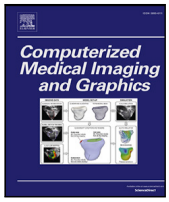
Computerized Medical Imaging and Graphics 116 (2024) 102418

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

journal homepage: www.elsevier.com/locate/compmedimag



2D/3D deformable registration for endoscopic camera images using self-supervised offline learning of intraoperative pneumothorax deformation

Tomoki Oya [a], Yuka Kadomatsu [b], Toyofumi Fengshi Chen-Yoshikawa [b], Megumi Nakao [c,*]

[a] Graduate School of Informatics, Kyoto University, Yoshida-Honmachi, Sakyo, Kyoto, 606-8501, Japan
[b] Nagoya University Hospital, 65 Tsurumai-cho, Showa-ku, Nagoya, 466-8550, Japan
[c] Graduate School of Medicine, Kyoto University, 53 Shogoin Kawahara-cho, Sakyo, Kyoto, 606-8507, Japan

ARTICLE INFO

Keywords:
2D/3D deformable registration
Self-supervised learning
Pneumothorax deformation
Endoscopic image
Thoracoscopic surgery

ABSTRACT

Shape registration of patient-specific organ shapes to endoscopic camera images is expected to be a key to realizing image-guided surgery, and a variety of applications of machine learning methods have been considered. Because the number of training data available from clinical cases is limited, the use of synthetic images generated from a statistical deformation model has been attempted; however, the influence on estimation caused by the difference between synthetic images and real scenes is a problem. In this study, we propose a self-supervised offline learning framework for model-based registration using image features commonly obtained from synthetic images and real camera images. Because of the limited number of endoscopic images available for training, we use a synthetic image generated from the nonlinear deformation model that represents possible intraoperative pneumothorax deformations. In order to solve the difficulty in estimating deformed shapes and viewpoints from the common image features obtained from synthetic and real images, we attempted to improve the registration error by adding the shading and distance information that can be obtained as prior knowledge in the synthetic image. Shape registration with real camera images is performed by learning the task of predicting the differential model parameters between two synthetic images. The developed framework achieved registration accuracy with a mean absolute error of less than 10 mm and a mean distance of less than 5 mm in a thoracoscopic pulmonary cancer resection, confirming improved prediction accuracy compared with conventional methods.

1. Introduction

Three-dimensional (3D) imaging is widely used in clinical medicine to obtain the patient-specific morphological structure of organs. The anatomy of organs, including vascular structures and tumors, is reconstructed as a 3D shape from computed tomography (CT) and magnetic resonance (MR) images before treatment or surgery. Alternatively, in endoscopic surgery, only two-dimensional (2D) images of target organs taken by an endoscopic camera are generally used to understand the intraoperative state of the organs. However, it is difficult to grasp the 3D anatomical structure of organs because of the high uncertainty caused by organ deformation and the limited 2D field of view during surgery.

To address this issue, image-guided surgery (Tokuno et al., 2020; Han et al., 2022; Buchs et al., 2013) using organ shapes derived from preoperative 3D-CT/MRI images has been attempted clinically. Intraoperative use of the organ shape has the potential to improve the accuracy of the surgery by visualizing internal vascular structures and the location of tumors. However, for soft organs that deform during treatment, such as the lungs, the 2D appearance, shape, and posture in the endoscopic images do not match the preoperative organ shapes. Clinical efforts have been made to manually align the preoperative organ shape to the appearance of the target organ in the camera image by partially adjusting the shape, position, and posture (Tokuno et al., 2020). Thus, automatic registration of the 3D shape to the occluded 2D camera image obtained during treatment is expected. Some studies tried to restore the partial shapes of organs by estimating the 3D information using a stereo endoscope or several surgical scenes from multiple angles (Geng and Xie, 2014; Lin et al., 2016). However, measurement-based approaches are limited to reconstructing visible regions and partial surfaces (Mountney et al., 2006; Penne et al., 2009; Kowalczuk et al., 2012; Hong et al., 2014; Zhao et al., 2016). Additionally, approaches that require additional 3D measurement devices are limited in their use to specialized surgical procedures or environments.

* Corresponding author.
E-mail address: nakao.megumi.6x@kyoto-u.ac.jp (M. Nakao).

https://doi.org/10.1016/j.compmedimag.2024.102418
Received 25 February 2024; Received in revised form 10 July 2024; Accepted 15 July 2024
Available online 19 July 2024

For a broader range of image-guided therapy, 2D/3D registration, i.e., deformable model registration with a single projection image, is expected to be feasible. 2D/3D registration, which predicts higher dimensional points from a 2D projection image, is an ill-posed problem without prior knowledge because of its high uncertainty. To address this issue, a data-driven organ deformation model representing morphological characteristics of organs with low-dimensional parameters has been reported (Wu et al., 2019; Wang et al., 2020; Nakao et al., 2022). Combined with deep learning, real time and accurate 2D/3D registration can be expected. Other methods have been used to improve prediction accuracy using bi-planar X-ray images (Ying et al., 2019; Kasten et al., 2020; Dong et al., 2023). However, most of these methods developed for image-guided radiotherapy assume a situation in which the entire projected image of the target organ can be obtained, and the registration performance for highly occluded camera images has not been examined. In addition, a practical solution is not given for the case where there is insufficient training data for learning the 2D/3D registration of highly occluded images.

Recent studies on registration to highly occluded camera images have used model-based registration methods (Koo et al., 2017; Adagolodjo et al., 2017; Özgür et al., 2018) that aim to align an organ shape generated from patient-specific 3D-CT images. Multiple viewpoint endoscopic images have also been used for model-based registration of the liver (Espinel et al., 2021). However, the calculation of organ deformations through parameter optimization in these studies poses a challenge because of the high computational cost and stability in solving optimization problems, leading to unsolved issues in real-time registration. Alternatively, for the issue of insufficient training data, data-driven learning has been explored using synthetic data generated from an organ deformation model (Brunet et al., 2019; Mendizabal et al., 2019; Pellicer-Valero et al., 2020; Pfeiffer et al., 2020). Although some studies have achieved accurate and real-time registration through model-based learning, most still rely on 3D measurement devices, and 2D/3D registration for a single viewpoint image has not been a focus.

This study proposes a deep learning framework for model-based 2D/3D registration for highly occluded camera images such as endoscopic images. The framework learns the registration of a patient-specific organ shape for a single-viewpoint endoscopic camera image using the correlation between the camera position/postures and surgical scenes. Because of the limited number of endoscopic images available for training, we use a synthetic image generated from the nonlinear deformation model (Maekawa et al., 2020; Nakao et al., 2021) that represents possible intraoperative pneumothorax deformations. In this paper, we call this model-based self-supervised training concept offline learning. To address the issue of prediction errors because of differences in input features between synthetic and real endoscopic images, image labels commonly obtained from both images are incorporated as input into the model training process, thereby reducing prediction errors.

With the exception of our study, few studies have applied offline learning to the 2D/3D registration problem of endoscopic images. Although the liver or brain with relatively small deformations has been the target, our target is intraoperative lung deformation with more than 50% volume changes (Nakao et al., 2021). The framework proposed in this paper can generate images reflecting prior knowledge from a patient's organ shape by changing view points, thus generating a wide variety of surgical scenes in which only a part of the shape is visible and learning 2D/3D registration tasks. In the experiments, we apply the proposed framework to thoracoscopic camera images for pulmonary resection and evaluated its registration performance while comparing with existing methods.

To the best of our knowledge, this is the first study to attempt deep learning-based 2D/3D deformable registration of an intraoperative pneumothorax state of the lung as a surgical guide for thoracoscopic pulmonary resection. The contributions of this paper are summarized in the following four points:

- 2D/3D model-based deformable registration framework for highly occluded endoscopic camera images
- a concept of self-supervised offline learning using a data-driven pneumothorax deformation model that provides a practical solution for the limited number of training datasets
- use of common image features obtained from synthetic and real images to reduce registration errors when transferring the learned model to real scenes
- performance analysis of the trained 2D/3D registration framework and its application to image-guided pulmonary resection in thoracoscopic surgery

2. Related work

2.1. 2D/3D model-based registration

2D/3D model-based registration has been widely studied as a parameter optimization problem (Markelj et al., 2012; Oliveira and Tavares, 2014; Wang et al., 2017; Liao et al., 2019; Ketcha et al., 2017). Miao et al. (2016) reported convolutional neural network (CNN)-based rigid-body registration, solving the problems of high computational cost and learning stability in parameter optimization. However, training CNNs requires a large volume of data, and supervised learning is challenging to implement for organs, particularly those undergoing deformation during surgery, for which collecting training data is difficult.

To address this issue, Wu et al. (2019) used a data-driven organ deformation model to generate a large volume of 2D/3D registered data, realizing the prediction of an intraoperative deflated lung shape with a 2D projection image. However, the shape prediction is represented in the form of point clouds, and it is difficult to establish correspondence between the vertex positions in the inflated and deflated states. Wang et al. (2020) reported a CNN-based framework for estimating the deformed shape of the lung; however, the results were validated using only synthetic 3D shapes, and the effectiveness in estimating respiratory deformations in real patients was not confirmed. The recently proposed graph convolutional network model has demonstrated the ability to perform 2D/3D deformable registration for digitally reconstructed radiographs of abdominal organs (Nakao et al., 2022). The network is capable of simultaneously learning the prediction of displacement maps and 3D deformation for abdominal organ meshes. Although the model can achieve 2D/3D registration with a clinically acceptable error, it requires the entire projection image, including the target organ, for registration. The accuracy of registration for occluded projection images remains unclear.

