Shape-guided Configuration-aware Learning for Endoscopic-image-based Pose Estimation of Flexible Robotic Instruments

Yiyao Ma*[1], Kai Chen*[1], Hon-Sing Tong[2], Ruofeng Wei[1], Yui-Lun Ng[2], Ka-Wai Kwok†[2,3,4], and Qi Dou†[1]

[1] Department of Computer Science and Engineering, The Chinese University of Hong Kong
[2] Agilis Robotics Limited
[3] Department of Mechanical Engineering, The University of Hong Kong
[4] Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong

Abstract. Accurate estimation of both the external orientation and internal bending angle is crucial for understanding a flexible robot state within its environment. However, existing sensor-based methods face limitations in cost, environmental constraints, and integration issues. Conventional image-based methods struggle with the shape complexity of flexible robots. In this paper, we propose a novel shape-guided configuration-aware learning framework for image-based flexible robot pose estimation. Inspired by the recent advances in 2D-3D joint representation learning, we leverage the 3D shape prior of the flexible robot to enhance its image-based shape representation. We first extract the part-level geometry representation of the 3D shape prior, then adapt this representation to the image by querying the image features corresponding to different robot parts. Furthermore, we present an effective mechanism to dynamically deform the shape prior. It aims to mitigate the shape difference between the adopted shape prior and the flexible robot depicted in the image. This more expressive shape guidance boosts the image-based robot representation and can be effectively used for flexible robot pose refinement. Extensive experiments on a general flexible robot designed for endoluminal surgery demonstrate the advantages of our method over a series of keypoint-based, skeleton-based and direct regression-based methods. Project homepage: `https://poseflex.github.io/`.

1 Introduction

Flexible robots made by compliant materials are increasingly prevalent for their flexibility, adaptability, and capacity to maneuver through narrow or irregular space which is otherwise challenging for traditional rigid robots. These characteristics make flexible robots become suited for a variety of applications, such

* indicates equal contributions.
† indicates corresponding authors.

as minimally invasive surgery [6, 33, 55] and dexterous manipulation [9, 26, 40]. Flexible robot pose estimation is an essential task that underpins the practicality and effectiveness of flexible robots. It involves determining the orientation and bending angle of the deformable robot body parts. This detailed depiction of the robot state is crucial for effective manipulation. However, the high deformability of flexible robots allows them to bend and stretch in intricate ways, which makes accurate flexible robot pose estimation extremely challenging.

Typically, existing flexible robot pose estimation approaches heavily depend on the use of advanced sensors. These methods employ various resistive [14,16,34] and capacitive [10, 28, 37] sensors to interpret the shape of the flexible robot by measuring the tension or pressure on the robot surface material. In addition, optical sensors [11, 27, 53] are used to measure changes in the optical fiber's spectrum profile, and magnetic sensors [3, 29] are used to measure changes in magnetic flux. From the recorded changes, the flexible robot deformation and movement are tracked and the pose parameters are estimated. However, these sensors can be costly and their usages are often constrained by specific environmental conditions. It is also not preferred to customize flexible robots to equip with extra sensors, because embedding the sensor within the robot body would reduce its lifespan, while affixing it to the robot surface causes severe occlusion. Moreover, attaching these heavy sensors to robots may introduce potential safety risks, making them unsuitable for critical scenarios such as minimally invasive surgery. Considering above limitations of existing sensor-based approaches, an image-based method for flexible robot pose estimation is highly desirable.

For image-based flexible robot pose estimation, the fundamental problem is how to perceive the robot shape from the image to associate it with the corresponding pose parameters. In this regard, some image-based approaches leverage keypoint detection techniques [4, 12, 17, 18, 46] to identify and localize robot joint positions. These extracted keypoints explicitly divide the robot into local components, which therefore could be used to effectively reconstruct the robot local shapes and recover its relevant pose parameters. However, these keypoint-based methods that have been widely adopted to objects in articulated structures (e.g., rigid robot arm or human body) are not applicable to typical flexible robots. As shown in Fig.1, flexible robots are often made of textureless materials without distinct texture patterns. Their shapes are close to continuum curves whose joints are hardly distinguishable. Detecting keypoints for these flexible robots is still challenging even for recent advanced keypoint detection algorithms [25, 39, 54]. To circumvent this problem, some methods [23, 24, 36] extract robot skeleton instead of keypoints to abstract the robot shape, and recover the pose parameters based on the length and curvature of the extracted skeleton line with numerical computation [36] or rendering-based optimization techniques [23]. However, this kind of holistic representation could not reflect part-level information expressively. Moreover, by explicitly modeling robot shape with keypoints or skeletons, the pose accuracy is unavoidably hindered by the precision of keypoint localization and completeness of the robot skeleton.

Alternatively, recent methods [35,42,50] extract discriminative robot features from the image, and based on which to directly regress the corresponding pose parameters. It is also empirically found in [42] that directly regressing robot pose from the robot feature implicitly learned from the image could achieve higher robustness and accuracy in robot manipulation scenarios. Nevertheless, how to effectively extract shape-aware robot features from the monocular image is the major challenge when this regression-based paradigm is applied to flexible robot pose estimation. As shown in Fig.1, on one hand, the non-rigid deformation nature of flexible robots usually makes the flexible robot exhibit very large shape variations in the image. The robot self-occlusion caused by extreme bending or stretching further increases the difficulties in perceiving the robot shape from the image. On the other hand, in the narrow workspace, the camera is required to be mounted with the robot base and the principal axis of the camera is approximately parallel to the approaching direction of the flexible robot. In this case, the flexible robot image would be dominated by the robot part close to the camera. As a result, it is difficult to precisely distinguish from the image the shape of robot parts away from the camera, which further hinders the estimation accuracy for the bending angle of the flexible robot.

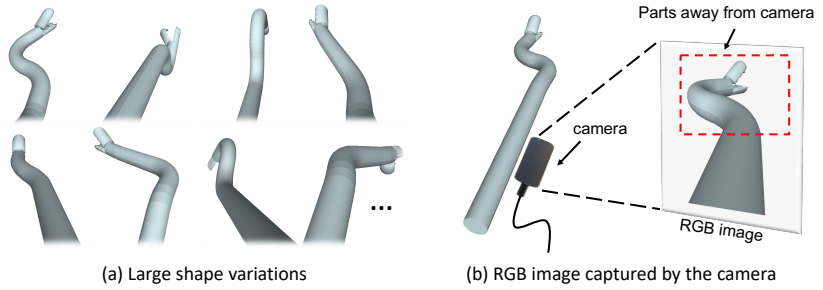

Fig. 1: Typical challenges for image-based flexible robot pose estimation. (a) Flexible robots exhibit textureless surfaces and diverse shapes. (b) Dominance of the robot part closer to the camera in the image can obscure robot parts further away.

