Autonomous Intelligent Navigation for Flexible Endoscopy Using Monocular Depth Guidance and 3-D Shape Planning

Yiang Lu[1†], Ruofeng Wei[2†], Bin Li[1], Wei Chen[1], Jianshu Zhou[1,4], Qi Dou[3,4], Dong Sun[2], and Yun-hui Liu[1,4]

Abstract— Recent advancements toward perception and decision-making of flexible endoscopes have shown great potential in computer-aided surgical interventions. However, owing to modeling uncertainty and inter-patient anatomical variation in flexible endoscopy, the challenge remains for efficient and safe navigation in patient-specific scenarios. This paper presents a novel data-driven framework with self-contained visual-shape fusion for autonomous intelligent navigation of flexible endoscopes requiring no priori knowledge of system models and global environments. A learning-based adaptive visual servoing controller is proposed to online update the eye-in-hand vision-motor configuration and steer the endoscope, which is guided by monocular depth estimation via a vision transformer (ViT). To prevent unnecessary and excessive interactions with surrounding anatomy, an energy-motivated shape planning algorithm is introduced through entire endoscope 3-D proprioception from embedded fiber Bragg grating (FBG) sensors. Furthermore, a model predictive control (MPC) strategy is developed to minimize the elastic potential energy flow and simultaneously optimize the steering policy. Dedicated navigation experiments on a robotic-assisted flexible endoscope with an FBG fiber in several phantom environments demonstrate the effectiveness and adaptability of the proposed framework.

I. INTRODUCTION

Robotic endoscope technologies have been widely deployed in a variety of diagnoses and treatments to ease medical devices' accessibility and alleviate physicians' burden. Before the future prevalence of wireless endoscopy using capsule devices or magnetic manipulation [1], [2], traditional flexible endoscopes, such as gastroscopes, colonoscopes, and bronchoscopes, are utilized in dominant endoscopic interventions for their cost-effectiveness and reliability [3]. Currently, endoscopic navigation based on offline trajectory and searching algorithms is well-studied [4], [5]. However, flexible endoscope operation on patient-specific and unstructured scenarios is still challenging considering unknown priori

environmental information, nonlinear system characteristics, and decision-making safety issues [2], [6].

Usually, endoscopic navigation leverages close visualization of intracorporeal scenes from the embedded camera, and utilizes lumen centralization and feature tracking for primary guidance [3], [6]–[8]. The lumen center and visual target can be determined by dark region segmentation [1], [7], depth estimation [8], or contours detection [2]. In real applications, dark region segmentation is impaired by complex lighting conditions, while versatile occlusions limit contours detection effectiveness [2]. Depth estimation methods with higher reliability can be mainly classified into multi-view stereo methods, learning-based approaches, and structured light solutions [6], [8]–[14]. Multi-view stereo methods can reconstruct 3-D scenes with depth estimation while requiring distinguishable features [6], [10]. Learning-based approaches demonstrate their empowered intelligence, among which self-supervised and unsupervised algorithms have been investigated for depth estimation of endoscopic scenes [8], [11], [13]. However, these methods either cannot compute the depth map in real time or adopt binocular images for estimation when perceiving the 3-D scene structure. Recently, vision transformer (ViT) shows great potential in image processing tasks, which extracts features without explicit downsampling and has a global receptive field through all stages [15], thus benefiting depth estimation.

After acquiring the effective visual target as guidance and feedback, various planning and control methodologies have been developed for visual servoing of flexible endoscopes [16]. For model-based methods, nonlinear structural properties and unknown disturbances can result in unknown deviations in the priori system modeling. Model-less and learning-based strategies show great potentials for this task, which update the robot behavior with eye-in-hand vision configuration by estimation and learning approaches [17]–[19]. Amongst, neural networks (NNs) were commonly utilized to learn the system models for visual servoing of soft and continuum manipulators [20], [21]. Deep reinforcement learning was also investigated for image-based control of colonoscopy navigation by devising an end-to-end policy [7]. Although they alleviate priori modeling, data exploration and offline training are required [19]. In addition to vision-based control, follow-the-leader deployment maintaining the endoscope shape while maneuvering the tip, is desired for flexible endoscopic navigation [6], [22].

To further reduce the patient's discomfort and prevent

[1]T Stone Robotics Institute, Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong, Hong Kong.

[2]Department of Biomedical Engineering, City University of Hong Kong, Hong Kong.

[3]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong.

[4]Hong Kong Center for Logistics Robotics, Hong Kong.

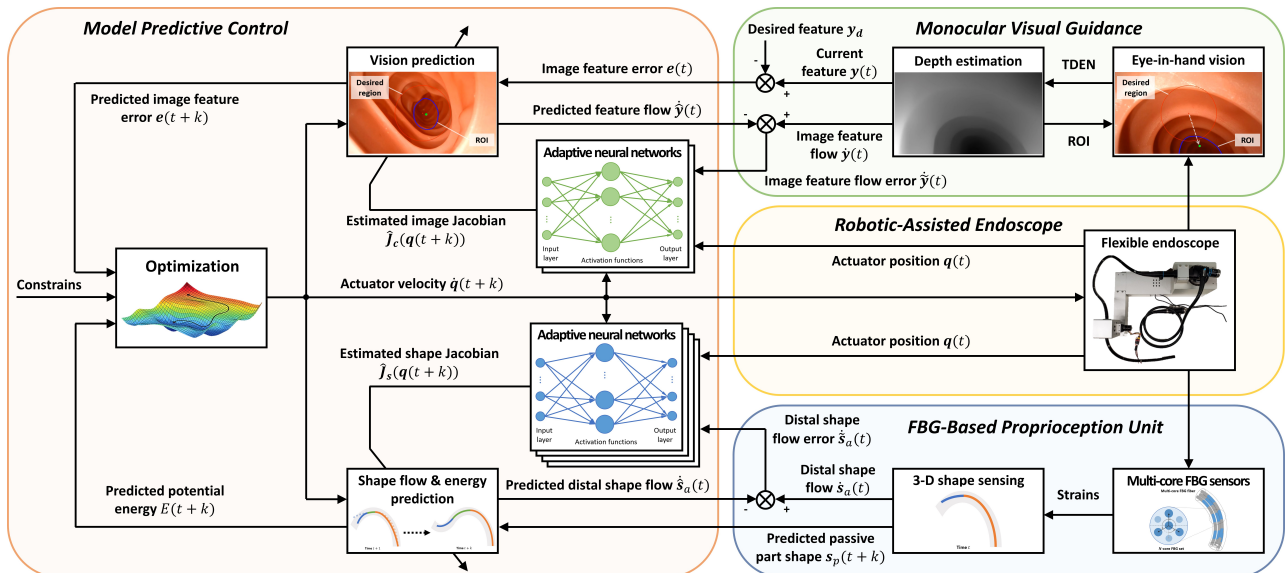[†]Y. Lu and R. Wei contributed equally to this work.