2.2. Registration for endoscopic images

For the past two decades, clinical efforts have been made to manually align organ shapes to highly occluded laparoscopic images for surgical guidance. However, some studies reported that rigid-body registration leads to inaccurate registration results (Konishi et al., 2005; Nicolau et al., 2011). To overcome this problem, researchers have tried to implement deformable registration using 3D measurements or 2D contours of organs as visual cues in laparoscopic images (Haouchine et al., 2015; Koo et al., 2017; Özgür et al., 2018; Modrzejewski et al., 2018; Koo et al., 2022). Koo et al. (2017) proposed simultaneous optimization of organ deformation and camera parameters. Özgür et al. (2018) employed a two-step registration approach to estimate the deformation of the organ shape and predict the camera parameters from the estimated shape. However, these methods require manual extraction of organ contours for registration, and fully automatic 2D/3D registration has not been achieved. A recent study proposed 2D/3D deformable registration for the liver with automated contour extraction (Koo et al., 2022); however, it takes considerable computation time for registration. Although accurate registration was achieved in

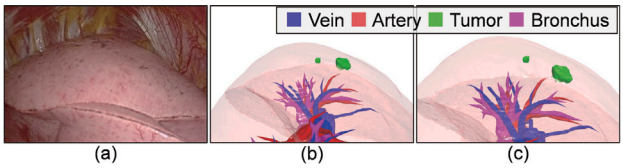


Fig. 1. Examples of endoscopic images and the deformed mesh of a lung: (a) deflated shape of a lung in an endoscopic view, (b) lung mesh with internal structures deformed by a pneumothorax deformation model, and (c) the mesh manually registered to the endoscopic image.

clinical cases, a significant error in some cases was still observed. Thus, optimization-based methods tend to have difficulties in achieving real-time performance and stability because of their high computational cost.

Recently, some studies have demonstrated the effectiveness of data-driven learning in deformable registration using synthetic data generated from an organ deformation model (Pfeiffer et al., 2019; Brunet et al., 2019; Pfeiffer et al., 2020). Pfeiffer et al. (2020) used an organ deformation model to simulate point clouds of the intraoperative surface of the liver obtained from 3D measurements during surgery and achieved registration of the organ shape with the point clouds. However, the method requires 3D measurements and is difficult to use for 2D/3D registration with synthetic images. Haouchine et al. (2022) proposed an augmented reality neurosurgical system to superimpose preoperative 3D meshes onto an intraoperative view of the brain surface by combining pose estimation and non-rigid refinement of the aligned vessel model. In our previous study, we explored data-driven learning using synthetic data generated from a kernel-based deformation model of intraoperative pneumothorax lungs (Oya et al., 2023). We used the 2D contour of the lung silhouette commonly obtained from synthetic and real thoracoscopic images as the input for the CNN model. Although real-time registration was achieved, the framework still faces the challenge of incorrect settings when predicting 3D shapes from contours, making it difficult to obtain learning gradients. The proposed framework in this paper presents a new design of offline learning using prior knowledge and aims to achieve better registration performance.

3. Methods

3.1. Problem definition

In endoscopic surgery, several conditions are assumed regarding the endoscopic image and intraoperative condition of the target organ. In this study, we assume the following conditions that are common in thoracoscopic surgeries.

- The initial shape of the organ is obtained from the patient-specific 3D-CT images measured before surgery
- The surgeon can generate endoscopic views of the organ mesh using 3D visualization software during preoperative planning (Tokuno et al., 2020)
- The organ deforms from the preoperative state; however, the changes are limited to a certain range and are statistically modeled (Maekawa et al., 2020)
- The insertion point of the endoscopic camera is also determined in advance; however, the camera position is not obtained during surgery
- The rotation around the thoracoscopic camera axis is given because it is clinically fixed by the surgeon's camera control.
- In the early stage of surgical procedures, a part of the organ's surface is observed in the endoscopic image, with more than 50% of the entire shape occluded

Using these surgical situations, we aim to achieve registration of the organ mesh by predicting the differential of the camera position and posture (i.e. camera parameters) and the shape between previously assumed scenes and the endoscopic image obtained during surgery.

Fig. 1 shows an example of endoscopic images and deformable organ mesh, including lung surfaces and internal vascular structures obtained from the patient's 3D-CT images. Fig. 1(a) shows a scene when observing the thoracic cavity with a thoracoscopic camera at the early stage of surgery, where a part of the upper lobe and inner chest wall is observed. Fig. 1(b) shows an example of reproducing the deflated shape of the lung from its preoperative inflated state by changing the deformation rate of the pneumothorax deformation model. The intraoperative appearance or thoracoscopic view can be simulated by adjusting camera parameters. Fig. 1(c) shows the result of manually adjusting the camera parameters to the corresponding surgical scene under surgeon supervision. After this mesh registration process, the lung surface is rendered transparent, and the tumor position, bronchus, and vascular structure are available as a surgical guide. Thus, automated 2D/3D registration allows the surgeon to identify the region of interests under the lung surface during surgery.

3.2. Proposed framework

Fig. 2 illustrates the offline learning framework for model-based registration for endoscopic camera images. As a pre-processing step, the framework generates a deformed shape M_D based on a pneumothorax deformation model from the initial shape M_I obtained from patient-specific 3D-CT images. This simulation can be performed in the preoperative planning, and expected surgical scenes are rendered manually by changing parameters for endoscopic views and the deformation model. We abbreviate the set of parameters as model parameters θ. The intraoperative appearance of the target lung is obtained from the endoscopic image I to be registered, and therefore, the purpose of this framework is to predict the differential of model parameters $\hat{\theta}$ between that of preoperatively assumed scenes and the given real scenes. In order to train CNNs, we use a set of rendered images (called the source) that can be obtained by rendering the deformed shape M_D and the image label I_L (called target) obtained from the endoscopic image I.

Although training CNNs requires a large volume of data, it is not easy to construct a large database of registration for a surgical scene. Thus, supervised learning using a few endoscopic images I directly as the training data is not expected to achieve sufficient prediction performance. To overcome this problem, we use the synthetic images simulating surgical scenes that can be rendered by changing the model parameters θ. Specifically, various deformed meshes M_D can be generated by randomly changing the deformation rate w. By rendering the deformed shape with randomly changing camera parameters, we obtain several pairs of projection images observed from different viewpoints. In offline learning, the rendered images generated using θ_s are used as the source, the image labels obtained using θ_t are used as the target, and CNN is trained to estimate the differential of the model parameters between the two states. For the source, we use synthetic images with prior knowledge about the positional relationship between the position and focus of the camera and lung mesh. Additionally, it is possible to generate various surgical scenes in which only a part of the shape is visible to train 2D/3D registration tasks. The details of the pneumothorax deformation model are described in Section 3.3.

The endoscopic images I subject to registration and synthetic virtual images generated from the deformed mesh used for training appear different. Therefore, in order for the model trained by offline learning to work correctly, the framework is designed to achieve both self-supervised learning and prediction using image features commonly obtained from the real and virtual images. Thus, in this study, the endoscopic image I is not directly given to the CNN; the label image I_L of the lung silhouette is instead fed as the input of CNN for the registration target. The label image I_L can be commonly generated

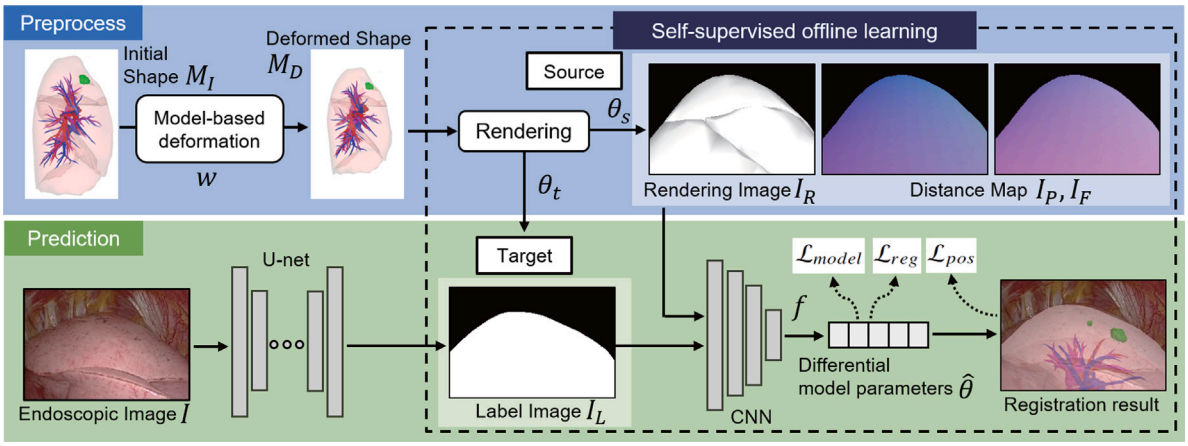


Fig. 2. Proposed offline learning framework. Offline learning was performed using synthetic images from a pneumothorax deformation model to learn the differential model parameters $\hat{\theta}$ between the source (I_R, I_P, I_F) and target (I_L).

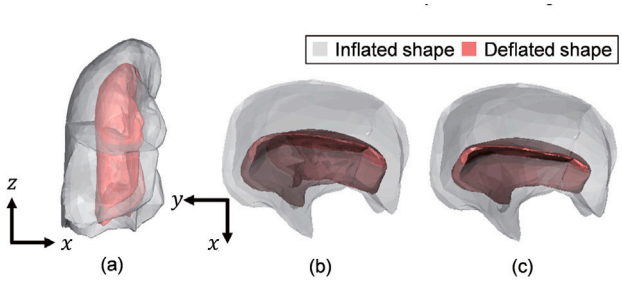


Fig. 3. Registered meshes of inflated/deflated lungs in 3D-CT coordinates. (a) Inflated (translucent) and deflated shape (opaque) in the case of $w = 1.0$, (b) axial views, and (c) the case of $w = 1.2$ used for deformation.

by rendering the deformable mesh in offline learning and by semantic segmentation in prediction.