In this paper, we introduce a novel shape-guided configuration-aware learning framework for flexible robot pose estimation. Recent advances in multimodal representation learning [1, 43, 44, 47] and 2D-3D object matching/generation [2,13,21] show that 2D-3D joint training can extract highly discriminative shape-aware features from a single 2D image. Inspired by these advancements, we propose to leverage the 3D shape prior of the flexible robot to enhance the robot feature's representation ability extracted from the 2D image, thereby improving the flexible robot pose's regression accuracy. We construct the 3D shape prior based on the flexible robot's pre-defined configuration parameters, assuming the robot adheres to the piecewise constant curvature model [45]. Given this configuration-aware flexible robot shape, we extract its part-level geometry

representation, based on which we enhance the image-based robot representation by querying image features corresponding to different robot parts. However, as the flexible robot's shape deviates from its initial state during motion, this static shape prior becomes less informative. In this regard, we further propose an effective mechanism to dynamically adjust the 3D shape prior based on the initial flexible robot pose. The dynamic shape prior provides more expressive geometry features for the robot, which further boosts the image-based flexible robot representation and can be effectively leveraged for flexible robot pose refinement. Extensive experiments on a general flexible robot prototype developed for endoluminal surgery demonstrate the advantages of our proposed method when compared with conventional keypoint-based, skeleton-based, and direct regression-based methods. We summarize our main contributions as follows:

- We present a shape-guided configuration-aware learning framework for flexible robot pose estimation. It leverages the 3D shape prior and the robot configuration information to enhance image-based flexible robot representation and significantly improve the image-based pose estimation accuracy.
- We introduce a configuration-aware shape deformation mechanism, which uses the initial flexible robot pose to transform the static shape prior to a dynamic one. We demonstrate that the dynamic shape prior is more expressive and can be effectively used for flexible robot pose refinement.
- Extensive evaluations on a flexible robot prototype developed for endoluminal surgery demonstrating the superiority of our method to conventional keypoint-based, skeleton-based, and regression-based approaches for image-based flexible robot pose estimation.

2 Related Works

2.1 Image-based Robot Pose Estimation

The classical approach to computing robot pose involves attaching fiducial markers to pre-defined locations along the robot's kinematic chain, which can be cumbersome and costly [5, 8, 51]. Recent keypoint-based methods regard representative robot joints as virtual markers to eliminate the need for real markers. For instance, DREAM [18] proposes to detect robot joints from image at first, followed by estimating pose using Perspective-n-Point algorithm [19] along with the forward kinematics and camera intrinsics. Afterwards, some recent works [17,39,46] take into account the robot mask or temporal information to improve the detection accuracy. Despite their remarkable performance on rigid robots, these methods encounter challenges when applied to flexible robots due to the continuous, textureless nature of flexible robots and the lack of distinguishable joints. Alternatively, recent methods such as [35, 42, 50] aim to directly regress pose parameters from images. However, effectively extracting shape-aware robot features from monocular images remains an open question for flexible robots. In this work, we propose a shape-guided approach for flexible robot pose estimation, yielding promising results for further research in this domain.

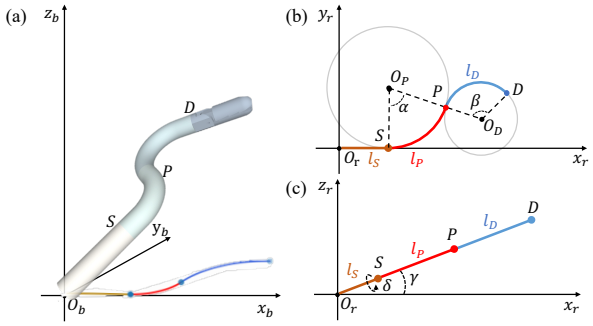

Fig. 2: Illustration of flexible robot pose. (a) Illustration of the robot arm in image coordinate system $x_b o_b r_b$. (b) Illustration of internal bending angles α, β in robot space $x_r o_r y_r$ plane. (c) Illustration of the yaw γ and roll δ in $x_r o_r z_r$ plane.

2.2 Flexible Robot Pose Estimation

Existing methods for flexible robot pose estimation mainly rely on attaching additional sensors to robots. These sensors discretize the continuous deformation of flexible robots into a set number of nodes, and then collect parameters such as force, torque, or spectral frequency response at these nodes for subsequent pose computation. For instance, resistive sensors [14, 16, 34] detect robot state through monitoring changes in their resistance due to bending or force and pressure. Capacitive sensors [10,28,37] achieve these in their generated electric fields. Optical sensors [11, 27, 53] leverage the changes in the spectrum profile of the reflected light within the optical fiber to compute robot deformation state. Magnetic sensors [3,29] measure magnetic flux changes arising from the displacement of a permanent magnet to detect robot movement. Recently, some works have explored leveraging vision information to complement sensor feedback for more robust pose estimation [22,36,38,41]. These approaches optimize sensor feedback using information derived from image features [7]. However, the deployment of most sensors often face limitations in cost, environment constraints, and integration issues. Different from them, our method relies solely on image data, and can be seamlessly deployed in various environments.

3 Method

3.1 Overview

Given a monocular RGB image $\mathcal{I}$ for the flexible robot arm, our objective is to estimate its corresponding pose parameters. Based on the widely used piecewise constant curvature model [45], the flexible robot arm could be divided into N_{part} segments. Taking a flexible robot arm with three segments (excluding the end effector) as an example, Fig. 2 illustrates the pose parameters that are required

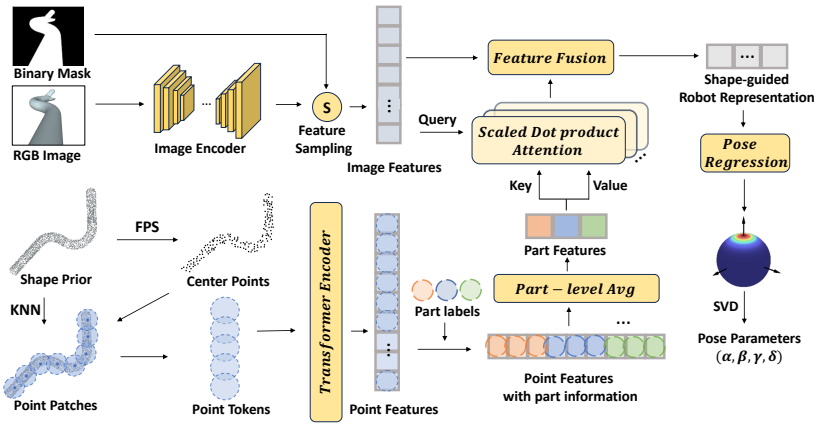

Fig. 3: Framework overview of the proposed shape-guided configuration-aware learning method for image-based flexible robot pose estimation.

to be estimated from the image, among which (α, β) denote the arc angles of SP and PD respectively, and (γ, δ) are the yaw and roll of the robot arm.

