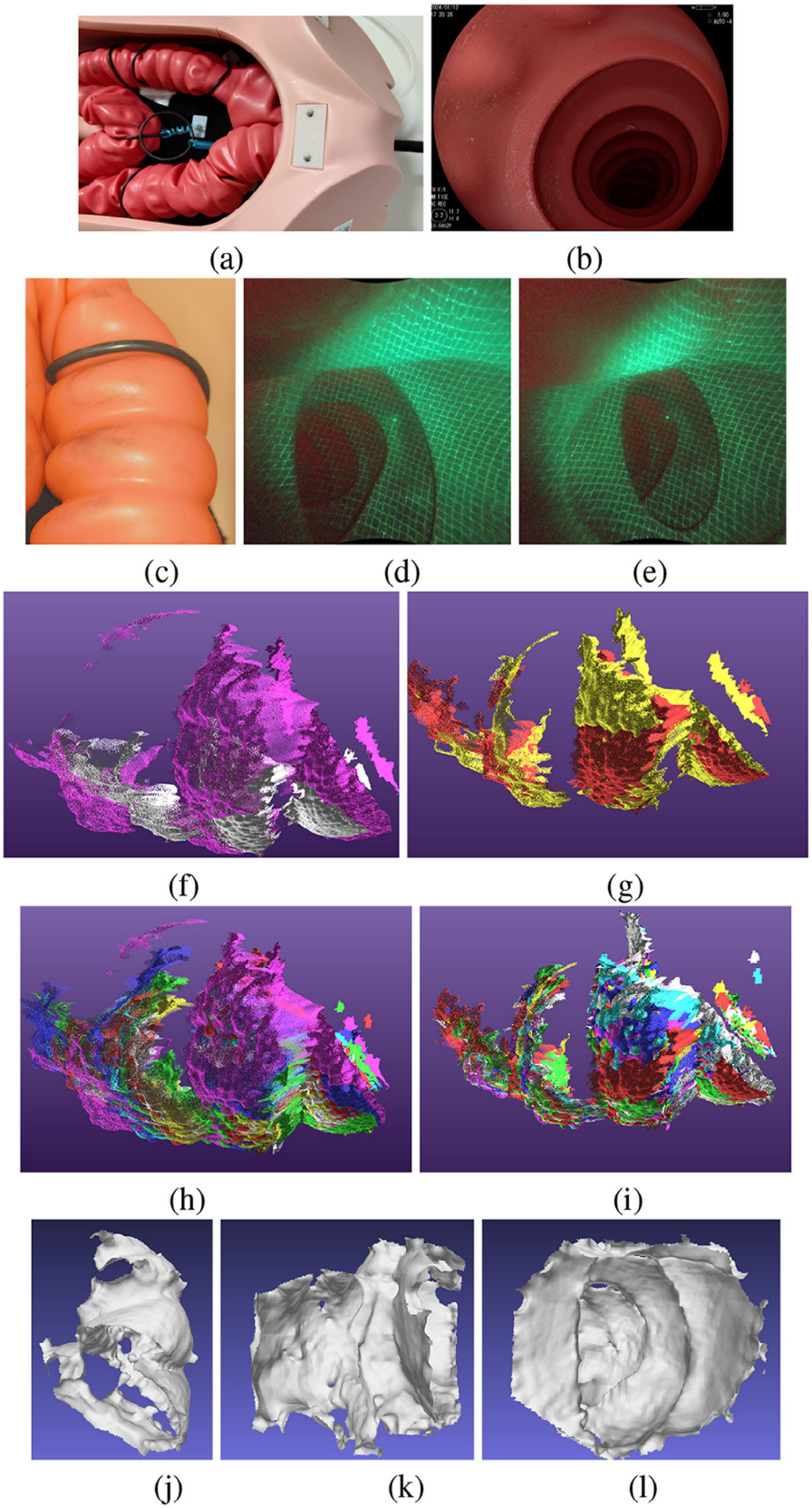

FIGURE 9 (a,b) Appearances of the target object (a wrinkled cylinder). (c) Measured part. (d,e) Pattern-projected images (frames 1 and 6). (f) Frames 1 and 6 before the multi-frame optimization. (g) Frames 1 and 6 after the multi-frame optimization. (h) Ten frames before the multi-frame optimization. (i) Ten frames after the multi-frame optimization. (j) Top view of the optimized shape. (k) Side view. (l) Front view.