Estimating the differential of model parameters between two scenes using only the silhouette of the target shape as a visual cue is an ill-posed problem. However, the camera parameters assumed in endoscopic surgery are expected to be distributed within a certain range. In addition, the lung mesh and camera parameters determined in advance by surgeons are available as prior distribution for prediction. Using this prior knowledge, we apply a task to CNN that relaxed the settings. Specifically, we train the CNN using the set of synthetic images, rendering image I_R that reflects the shading and distance maps I_P and I_F, which reflects the distance information obtained from the positional relationship between the position and focus of the camera and lung mesh, respectively. The details of the source and target images used in offline learning are described in Section 3.4.

3.3. Pneumothorax deformation model

The pneumothorax deformation model used in this study was constructed using a data-driven modeling approach and represents the mean deformation in the intraoperative pneumothorax state optimized for a given initial lung shape. The process of obtaining the deformed shape is based on a previous study (Maekawa et al., 2020; Nakao et al., 2021) as follows.

STEP1 Obtain a 3D mesh of the whole lung from the 3D-CT image taken before surgery, and a pair of meshes with partial shapes of both inflated and deflated states of the lung using the intraoperative cone-beam CT taken from the same patient.

STEP2 Register the two partial lung meshes to the whole lung mesh in 3D-CT coordinates using deformable mesh registration (Nakao et al., 2019). The displacement vectors are determined by the inflated/deflated state of the meshes with point-to-point correspondence.

STEP3 Perform STEP1 and STEP2 for all patient data, and train a kernel-based deformation model. The learned kernel model can reconstruct the displacement vectors from the inflated to deflated state and works as the pneumothorax deformation model.

Fig. 3 shows the inflated and deflated lung shapes for the same patient. Applying the model-based deformation to the inflated shape obtained from 3D-CT images, the mean lung deformation estimated for the given mesh is represented as shown in Fig. 3(a) and (b). Significant deformation with volume changes is confirmed in both the coronal and axial views. By changing the deformation rate w, the amount of deformation can be controlled as shown in Fig. 3(c) based on Eq. (1).

$$v_i' = v_i + w \cdot u_i \tag{1}$$

where v_i' and v_i are the vertex position of the deformed and initial mesh, respectively. u_i is the displacement vector calculated for the vertex of the initial shape. Deformation of the entire mesh is calculated by transforming all vertices $v_i \in \mathcal{V}$ $(i = 1, 2, \ldots, n$, where n is the number of vertices) of the composing M_I. A set of displacement vectors u_i $(i = 1, 2, \ldots, n)$ forms a spatially non-uniform displacement field, which represents a nonlinear pneumothorax deformation based on the patient-specific inflated lung shape.

w is the scalar deformation rate that controls the amount of nonlinear shape change in the pneumothorax deformation model. $w = 1.0$ represents the mean deformation of the population, and $w = 0.0$ represents the given initial mesh in the inflated state. The surgeon can control the volume by changing the air pressure in the lungs at the beginning of the surgical procedure. We confirmed that the fluctuation of the volume change in the deflated state was approximately $\pm 10\%$ in past thoracoscopic surgery (Nakao et al., 2021). The vertex displacements in this fluctuation were no greater than 5 mm, indicating a relatively small deformation. Therefore, w was introduced as a weight parameter for linear interpolation to sufficiently represent this fluctuation between patients. Specifically, with u_i representing the spatially nonlinear deformation, w is fixed as 1.0 for the source and changed between 0.9 and 1.1 for the target, to simulate the expected volumetric variability of the deflated lung.

3.4. Offline learning using synthetic images

In this section, we describe the details of offline learning, which learns a model that estimates the differential model parameters θ_s, θ_t for

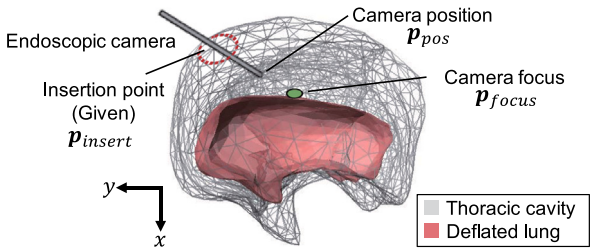


Fig. 4. Positional relationship between the endoscopic camera and organ shape during surgery. Our framework assumes a rigid endoscope, and the camera focus is determined by the camera tip position and axis.

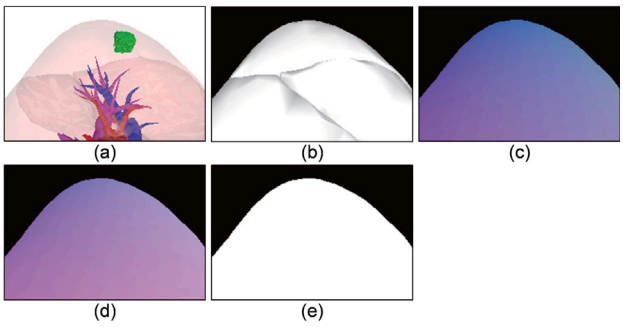


Fig. 5. Synthetic images generated from a pneumothorax deformation model: (a) lung surfaces and internal vascular structures, (b) rendered image with shading (I_R), (c) distance map defined between the camera position and each vertex of the mesh (I_P), (d) distance map defined between the camera focus and each vertex of the mesh (I_F), and (e) binary label for the target (I_L).

the source (I_R, I_P, I_F) and target I_L. We adopt the label image I_L of the lung silhouette as the target, which can be commonly generated from the endoscopic image and deformed mesh M_D to reduce registration errors when transferring the learned model to prediction in offline learning.

Fig. 4 shows the positional relationship between the endoscopic camera and organ mesh during surgery observed from the axial view. Because a rigid scope is used in endoscopic surgery, the focus is uniquely determined from the camera tip position and camera axis. The rendered image or the visual appearance of the mesh, including the translation, scaling, and rotation, is uniquely determined by the camera position, focus, and up vector. The camera model was customized to the thoracoscopic environment to achieve stable training for the ill-posed 2D/3D registration problem. As we assumed that the rotation around the thoracoscopic camera axis (i.e., the camera up vector for rendering), the state of the camera is defined by the following equations:

$$p_{pos} = p_{insert} + d \cdot e_{axis} \tag{2}$$

$$p_{focus} = p_{pos} + c_0 \cdot e_{axis} \tag{3}$$

where p_{pos} is the camera tip position, p_{focus} is the camera focus, and p_{insert} is the insertion point determined in the preoperative planning for thoracoscopic surgery. d is the insertion depth and e_{axis} is a 3D unit vector that determines the thoracoscopic camera orientation or camera axis. c_0 is a constant coefficient obtained from the hardware specification of thoracoscopy. In this camera model, the set of four dimensional variables $\theta = (p_{pos}, w)$ is the target model parameter for registration, meaning that the insertion depth d and camera pose e_{axis} are predicted while restricting rotation around the camera axis. p_{focus} is then determined uniquely using the estimated p_{pos} along with the camera axis e_{axis}. Based on this scheme, when the camera tip position and camera focus are updated, the visual appearance (i.e., the translation, scaling, and rotation) of the object change. Because the estimated viewpoint is represented by the 4×4 modelview matrix M that defines the 3D transformations including translation, scaling, or rotation, we can evaluate the registration error between the ground truth and aligned mesh. A perspective projection with an angle of view of 55.79° is used to generate the rendered image based on the thoracoscopic hardware specification.

Fig. 5 shows a set of synthetic images that can be generated from a pneumothorax deformation model. Unlike the rendered image shown in Fig. 5(a), the label image in Fig. 5(e), which is used as the target in this study, is commonly obtained from the deformed mesh and endoscopic images. The target label images can be generated from rendering of the mesh in offline training and from endoscopic images using semantic segmentation in the prediction. Although similar label images could be used for the source, the task of predicting the differential model parameters from only two binary label images (source and target) is incorrectly posed, and the estimation error is expected to be large. Therefore, we aim to improve the error by adding prior knowledge of the shading (Fig. 5(b)) and distance information (Fig. 5(c), (d)) to the source.

First, the rendering image I_R shown in Fig. 5(b) is obtained by simulating a point light source at the camera positions and rendering the deformed mesh with shading. This image visualizes the orientation of the surface relative to the camera axis and the boundary of lung lobes because of shading. Therefore, it is possible to learn prior knowledge about the direction of the camera axis and the organ mesh.

Next, the distance maps I_P, I_F shown in Fig. 5(c) and (d) are obtained as color maps for the label image by calculating the 3D vector between the camera position and vertex, and between the camera focus and vertex of the organ mesh, and then converting the 3D vector to RGB values. Here, the 2D map I_P or I_F introduced in this study is not a conventional depth map representing the 1D scalar distance between the camera position and the surface point in the depth direction. I_P is the 2D distribution of 3D vectors from the camera position to the visible vertex of the target mesh. Thus, the RGB color directly represents the direction (or 3D flow) from surface points to the camera.

The distance between the organ mesh and camera is reflected in the rendering image through perspective projection; however, the appearance change is often minute, and it is difficult to grasp the distance and direction from the label image shown in Fig. 5(e). In spite of the fact, the partial appearance is greatly affected by the initial camera position and posture because of the small distance to the surface of the target lobes. In particular, we focus on the fact that the same changes in the registration parameters have different effects on 2D appearance, depending on the initial camera focus. To feed initial camera conditions to the CNN, we incorporated the two distance maps into our offline learning. We note that the distance map and shading are generated for the source but not generated for the target intraoperative camera image; the only information needed for inference during surgery is the binary label of the lung region.

Based on the above design, we define the model f to be learned in offline learning as in Eq. (4)

$$\hat{\theta} = f_\phi(I_R, I_P, I_F, I_L) \tag{4}$$

where ϕ are the weights of the network and $\hat{\theta}$ are the differential model parameters, the output. We used a typical CNN with eight convolutional layers (conv2d with 3×3 kernel size and ReLU for activation function), max pooling, and fully connected layers for output of the parameters. Considering the usage of GPU memory in CNN training, the thoracoscopic image and synthetic image were both sized at 240×160 pixel. The output dimension is four (i.e., the camera position and scalar deformation weight).