Precisely regressing both the global orientation and the internal bending angles for the flexible robot arm requires a highly expressive robot feature from the image. In this regard, we propose to leverage a 3D shape prior that corresponds to the flexible robot arm at home position to enhance the shape representation ability of robot features extracted from the 2D image. It is achieved via a shape-guided configuration-aware learning approach presented in Section 3.2. As the flexible robot's movement, its shape gradually deviates from the one at home position, which makes the guidance from the shape prior becomes less effective and limits the final pose accuracy. To eliminate this problem, based on the initial flexible robot pose from Section 3.2, we further present a method in Section 3.3 to dynamically adjust the 3D shape prior for the flexible robot arm and use this more expressive and precise shape prior for flexible robot pose refinement. To facilitate regression-based flexible robot pose estimation/refinement, we resort to the matrix Fisher distribution to parameterize the flexible robot pose in Section 3.4. It can not only lead to a continuous learning space for pose regression, but also recover the pose uncertainty, which potentially could be leveraged for vision-based flexible robot control and manipulation.

3.2 Pose Estimation with Configuration-aware Shape Guidance

A straightforward approach for image-based pose estimation involves extracting 2D features from images and directly regressing the pose parameters. However, 2D features are insufficient for conveying the detailed information of various robot parts, leading to suboptimal performance in flexible robot pose estimation.

To address this limitation, we propose utilizing a 3D shape prior of the flexible robot to enhance the image features at the part level. Fig. 3 shows the overview

of our proposed method. In detail, we first parse the input RGB image with an image encoder and sample image features within the area of binary mask. Simultaneously, we consider the 3D shape prior as a point cloud $X \in \mathbb{R}^{N\times 3}$ and utilize the Farthest Point Sampling (FPS) to uniformly sample M center points P_c from it. Then, we employ the K-Nearest Neighbors (KNN) algorithm to select K nearest points for each center point, resulting in point patches P_p:

$$P_c = FPS(X),\ P_c \in \mathbb{R}^{M\times 3}, \tag{1}$$

$$P_p = KNN(X, P_c),\ P_p \in \mathbb{R}^{M\times K\times 3}. \tag{2}$$

Subsequently, we utilize a PointNet architecture [31], which comprises multiple MLPs and max pooling layers, to embed each patch in P_p into point tokens P_t. Additionally, to preserve location information, the coordinates of each P_c are encoded using an MLP to generate positional embeddings (PE).

Afterwards, we utilize Transformer blocks to encode P_t along with their PE, and obtain encoded point features f_p. The Transformer blocks are effective in capturing the correlation among the information within the patches, facilitating both local and global feature representation. Since the configuration-embedded shape prior has part label part_j for each point p_i, we take the average of the f_p within same part as its part feature:

$$f_{\text{part}_j} = \frac{1}{|s(\text{part}_j)|} \sum_{i\in s(\text{part}_j)} f_{p_i}, \tag{3}$$

where $s(\text{part}_j)$ represents the indices of points within part_j, and $|\cdot|$ is cardinality. This process yields a feature representation that captures highly representative part-level information. After that, we leverage the strong expressive ability of multi-head attention to model the high-level similarity between the image features f_I and the concatenated part features f_{part}, and based on the similarity to adapt shape prior information to image. In specific, we query the f_I with the f_{part} as keys and values to obtain a part-aware image descriptor f_d:

$$f_d^{(h)} = \sigma\left(\frac{f_I^{(h)} W_Q^{(h)} (f_{\text{part}}^{(h)} W_K^{(h)})^T}{\sqrt{d}}\right) f_{\text{part}}^{(h)} W_V^{(h)}, \tag{4}$$

where $h = 1, 2, ...H$ represents the index of multi-head attention. $W_Q^{(h)}$, $W_K^{(h)}$, and $W_V^{(h)}$ are learnable projection matrices for query, key and value, respectively. $\sigma(\cdot)$ is a softmax function to normalize the similarity value by row.

In each head, we compute the similarity between $f_I^{(h)}$ and $f_{\text{part}}^{(h)}$ in the projected embedding space, and multiply this similarity with $f_{\text{part}}^{(h)}$ to get the semantic feature. By concatenating the H attention blocks together, we can obtain the comprehensive transferred descriptor:

$$f_d = \text{concat}(f_d^{(1)}, f_d^{(2)}, ..., f_d^{(H)}). \tag{5}$$

This descriptor contains the distinctive characteristics of each part, enhancing the part-level representation of the image features. Finally, the original f_I fused

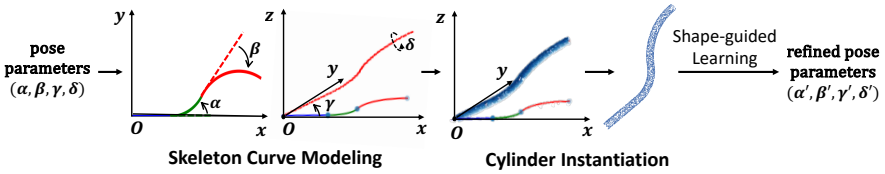

Fig. 4: Illustration of the pose refinement scheme. Based on the initial flexible robot pose, we deform the robot shape prior via skeleton curve modeling and cylinder instantiation.

with the f_d are subsequently regressed to pose parameters. This fusion process ensures that both the global image features and the detailed part-level information contribute to accurate pose estimation.

3.3 Pose Refinement with Configuration-aware Shape Deformation

When the flexible robot undergoes manipulation, its shape deviates from its initial home position, causing the static shape prior to become less representative of the current robot shape. To address this, we propose an effective mechanism, as shown in Fig. 4, to dynamically deform the 3D shape prior based on the pose prediction, ensuring its alignment with the current robot shape. Specifically, we introduce a function Λ that generates the 3D shape $\mathcal{G}$ using the provided pose and configuration parameters:

$$\mathcal{G} = \Lambda(\hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{\delta}, \mathbf{w}), \tag{6}$$

where $\hat{\cdot}$ represents the predicted pose parameters, and $\mathbf{w}$ denotes the pre-defined flexible robot configurations, including parameters such as part length l_{part_j}, number of control points N_c, robot radius r, etc.