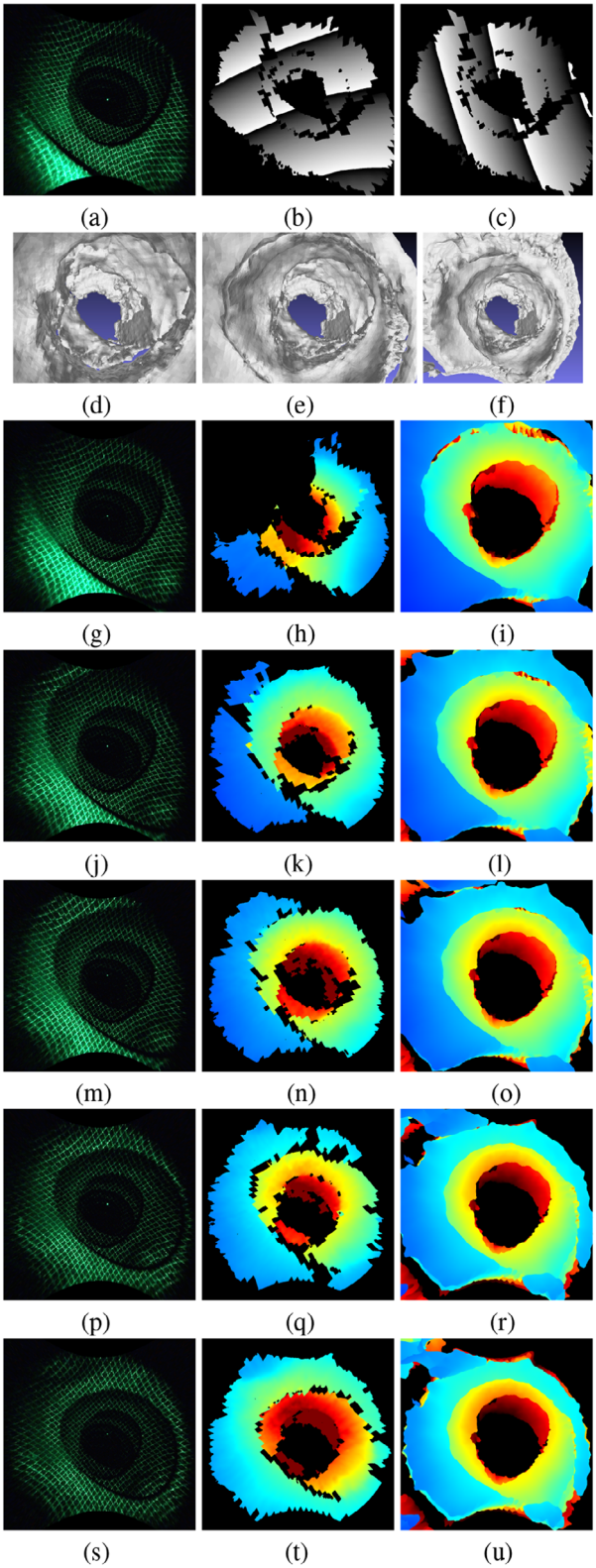

FIGURE 10 Measurement of training-phantom for colonoscopy with free motions of both the projector and the camera. The number of images in the sample sequence is 70. (a–c) Sample of projector *xy*-coordinate image estimation. (b,c) are projector coordinates of (a). (d–f) Surface extracted from optimized neural-represented SDF. (g–u) Captured images (the left column), non-optimized frame-wise depth images (the middle column), depth images rendered from the shape shown in (d–f) (the right column).