3.5. Loss function

This section describes three loss functions that achieve learning of the model f, which estimates the differential of the model parameters for the input source and target images.

First, we introduce the model parameter loss $\mathcal{L}_{model}$, which calculates the error between the true value of the model parameter differential θ and the predicted value $\hat{\theta}$, and is defined as follows.

$$\mathcal{L}_{model} = \frac{1}{m}\sum_{j=1}^{m} \|\theta_j - \hat{\theta}_j\|_2^2 \tag{5}$$

where m is the dimension of the model parameters. This loss works to maintain the differential of the predicted model parameters as their true values.

Next, in order to achieve accurate registration of the organ mesh, the error between the true value and predicted shape's position and posture should also be small. The proposed method introduces $\mathcal{L}_{pos}$ defined in Eq. (6) as the loss of positional error between the transformed vertices.

$$\mathcal{L}_{pos} = \frac{1}{n}\sum_{i=1}^{n} \|M v_i - \hat{M}\hat{v}_i\|_1 \tag{6}$$

where n is the number of vertices of the deformed mesh and $v_i, \hat{v}_i (i = 1, 2, \ldots, n)$ are the target and estimated position of each vertex in the deformed mesh. M and $\hat{M}$ are, respectively, the target and estimated modelview matrix. This loss works to keep the vertex positions of the estimated shape in the camera space closer to the vertex positions of the target shape.

Finally, to avoid overfitting, we introduce $\mathcal{L}_{reg}$ defined in Eq. (7) as a L2 regularization term defined by

$$\mathcal{L}_{reg} = \sum_{j=1}^{m} \|\hat{\theta}_j\|_2^2 \tag{7}$$

The total loss function $\mathcal{L}$ is defined as a weighted linear sum of three loss functions as in Eq. (8).

$$\mathcal{L} = \mathcal{L}_{model} + \alpha\mathcal{L}_{pos} + \beta\mathcal{L}_{reg} \tag{8}$$

4. Experiments

To evaluate the performance of the proposed method for 2D/3D registration on a single thoracoscopic image, we conducted two experiments: (1) analysis of the proposed offline learning and (2) evaluation of the registration performance on a thoracoscopic image with the existing methods. We used a PC (CPU: Intel Core i7-9700K 3.60 GHz, Memory: 32 GB) equipped with a NVIDIA GeForce RTX 2070 for training. The overall framework was implemented with Python 3.9 and TensorFlow GPU 2.6. We trained the CNN with a batch size of 32 and 300 epochs, using the Adam optimizer with a learning rate of 1.0×10^{-3}.

4.1. Dataset and preprocessing

In this study, we used preoperative 3D-CT images and surgical videos of the 16 thoracoscopic lung cancer surgery cases provided by the Department of Thoracic Surgery, Kyoto University Hospital and Nagoya University Hospital. We confirmed scenes in which both the surface of the lung and the thoracic cavity were visible at the beginning of surgery after insertion of the thoracoscope, and obtained 160 images from approximately 10 frames in each case that had different appearances.

After extracting the lung regions from the preoperative 3D-CT of the inflated state for each case, the upper, middle, and lower lobes and, in some cases, the resection area that was distinguished from the lung lobes, were extracted. A series of extraction processes were performed automatically using Synapse VINCENT from Fujifilm Corporation. After obtaining a triangle mesh with 1000 faces from each of the three lung lobe regions or the four regions that were distinguished from the lung lobes and resection region, we obtained the deformed mesh using a pneumothorax deformation model constructed in a previous study (Maekawa et al., 2020; Nakao et al., 2021) and by setting $w = 1.0$ in Eq. (1).

4.2. Analysis of offline learning

First, we conducted quantitative experiments to verify the estimation performance and effectiveness of the additional maps in the offline learning. We used four error metrics for registration accuracy: dice similarity coefficient (DSC), mean distance (MD), Hausdorff distance (HD), and mean absolute error (MAE). DSC is an evaluation of the similarity between two projected images of estimated and target shapes projected in 2D. MD and HD are the mean and maximum bidirectional distance defined by the nearest vertex of the estimated and target shapes, respectively. These metrics quantify the 3D difference between the shapes. MAE is the average distance between each vertex in the estimated and target shapes, which is also known as the average distance (ADD) metric (Hinterstoisser et al., 2012), calculated in Eq. (9).

$$MAE = \frac{1}{n}\sum_{i=1}^{n} \|M v_i - \hat{M}\hat{v}_i\|_1 \tag{9}$$

where n is the number of vertices of the deformed mesh and $v_i, \hat{v}_i (i = 1, 2, \ldots n)$ are the target and estimated position of each vertex in the mesh. M and $\hat{M}$ are, respectively, the target and estimated modelview matrix.

A large number of training data sets can be generated in offline learning, and registration performance for synthetic images is expected to improve with the training data. However, in the segmentation field, the segmentation accuracy for real data decreases with the increase in the data using a large number of synthetic images in training (Tang et al., 2019). Therefore, it is necessary to find the volume of training data that is expected to achieve sufficient accuracy for real data as well, among the multiple variations of training data.

In this experiment, we conducted training and evaluation using 3200 synthetic images as training data, 640 images as validation data, and 640 images as test data. Synthetic images were randomly generated with a range of ±10 mm and ±20 mm from the center of each parameter for all 3D camera parameters to be estimated. One case with one range generated 100 data sets for training and 20 data sets for validation/test data. Before performance analysis, we determined the hyperparameters $\alpha = 1.0\times10^{-3}$ and $\beta = 1.0\times10^{-4}$ based on a grid search by changing the order of each parameter from 1.0×10^{-4} to 1.0×10^{-1} using the validation dataset. We then confirmed the registration performance of the proposed framework.

4.2.1. Performance analysis

First, we confirmed the registration performance of the proposed framework in 2D and 3D. As shown in the experimental results, the proposed method was able to achieve 7.67 mm registration on average with respect to synthetic images. The MD was 3.73 mm on average, which is close to the acceptable error of 2 - 3 mm for MD according to the American Association of Physicists in Medicine guideline for image registration and fusion (Brock et al., 2017), although this is only for synthetic images. Fig. 6(a) shows the registration results for synthetic images. Case 12 shows the average registration error for synthetic images, and Case 11 shows the maximum registration error. The initial (Fig. 6(a) left) compares the results of the 2D projections of the source (red) and target (blue) shapes, where the source is the average shape generated with $w = 1.0$. The prediction (Fig. 6(a) right) compares the results of 2D projections of the estimated (red) and target shapes. Gray areas are where the source or estimated shape overlaps with the target shape, and the values of MAE and DSC are described below each image. The proposed method increased the DSC. In Case 12, both MAE and DSC were significantly improved from the initial error, and the camera

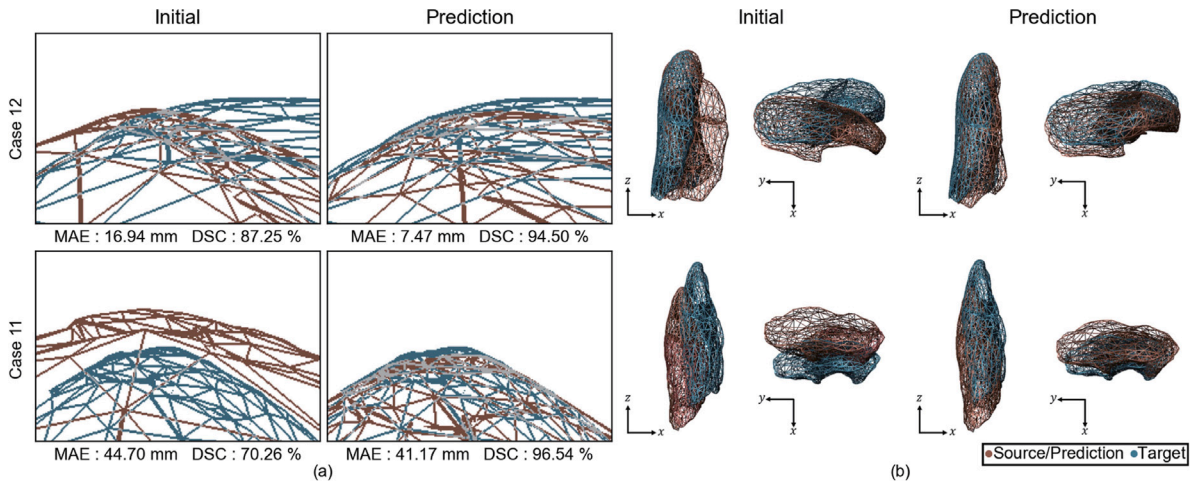


Fig. 6. Examples of registration for synthetic images in offline learning. (a) Appearance in the endoscopic view. Case 12: average registration error, Case 11: maximum registration error. (b) 3D visualization of the source/estimated (red) and target (blue) shapes in each case.

Table 1
Effectiveness of each component for model parameter estimation in offline learning of 2D/3D registration.

	MAE [mm]	DSC [%]	MD [mm]	HD [mm]
Initial	14.15 ± 8.39	93.97 ± 5.13	6.82 ± 3.78	18.79 ± 10.20
w/o shading	7.85 ± 6.27	97.79 ± 2.44	3.85 ± 2.81	10.58 ± 7.60
w/o position distance map	7.83 ± 6.21	**97.87 ± 1.97**	3.82 ± 2.77	10.52 ± 7.55
w/o focus distance map	7.99 ± 6.34	97.87 ± 2.25	3.90 ± 2.82	10.72 ± 7.75
w/o positional loss ($\mathcal{L}_{pos}$)	7.74 ± 6.29	97.87 ± 2.16	3.80 ± 2.81	10.31 ± 7.58
w/o regularization loss ($\mathcal{L}_{reg}$)	7.84 ± 6.35	97.75 ± 2.21	3.84 ± 2.83	10.52 ± 7.82
w/o deformation rate	7.84 ± 6.40	97.86 ± 2.21	3.82 ± 2.85	10.42 ± 7.72
Proposed	**7.67 ± 6.37**	97.86 ± 2.08	**3.73 ± 2.82**	**10.28 ± 7.66**

posture and lung shape were modified to be more correct. Alternatively, in Case 11, the DSC value was improved; however, MAE was improved only slightly, resulting in a distance of more than 40 mm from the target shape. Fig. 6(b) shows the results of 3D visualization of the source, estimated and target shape from the coronal and axial views. In the estimation results of Case 11, the axial deviation is resolved from the initial state; however, the coronal deviation is still large. This shows that the result with a relatively good DSC and a large MAE value indicates the presence of a registration error in the depth direction in the endoscopic views.