Without loss of generality, the flexible robot can be approximated as a union of cylinders, with each cylinder centered along the 3D skeleton curve [20,23]. This process involves initially determining the skeleton curve using a set of control points $\{\mathbf{c_i}|\mathbf{c_i} \in \mathbb{R}^3;\ i = 1, 2, ..N_c\}$ and subsequently sampling a cylinder between each pair of points $(\mathbf{c_{i-1}}, \mathbf{c_i})$. In the first step, to explicitly model the skeleton curve for part_j, we parameterize each control point in the base coordinate system:

$$\mathbf{c_i^{b_j}} = (x_i^{b_j}, y_i^{b_j}, z_i^{b_j}) = (\frac{l_{\text{part}_j}}{\theta^j} \sin\theta_i^j, \frac{2l_{\text{part}_j}}{\theta^j} \sin^2(\frac{\theta_i^j}{2}), \frac{l_{\text{part}_j}}{\theta^j} \sin\theta_i^j \tan\gamma), \tag{7}$$

where i denotes the index of points within part_j, b_j is the base coordinate system specific to each part_j, l_{part_j} represents the pre-defined constant length for part_j, and $\theta^j \in \{\alpha, \beta\}$ represents the internal bending angles for part_j. Each θ_i^j is sampled from a uniform distribution $U(0, \theta^j)$. Afterwards, we rotate and transform

these control points to robot coordinate based on the position of part_{j-1}:

$$\mathbf{c_i^{r_j}} = R_{\text{part}_{j-1}} \cdot (x_i^{b_j}, y_i^{b_j}, z_i^{b_j}) + T_{\text{part}_{j-1}}, \tag{8}$$

where $R_{\text{part}_{j-1}}$ and $T_{\text{part}_{j-1}}$ represent the rotation matrix and translation vector from the origin to $\overline{c}_{j-1}$, which is the last control point of part_{j-1}:

$$R_{\text{part}_{j-1}} = \begin{bmatrix} \cos\theta^{j-1} & -\sin\theta^{j-1} & 0 \\ \sin\theta^{j-1} & \cos\theta^{j-1} & 0 \\ 0 & 0 & 1 \end{bmatrix}, T_{\text{part}_{j-1}} = (\overline{x}_{j-1}, \overline{y}_{j-1}, \overline{z}_{j-1})^T. \tag{9}$$

Subsequently, each point on skeleton will be rotated by ($\mathbf{e}$, δ) using Rodrigues' rotation formula, where $\mathbf{e} = (cos\gamma, 0, sin\gamma)$ is the axis of robot's root direction:

$$\mathbf{c_i} = (\cos\delta)\mathbf{c_i^{r_j}} + (\sin\delta)(\mathbf{e} \times \mathbf{c_i^{r_j}}) + (1 - \cos\delta)(\mathbf{e} \cdot \mathbf{c_i^{r_j}})\mathbf{e}. \tag{10}$$

Finally, we can randomly sample N_{cyl} points from the cylinder composed of two adjacent points ($\mathbf{c_i}$, $\mathbf{c_{i-1}}$), and obtain $\mathcal{G}$. With the deformation function $\Lambda(\cdot)$, we can derive a unique shape prior for each robot pose. These dynamic shape priors offer more comprehensive geometry features for the robot, thereby enhancing its image-based representation and improving the accuracy of pose prediction. Step-by-step derivation will be shown in supplemental materials.

3.4 Pose Regression with Probabilistic Representation

Given the enhanced flexible robot feature from the 2D image, we regress the corresponding pose parameters based on a probabilistic model that is widely used for rotation regression [15,30,49]. Specifically, for each degree of freedom of the flexible robot pose, we represent it with an independent matrix Fisher distribution $\mathcal{M}(\mathbf{R}; \mathbf{A})$, which defines a probability density function over the rotation matrix $\mathbf{R}$ in the form of:

$$p(\mathbf{R}) = \mathcal{M}(\mathbf{R}; \mathbf{A}) = \frac{1}{n(\mathbf{A})}\exp(\text{tr}(\mathbf{A}^T\mathbf{R})), \tag{11}$$

where $\mathbf{A}$ is the distribution parameters, and $n(\mathbf{A})$ is a normalizing constant. We consider each pose parameter of the flexible robot arm separately with an independent matrix Fisher distribution. In this way, the adopted probabilistic representation can be easily scaled to flexible robot arms with different degrees of freedom. Note that the adopted matrix Fisher distribution over-parameterizes the pose parameter. It is a continuous representation that offers significant advantages in network regression and surpasses the conventional Euler-based or quaternion-based representations, which are prone to encountering disconnected local minimal problems. Moreover, from the regressed distribution parameters $\mathbf{A}$, we could not only recover the pose value but also the corresponding uncertainty which indicates the confidence of the prediction.

During training, instead of minimizing the mean squared error or the L_1/L_2 distance between the predicted and ground-truth pose parameters, we leverage

the negative log likelihood loss to learn the probabilistic model. The loss function is defined as:

$$L(\mathbf{Y}, \mathbf{A}) = -\log(\mathcal{M}(\mathbf{Y}; \mathbf{A})), \tag{12}$$

where $\mathbf{Y}$ denotes the ground-truth rotation matrix. During inference, given the regressed distribution matrix $\mathbf{A}$, the mode and dispersion of the distribution can be obtained by computing singular value decomposition (SVD) on $\mathbf{A}$, which is $\mathbf{A} = \mathbf{USV}^T$. The pose value could be derived from $\mathbf{U}$ and $\mathbf{V}$, and the singular values in $\mathbf{S}$ reveal the concentration of the distribution, indicating the reliability of the predicted pose.

4 Experiments

In this section, we will answer the following key questions through our conducted experiments. (1) Does the proposed method surpass existing image-based approaches, including keypoint-based, skeleton-based, and regression-based methods? (2) How effective is the proposed configuration-aware shape guidance in enhancing the accuracy of flexible robot pose estimation? (3) Given the shape prior of the flexible robot arm, does the proposed method, which utilizes shape guidance through 3D-2D joint training, outperform other alternative designs? (4) Is the proposed method generalizable and can be applied to flexible robots with different configurations? (5) Is the proposed method robust enough under various challenging surgical environments?