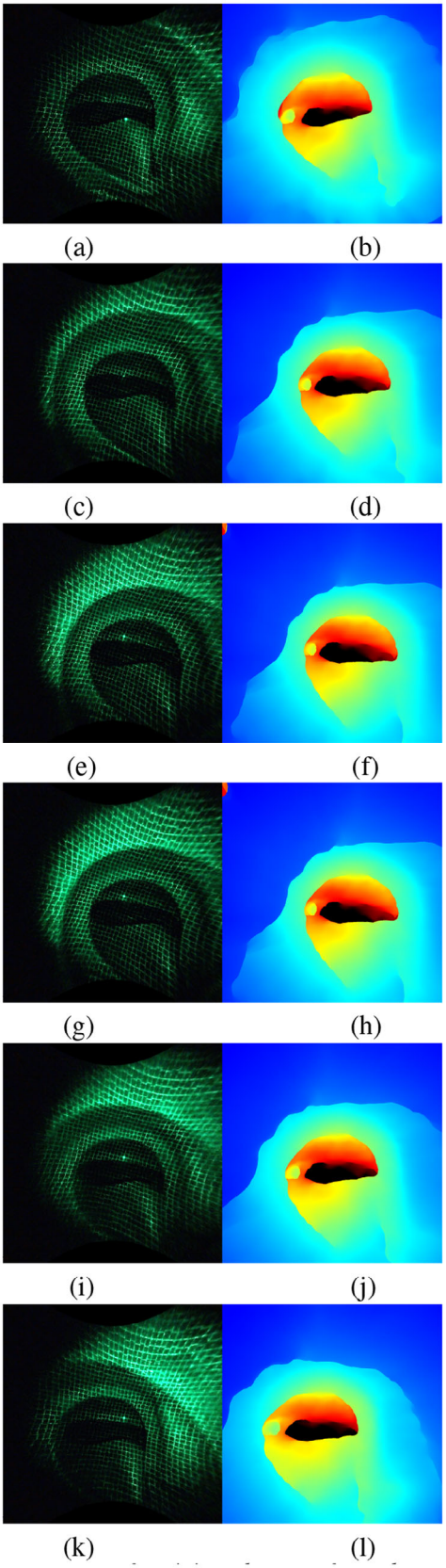

FIGURE 11 Measurement of training-phantom for colonoscopy with free motions of both the projector and the camera. The number of images in the sample sequence is 51. Captured images (the left column), depth images rendered from the globally-optimized reconstructed shape (the right column).