Table 2
Registration accuracy with respect to number of training data sets used in offline learning.

	MAE [mm]	DSC [%]	MD [mm]	HD [mm]
Initial	16.31 ± 7.30	92.28 ± 4.86	7.75 ± 3.19	21.65 ± 8.97
160	14.21 ± 6.93	92.81 ± 5.10	6.97 ± 3.06	20.12 ± 9.01
800	11.57 ± 6.74	95.72 ± 3.10	5.56 ± 2.91	15.54 ± 8.24
1600	11.24 ± 6.74	95.89 ± 3.04	5.44 ± 2.95	14.96 ± 8.29
3200	9.97 ± 5.96	96.01 ± 2.91	4.84 ± 2.62	13.65 ± 7.45
4800	10.02 ± 5.96	**96.35 ± 2.46**	4.90 ± 2.63	13.80 ± 7.56
6400	**9.74 ± 6.02**	96.35 ± 2.69	**4.75 ± 2.65**	**13.45 ± 7.56**

4.2.2. Ablation study

We then conducted an ablation study to verify the effectiveness of each component introduced in the proposed model. The framework comprises rendered images with shading, two distance map inputs as prior knowledge, an inter-vertex error $\mathcal{L}_{pos}$, and a regularization term $\mathcal{L}_{reg}$. Additionally, we change the deformation rate w of a pneumothorax deformation model for the target during training so w can be considered one of the components of the framework. Because it is difficult to validate the effectiveness of all of these elements in all combinations, we trained a model without each element and compared the registration accuracy with that of the proposed model.

Table 1 summarizes the mean ± standard deviation of each error metric in the compared models, indicating the effectiveness of each component of the proposed framework. "Initial" refers to a non-registered state for the synthetic source/target images in offline learning. "w/o shading" uses label images instead of rendering images with shading, and uses label images and distance maps as the source and label images as the target. "w/o position distance map" and "w/o focus distance map" are models that use a rendered image and one distance map as the source and the label image as target for the input, without input of a distance map between the camera position or focus and organ mesh as prior knowledge. "w/o positional loss" and "w/o regularization loss" are models in which the positional loss $\mathcal{L}_{pos}$ and regularization term $\mathcal{L}_{reg}$ are excluded from the overall loss function. "w/o deformation rate" is a model in which the camera parameters are estimated in 3D, the deformation rate w is not estimated, and the mean deformation $w = 1.0$ is used as the deformed mesh.

These results confirm the effectiveness of each component of the proposed framework. In particular, the effects of the distance map of the camera focus and the regularization term $\mathcal{L}_{reg}$ are notable. The error was larger when the distance map of the camera focus was ablated than the distance map of camera position. It is possible that the camera position is set at a distance from the organ mesh and the focus is set at a position on the surface or inside the organ mesh. Therefore, it is possible that the distance map, which visualizes the distance of each point from a point in the organ mesh, is more useful information for estimation. Additionally, because $\mathcal{L}_{pos}$ calculates the error per vertex, $\mathcal{L}_{reg}$ is necessary for stable training of the network. In addition, without estimating the weights w of a pneumothorax deformation model, the variance of the error becomes large. Therefore, stable estimation requires the estimation of the deformation rate as well.

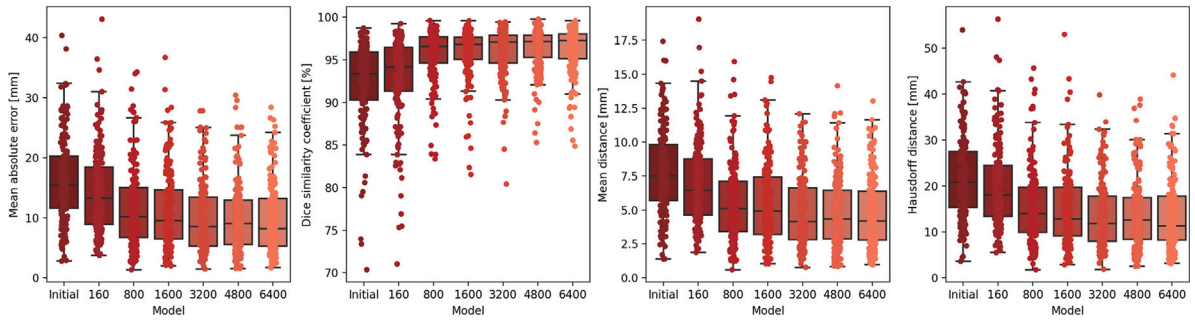


Fig. 7. Box plots of MAE, DSC, MD, and HD to evaluate the registration accuracy with respect to number of training data sets. "Initial" indicates the registration error using the deflated mesh with mean deformation ($w = 1.0$) and a viewpoint set by the surgeon's manual operations.

4.3. Evaluation

Next, we verified the 2D/3D deformable registration performance to real thoracoscopic images for the model trained on synthetic images that can be generated from a pneumothorax deformation model. A standard U-Net (Ronneberger et al., 2015) was used for the semantic segmentation of the lung region in a total 160 thoracoscopic images. The U-Net was trained in our past study using the other training dataset, which was different from the 160 images prepared in this experiment. We extracted 30 to 50 frames from the surgical movies of 25 patients who underwent thoracoscopic pulmonary resections in the Department of Thoracic Surgery, Kyoto University Hospital from 2018 to 2019. We manually generated 1792 binary labels of the lung regions and trained the U-Net model. Leave-one-out cross-validation using 24 cases for each training showed that the mean Dice coefficient was 0.956, the minimum was 0.768, and the maximum was 0.992.

The estimation error in tumor position is important for clinical use. However, as it is difficult to obtain its real-world 3D position during surgery for this experiment, we evaluated the registration error using lung meshes. The patient-specific deformable lung mesh is clinically available in preoperative planning (Fig. 1), and surgeons can interactively change the viewpoint using a thoracoscopic rendering mode in the visualization software (Tokuno et al., 2020). The ground truth was obtained by the surgeon's manual operations of the deformable lung meshes while watching a given thoracoscopic scene, retrospectively. Specifically, the camera tip position p_{pos}, camera focus p_{focus}, and deformation rate w were manually adjusted to make the viewpoint as close as possible to the given specific time frame of the surgical video. As a summary, we define the initial state and ground truth as follows.

Initial: the mean deformation state by setting $w = 1.0$ and the viewpoint estimated by a surgeon with manual operations based on his/her surgical knowledge and experience. The camera position p_{pos} was set without referring to surgical videos or thoracoscopic images to mimic preoperative planning. If the surgeon provided multiple visual appearances to be assumed, we used the average values of the parameters.

Ground truth: the modified deformation and viewpoint manually adjusted by the surgeon while viewing a given thoracoscopic video frame (i.e., surgical scene). All of the parameters (i.e., the camera position, focus, and deformation rate w excluding rotation around the camera axis) were updated to obtain the ground truth.

4.3.1. Registration performance

In Section 4.2, we set the number of training data sets to that expected to achieve sufficient accuracy for real data, among multiple variations of training data. In this experiment, we verified whether this number of training data sets in offline learning was appropriate or not by confirming the registration accuracy for thoracoscopic images using models with multiple variations of the number of training data sets.

As in Section 4.2, we used four error metrics: MAE, DSC, MD, and HD. In this experiment, the available number of thoracoscopic image data sets is 160. Therefore, we conducted training and evaluation using 160, 800, 1600, 3200, 4800, and 6400 synthetic images generated at $1\times, 5\times, 10\times, 20\times, 30\times$, and $40\times$ that number for the training data (20% for validation) and 160 synthetic images as the source and 160 thoracoscopic images as the target for the evaluation data. A synthetic image was randomly generated with a range of ± 10 mm and ± 20 mm for the training and ± 20 mm for the evaluation, respectively, from the center of each parameter in all 3D camera parameters to be estimated. We used the same parameter sets described in Section 4.2.

Fig. 7 shows the registration accuracy of the model trained for each data set in offline learning of 160 thoracoscopic images. Although the accuracy of DSC improves as the number of data sets increases, the accuracy of the other error metrics increases only slightly from approximately 3200 data sets and is within a certain range. Table 2 summarizes the mean ± standard deviation of error metrics in the compared models. As shown in the accuracy improvement from 1600 to 3200 data sets and from 3200 to 4800 or 4800 to 6400 data sets, the accuracy improvement was small, even with the same number of data sets and the same $2\times$ increase. Thus, the error is within a certain range for an increase in the number of data sets using offline learning.

Fig. 8 shows examples of registration errors for a model with 3200 data sets (the proposed model). Case 15 shows a registration error of approximately the 25% percentile, Case 14 shows a registration error of approximately the median, Case 11 shows a registration error of approximately the 75% percentile, and Case 8 shows a registration error at the exact tumor position. Fig. 8(a) shows the result of the augmented endoscopic image and visualization of the estimated (red) and target (blue) shapes in the thoracoscopic view. In the registration results, the tumor position in green, artery in red, vein in blue, and bronchus in purple are indicated. In the visualization of the estimated and target shapes in the endoscopic field of view, the overlap of the two shapes is shown in gray. The values of the four error metrics are noted below the image of each case. Fig. 8(b) shows the 3D visualization of the estimated and target shapes from the coronal and axial views. Fig. 8(a), (b) shows that Case 15 achieved highly accurate registration in both 2D and 3D. Although the lobes are far apart and there are still some errors, all error metrics are highly accurate. Case 14 is well registered; however, there is misregistration derived from 3D rotation misalignment. As a result, the shape difference of the lower lobe became larger, and the HD was more than 10 mm. Thus, the results with relatively small MAE but large HD derived from the rotational component of the registration error. In Case 11, DSC was high; however, both MAE and HD were large. As shown in the coronal view, the deviation in the vertical direction, i.e., in the axial direction, is large. In addition, there was a horizontal misalignment in the appearance from the axial view, resulting in a registration of more than 10 mm MAE.