4.1 Experimental Settings

Datasets and Evaluation Metrics. Since there are no publicly available benchmark datasets for flexible robot pose estimation, to compare different flexible robot pose estimation approaches, we utilize a flexible robotic system developed by Agilis Robotics which is designed to perform minimally invasive surgery in the bladder and gastrointestinal (GI) tract through natural orifices. Without loss of generality, we used the flexible robot prototype in the GI scenario and collected 12473 image-pose pairs in ex vivo environments. The ground-truth flexible robot pose is obtained from the robot motor signal and is post-processed with careful manual rectification to ensure its reliability. In order to comprehensively evaluate all methods, we used different degrees of freedom to control the flexible robot arm during data collection. Specifically, the endoscopic image dataset could be divided into 3 groups G_1, G_2, and G_3. In G_1, only δ changes. In G_2, both γ and δ change. In G_3, all pose parameters α, β, γ, δ change simultaneously. We divided the dataset into training, validation, and testing sets. After that, we used 9473 images for model training and validation, and used 3000 images for testing. We evaluated the model on testing images of G_1, G_2, and G_3, respectively. Please refer to the supplementary materials for more implementation details.

Following the prior work of [49], we reported both average and median angular errors for each of the predicted pose parameters. Meanwhile, we reported

Table 1: Comparison of our methods with the state-of-the-art methods on G_1. The initial pose of PoseRefine. comes from the results of PoseEst.

Metrics		KP [18]	SKL [36]	DR [42]	SimPS [35]	PoseEst.	PoseRefine.
δ	Mean (°) ↓	49.34	41.86	23.18	7.14	5.86	**4.08**
	Med. (°) ↓	41.16	22.77	18.93	4.22	5.07	**2.82**
	Acc5° (%) ↑	9.0	14.0	19.8	57.1	49.3	**70.1**
	Acc10° (%) ↑	17.5	30.1	34.4	79.8	86.0	**93.9**

Table 2: Comparison of our methods with the state-of-the-art methods on G_2.

Metrics		KP [18]	SKL [36]	DR [42]	SimPS [35]	PoseEst.	PoseRefine.
γ	Mean (°) ↓	17.19	16.08	2.60	**2.47**	2.93	2.80
	Med. (°) ↓	14,68	19.97	**1.77**	1.79	2.00	2.37
δ	Mean (°) ↓	44.41	48.67	22.73	7.29	4.77	**3.47**
	Med. (°) ↓	34.45	29.58	18.23	4.82	4.30	**2.51**
	Acc5° (%) ↑	4.1	12.8	11.4	51.4	57.9	**74.4**
	Acc10° (%) ↑	10.1	26.7	32.8	77.3	92.7	**97.1**

the prediction accuracy with respect to 5° and 10°, which measures the ratio of predictions whose prediction error is smaller than 5° or 10°.

Competing Methods. We compared our method with four competing methods. These competing methods could be categorized into three groups:

- Keypoint-based method KP [18]: We pre-defined six keypoints on the flexible robot body. Then, we used the recent stat-of-the-art method [48] for image-based keypoint localization. Based on the extracted keypoints, we parsed the flexible robot shape and computed the corresponding pose parameters.
- Skeleton-based method SKL [36]: We extracted the robot skeleton from its binary mask using the fast skeletonization algorithm [52]. Then, the skeleton was robustly fitted with the Bezier curve for computing the pose parameters.
- Regression-based: Furthermore, we compared our method with two recent state-of-the-art regression-based approaches DR [42] and SimPS [35].

For competing methods KP and SKL, we lifted the 2D keypoint and 2D skeleton to 3D with the depth map predicted by the depth prediction Transformer [32]. It aims to recover the flexible robot shape in the 3D space, which is necessary for computing the corresponding pose parameters. Please refer to the supplementary material for more implementation details of the competing methods.

4.2 Main Results

Table 1-3 present the comparative results with four competing methods on G_1, G_2, and G_3 respectively. For our proposed method, we report both the pose estimation (PoseEst.) result with configuration-aware shape guidance (Section 3.2) and the pose refinement (PoseRefine.) result after configuration-aware shape deformation (Section 3.3). The keypoint-based baseline KP achieves poor performance in all three evaluation scenarios. It is because keypoints of the flexible

Table 3: Comparison of our methods with the state-of-the-art methods on G_3.

Metrics		KP [18]	SKL [36]	DR [42]	SimPS [35]	PoseEst.	PoseRefine.
α	Mean (°) ↓	14.28	14.85	11.78	9.94	8.76	**7.08**
	Med. (°) ↓	10.96	13.93	7.97	5.21	4.78	**4.16**
β	Mean (°) ↓	20.92	19.59	4.83	5.65	4.79	**4.66**
	Med. (°) ↓	18.86	18.59	3.88	4.92	3.90	**3.28**
γ	Mean (°) ↓	26.20	29.11	2.33	2.30	2.05	**1.62**
	Med. (°) ↓	28.79	31.69	1.80	1.76	1.58	**1.19**
δ	Mean (°) ↓	45.39	56.60	27.49	17.96	14.58	**11.68**
	Med. (°) ↓	21.75	27.62	22.82	10.64	5.94	**4.66**
	Acc5° (%) ↑	11.9	13.4	13.9	27.3	45.9	**53.0**
	Acc10° (%) ↑	27.4	23.8	15.0	48.2	59.5	**71.2**

robot arm are hard to distinguish from the image. The large keypoint localization errors result in poor accuracy. Alternatively, when we resort to the robot skeleton rather than keypoints to parse the flexible robot shape, the corresponding pose accuracy gets obvious improvement (as shown in Table 1-2) when compared with KP. However, extracting a complete robot skeleton from the image remains challenging, particularly when dealing with flexible robots that have a high degree of freedom during motion. As shown in Fig. 5, it is difficult to obtain an accurate flexible robot pose from the extracted robot skeleton. Similar to the observation in [42], we also found that two regression-based baselines consistently outperform the conventional keypoint- and skeleton-based methods. Nevertheless, lacking an effective mechanism to model the flexible robot shape variation, these two methods still fall short in flexible robot pose estimation. In contrast, our method leverages the informative shape guidance to enhance the image-based flexible robot shape representation, which significantly improves the accuracy of regression-based flexible robot pose estimation. Moreover, by deforming the flexible robot shape with the initial pose parameters, our method manages to further improve the pose accuracy with more precise and expressive shape guidance. Please refer to the supplementary for more qualitative results.