Fig. 1: The proposed data-driven framework for autonomous navigation of flexible endoscopes, including robotic-assisted endoscope, monocular visual guidance, FBG-based proprioception unit, and learning-based model predictive control.

damage to anatomical tissue resulting from unexpected inter-actions, force sensing and haptic guidance were deployed for monitoring the cognitive load during autonomous navigation of flexible endoscopy [4], [23]. Alternatively, deflection and external force in confined spaces can be estimated based on the principle of the minimum total potential energy [24], [25]. Fiber Bragg grating (FBG) sensors have been widely utilized for tip localization and shape sensing of continuum medical robots [20], [26]–[33], because of small size, high biocompatibility, and high sampling rate without line-of-sight constrains. With the working principle of strain measurements, FBGs can be also used to acquire the elastic potential energy of flexible endoscopes from distributed bending/torsion signals [26], [29]. However, studies about energy prediction and 3-D shape planning to avoid excessive contact forces on surrounding anatomy during flexible endoscope navigation have not been reported.

In this paper, we propose a novel data-driven framework as shown in Fig. 1, which can automatically navigate flexible endoscopes without priori data exploration and offline trajectory. To the best of our knowledge, this is the first work in consideration of minimizing potential energy flow for planning and autonomous navigation of unmodeled endoscopes, by incorporating learning-based monocular visual guidance and control, together with FBG-based proprioception of 3-D configuration. The proposed method can effectively online learn the vision-motor behavior with the convergences of image tracking error and learning parameters, and adaptively compensate for the modeling uncertainty with disturbances. Our main contributions are summarized as follows:

1) Design and implementation of a ViT-based depth estimation network for monocular endoscopic guidance, and a learning-based adaptive visual servoing strategy to online update the eye-in-hand vision-motor configuration of the flexible endoscope, and simultaneously track the image region of interest (ROI).

2) Development of an energy-motivated shape planning algorithm by leveraging FBG-based proprioception, which can not only measure the endoscope 3-D configuration but also monitor and minimize its potential energy flow to improve the navigation performance.

3) Integration of depth guidance with learning-based shape flow and energy prediction into a model predictive control (MPC) framework for steering policy optimization of autonomous endoscope navigation.

4) Experimental validations on a robotic-assisted colonoscope system embedded with FBG sensors in several phantoms, the results of which demonstrate the feasibility and adaptability of the proposed framework.

II. MONOCULAR DEPTH-GUIDED VISUAL SERVOING

This section introduces the depth-guided visual servoing strategy, including ViT-based depth estimation and image-based data-driven control of flexible endoscopes.

A. ViT-Based Monocular Depth Guidance

Our ViT-based depth estimation network (VDEN) is designed on an encoder-decoder structure using ViT as the encoder basic block as shown in Fig. 2. We rearrange the embeddings from the encoder into image-like feature maps and combine them from a three-stage decoder into the final dense depth. The standard transformer receives as input a 1-D sequence of token embeddings, so we first map the endoscopic image $\mathcal{X} \in \mathbb{R}^{H \times W \times C}$ into a stacked sequence $\mathcal{K}_0 \in \mathbb{R}^{(N_p+1) \times D}$ in matrix form, where (H, W) denotes the resolution of the image, and C is the number of channels. Specifically, the image $\mathcal{X}$ is divided into N_P patches with the size of $P \times P \times C$, where (P, P) represents the resolution of each patch and $N_P = H \cdot W / P^2$. Then, these patches are flattened into vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n \in \mathbb{R}^{1 \times (P^2 \cdot C)}$, $n \in \{1, 2, ..., N_P\}$. To obtain the spatial position of each patch, we concatenate these patches with a position embedding and
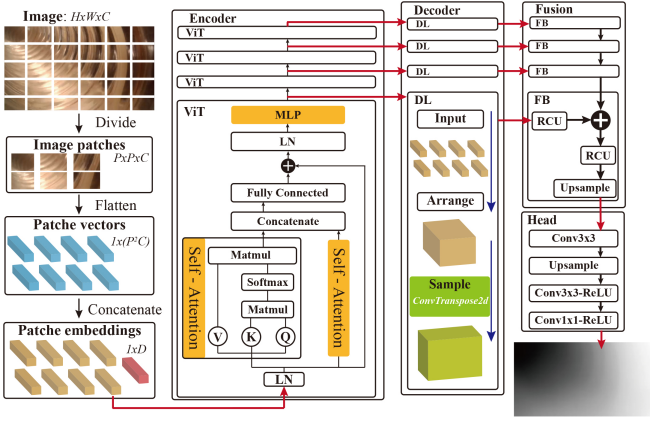
Fig. 2: The proposed ViT-based depth estimation for monocular visual guidance and feedback with endoscopic image as input and depth map as output.

a special projection embedding, and refer to the output of this concatenation as the patch embeddings. Thus, these patch embeddings can be calculated from $\mathcal{K}_0$ as:

$$\mathcal{K}_0 = \begin{bmatrix} \boldsymbol{x}_0^\mathsf{T} & \boldsymbol{E}^\mathsf{T}\boldsymbol{x}_1^\mathsf{T} & \boldsymbol{E}^\mathsf{T}\boldsymbol{x}_2^\mathsf{T} & \cdots & \boldsymbol{E}^\mathsf{T}\boldsymbol{x}_{N_P}^\mathsf{T} \end{bmatrix}^\mathsf{T} + \boldsymbol{E}_{pos} \quad (1)$$

where $\boldsymbol{x}_0$ is a learned embedding to aggregate features into the image level, $\boldsymbol{E} \in \mathbb{R}^{(P^2 \cdot C) \times D}$ is the projection embedding, and $\boldsymbol{E}_{pos} \in \mathbb{R}^{(N_p+1) \times D}$ is the position embedding with D being the feature dimension of each patch embedding.

The patch embeddings are input into ViT-based encoder with L ViT layers, and they are converted to new feature maps $\mathcal{K}_l$, $l \in \{1, ..., L\}$, which are the output of the l-th ViT layer. Here, we adopt the transformer encoder designed by [34]. Each encoder layer has multi-head self-attention and multi-layer perception. Therefore, the global image features can be efficiently extracted from the encoder. Afterwards, we build the decoder [15], each layer of which for depth estimation consists of three parts:

$$\textbf{Input} : \mathbb{R}^{(N_p+1) \times D} \rightarrow \mathbb{R}^{N_p \times D}$$
$$\textbf{Arrange} : \mathbb{R}^{N_p \times D} \rightarrow \mathbb{R}^{\frac{H}{P} \times \frac{W}{P} \times D} \quad (2)$$
$$\textbf{Sample} : \mathbb{R}^{\frac{H}{P} \times \frac{W}{P} \times D} \rightarrow \mathbb{R}^{\frac{H}{r} \times \frac{W}{r} \times D'}$$

where r is the output size ratio of the feature map w.r.t. the input image and D' is the dimension of the decoder output.