We measured the computation time for the whole registration process. The average computation time was 65.4 ms (15.3 frames per second), demonstrating the real-time performance for intraoperative registration. Our model was also applied to Case 8 in which tumor

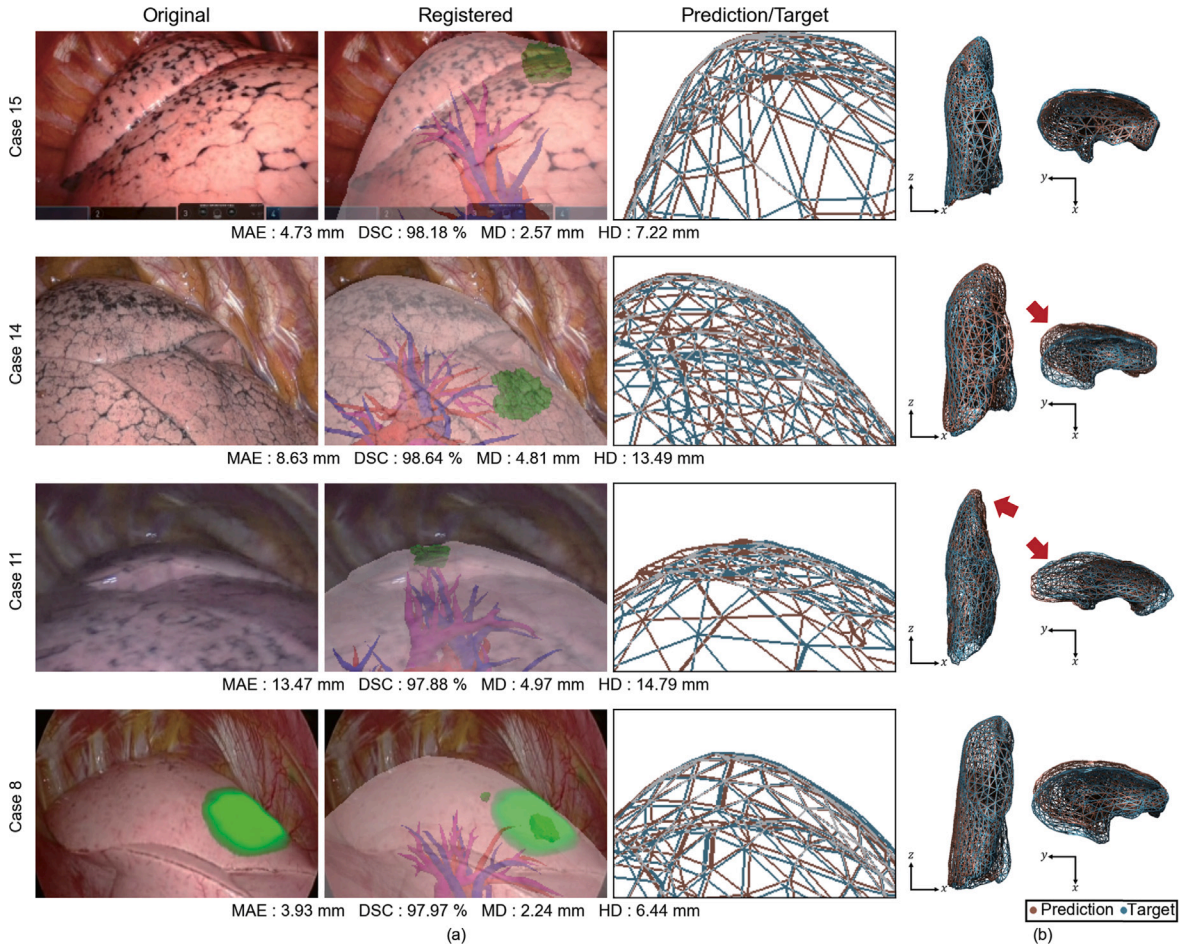


Fig. 8. 2D/3D registration results for endoscopic images. (a) Original surgical scene (left), overlaid image with registered mesh (center), prediction/target meshes (right) in endoscopic views, and (b) the estimated (red) and target (blue) shapes. Case 15 with approximately 25% percentile error, Case 14 with approximately median error, and Case 11 with approximately 75% percentile error. Case 8 underwent VALMAP surgery.

location was visualized by Virtual-Assisted Lung Mapping (VAL-MAP) using indigo carmine (ICG) (Sato et al., 2014). ICG VAL-MAP is a bronchoscopic marking technique where a surgeon performs tumor resection using the dye marking as a surgical guide. The overlaid shape of the tumor located inside the registered mesh is fully included in the area illuminated fluorescent green in the thoracoscopic images. This means that the proposed method achieved accurate registration, and all error metrics also showed high accuracy.

4.3.2. Method comparison

To confirm the 2D/3D deformable registration accuracy for thoracoscopic images, we conducted quantitative comparisons with conventional 2D/3D deformable registration methods. There are only a few methods for 2D/3D deformable registration for single endoscopic images and highly occluded camera images, such as thoracoscopic images, and no studies have been reported for the lung. Although there are differences among conventional methods in the degree of deformation freedom and input images, we selected three conventional concepts for comparison. We re-implemented the programs used in the model comparisons, reproducing the nature of the concept and input data for the network model of the other methods. The only difference between the models is the input data, and the model configuration and parameters are the same to ensure fairness and reproducibility of the experiments.

Direct: 2D/3D registration that directly uses the real thoracoscopic image as an input to the CNN without annotation or segmentation. We chose this model as a baseline because it is a straightforward approach that has been attempted manually or automatically in conventional studies for surgical guidance (Tokuno et al., 2020). In this model, the differential parameters were estimated directly from the rendered image (source) and the thoracoscopic image (target).

Contour: Contour-based 2D/3D registration that minimizes the contour difference from the intraoperative endoscopic image by projecting a deformed mesh simulating organ deformation in 2D. This idea has been used in Koo et al. (2022). To implement the nature of this concept, we estimated the differential of parameters from the rendering image (source) and the thoracoscopic image (target) with contours added to the lung regions.

Virtual: 2D/3D registration using our previous offline learning (Oya et al., 2023). This 2D/3D deformable registration method learned using synthetic images that can be generated from a pneumothorax deformation model. In this model, only binary labels were used as the source in virtual training.

As in Section 4.2, we used four error metrics: MAE, DSC, MD, and HD. We conducted cross-validation on all 16 cases, using a synthetic image as the source and a real image as the target. We used 150 data sets in 15 cases for the training data (20% for validation) and 10 data sets in 1 case for the evaluation data. We trained with a batch size of 30. Synthetic images were generated randomly with a range of ± 20 mm from the center of each parameter for all 3D camera parameters to be estimated. One case generated 10 data sets. In addition, we trained the Virtual and proposed model in the same conditions as in Section 4.2. For the proposed model, besides the rendering image, two distance maps and a label image (target), which is a silhouette of the lung

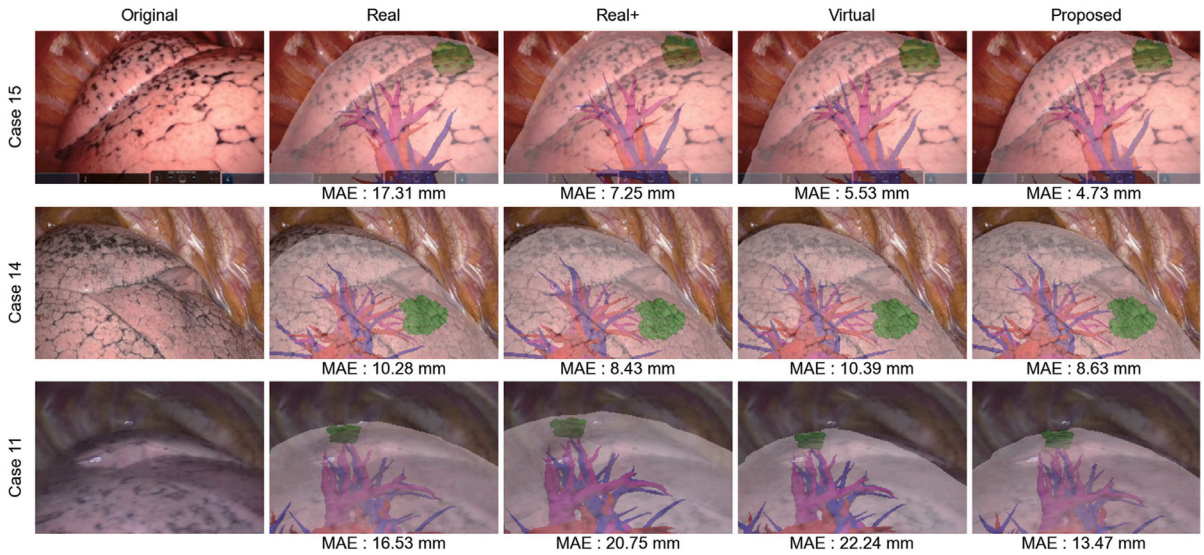


Fig. 9. Comparison of the registration results of conventional methods (Direct, Contour, Virtual) and the proposed method: Case 15 with approximately 25% percentile error, Case 14 with approximately median error, and Case 11 with approximately 75% percentile error.