4.3 Ablation Study

Effectiveness of shape guidance. In this experiment, we studied the effectiveness of the proposed shape guidance for flexible robot pose estimation. First, we compared the performance with and without using the shape guidance. The result is depicted in Fig. 6(a). The leverage of the shape guidance can reduce the prediction error for most pose parameters. In addition, the model with shape guidance consistently achieves higher accuracy with respect to different error thresholds. Second, we simplify the adopted shape guidance by omitting the step of using the robot configurations to aggregate part-level robot representations. Fig. 6(b) presents the comparative results. Removing the robot configuration information from the shape guidance would consistently degrade the pose accuracy. These results demonstrate the advantage of the proposed configuration-aware shape guidance for flexible robot pose estimation.

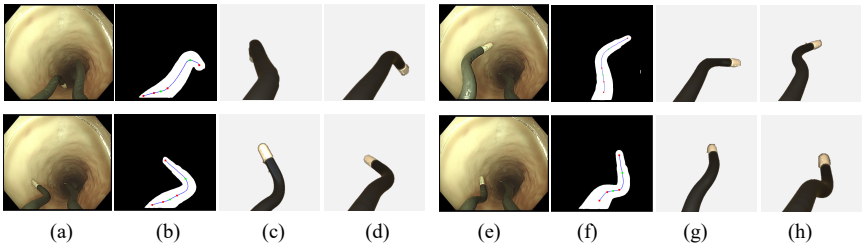

Fig. 5: Qualitative comparison with skeleton-based baseline. (a)(e) are input RGB images; (b)(f) depict the extracted flexible robot skeleton; (c)(f) exhibits pose estimation results of SKL [36]; (d)(h) present our results.

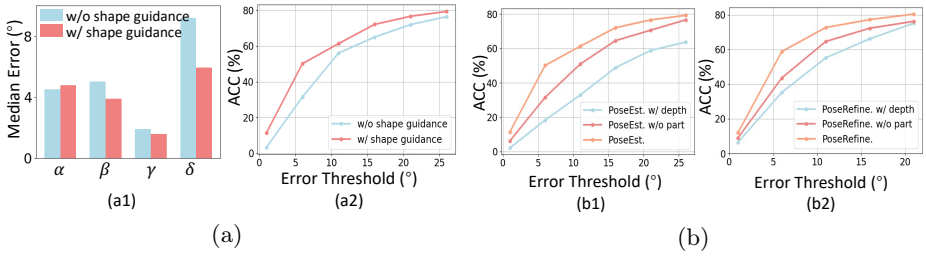

Fig. 6: Ablation on shape guidance (a) and comparison with depth-based counterparts (b). (a1) The median error of the model w/ or w/o shape guidance. (a2) The accuracy of δ under different error threshold. (b1) The accuracy of PoseEst. w/ or w/o depth. (b2) The accuracy of PoseRefine. w/ or w/o depth. For a fair comparison, we conduct our method w/o part-level feature integration.

Comparison with depth-based counterparts. In this experiment, we compared our method with the depth-based counterpart. Specifically, we recover the depth map from the image with a pre-trained depth prediction Transformer [32], lift the flexible robot to 3D, and extract the geometry feature from the robot point cloud for pose estimation. We evaluate the above counterpart in both pose estimation and pose refinement stages. As shown in Fig. 6(b), in both stages, the depth-based counterpart is inferior to ours. We suppose that it is primarily attributed to the depth noise. Although the pre-trained model can predict the depth map with rich boundaries and details [†], the recovered flexible robot shape still suffer severe shape distortions, which limit the pose accuracy. In contrast, our method is suited to the scenario where the depth data is scarce. Without relying on the depth data, our method achieves the highest pose accuracy.

Generalizability for robots with diverse configurations. To assess the generalizability of our method to different flexible robots, we redesigned our robot prototype based on the definition of robot configurations. We made modifications on the robot arm by varying the arm thickness (Thick.), arm length (Len.),

[†] Please refer to the supplementary material to check the depth map.

Table 4: Quantitative evaluation on flexible robots with diverse configurations and results under different environmental changes.

Methods	Diverse robot configurations						Diverse environmental changes					
	Thick.±20%		Len.±20%		Num.+1		Lighting		Occlusion		Scope Rot.	
	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.
SimPS [35]	12.78	8.58	12.56	8.42	19.44	12.39	15.20	11.55	23.26	12.51	19.33	10.16
PoseEst.	**8.24**	**5.70**	**7.24**	**5.29**	**12.88**	**9.42**	**10.49**	**9.43**	**18.68**	**11.01**	**12.67**	**10.14**

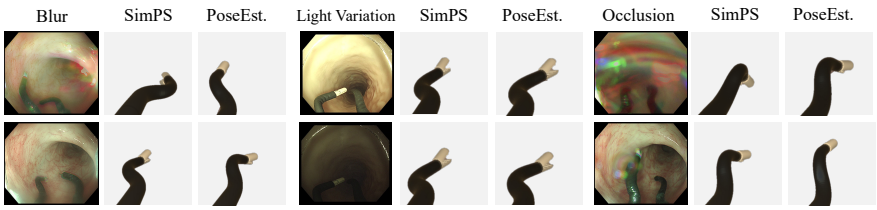

Fig. 7: Qualitative results under environmental changes.

and the number of segments (Num.). Experiment results in Tab. 4 demonstrate that our methods can smoothly adapt to robots with diverse configurations and surpass the most competitive baseline in the main result.

Robustness to challenging environments. To evaluate the robustness of the proposed methods, we further conducted experiments under challenging visual conditions that typically present in surgery. These conditions encompassed too bright or dark lighting conditions (Lighting), visual occlusions caused by flushing water and bubbles (Occlusion), and image blur caused by robot motion (Scope Rot.). As shown in Fig. 7 and Tab. 4, with the help of 3D shape guidance, our method keeps commendable performance in these challenging scenarios.

5 Conclusion

In this paper, we present an image-based approach for flexible robot pose estimation. We study to leverage the 3D shape prior and the configuration information of the flexible robot arm to improve the image-based shape representation for the soft arm, which significantly improves the flexible robot pose estimation accuracy. Moreover, by deforming the shape prior based on the initial flexible robot pose, we manage to further improve the image-based flexible robot representation with more expressive shape guidance for flexible robot pose refinement. Extensive experiments on surgical flexible robots demonstrate the superiority of our method over existing approaches. Future work will focus on exploring an effective solution of using cost-effective synthetic data for model training, and integrating the proposed methods into the flexible robot platform for vision-based flexible robot control and manipulation.