After the decoder, we combine the extracted feature maps using a residual convolution unit-based fusion block and gradually upsample the map in each fusion stage. Finally, a depth estimation output head is used to produce the depth maps. In addition, we adopt the scale-invariant trimmed loss [35] to train the whole model. The region of interest (ROI) Ω is determined from the largest connected component of the deepest 5 % pixels, as the navigation guidance.

B. Eye-in-Hand Vision-Motor Configuration

Given the current ROI Ω with its center of mass $\boldsymbol{p} \in \mathbb{R}^2$ from VDEN represented by the blue contour and green point in the monocular visual guidance module of Fig. 1, respectively, a learning-based adaptive controller for online update of the flexible endoscope model together with eye-in-hand visual model, and simultaneous visual servoing, is designed to drive the image central region towards the current

ROI Ω. Therefore, the mass center $\boldsymbol{p}_d \in \mathbb{R}^2$ of the desired region Ω_d (red contour) is located at the image center, which can be also chosen manually by the clinician. we implement a region-based visual servoing method here, the details of which can be found in a previous work [36]. We accordingly define the composed image feature as $\boldsymbol{y} \in \mathbb{R}^2$, which represents the smoothed mass center coordinates of Ω to ensure its continuity through winding number calculation.

To derive the visual servoing controller, we first define the vision-motor model with the task-space velocity, i.e., eye-in-hand camera velocity $\boldsymbol{v}_c \in \mathbb{R}^6$, and the motor position input of the endoscope system as $\boldsymbol{q} = \begin{bmatrix} q_1 & q_2 & q_3 \end{bmatrix}^\mathsf{T} \in \mathbb{R}^3$, where q_1 and q_2 represent the motor positions corresponding to deflections of the distal continuum mechanism, and q_3 is the motor position for insertion motion. The differential kinematics of the flexible endoscope that relates the camera velocity $\boldsymbol{v}_c$ to the motor velocity $\dot{\boldsymbol{q}}$, is given as $\boldsymbol{v}_c = \mathcal{J}_r(\boldsymbol{q})\dot{\boldsymbol{q}}$, where $\mathcal{J}_r(\boldsymbol{q}) \in \mathbb{R}^{6 \times 3}$ denotes the Jacobian matrix of the endoscope system. The eye-in-hand visual model of the flexible endoscope mapping the image feature flow $\dot{\boldsymbol{y}}$ to the camera motion $\boldsymbol{v}_c$ can be formulated as $\dot{\boldsymbol{y}} = \mathcal{L}_y \boldsymbol{v}_c$, where $\mathcal{L}_y \in \mathbb{R}^{2 \times 6}$ represents the vision interaction matrix for the image feature $\boldsymbol{y}$. From above two mappings, we can derive the vision-motor configuration of the endoscope system as

$$\dot{\boldsymbol{y}} = \mathcal{J}_c(\boldsymbol{q})\dot{\boldsymbol{q}} = \mathcal{L}_y \mathcal{J}_r(\boldsymbol{q})\dot{\boldsymbol{q}} \quad (3)$$

where $\mathcal{J}_c(\boldsymbol{q}) = \mathcal{L}_y \mathcal{J}_r(\boldsymbol{q}) \in \mathbb{R}^{2 \times 3}$ denotes the combined image Jacobian matrix of flexible endoscope mapping its motor velocity $\dot{\boldsymbol{q}}$ to the image feature velocity $\dot{\boldsymbol{y}}$ (i.e., image feature flow). Note that the endoscope Jacobian $\mathcal{J}_r(\boldsymbol{q})$ is assumed to be unknown due to its nonlinear structural properties, hence we approximate the image Jacobian $\mathcal{J}_c(\boldsymbol{q})$ without any priori identification of it.

C. Learning-Based Adaptive Visual Servoing

To this end, we employ two adaptive NNs $\mathcal{W}_i \boldsymbol{\theta}_i(\boldsymbol{q}), i \in \{1, 2\}$ as shown in Fig. 1, to learn the rows of image Jacobian $\mathcal{J}_c(\boldsymbol{q})$, with motor position $\boldsymbol{q}$ as input given by

$$\mathcal{J}_c(\boldsymbol{q}) = \begin{bmatrix} \mathcal{W}_1 \boldsymbol{\theta}_1(\boldsymbol{q}) & \mathcal{W}_2 \boldsymbol{\theta}_2(\boldsymbol{q}) \end{bmatrix}^\mathsf{T} \quad (4)$$

where $\mathcal{W}_i \in \mathbb{R}^{3 \times \xi}, i \in \{1, 2\}$, represents the ideal weight matrix of the i-th NN with ξ neurons, and $\boldsymbol{\theta}_i(\boldsymbol{q}) \in \mathbb{R}^\xi, i \in \{1, 2\}$, denotes the corresponding vector of the activation functions. With online learning of the image Jacobian in (4), we can linearly parameterize the image feature flow $\dot{\boldsymbol{y}}$ as

$$\dot{\boldsymbol{y}} = \underbrace{\begin{bmatrix} \mathcal{W}_1 \boldsymbol{\theta}_1(\boldsymbol{q}) & \mathcal{W}_2 \boldsymbol{\theta}_2(\boldsymbol{q}) \end{bmatrix}^\mathsf{T}}_{\mathcal{J}_c(\boldsymbol{q})} \dot{\boldsymbol{q}} = \boldsymbol{\Theta}^\mathsf{T}(\boldsymbol{q})\mathcal{Q}^\mathsf{T}(\dot{\boldsymbol{q}})\overline{\mathcal{W}} \quad (5)$$

where $\boldsymbol{\Theta}(\boldsymbol{q}) = \text{diag}(\boldsymbol{\theta}_1(\boldsymbol{q}), \boldsymbol{\theta}_2(\boldsymbol{q})) \in \mathbb{R}^{2\xi \times 2}$ and $\boldsymbol{Q}(\dot{\boldsymbol{q}}) = \text{diag}(\dot{\boldsymbol{q}}, \dot{\boldsymbol{q}}, \cdots, \dot{\boldsymbol{q}}) \in \mathbb{R}^{6\xi \times 2\xi}$ are diagonal block matrices grouping the activation functions $\boldsymbol{\theta}_i(\boldsymbol{q})$, $i \in \{1, 2\}$, and 2ξ motor motion $\dot{\boldsymbol{q}}$, respectively, and $\overline{\mathcal{W}} \in \mathbb{R}^{6\xi}$ denotes a vector stacking the columns of the ideal weights $\mathcal{W}_i$, $i \in \{1, 2\}$.