Table 3
Comparison of the registration error with conventional and proposed methods.

	MAE [mm]	DSC [%]	MD [mm]	HD [mm]
Initial	16.31 ± 7.30	92.28 ± 4.86	7.75 ± 3.19	21.65 ± 8.97
Direct	15.50 ± 6.77	92.39 ± 5.61	7.44 ± 3.01	20.53 ± 8.31
Contour	15.84 ± 7.32	90.80 ± 6.85	7.67 ± 3.14	22.62 ± 9.15
Virtual	11.06 ± 6.93	95.20 ± 4.29	5.34 ± 3.03	15.27 ± 9.11
Proposed	**9.97 ± 5.96**	**96.01 ± 2.91**	**4.84 ± 2.62**	**13.65 ± 7.45**

region, were obtained from the thoracoscopic image using U-net as input for comparison.

Table 3 summarizes the mean ± standard deviations of each error metric of the registration for each method. The registration errors of the proposed method are significantly smaller than those of existing methods for all error metrics (ANOVA; $p < 0.05$). The proposed method is able to align with high overall accuracy, and the standard deviation of the proposed method was smaller than that of conventional methods, which suggests that the proposed method is a more stable registration than conventional methods. These results also indicate that each component of the model has a small but positive effect on registration accuracy.

Fig. 9 shows examples of registration errors. Case 15 shows an approximately 25% percentile registration error for the proposed method, Case 14 shows an approximately median error, and Case 11 shows an approximately 75% percentile error. The results are shown as an augmented endoscopic image. The MAE value is noted below each image. In the registration results, the tumor position in green, artery in red, vein in blue, and bronchus in purple are indicated. In Case 15, the tumor position in green is estimated to be outside of the organ region in the thoracoscopic image in other methods; however, the proposed method can improve the position. Although the overlay image of the Direct model may appear to show better performance, as the error metric shows, the result had a large deviation in the depth direction, resulting in the outside of the organ region. In addition, the proposed method achieves a registration error of less than 5.0 mm, which confirms the effectiveness of the proposed method. In Case 14, the proposed method improved the matching between the contour of the lung region in the thoracoscopic image and the contour of the weighted rendered image. The MAE of the contour-based method was slightly lower than that of the proposed method. In Case 11, the distance between the camera and organ was large and the small part of the upper lobe was observed. Therefore, it was difficult to estimate the distance between the camera and organ from label images only, resulting in an error that was larger in the Virtual model. In contrast, Case 11 achieved a registration closer to the target shape registration than the other methods, although there were some contour errors.

Difficulties in the 2D/3D deformable registration include the fact that changes in camera parameters do not always reflect unique changes in 2D image features. The proposed method achieves more stable registration than conventional methods. It is possible that the rendering with shading and the distance maps introduced in the proposed method are effective at estimating such information.

5. Discussion

We proposed an offline learning method that uses synthetic images that can be generated from a pneumothorax deformation model and prior knowledge obtained from synthetic images to enable data-driven learning. This is the first study to achieve effective 2D/3D deformable registration for thoracoscopic camera images in thoracoscopic pulmonary cancer resection. The difficulty of 2D/3D registration for highly occluded images is due to the increased uncertainty or underdetermined factors. Although shape recovery for arbitrary general objects is obviously difficult, we present a practical approach to addressing the problem by introducing prior knowledge about collapsed lungs and their visual appearance in thoracoscopic surgery from the following two perspectives: (1) the use of a spatially nonlinear deformation atlas to reduce the registration parameters to be estimated, and (2) a conditional deep learning model that encodes initial camera states as two distance maps. The results showed the usefulness of patient-specific offline learning that needs only a single binary mask during inference. The same mask can be obtained from the rendered image for offline learning, enabling generation of training data for learning 2D/3D registration. The model was trained for each patient's mesh generated from preoperative 3D-CT images.

The average registration errors, which evaluate the 3D errors between the shapes, were improved by each component of the proposed model. However, the introduction of some components increased the variances of the error metrics. In particular, the introduction of the distance map between the camera position and organ mesh increased the variance in all error metrics. In this experiment, the camera positions of the source and target were randomly generated, so there were combinations of the source and target whose camera positions were far apart. Because the distance map is based on the information for

the source, the effectiveness of the distance map can be influenced by the differential camera positions of the source and target. There is room for improvement in the generation of source and target images because real surgical environments are expected to have similar scenes as the intraoperative scene, such as the same direction of the camera axis. The largest registration error was observed in Case 11 for both the registration to the synthetic image and thoracoscopic image. The training data for the proposed model comprises 200 data sets for each case, and the camera position and focus change depending on the surgical region, which may explain why there is little effective data for training Case 11.

Our experiments had certain limitations. For example, the pneumothorax deformation model used in this study is limited in its ability to produce synthetic images that completely match the appearance of thoracoscopic images. The scalar parameter w was used to represent the variation among patients. It was intended to suppress the increase in the number of parameters to be learned in offline learning and to improve the stability of learning. We believe that the introduction of high-dimensional atlas parameters to represent local deformation and shape variations of organs, expansion of the applicable range of offline learning, and improvement of accuracy are important topics for future work. We also assumed that the insertion point was determined in the preoperative planning for thoracoscopic surgery. The actual insertion point during surgery could contain substantial setup noise from the plan. However, changing both the insertion point and camera position resulted in an ill-posed setup, which significantly increased the estimation error. Because we performed rendering using a perspective projection, we could assume a higher probability of a one-to-one correspondence between the camera parameter sets and rendered image, which makes stable inference possible. We performed registration to thoracoscopic images of 16 right lung cases. It would be desirable to verify the registration accuracy of the left lung with the same number of cases as the right lung.

Because we did not aim to improve the basic architecture of the deep learning model, we used a general CNN and U-Net, considering the simplicity and reproducibility of the implementation required for the experiments. We also noted that there were no significant differences in the registration performance compared with using the Vision Transformer. Patient-specific training could be improved if additional fine tuning was performed with sufficient population datasets. However, the accuracy was not improved in our 16 cases, meaning that there was too much variety of the lung silhouettes to train the model with stability in the thoracoscopic images. As future work, it would be interesting to extend the idea and consider learning that merges real and offline data while maintaining consistency between the real image features obtained during surgery and those obtained during offline learning. For instance, using multiple images as input for prediction could improve the accuracy.

6. Conclusion

In this paper, we proposed an offline learning framework for 2D/3D deformable registration problems of single highly occluded camera images such as endoscopic images. The proposed framework uses synthetic images that can be generated from a pneumothorax deformation model that represent the organ deformation during surgery, and improves the prediction accuracy with prior knowledge using images commonly obtained from both synthetic and real images as input to the model.

We performed registration accuracy on thoracoscopic images extracted from surgical videos of 16 right lung thoracoscopic surgery cases, and confirmed that the proposed method can achieve a registration accuracy of less than 10 mm for MAE and 5 mm for MD, which are greater than those of conventional methods.

CRediT authorship contribution statement

Tomoki Oya: Writing – original draft, Visualization, Methodology, Formal analysis, Conceptualization. **Yuka Kadomatsu:** Validation, Data curation. **Toyofumi Fengshi Chen-Yoshikawa:** Validation, Supervision, Project administration, Funding acquisition. **Megumi Nakao:** Writing – review & editing, Writing – original draft, Supervision, Software, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Megumi Nakao report financial support was provided by the Japan Society for the Promotion of Science.

Data availability

The authors do not have permission to share data.

Acknowledgments

This research was supported by Grant-in-Aid for Scientific Research (B) 19H04484 and 21H03020 from the Japan Society for the Promotion of Science. We thank Ashleigh Cooper, PhD, from Edanz (https://jp.edanz.com/ac) for editing a draft of this manuscript.

References

Adagolodjo, Y., Trivisonne, R., Haouchine, N., Cotin, S., Courtecuisse, H., 2017. Silhouette-based pose estimation for deformable organs application to surgical augmented reality. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems. IROS, pp. 539–544.

Brock, K.K., Mutic, S., McNutt, T., Li, H., Kessler, M., 2017. Use of image registration and fusion algorithms and techniques in radiotherapy: Report of the AAPM radiation therapy committee task group no. 132. Med. Phys. 44, e43–e76.

Brunet, J.N., Mendizabal, A., Petit, A., Golse, N., Vibert, E., Cotin, S., 2019. Physics-based deep neural network for augmented reality during liver surgery. In: International Conference on Medical Image Computing and Computer Assisted Intervention. MICCAI, pp. 137–145.

Buchs, N.C., Volonte, F., Pugin, F., Toso, C., Fusaglia, M., Gavaghan, K., Majno, P.E., Peterhans, M., Weber, S., Morel, P., 2013. Augmented environments for the targeting of hepatic lesions during image-guided robotic liver surgery. J. Surg. Res. 184 (2), 825–831.

Dong, G., Dai, J., Li, N., Zhang, C., He, W., Liu, L., Chan, Y., Li, Y., Xie, Y., Liang, X., 2023. 2D/3D non-rigid image registration via two orthogonal X-ray projection images for lung tumor tracking. Bioengineering 10 (2), 144.

Espinel, Y., Calvet, L., Botros, K., Buc, E., Tilmant, C., Bartoli, A., 2021. Using multiple images and contours for deformable 3D-2D registration of a preoperative CT in laparoscopic liver surgery. In: International Conference on Medical Image Computing and Computer Assisted Intervention. MICCAI, pp. 657–666.

Geng, J., Xie, J., 2014. Review of 3-D endoscopic surface imaging techniques. IEEE Sens. J. 14 (4), 945–960.

Han, R., Jones, C., Lee, J., Wu, P., Vagdargi, P., Uneri, A., Helm, P., Luciano, M., Anderson, W., Siewerdsen, J., 2022. Deformable MR-CT image registration using an unsupervised, dual-channel network for neurosurgical guidance. Med. Image Anal. 75, 102292.