Acknowledgements. This work was supported in part by Hong Kong Innovation and Technology Commission under Project No. PRP/026/22FX, in part by Agilis Robotics and its subsidiaries, Agilis Robotics Limited and Agilis Robotics Limited (Guangzhou), and in part by a grant from the NSFC/RGC Joint Research Scheme sponsored by the Research Grants Council of the Hong Kong Special Administrative Region, China and the National Natural Science Foundation of China (Project No. N_CUHK410/23).

References

1. Afham, M., Dissanayake, I., Dissanayake, D., Dharmasiri, A., Thilakarathna, K., Rodrigo, R.: Crosspoint: Self-supervised cross-modal contrastive learning for 3d point cloud understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9902–9912 (2022)
2. Arsomngern, P., Nutanong, S., Suwajanakorn, S.: Learning geometric-aware properties in 2d representation using lightweight cad models, or zero real 3d pairs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21371–21381 (2023)
3. Baaij, T., Holkenborg, M.K., Stölzle, M., van der Tuin, D., Naaktgeboren, J., Babuška, R., Della Santina, C.: Learning 3d shape proprioception for continuum soft robots with multiple magnetic sensors. Soft Matter **19**(1), 44–56 (2023)
4. Bilić, I., Marić, F., Marković, I., Petrović, I.: A distance-geometric method for recovering robot joint angles from an rgb image. arXiv preprint arXiv:2301.02051 (2023)
5. Cartucho, J., Wang, C., Huang, B., S. Elson, D., Darzi, A., Giannarou, S.: An enhanced marker pattern that achieves improved accuracy in surgical tool tracking. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization **10**(4), 400–408 (2022)
6. Chautems, C., Tonazzini, A., Boehler, Q., Jeong, S.H., Floreano, D., Nelson, B.J.: Magnetic continuum device with variable stiffness for minimally invasive surgery. Advanced Intelligent Systems **2**(6), 1900086 (2020)
7. Chin, K., Hellebrekers, T., Majidi, C.: Machine learning for soft robotic sensing and control. Advanced Intelligent Systems **2**(6), 1900171 (2020)
8. Garrido-Jurado, S., Muñoz-Salinas, R., Madrid-Cuevas, F.J., Marín-Jiménez, M.J.: Automatic generation and detection of highly reliable fiducial markers under occlusion. Pattern Recognition **47**(6), 2280–2292 (2014)
9. Gu, G., Zhang, N., Xu, H., Lin, S., Yu, Y., Chai, G., Ge, L., Yang, H., Shao, Q., Sheng, X., et al.: A soft neuroprosthetic hand providing simultaneous myoelectric control and tactile feedback. Nature biomedical engineering **7**(4), 589–598 (2023)
10. Ha, K.H., Zhang, W., Jang, H., Kang, S., Wang, L., Tan, P., Hwang, H., Lu, N.: Highly sensitive capacitive pressure sensors over a wide pressure range enabled by the hybrid responses of a highly porous nanocomposite. Advanced Materials **33**(48), 2103320 (2021)
11. He, Y., Gao, L., Bai, Y., Zhu, H., Sun, G., Zhu, L., Xu, H.: Stretchable optical fibre sensor for soft surgical robot shape reconstruction. Optica Applicata **51**(4) (2021)
12. Heindl, C., Zambal, S., Ponitz, T., Pichler, A., Scharinger, J.: 3d robot pose estimation from 2d images. arXiv preprint arXiv:1902.04987 (2019)

13. Jing, L., Vahdani, E., Tan, J., Tian, Y.: Cross-modal center loss for 3d cross-modal retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3142–3151 (2021)
14. Katzschmann, R.K., Thieffry, M., Goury, O., Kruszewski, A., Guerra, T.M., Duriez, C., Rus, D.: Dynamically closed-loop controlled soft robotic arm using a reduced order finite element model with state observer. In: 2019 2nd IEEE international conference on soft robotics (RoboSoft). pp. 717–724. IEEE (2019)
15. Khatri, C., Mardia, K.V.: The von mises–fisher matrix distribution in orientation statistics. Journal of the Royal Statistical Society Series B: Statistical Methodology **39**(1), 95–106 (1977)
16. Kim, S.Y., Choo, Y., Bilodeau, R.A., Yuen, M.C., Kaufman, G., Shah, D.S., Osuji, C.O., Kramer-Bottiglio, R.: Sustainable manufacturing of sensors onto soft systems using self-coagulating conductive pickering emulsions. Science robotics **5**(39), eaay3604 (2020)
17. Lambrecht, J., Grosenick, P., Meusel, M.: Optimizing keypoint-based single-shot camera-to-robot pose estimation through shape segmentation. In: 2021 IEEE International Conference on Robotics and Automation (ICRA). pp. 13843–13849. IEEE (2021)
18. Lee, T.E., Tremblay, J., To, T., Cheng, J., Mosier, T., Kroemer, O., Fox, D., Birchfield, S.: Camera-to-robot pose estimation from a single image. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). pp. 9426–9432. IEEE (2020)
19. Lepetit, V., Moreno-Noguer, F., Fua, P.: Epnp: An accurate o (n) solution to the pnp problem. International journal of computer vision **81**(2), 155–166 (2009)
20. Li, S., Hao, G.: Current trends and prospects in compliant continuum robots: A survey. In: Actuators. vol. 10, p. 145. MDPI (2021)
21. Lin, M.X., Yang, J., Wang, H., Lai, Y.K., Jia, R., Zhao, B., Gao, L.: Single image 3d shape retrieval via cross-modal instance and category contrastive learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11405–11415 (2021)
22. Loo, J.Y., Ding, Z.Y., Baskaran, V.M., Nurzaman, S.G., Tan, C.P.: Robust multimodal indirect sensing for soft robots via neural network-aided filter-based estimation. Soft Robotics **9**(3), 591–612 (2022)
23. Lu, J., Liu, F., Girerd, C., Yip, M.: Image-based pose estimation and shape reconstruction for robot manipulators and soft, continuum robots via differentiable rendering. In: ICRA 2023-IEEE International Conference on Robotics and Automation (2023)
24. Lu, J., Richter, F., Lin, S., Yip, M.C.: Tracking snake-like robots in the wild using only a single camera. arXiv preprint arXiv:2309.15700 (2023)
25. Lu, J., Richter, F., Yip, M.C.: Pose estimation for robot manipulators via keypoint optimization and sim-to-real transfer. IEEE Robotics and Automation Letters **7**(2), 4622–4629 (2022)
26. Mair, L.O., Adam, G., Chowdhury, S., Davis, A., Arifin, D.R., Vassoler, F.M., Engelhard, H.H., Li, J., Tang, X., Weinberg, I.N., et al.: Soft capsule magnetic millirobots for region-specific drug delivery in the central nervous system. Frontiers in Robotics and AI **8**, 702566 (2021)
27. Monet, F., Sefati, S., Lorre, P., Poiffaut, A., Kadoury, S., Armand, M., Iordachita, I., Kashyap, R.: High-resolution optical fiber shape sensing of continuum robots: A comparative study. In: 2020 IEEE international conference on robotics and automation (ICRA). pp. 8877–8883. IEEE (2020)