Since the real image Jacobian $\mathcal{J}_c(\boldsymbol{q})$ is unknown, we estimate it denoted by $\widehat{\mathcal{J}}_c(\boldsymbol{q}) \in \mathbb{R}^{2 \times 3}$ through two estimated adaptive NNs $\widehat{\mathcal{W}}_i \boldsymbol{\theta}_i(\boldsymbol{q}), i \in \{1, 2\}$, given by

$$\widehat{\mathcal{J}}_c(\boldsymbol{q}) = \begin{bmatrix} \widehat{\mathcal{W}}_1 \boldsymbol{\theta}_1(\boldsymbol{q}) & \widehat{\mathcal{W}}_2 \boldsymbol{\theta}_2(\boldsymbol{q}) \end{bmatrix}^\mathsf{T} \quad (6)$$

where $\widehat{\mathcal{W}}_i \in \mathbb{R}^{3\times\xi}, i \in \{1,2\}$, represents the estimated NN weight matrix. Therefore, we can derive the prediction error of image feature flow $\dot{\boldsymbol{y}} \in \mathbb{R}^2$, i.e., the image flow error as

$$\dot{\tilde{\boldsymbol{y}}} = \dot{\boldsymbol{y}} - \widehat{\mathcal{J}}_c(\boldsymbol{q})\dot{\boldsymbol{q}} = \boldsymbol{\Theta}^{\mathsf{T}}(\boldsymbol{q})\mathcal{Q}^{\mathsf{T}}(\dot{\boldsymbol{q}})\widetilde{\mathcal{W}} \tag{7}$$

where $\widetilde{\mathcal{W}} = \overline{\mathcal{W}} - \widehat{\mathcal{W}} \in \mathbb{R}^{6\xi}$ denotes the NN weights estimation error in vector form, which can be updated from a modified composite adaptation algorithm [37] as

$$\dot{\widehat{\overline{\mathcal{W}}}} = \boldsymbol{\Gamma}_W^{-1}\left[\mu_e\mathcal{Q}(\dot{\boldsymbol{q}})\boldsymbol{\Theta}(\boldsymbol{q})\boldsymbol{e} + \mu_y\mathcal{Q}(\dot{\boldsymbol{q}})\boldsymbol{\Theta}(\boldsymbol{q})\dot{\tilde{\boldsymbol{y}}}\right] \tag{8}$$

where μ_e and μ_y are positive constants, $\boldsymbol{\Gamma}_W \in \mathbb{R}^{6\xi\times6\xi}$ is a positive definite and diagonal gain matrix, and $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{y}_d \in \mathbb{R}^2$ is denoted as the image feature error (tracking error).

Finally, we can design a velocity controller to drive the image feature $\boldsymbol{y}$ towards the desired one $\boldsymbol{y}_d$ by using the learned image Jacobian $\widehat{\mathcal{J}}_c(\boldsymbol{q})$ as $\dot{\boldsymbol{q}} = -\mu_c\widehat{\mathcal{J}}_c^+(\boldsymbol{q})\boldsymbol{e}$, where μ_c is a positive constant and $\widehat{\mathcal{J}}_c^+(\boldsymbol{q}) \in \mathbb{R}^{3\times2}$ denotes the Moore–Penrose pseudo-inverse of $\widehat{\mathcal{J}}_c(\boldsymbol{q})$. Under this control algorithm, the closed-loop system stability, together with the convergences of image feature error and image flow error to zeros, i.e., $\lim_{t\to\infty} \boldsymbol{e} = 0$ and $\lim_{t\to\infty} \dot{\tilde{\boldsymbol{y}}} = 0$, can be guaranteed, and a similar proof of which can be found in [38] for shape servoing tasks. Alternatively, we can apply a learning-based MPC strategy to predict the future behavior of the flexible endoscope over a future time horizon Φ as

$$\underset{\dot{\boldsymbol{q}}(t+k+1)}{\arg\min} \sum_{k=0}^{\Phi} \eta_k \left\|\boldsymbol{y}(t+k+1) - \boldsymbol{y}_d\right\|_2^2 \tag{9}$$

s.t. $\dot{q}_3(t+k) \geq 0$ as insertion motion with the predicted feature $\boldsymbol{y}(t+k+1)$ for $k \in \{0,1,2,...,\Phi\}$ being

$$\boldsymbol{y}(t+k+1) = \boldsymbol{y}(t+k) + \widehat{\mathcal{J}}_c(\boldsymbol{q}(t+k))\dot{\boldsymbol{q}}(t+k)\Delta t \tag{10}$$

where Δt is the iteration interval, and the image Jacobian $\widehat{\mathcal{J}}_c(\boldsymbol{q}(t+k))$, $k \in \{0,1,2,...,\Phi\}$, can be predicted through the adaptive NNs as shown in Fig. 1.

III. Energy-Motivated 3-D Shape Planning and Autonomous Navigation Framework

In this section, we describe an energy-motivated shape planning algorithm by using FBG sensors, and a learning-based MPC framework for flexible endoscope navigation.

A. FBG-based Proprioception of Elastic Potential Energy

To avoid unnecessary and excessive interactions with surrounding anatomy during endoscopic navigation, we design an optimization-based planning method for minimizing the elastic potential energy flow of the system. By solely embedding a multi-core FBG fiber in the flexible endoscope, a robust and accurate filtering-based algorithm was previously reported [29] to acquire the endoscope 3-D shape, which is self-contained, independent of external sensors, and can maintain high sensing quality against perturbations. At current time t, we construct the entire endoscope shape by two part including the active part $\boldsymbol{s}_a \in \mathbb{R}^{m_a}$ of the distal continuum mechanism, and passive part $\boldsymbol{s}_p \in \mathbb{R}^{m_p}$ with a length of $l_p(t)$ as shown in Fig. 3 (a). In addition to shape sensing as shown in Fig. 1, we utilized the bending and
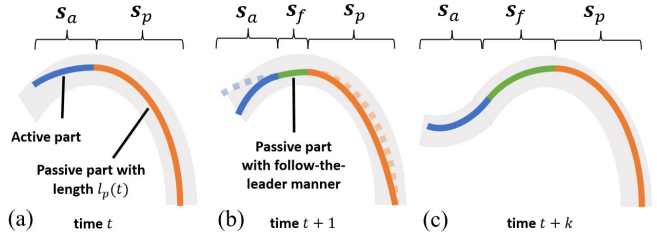

Fig. 3: Shape prediction of flexible endoscope for elastic potential energy minimization: (a) current shape, (b) shape at time $t+1$, and (c) shape at time $t+k$.

torsion signals from the distributed FBG sensors to approximate the elastic potential energy of the flexible endoscope formulated by $\mathcal{E}(t) = \mathcal{E}(\boldsymbol{s}_a(t), \boldsymbol{s}_p(t)) : \mathbb{R}^{m_a} \times \mathbb{R}^{m_p} \mapsto \mathbb{R}$, following the working principal of strain measurements.