Haouchine, N., Cotin, S., Peterlík, I., Dequidt, J., Sanz-Lopez, M., Kerrien, E., Berger, M.O., 2015. Impact of soft tissue heterogeneity on augmented reality for liver surgery. IEEE Trans. Vis. Comput. Graphics 21, 584–597.

Haouchine, N., Juvekar, P., Nercessian, M., Wells III, W.M., Golby, A., Frisken, S., 2022. Pose estimation and non-rigid registration for augmented reality during neurosurgery. IEEE Trans. Biomed. Eng. 69 (4), 1310–1317.

Hinterstoisser, S., Lepetit, V., Ilic, S., Holzer, S., Bradski, G., Konolige, K., Navab, N., 2012. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In: The Asian Conference on Computer Vision. ACCV, pp. 5–9.

Hong, D., Tavanapong, W., Wong, J., Oh, J., de Groen, P.C., 2014. 3D reconstruction of virtual colon structures from colonoscopy images. Comput. Med. Imaging Graph. 38 (1), 22–23.

Kasten, Y., Doktofsky, D., Kovler, I., 2020. End-to-end convolutional neural network for 3D reconstruction of knee bones from bi-planar X-ray images. In: Machine Learning for Medical Image Reconstruction. Springer International Publishing, Cham, pp. 123–133.

Ketcha, M.D., De Silva, T., Uneri, A., Jacobson, W., Goerres, J., Kleinszig, G., Vogt, S., Wolinsky, J.P., Siewerdsen, J.H., 2017. Multi-stage 3D-2D registration for correction of anatomical deformation in image-guided spine surgery. Phys. Med. Biol. 62, 4604–4622.

Konishi, K., Hashizume, M., Nakamoto, M., Kakeji, I., Taketomi, A., Sato, Y., Tamura, S., Maehara, Y., 2005. Augmented reality navigation system for endoscopic surgery based on three-dimensional ultrasound and computed tomography: Application to 20 clinical cases. In: International Congress Series, vol. 1281, pp. 537–542.

Koo, B., Özgür, E., Le Roy, B., Buc, E., Bartoli, A., 2017. Deformable registration of a preoperative 3D liver volume to a laparoscopy image using contour and shading cues. Med. Image Comput. Comput. Assist. Interv. 326–334.

Koo, B., Robu, M.R., Allam, M., Pfeiffer, M., Thompson, S., Gurusamy, K., Davidson, B., Speidel, S., Hawkes, D., Stoyanov, D., Clarkson, M.J., 2022. Automatic, global registration in laparoscopic liver surgery. Int. J. Comput. Assist. Radiol. Surg. 17, 167–176.

Kowalczuk, J., Meyer, A., Carlson, J., Psota, E., Buettner, S., Pèrez, L., Farritor, S., Oleynikov, D., 2012. Real-time three-dimensional soft tissue reconstruction for laparoscopic surgery. Surg. Endosc. 26, 3413–3417.

Liao, H., Lin, W.A., Zhang, J., Luo, J., Zhou, S.K., 2019. Multiview 2D/3D rigid registration via a point-of-interest network for tracking and triangulation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 12630–12639.

Lin, B., Sun, Y., Qian, X., Goldgof, D., Gitlin, R., You, Y., 2016. Video-based 3D reconstruction laparoscope localization and deformation recovery for abdominal minimally invasive surgery: A survey. Int. J. Med. Robot. Comput. Assist. Surg. 12 (2), 158–178.

Maekawa, H., Nakao, M., Mineura, K., Chen-Yoshikawa, T., Matsuda, T., 2020. Model-based registration for pneumothorax deformation analysis using intraoperative cone-beam CT images. In: 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society. EMBC, pp. 5818–5821.

Markelj, P., Tomaževič, D., Likar, B., Pernuš, F., 2012. A review of 3D/2D registration methods for image-guided interventions. Med. Image Anal. 16 (3), 642–661.

Mendizabal, A., Màrquez-Neila, P., Cotin, S., 2019. Simulation of hyperelastic materials in real-time using deep learning. Med. Image Anal. 59, 101569.

Miao, S., Wang, Z.J., Liao, R., 2016. A CNN regression approach for real-time 2D/3D registration. IEEE Trans. Med. Imaging 35 (5), 1352–1363.

Modrzejewski, R., Collins, T., Bartoli, A., Hostettler, A., Marescaux, J., 2018. Soft-body registration of preoperative 3D models to intra-operative RGBD partial body scans. In: International Conference on Medical Image Computing and Computer Assisted Intervention. MICCAI, pp. 39–46.

Mountney, P., Stoyanov, D., Davison, A., Yang, G.Z., 2006. Simultaneous stereoscope localization and soft-tissue mapping for minimal invasive surgery. In: International Conference on Medical Image Computing and Computer Assisted Intervention. MICCAI, pp. 347–354.

Nakao, M., Maekawa, H., Mineura, K., Chen-Yoshikawa, T., Date, H., Matsuda, T., 2021. Kernel-based modeling of pneumothorax deformation using intraoperative cone-beam CT images. In: Medical Imaging 2021: Image-Guided Procedures, Robotic Interventions, and Modeling. SPIE, p. 115980P.

Nakao, M., Nakamura, M., Matsuda, T., 2022. Image-to-graph convolutional network for 2D/3D deformable model registration of low-contrast organs. IEEE Trans. Med. Imaging 41 (12), 3747–3761.

Nakao, M., Tokuno, J., Chen-Yoshikawa, T.F., Date, H., Matsuda, T., 2019. Surface deformation analysis of collapsed lungs using model-based shape matching. Int. J. Comput. Assist. Radiol. Surg. 14 (10), 1763–1774.

Nicolau, S., Soler, L., Marescaux, J., 2011. Augmented reality in laparoscopic surgical oncology. Surg. Oncol. 20 (3), 189–201.

Oliveira, F.P., Tavares, J.M.R., 2014. Medical image registration: A review. C. Methods Biomech. Biomed. Eng. 17 (2), 73–93.

Oya, T., Nakao, M., Matsuda, T., 2023. Shape reconstruction from thoracoscopic images using self-supervised virtual learning. p. 2301.10863, arXiv.

Özgür, E., Koo, B., Le Roy, B., Buc, E., Bartoli, A., 2018. Preoperative liver registration for augmented monocular laparoscopy using backward–forward biomechanical simulation. Int. J. Comput. Assist. Radiol. Surg. 13, 1629–1640.

Pellicer-Valero, O.J., Rupèrez, M.J., Martìnez-Sanchis, S., Martìn-Guerrero, J.D., 2020. Real-time biomechanical modeling of the liver using machine learning models trained on finite element method simulations. Expert Syst. Appl. 143, 113083.

Penne, J., Höller, K., Stürmer, M., Schrauder, T., Schneider, A., Engelbrecht, R., Feuß ner, H., Schmauss, B., Hornegger, J., 2009. Time-of-flight 3-D endoscopy. In: International Conference on Medical Image Computing and Computer Assisted Intervention. MICCAI, pp. 467–474.

Pfeiffer, M., Riediger, C., Leger, S., Kühn, J.P., Seppelt, D., Hoffmann, R.T., Weitz, J., Speidel, S., 2020. Non-rigid volume to surface registration using a data-driven biomechanical model. In: International Conference on Medical Image Computing and Computer Assisted Intervention. MICCAI, pp. 724–734.

Pfeiffer, M., Riediger, C., Weitz, J., Speidel, S., 2019. Learning soft tissue behavior of organs for surgical navigation with convolutional neural networks. Int. J. Comput. Assist. Radiol. Surg. 14, 1147–1155.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networksfor biomedical image segmentation. In: Medical Image Computing and Computer Assisted Intervention. MICCAI, pp. 234–241.

Sato, M., Omasa, M., Chen, F., Sato, T., Sonobe, M., Bando, T., Date, H., 2014. Use of virtual assisted lung mapping (VAL-MAP), a bronchoscopic multispot dye-marking technique using virtual images, for precise navigation of thoracoscopic sublobar lung resection. J. Thorac. Cardiovasc. Surg. 147 (6), 1813–1819.

Tang, Z., Chen, K., Pan, M., Wang, M., Song, Z., 2019. An augmentation strategy for medical image processing based on statistical shape model and 3D thin plate spline for deep learning. IEEE Access 7, 133111–133121.

Tokuno, J., Chen-Yoshikawa, T.F., Nakao, M., Ikeda, M., Matsuda, T., Date, H., 2020. Resection process map: A novel dynamic simulation system for pulmonary resection. J. Thorac. Cardiovasc. Surg. 159 (3), 1130–1139.

Wang, J., Schaffert, R., Borsdorf, A., Heigl, B., Huang, X., Hornegger, J., Maier, A., 2017. Dynamic 2-D/3-D rigid registration framework using point-to-plane correspondence model. IEEE Trans. Med. Imaging 36 (9), 1939–1954.

Wang, Y., Zhong, Z., Hua, J., 2020. DeepOrganNet: On-the-fly reconstruction and visualization of 3D / 4D lung models from single-view projections by deep deformation network. IEEE Trans. Vis. Comput. Graphics 26 (1), 960–970.

Wu, S., Nakao, M., Tokuno, J., Chen-Yoshikawa, T., Matsuda, T., 2019. Reconstructing 3D lung shape from a single 2D image during the deaeration deformation process using model-based data augmentation. In: 2019 IEEE EMBS International Conference on Biomedical Health Informatics. BHI, pp. 1–4.

Ying, X., Guo, H., Ma, K., Wu, J., Weng, Z., Zheng, Y., 2019. X2CT-GAN: Reconstructing CT from bi-planar X-rays with generative adversarial networks. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 10611–10620.

Zhao, Q., Price, T., Qian, X., Pizer, S., Niethammer, M., Alterovitz, R., Rosenman, J., 2016. The endoscopogram: A 3D model reconstructed from endoscopic video frames. In: International Conference on Medical Image Computing and Computer Assisted Intervention. MICCAI, pp. 439–447.