28. Navarro, S.E., Nagels, S., Alagi, H., Faller, L.M., Goury, O., Morales-Bieze, T., Zangl, H., Hein, B., Ramakers, R., Deferme, W., et al.: A model-based sensor fusion approach for force and shape estimation in soft robotics. IEEE Robotics and Automation Letters **5**(4), 5621–5628 (2020)
29. Ozel, S., Skorina, E.H., Luo, M., Tao, W., Chen, F., Pan, Y., Onal, C.D.: A composite soft bending actuation module with integrated curvature sensing. In: 2016 IEEE International Conference on Robotics and Automation (ICRA). pp. 4963–4968. IEEE (2016)
30. Prentice, M.J.: Orientation statistics without parametric assumptions. Journal of the Royal Statistical Society Series B: Statistical Methodology **48**(2), 214–222 (1986)
31. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 652–660 (2017)
32. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12179–12188 (2021)
33. Ranzani, T., Cianchetti, M., Gerboni, G., De Falco, I., Menciassi, A.: A soft modular manipulator for minimally invasive surgery: design and characterization of a single module. IEEE Transactions on Robotics **32**(1), 187–200 (2016)
34. Shih, B., Christianson, C., Gillespie, K., Lee, S., Mayeda, J., Huo, Z., Tolley, M.T.: Design considerations for 3d printed, soft, multimaterial resistive sensors for soft robotics. Frontiers in Robotics and AI **6**, 30 (2019)
35. Souipas, S., Nguyen, A., Laws, S.G., Davies, B.L., y Baena, F.R.: Simps-net: Simultaneous pose & segmentation network of surgical tools. IEEE Transactions on Medical Robotics and Bionics (2023)
36. Tanaka, K., Minami, Y., Tokudome, Y., Inoue, K., Kuniyoshi, Y., Nakajima, K.: Continuum-body-pose estimation from partial sensor information using recurrent neural networks. IEEE Robotics and Automation Letters **7**(4), 11244–11251 (2022)
37. Teyssier, M., Parilusyan, B., Roudaut, A., Steimle, J.: Human-like artificial skin sensor for physical human-robot interaction. In: 2021 IEEE International Conference on Robotics and Automation (ICRA). pp. 3626–3633. IEEE (2021)
38. Thuruthel, T.G., Shih, B., Laschi, C., Tolley, M.T.: Soft robot perception using embedded soft sensors and recurrent neural networks. Science Robotics **4**(26), eaav1488 (2019)
39. Tian, Y., Zhang, J., Yin, Z., Dong, H.: Robot structure prior guided temporal attention for camera-to-robot pose estimation from image sequence. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8917–8926 (2023)
40. Toshimitsu, Y., Wong, K.W., Buchner, T., Katzschmann, R.: Sopra: Fabrication & dynamical modeling of a scalable soft continuum robotic arm with integrated proprioceptive sensing. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 653–660. IEEE (2021)
41. Truby, R.L., Della Santina, C., Rus, D.: Distributed proprioception of 3d configuration in soft, sensorized robots via deep learning. IEEE Robotics and Automation Letters **5**(2), 3299–3306 (2020)
42. Valassakis, E., Dreczkowski, K., Johns, E.: Learning eye-in-hand camera calibration from a single image. In: Conference on Robot Learning. pp. 1336–1346. PMLR (2022)

43. Wang, Y., Chen, X., Cao, L., Huang, W., Sun, F., Wang, Y.: Multimodal token fusion for vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12186–12195 (2022)
44. Wang, Y., Ye, T., Cao, L., Huang, W., Sun, F., He, F., Tao, D.: Bridged transformer for vision and point cloud 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12114–12123 (2022)
45. Webster III, R.J., Jones, B.A.: Design and kinematic modeling of constant curvature continuum robots: A review. The International Journal of Robotics Research **29**(13), 1661–1683 (2010)
46. Xu, H., Runciman, M., Cartucho, J., Xu, C., Giannarou, S.: Graph-based pose estimation of texture-less surgical tools for autonomous robot control. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 2731–2737. IEEE (2023)
47. Xu, P., Zhu, X., Clifton, D.A.: Multimodal learning with transformers: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
48. Yang, J., Gao, M., Li, Z., Gao, S., Wang, F., Zheng, F.: Track anything: Segment anything meets videos. arXiv preprint arXiv:2304.11968 (2023)
49. Yin, Y., Cai, Y., Wang, H., Chen, B.: Fishermatch: Semi-supervised rotation regression via entropy-based filtering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11164–11173 (2022)
50. Yoshimura, M., Marinho, M.M., Harada, K., Mitsuishi, M.: Single-shot pose estimation of surgical robot instruments' shafts from monocular endoscopic images. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). pp. 9960–9966. IEEE (2020)
51. Zhang, L., Ye, M., Chan, P.L., Yang, G.Z.: Real-time surgical tool tracking and pose estimation using a hybrid cylindrical marker. International journal of computer assisted radiology and surgery **12**, 921–930 (2017)
52. Zhang, T.Y., Suen, C.Y.: A fast parallel algorithm for thinning digital patterns. Communications of the ACM **27**(3), 236–239 (1984)
53. Zhang, Z., Wang, X., Wang, S., Meng, D., Liang, B.: Shape detection and reconstruction of soft robotic arm based on fiber bragg grating sensor array. In: 2018 IEEE International Conference on Robotics and Biomimetics (ROBIO). pp. 978–983. IEEE (2018)
54. Zhong, X., Zhu, W., Liu, W., Yi, J., Liu, C., Wu, Z.: G-sam: A robust one-shot keypoint detection framework for pnp based robot pose estimation. Journal of Intelligent & Robotic Systems **109**(2), 28 (2023)
55. Zhu, J., Lyu, L., Xu, Y., Liang, H., Zhang, X., Ding, H., Wu, Z.: Intelligent soft surgical robots for next-generation minimally invasive surgery. Advanced Intelligent Systems **3**(5), 2100011 (2021)