To minimize the future potential energy flow, we predict the elastic potential energy $\mathcal{E}(t+k+1)$, $k \in \{0,1,2,...,\Phi\}$, by dividing the entire endoscope shape into three parts: the active part $\boldsymbol{s}_a(t+k+1)$, passive part with follow-the-leader manner $\boldsymbol{s}_f(t+k+1) \in \mathbb{R}^{m_f}$, where $m_f(t+k+1)$ is dependent on the insertion length at each time instant, and passive part $\boldsymbol{s}_p(t+k+1)$ with the length of $l_p(t)$, as shown in Fig. 3 (b). By taking advantage of state prediction in the Kalman filter, the sensing algorithm is hereby improved to predict the passive part shape $\boldsymbol{s}_p(t+k+1)$ as shown in the FBG-based sensing module of Fig. 1, which is deformed due to its interactions with anatomical tissues during endoscope insertion. Using FBG-based proprioception of the endoscope 3-D configuration, the active part shape $\boldsymbol{s}_a(t+k+1)$ can be predicted through a learning-based approach as

$$\boldsymbol{s}_a(t+k+1) = \boldsymbol{s}_a(t+k) + \widehat{\mathcal{J}}_s(\boldsymbol{q}'(t+k))\dot{\boldsymbol{q}}'(t+k)\Delta t \tag{11}$$

where $\boldsymbol{q}' = \begin{bmatrix} q_1 & q_2 \end{bmatrix}^{\mathsf{T}} \in \mathbb{R}^2$ denotes two motors' positions for distal deflections, $\widehat{\mathcal{J}}_s(\boldsymbol{q}') \in \mathbb{R}^{m_a\times2}$ is the estimated shape Jacobian corresponding to the actual one $\mathcal{J}_s(\boldsymbol{q}') \in \mathbb{R}^{m_a\times2}$, which relates the active part shape flow $\dot{\boldsymbol{s}}_a$ to $\dot{\boldsymbol{q}}'$. In a similar way updating the image Jacobian $\widehat{\mathcal{J}}_c(\boldsymbol{q}(t+k))$, we can learn and predict the shape Jacobian $\widehat{\mathcal{J}}_s(\boldsymbol{q}'(t+k))$, from m_a adaptive NNs [38] as shown in Fig. 1.

B. Potential Energy Prediction and Shape Planning

By making the assumption of follow-the-leader behavior [6], [22], we can predict the passive part shape $\boldsymbol{s}_f(t+k+1)$, $k \in \{0,1,2,...,\Phi\}$, following the current shape $\boldsymbol{s}_a(t)$ and the predicted ones $\boldsymbol{s}_a(t+k+1)$ of the endoscope active part as shown in Fig. 3 (b), given by

$$\boldsymbol{s}_f(t+k+1) = \begin{bmatrix} \boldsymbol{0}_{m_i\times m_o} & \boldsymbol{I}_{m_i} & \boldsymbol{0}_{m_i\times m_f} \\ \boldsymbol{0}_{m_f\times m_o} & \boldsymbol{0}_{m_f\times m_i} & \boldsymbol{I}_{m_f} \end{bmatrix} \begin{bmatrix} \boldsymbol{s}_a(t+k) \\ \boldsymbol{s}_f(t+k) \end{bmatrix} \tag{12}$$

where $\boldsymbol{I}_a \in \mathbb{R}^{a\times a}$ and $\boldsymbol{0}_{a\times b} \in \mathbb{R}^{a\times b}$, $a \in \{m_i, m_f\}$ and $b \in \{m_i, m_f, m_o\}$, denote the identity and zero matrices, respectively, and $m_o = m_a - m_i$ with m_i being the endoscope insertion length. Consequently, the elastic potential energy of the flexible endoscope $\mathcal{E}(t+k')$, $k' \in \{1,2,...,\Phi\}$, can be predicted from the entire endoscope shape (three parts) predicted at time $t+k'$ as shown in Fig. 3 (c), given by

$$\mathcal{E}(t+k') = \mathcal{E}\left(\boldsymbol{s}_a(t+k'), \boldsymbol{s}_f(t+k'), \boldsymbol{s}_p(t+k')\right)$$
$$: \mathbb{R}^{m_a} \times \mathbb{R}^{m_f} \times \mathbb{R}^{m_p} \mapsto \mathbb{R} \tag{13}$$
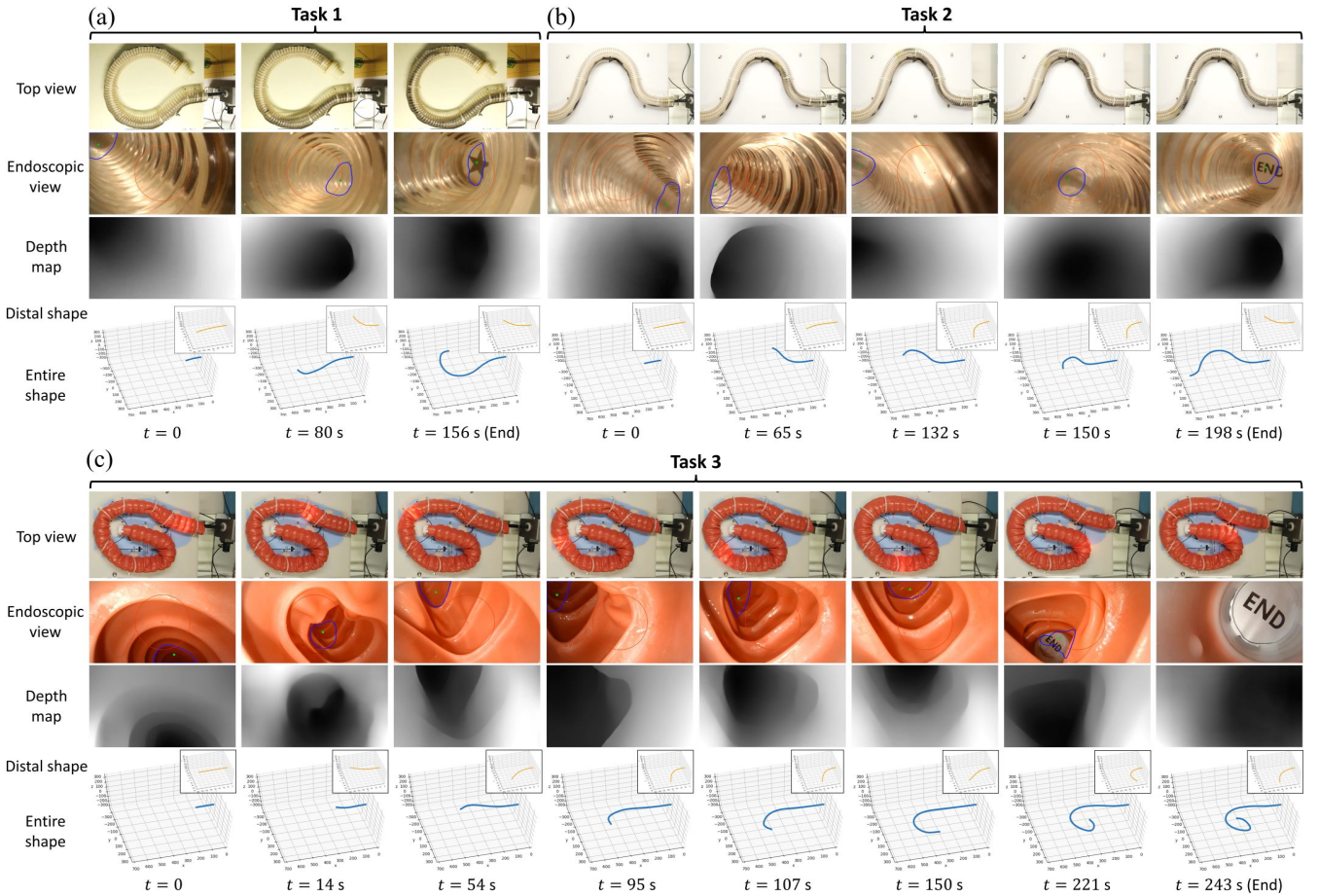
Fig. 4: Snapshots in Task 1, 2, and 3: top view, endoscopic view, depth map, and FBG shapes of entire colonoscope and distal part.