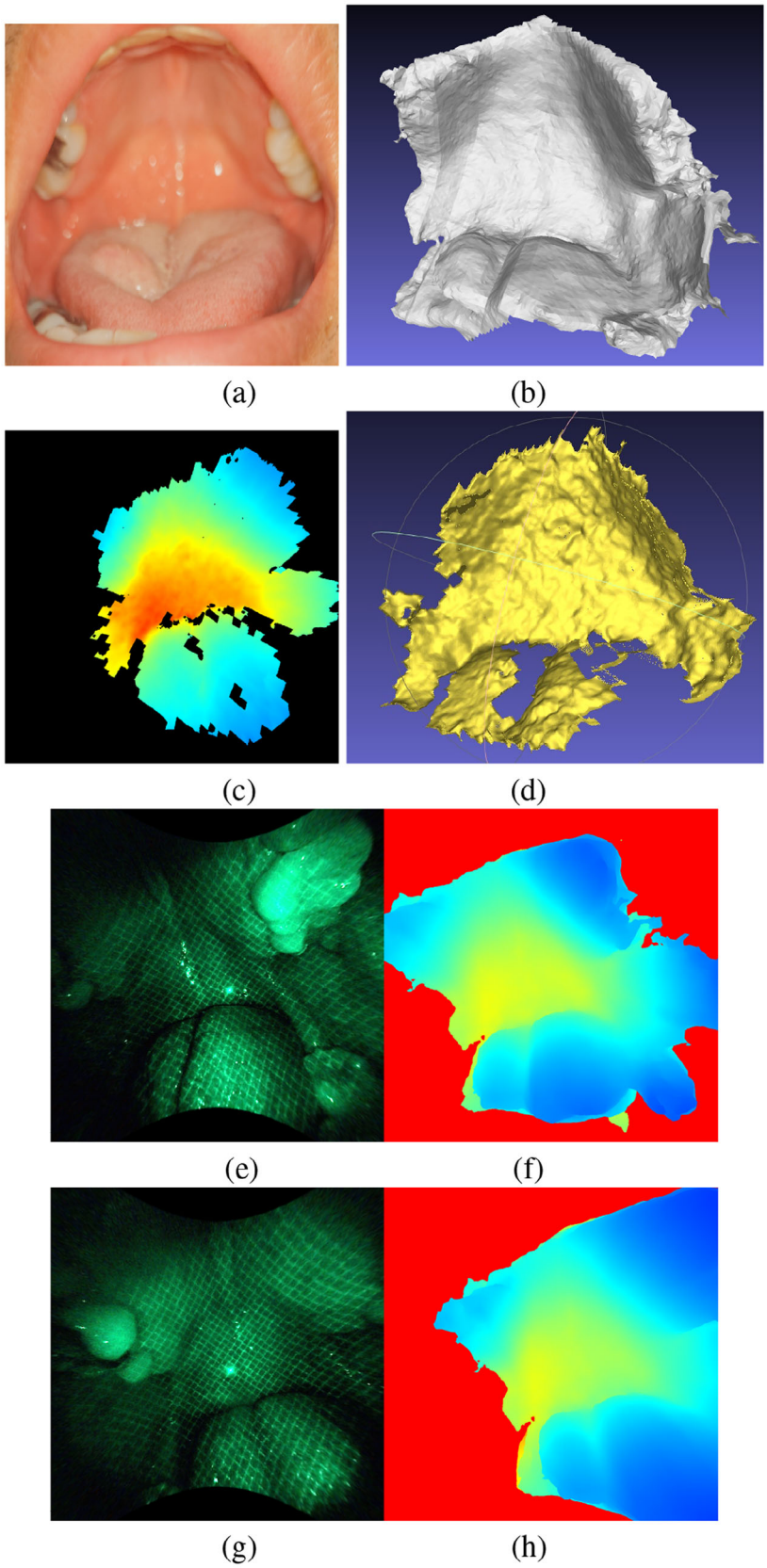

FIGURE 12 Measurement of a shape inside a human's mouth. The number of images in the sample sequence is 39. (a) The appearance. (b) The optimized shape from the neural-represented signed distance field optimized. (c) Depth image of a single frame (no optimization). (d) Shape from depth image (c). (e–h) Captured images (the left column), depth images rendered from the globally-optimized reconstructed shape (the right column).

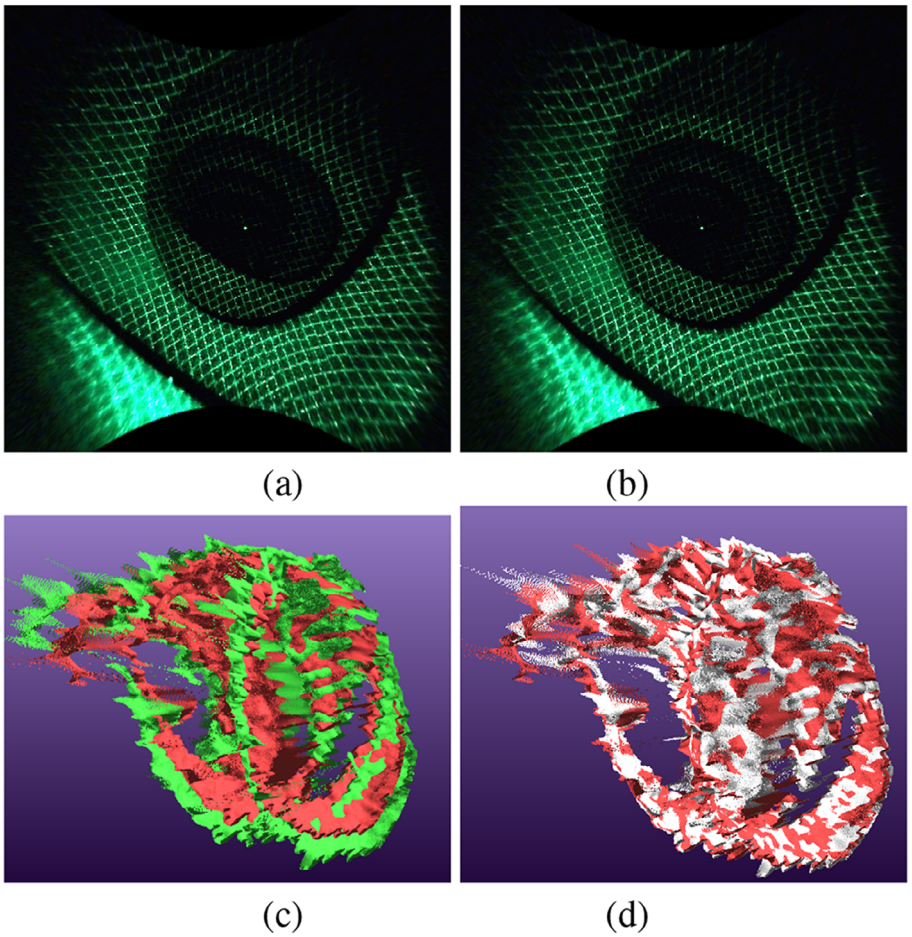

FIGURE 13 Shape consistency evaluation for Figure 10 sample. (a,b) Frames 10 and 11 of Figure 10 sequence. (c) ICP registration result between frames (a) and (b) where each frames reconstructed using pose parameters before global optimization (frame-wise pose calibration). (d) ICP registration result of shapes generated from globally optimized poses. RMSEs of ICP registrations for (a) and (b) were 0.0138 and 0.00314, respectively.

camera and projector poses are not optimized globally. The right column shows optimized the neural shape rendered ad depth images from the viewpoints of optimized camera positions.