Therefore, a local shape trajectory with the smallest cost of the potential energy flow can be obtained by solving the following optimization problem as

$$\underset{\dot{\boldsymbol{q}}(t+k+1)}{\arg\min} \sum_{k=0}^{\Phi} \lambda_k \left\| \mathcal{E}(t+k+1) - \mathcal{E}(t+k) \right\|_2^2 \qquad (14)$$

s.t. (11), (12), (13), and $\dot{\boldsymbol{q}}_3(t+k) \geq 0$. Note that the proposed shape planning algorithm is generic and feasible for other shape sensing techniques [27] rather than only for FBGs.

C. Learning-Based Model Predictive Control

In a combination of the depth-guided visual control and energy-motivated shape planning described above, we derive a data-driven framework for autonomous navigation of flexible endoscopes as shown in Fig. 1, based on the development of a learning-based MPC strategy with the following formulation:

$$\underset{\dot{\boldsymbol{q}}(t+k+1)}{\arg\min} \sum_{k=0}^{\Phi} \left\{ \eta_k \left\| \boldsymbol{y}(t+k+1) - \boldsymbol{y}_d \right\|_2^2 \right.$$
$$\left. + \lambda_k \left\| \mathcal{E}(t+k+1) - \mathcal{E}(t+k) \right\|_2^2 \right\} \qquad (15)$$

s.t. (10) ∼ (13), and $\dot{\boldsymbol{q}}_3(t+k) \geq 0$, where the first term is to control the image ROI, the second term is for potential energy flow minimization, η_k and λ_k, $k \in \{0, 1, 2, ..., \Phi\}$, denote the weights w.r.t. these two tasks. Finally, the steering policy over a time horizon $\dot{\boldsymbol{q}}(t+k+1)$, $k \in \{0, 1, 2, ..., \Phi\}$,

can be optimized by satisfying the above objective function s.t. the constraints, and then input $\dot{\boldsymbol{q}}(t+k+1)$ to the motors of the endoscope system, which can impose the desired image ROI and simultaneously minimize the potential energy flow.

IV. EXPERIMENTS

In this section, the experimental setup is firstly described, and the results are accordingly presented and discussed.

A. Setup

The training dataset for the proposed depth estimation is formulated with MIX 5 [35] and endoscopic images [13]. The endoscopic images were constructed from virtual capsule endoscopy, which was developed in Unity and had 21887 frames with corresponding ground truth depth maps. Our VDEN model was implemented in PyTorch and trained using Adam solver for 60 epochs with the batch size of 6 and the learning rate was set as 1e-5. In the experiments, the original image was cropped to $384 \times 384 \times 3$ as input. The resolution $P = 16$, the number L of ViT layers was 24, and the dimensions $D = 768$ and $D^{'} = 256$ with the ratio $r = 2$.

We performed the experiments on a robotic-assisted flexible endoscope system previously detailed in [29], which was integrated with a colonoscope, a handle drive module for distal deflections, a dual-wheel friction module for endoscope insertion, a multi-core FBG fiber, and an eye-in-hand
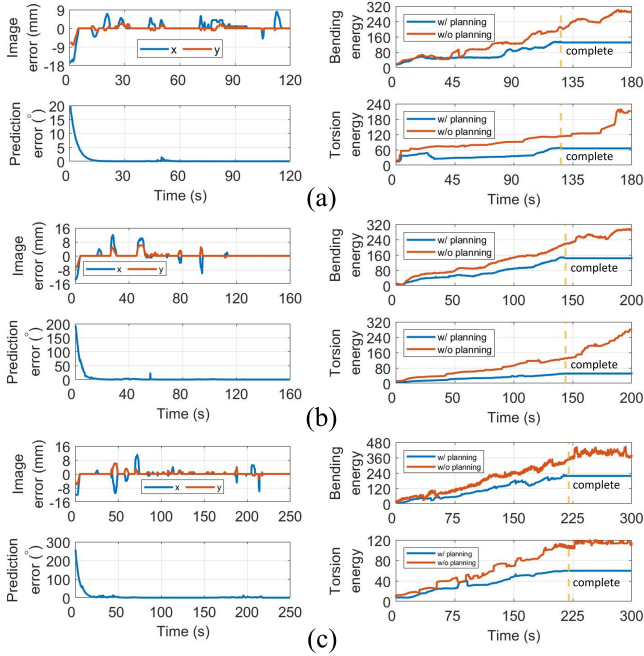
Fig. 5: Image tracking error, shape flow prediction error, bending and torsion potential energy using the methods with and without shape planning in (a) Task 1, (b) Task 2, and (c) Task 3.