By comparing the left and the right column of Figure 10i–w, we can see that the consistent depth images were obtained, while the frame-wise depth images have a lot of holes and frequent motion inconsistency.

Figure 11 shows results of another image sequence of colon training-phantom with 51 frames. Although there is some abrupt motion, the optimized pose parameters tracks the proper viewpoints.

Figure 12 shows results where shape inside human mouth was captured with 39 frames. Compared to (c) and (d) that is frame-wise shape estimation, the optimized shape (b) has improved shape. (e)–(h) shows depth images of the optimized scene and the camera positions. We can see that the consistent shapes and motion could be obtained.

To show effects of simultaneous optimization of projector and camera poses, we reconstructed shapes of two consequent frames (frames 10 and 11 of the image sequence shown Figure 10) using the projector xy-coordinate images of each of the frames. The source frames of them are shown in Figure 13a,b. We used projector-to-camera poses before and after the optimization to reconstruct these frames. Figure 13c shows the shapes generated projector-to-camera pose that is evaluated by a frame-wise calibration. Figure 13d shows the shapes from the optimized pose. Figure 13c clearly shows shape inconsistency between the two frames, whereas frames shown in Figure 13d were consistent. The alignment errors (RMSE) after ICP registration were 0.0138 and 0.00314 for (c) and (d), respectively.

6 | CONCLUSION

In the paper, a novel multi-frame auto-calibration for active-stereo scanning with freely moved projector-camera system is proposed, where an optimization with a differentiable renderer estimates the projector and camera poses as well as consistent shape of multiple scans, even if initial positions are largely apart from the GT. In the method, active-stereo observation process is directly modelled 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 properly with the scanned data using a remote surgery system. In the future, developing

AR/VR system with realtime process for actual remote surgery is planned.

ACKNOWLEDGEMENTS

This work was supported by JST Startup JPMJSF23DR and JSPS/KAKENHI JP20H00611, JP23H03439 and NEDO(JPNP20006) in Japan.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data used in this paper is not publicly available.

ORCID

Ryo Furukawa https://orcid.org/0000-0002-2063-1008

REFERENCES

1. 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: Proceedings of the OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis, pp. 139–150. Springer, Cham (2018)
2. Furukawa, R., Nagamatsu, G., Oka, S., Kotachi, T., Okamoto, Y., Tanaka, S., Kawasaki, H.: Simultaneous shape and camera-projector parameter estimation for 3d endoscopic system using cnn-based grid-oneshot scan. Healthcare Technol. Lett. 6(6): 249–254 (2019)
3. Furukawa, R., Oka, S., Kotachi, T., Okamoto, Y., Tanaka, S., Sagawa, R., Kawasaki, H.: Fully auto-calibrated active-stereo-based 3d endoscopic system using correspondence estimation with graph convolutional network. In: Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 4357–4360. IEEE, Piscataway, NJ (2020)
4. Besl, P.J., McKay, N.D.: Method for registration of 3-d shapes. IEEE Trans. Pattern Anal. Mach. Intell. 14(2), 239–256 (1992)
5. Curless, B., Levoy, M.: A volumetric method for building complex models from range images. In: Proceedings of the SIGGRAPH 96, pp. 303–312. Association for Computing Machinery, New York, NY (1996)
6. Furukawa, R., Sagawa, R., Oka, S., Tanaka, S., Kawasaki, H.: Single and multi-frame auto-calibration for 3d endoscopywith differential rendering. In: Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 1–5. IEEE, Piscataway, NJ (2023)
7. Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W.: Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In: Proceedings of the 35th International Conference on Neural Information Processing Systems, pp. 27171–27183. Curran Associates Inc., Red Hook, NY (2021)
8. Mahmoud, N., Collins, T., Hostettler, A., Soler, L., Doignon, C., Montiel, J.M.M.: Live tracking and dense reconstruction for handheld monocular endoscopy. IEEE Trans. Med. Imaging 38(1), 79–89 (2018)
9. Chen, L., Tang, W., John, N.W., Wan, T.R., Zhang, J.J.: SLAM-based dense surface reconstruction in monocular minimally invasive surgery and its application to augmented reality. Comput. Methods Programs Biomed. 158, 135–146 (2018)
10. Leonard, S., Sinha, A., Reiter, A., Ishii, M., Gallia, G.L., Taylor, R.H., Hager, G.D.: Evaluation and stability analysis of video-based navigation system for functional endoscopic sinus surgery on in vivo clinical data. IEEE Trans. Med. Imaging 37(10), 2185–2195 (2018)
11. 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 Rob. Autom. Lett. 3(4), 4068–4075 (2018)
12. Lamarca, J., Parashar, S., Bartoli, A., Montiel, J.: DefSLAM: Tracking and mapping of deforming scenes from monocular sequences. IEEE Trans. Rob. 37(1), 291–303 (2020)
13. Zhou, H., Jayender, J.: EMDQ-SLAM: Real-time high-resolution reconstruction of soft tissue surface from stereo laparoscopy videos. In: Proceedings of the 24th International Conference MICCAI 2021, pp. 331–340. Springer, Berlin (2021)
14. Liao, J., Cai, L.: A calibration method for uncoupling projector and camera of a structured light system. In: Proceedings of the 2008 IEEE/ASME International Conference on Advanced Intelligent Mechatronics, pp. 770–774. IEEE, Piscataway, NJ (2008)
15. Yamauchi, K., Saito, H., Sato, Y.: Calibration of a structured light system by observing planar object from unknown viewpoints. In: Proceedings of the 19th International Conference on Pattern Recognition, pp. 1–4. IEEE, Piscataway, NJ (2008)
16. Bouaziz, S., Tagliasacchi, A., Pauly, M.: Sparse iterative closest point. Comput. Graphics Forum 32(5), 1–11 (2013)
17. Yang, J., Li, H., Jia, Y.: Go-ICP: Solving 3D Registration Efficiently and Globally Optimally. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1457–1464. IEEE, Piscataway, NJ (2013)
18. Combès, B., Prima, S.: An efficient EM-ICP algorithm for symmetric consistent non-linear registration of point sets. In: Proceedings of the 13th International Conference on Medical Image Computing and Computer-Assisted Intervention: Part II, pp. 594–601. Springer, Berlin (2010)
19. Sinko, M., Kamencay, P., Hudec, R., Benco, M.: 3d registration of the point cloud data using icp algorithm in medical image analysis. In: Proceedings of the 2018 ELEKTRO, pp. 1–6. IEEE, Piscataway, NJ (2018)
20. Henderson, P., Ferrari, V.: Learning to generate and reconstruct 3d meshes with only 2d supervision. arXiv:1807.09259 (2018)
21. Palazzi, A., Bergamini, L., Calderara, S., Cucchiara, R.: End-to-end 6-DOF object pose estimation through differentiable rasterization. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops, pp. 702–715. Springer, Cham (2018)
22. Kato, H., Ushiku, Y., Harada, T.: Neural 3d mesh renderer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3907–3916. IEEE, Piscataway, NJ (2018)
23. Liu, H.-T.D., Tao, M., Jacobson, A.: Paparazzi: surface editing by way of multi-view image processing. ACM Trans. Graph. 37(6), 221–1 (2018)
24. Gadelha, M., Maji, S., Wang, R.: 3d shape induction from 2d views of multiple objects. In: Proceedings of the 2017 International Conference on 3D Vision (3DV), pp. 402–411. IEEE, Piscataway, NJ (2017)
25. Gwak, J., Choy, C.B., Chandraker, M., Garg, A., Savarese, S.: Weakly supervised 3d reconstruction with adversarial constraint. In: Proceedings of the 2017 International Conference on 3D Vision (3DV), pp. 263–272. IEEE, Piscataway, NJ (2017)
26. Henzler, P., Mitra, N.J., Ritschel, T.: Escaping plato's cave: 3D shape from adversarial rendering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9984–9993. IEEE, Piscataway, NJ (2019)
27. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NERF: Representing scenes as neural radiance fields for view synthesis. Commun. ACM 65(1), 99–106 (2021)
28. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Trans. Graphics (ToG) 41(4), 1–15 (2022)

How to cite this article: Furukawa, R., Kawasaki, H., Sagawa, R.: Incremental shape integration with inter-frame shape consistency using neural SDF for a 3D endoscopic system. Healthc. Technol. Lett. 12, e70001 (2025). https://doi.org/10.1049/htl2.70001