	Methods	T_{in} (s)	L_{et} (mm)	$\|e\|$ (mm)
Task 1	w/o planning	179 ± 26	1752.4 ± 449.3	1.24 ± 0.27
	w/ planning	125 ± 11	1181.0 ± 153.1	0.70 ± 0.29
Task 2	w/o planning	204 ± 13	1668.3 ± 141.4	1.25 ± 0.25
	w/ planning	159 ± 37	1292.4 ± 173.2	0.63 ± 0.23
Task 3	w/o planning	334 ± 22	1935.5 ± 397.1	1.47 ± 0.21
	w/ planning	238 ± 17	1485.6 ± 119.9	0.87 ± 0.27

accuracy of 90.88 % in Task 1, 91.08 % in Task 2, and 82.25 % in Task 3, where the overlap degree of the estimated and real deep regions in each frame was examined. Moreover, sharp boundaries were well kept, and the estimated depth increased smoothly from near to far scenes as demonstrated in the depth maps of Fig. 4, despite in unknown environments with reflective and translucent surfaces. We can observe the estimated depth maps during all tasks were particularly stable, hence helpful for autonomous endoscope navigation.

As indicated in the endoscopic views of Fig. 4 as well as the image tracking errors in the top-left images of Fig. 5, the learning-based eye-in-hand visual servoing can robustly and adaptively track the image ROI (blue contour) with average image errors of 0.70, 0.63, and 0.87 mm in Task 1, 2, and 3, respectively, and complete the navigation tasks. It is noticed that the shape flow prediction errors converged to zeros as shown in the bottom-left plots of Fig. 5 (a), (b), and (c), which indicate that the NN weights were learned from their mapping in [38]. Besides, the bending and torsion potential energy in the right plots of Fig. 5 (a), (b), and (c), demonstrate that our planning algorithm can finish the navigation tasks using much less elastic potential energy as compared with that without shape planning, which can prevent unnecessary and excessive elastic contact forces on surrounding anatomy. From Table I, the proposed method costs less time T_{in} with shorter endpoint trajectory L_{et} for navigation compared with the method without shape planning, which also outperforms the manually-controlled steering procedure (more than 15 min). These results of phantom experiments support our claims that the framework can perform robust compliance to deformable environments, online learn the time-varying model, and simultaneously steer the flexible endoscope for efficient and safe navigation.

V. CONCLUSIONS

In this paper, we propose a novel data-driven autonomous navigation framework for flexible endoscopy, by leveraging monocular depth guidance and energy-motivated shape planning. The method can online learn the eye-in-hand vision-motor configuration of flexible endoscope without any priori knowledge of system model and environmental information, and also minimize the system potential energy flow. The results of several phantom experiments show the feasibility and adaptability of the framework.

In the future, we will optimize the framework with advanced perception and control approaches to improve its performance, which will be evaluated in more realistic scenes. The system issues about learning approximation errors and colliding with anatomical tissues will be taken into consideration as well.

camera. The system has a distal active part of 120 mm length and a passive working length of 880 mm, hence the FBG-based proprioception unit has 1 m sensing length in total. The parameters in the MPC framework were set as follows: $\mu_e = 0.01$, $\mu_y = 0.2$, $\mathbf{\Gamma}_W^{-1} = \mathbf{I}_{54} \in \mathbb{R}^{54 \times 54}$, $\eta_k = 1/2^k$, $\lambda_k = 1/2^{k+1}$, $k \in \{0, 1, 2, ..., \Phi\}$, and $\Phi = 20$, which were tuned according to the initial experiments with convergences of control and prediction errors to zeros. As same with those in [38], 9 neurons (i.e., $\xi = 9$) and radial basis functions (RBFs) were employed for NNs learning of both vision and shape prediction, and so were the center and width values.

B. Results and Discussion

To evaluate the performance of the proposed framework, three tasks compared with the method using vision guidance but without shape planning, were performed in two feature-less phantoms with different shapes and one colonoscopy phantom (Kyoto Kagaku M40) as shown in Fig. 4. The phantoms have diameters of 50 mm and lengths of more than 1 m to mimic the endoscopic scenarios. We conducted each task 8 times and recorded the results with snapshots in Fig. 4, including the top view of endoscope, endoscopic view, depth map, and FBG-based shapes of entire colonoscope and distal part. Four qualitative results during one trial of each task are plotted in Fig. 5 (a), (b), and (c), respectively, including image tracking error e, 2-norm prediction error of shape flow $\|\dot{\tilde{s}}_a\|$, bending potential energy $\mathcal{E}_b$, and torsion potential energy $\mathcal{E}_t$. Moreover, we present three quantitative metrics referring to [7], in terms of insertion time T_{in}, endpoint trajectory length L_{et}, and average 2-norm of image tracking error $\|e\|$ of three tasks in Table I.

For the performance of depth estimation, our VDEN model running at 19 FPS, can provide reasonable depth maps with

REFERENCES

[1] J. W. Martin, B. Scaglioni, J. C. Norton, V. Subramanian, A. Arezzo, K. L. Obstein, and P. Valdastri, "Enabling the future of colonoscopy with intelligent and autonomous magnetic manipulation," *Nature machine intelligence*, vol. 2, no. 10, pp. 595–606, 2020.

[2] J. M. Prendergast, G. A. Formosa, M. J. Fulton, C. R. Heckman, and M. E. Rentschler, "A real-time state dependent region estimator for autonomous endoscope navigation," *IEEE Transactions on Robotics*, vol. 37, no. 3, pp. 918–934, 2020.

[3] N. van der Stap, F. van der Heijden, and I. A. Broeders, "Towards automated visual flexible endoscope navigation," *Surgical Endoscopy*, vol. 27, no. 10, pp. 3539–3547, 2013.

[4] H.-E. Huang, S.-Y. Yen, C.-F. Chu, F.-M. Suk, G.-S. Lien, and C.-W. Liu, "Autonomous navigation of a magnetic colonoscope using force sensing and a heuristic search algorithm," *Scientific reports*, vol. 11, no. 1, pp. 1–15, 2021.

[5] A. Favaro, A. Segato, F. Muretti, and E. De Momi, "An evolutionary-optimized surgical path planner for a programmable bevel-tip needle," *IEEE Transactions on Robotics*, vol. 37, no. 4, pp. 1039–1050, 2021.

[6] C. Girerd, A. V. Kudryavtsev, P. Rougeot, P. Renaud, K. Rabenorosoa, and B. Tamadazte, "Slam-based follow-the-leader deployment of concentric tube robots," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 548–555, 2020.

[7] A. Pore, M. Finocchiaro, D. Dall'Alba, A. Hernansanz, G. Ciuti, A. Arezzo, A. Menciassi, A. Casals, and P. Fiorini, "Colonoscopy navigation using end-to-end deep visuomotor control: A user study," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 9582–9588.

[8] R. Wei, B. Li, H. Mo, B. Lu, Y. Long, B. Yang, Q. Dou, Y. Liu, and D. Sun, "Stereo dense scene reconstruction and accurate localization for learning-based navigation of laparoscope in minimally invasive surgery," *IEEE Transactions on Biomedical Engineering*, 2022.

[9] X. Maurice, C. Albitar, C. Doignon, and M. de Mathelin, "A structured light-based laparoscope with real-time organs' surface reconstruction for minimally invasive surgery," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2012, pp. 5769–5772.

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

[11] X. Liu, A. Sinha, M. Ishii, G. D. Hager, A. Reiter, R. H. Taylor, and M. Unberath, "Dense depth estimation in monocular endoscopy with self-supervised learning methods," *IEEE transactions on medical imaging*, vol. 39, no. 5, pp. 1438–1447, 2019.

[12] B. Li, B. Lu, Y. Lu, Q. Dou, and Y.-H. Liu, "Data-driven holistic framework for automated laparoscope optimal view control with learning-based depth perception," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 12 366–12 372.

[13] K. B. Ozyoruk, G. I. Gokceler, T. L. Bobrow, G. Coskun, K. Incetan, Y. Almalioglu, F. Mahmood, E. Curto, L. Perdigoto, M. Oliveira *et al.*, "Endoslam dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos," *Medical image analysis*, vol. 71, p. 102058, 2021.

[14] R. Wei, B. Li, H. Mo, F. Zhong, Y. Long, Q. Dou, Y.-H. Liu, and D. Sun, "Distilled visual and robot kinematics embeddings for metric depth estimation in monocular scene reconstruction," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 8072–8077.

[15] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12 179–12 188.

[16] A. A. Nazari, K. Zareinia, and F. Janabi-Sharifi, "Visual servoing of continuum robots: Methods, challenges, and prospects," *The International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 18, no. 3, p. e2384, 2022.

[17] T. George Thuruthel, Y. Ansari, E. Falotico, and C. Laschi, "Control strategies for soft robotic manipulators: A survey," *Soft robotics*, vol. 5, no. 2, pp. 149–163, 2018.

[18] G. Fang, X. Wang, K. Wang, K.-H. Lee, J. D. Ho, H.-C. Fu, D. K. C. Fu, and K.-W. Kwok, "Vision-based online learning kinematic control for soft robots using local gaussian process regression," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1194–1201, 2019.

[19] X. Wang, Y. Li, and K.-W. Kwok, "A survey for machine learning-based control of continuum robots," *Frontiers in Robotics and AI*, p. 280, 2021.

[20] X. Wang, G. Fang, K. Wang, X. Xie, K.-H. Lee, J. D. Ho, W. L. Tang, J. Lam, and K.-W. Kwok, "Eye-in-hand visual servoing enhanced with sparse strain measurement for soft continuum robots," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 2161–2168, 2020.

[21] G. Fang, M. C. Chow, J. D. Ho, Z. He, K. Wang, T. Ng, J. K. Tsoi, P.-L. Chan, H.-C. Chang, D. T.-M. Chan *et al.*, "Soft robotic manipulator for intraoperative mri-guided transoral laser microsurgery," *Science Robotics*, vol. 6, no. 57, p. eabg5575, 2021.

[22] C. Culmone, S. F. Yikilmaz, F. Trauzettel, and P. Breedveld, "Follow-the-leader mechanisms in medical devices: A review on scientific and patent literature," *IEEE Reviews in Biomedical Engineering*, 2021.

[23] R. Reilink, S. Stramigioli, A. M. Kappers, and S. Misra, "Evaluation of flexible endoscope steering using haptic guidance," *The International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 7, no. 2, pp. 178–186, 2011.

[24] N. Sarli and N. Simaan, "Minimal visual occlusion redundancy resolution of continuum robots in confined spaces," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 6448–6454.

[25] H. Kim, J. M. You, M. Hwang, K.-U. Kyung, and D.-S. Kwon, "Sigmoidal auxiliary tendon-driven mechanism reinforcing structural stiffness of hyper-redundant manipulator for endoscopic surgery," *Soft Robotics*, 2022.

[26] H. Wang, R. Zhang, W. Chen, X. Liang, and R. Pfeifer, "Shape detection algorithm for soft manipulator based on fiber bragg gratings," *IEEE/ASME Transactions on Mechatronics*, vol. 21, no. 6, 2016.

[27] C. Shi, X. Luo, P. Qi, T. Li, S. Song, Z. Najdovski, T. Fukuda, and H. Ren, "Shape sensing techniques for continuum robots in minimally invasive surgery: A survey," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 8, pp. 1665–1678, 2016.

[28] F. Alambeigi, S. A. Pedram, J. L. Speyer, J. Rosen, I. Iordachita, R. H. Taylor, and M. Armand, "Scade: Simultaneous sensor calibration and deformation estimation of fbg-equipped unmodeled continuum manipulators," *IEEE Transactions on Robotics*, vol. 36, no. 1, 2019.

[29] Y. Lu, B. Lu, B. Li, H. Guo, and Y.-H. Liu, "Robust three-dimensional shape sensing for flexible endoscopic surgery using multi-core fbg sensors," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, 2021.

[30] S. Sefati, R. Hegeman, I. Iordachita, R. H. Taylor, and M. Armand, "A dexterous robotic system for autonomous debridement of osteolytic bone lesions in confined spaces: Human cadaver studies," *IEEE Transactions on Robotics*, 2021.

[31] Y. Lu, W. Chen, Z. Chen, J. Zhou, and Y.-h. Liu, "Fbg-based variable-length estimation for shape sensing of extensible soft robotic manipulators," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 01–08.

[32] H. Cao, W. Chen, Y. Lu, J. Huang, J. Zhou, and Y. Liu, "An end-to-end proprioception framework for soft continuum robot," in *2022 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE, 2022, pp. 141–147.

[33] Y. Lu, W. Chen, B. Li, B. Lu, J. Zhou, Z. Chen, and Y.-H. Liu, "A robust graph-based framework for 3-d shape reconstruction of flexible medical instruments using multi-core fbgs," *IEEE Transactions on Medical Robotics and Bionics*, 2023.

[34] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2020.

[35] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[36] B. Yang, W. Chen, Z. Wang, Y. Lu, J. Mao, H. Wang, and Y.-H. Liu, "Adaptive fov control of laparoscopes with programmable composed constraints," *IEEE Transactions on Medical Robotics and Bionics*, vol. 1, no. 4, pp. 206–217, 2019.

[37] J.-J. E. Slotine, W. Li *et al.*, *Applied nonlinear control*. Prentice hall Englewood Cliffs, NJ, 1991, vol. 199, no. 1.

[38] Y. Lu, W. Chen, B. Lu, J. Zhou, Z. Chen, Q. Dou, and Y.-H. Liu, "Robust data-driven 3-d shape servoing of unmodeled continuum robots using fbg sensors in unstructured environments," *arXiv preprint arXiv:2209.05095*, 2022.