2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

Graph-based Pose Estimation of Texture-less Surgical Tools for Autonomous Robot Control

Haozheng Xu, Mark Runciman, João Cartucho, Chi Xu and Stamatia Giannarou

***Abstract*— In Robot-assisted Minimally Invasive Surgery (RMIS), the estimation of the pose of surgical tools is crucial for applications such as surgical navigation, visual servoing, autonomous robotic task execution and augmented reality. A plethora of hardware-based and vision-based methods have been proposed in the literature. However, direct application of these methods to RMIS has significant limitations due to partial tool visibility, occlusions and changes in the surgical scene. In this work, a novel keypoint-graph-based network is proposed to estimate the pose of texture-less cylindrical surgical tools of small diameter. To deal with the challenges in RMIS, keypoint object representation is used and for the first time, temporal information is combined with spatial information in keypoint graph representation, for keypoint refinement. Finally, stable and accurate tool pose is computed using a PnP solver. Our performance evaluation study has shown that the proposed method is able to accurately predict the pose of a textureless robotic shaft with an ADD-S score of over 98%. The method outperforms state-of-the-art pose estimation models under challenging conditions such as object occlusion and changes in the lighting of the scene.**

I. INTRODUCTION

Robot-assisted Minimally Invasive Surgery (RMIS) has become an important area of research, since it aims to develop technology to assist surgeons during complex tasks and reduce surgical workload. One of the ultimate goals of RMIS is to develop platforms capable of autonomous task execution, which requires robots to accurately track the position of the surgical tools. This tracking can be achieved using the raw endoscopic video as input or using kinematic input.

Most robots in RMIS are cable-driven robots, and therefore the kinematic input is not always exact since the kinematic data reflects the motor's position and not the actual position of the joints, which are connected to the motor via a cable. Moreover, some soft robots have many desirable qualities for RMIS, such as compliance and the ability to change shape, stiffness, and volume [1]. Soft robotic devices can be difficult to control [2] and sensing modalities are often difficult to integrate [3]. An alternative to track surgical instruments is the use of external hardware, such as OptiTrack™systems, depth cameras, and electromagnetic trackers. Adding extra hardware into the operating theatre is impractical.

*This work was supported by the Royal Society [URF\R\2 01014] , EPSRC [EP/W004798/1] and the NIHR Imperial Biomedical Research Centre.

[1] Hamlyn Centre for Robotic Surgery, Department of Surgery and Cancer, Imperial College London, United Kingdom {haozheng.xu19, stamatia.giannarou}@imperial.ac.uk

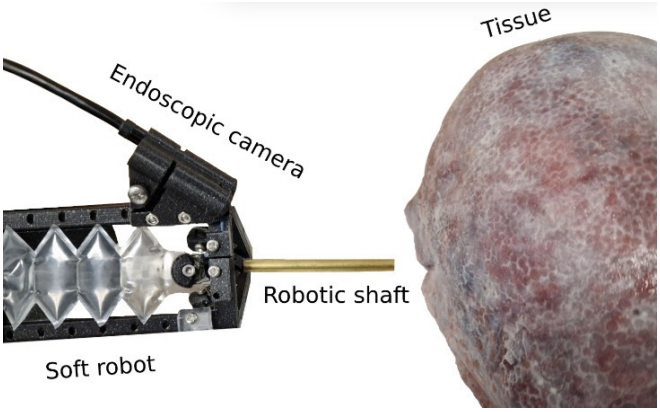

Fig. 1: Experimental setup for autonomous tissue scanning with imaging probes held by a soft robot.

The most practical solution for surgery is to estimate the surgical tool's pose directly from the raw endoscopic video, hence vision-based methods have been developed. These methods can be categorized into marker-based methods and marker-less methods. In RMIS, most of the surgical tools are cylindrical objects, since they are usually inserted through trocars to reach the target anatomy. Therefore, various cylindrical markers have been designed [4][5], which can be printed and wrapped around these cylindrical tools. These marker-based methods rely on detecting visual features such as corners or blobs on the marker. A limitation of these methods is that the marker must be always visible in the camera's Field of View (FoV). In addition, they require the markers to be sterile and to be attached to the surgical tools before surgery, which is impractical. This issue has steered the research into the development of marker-less methods, which focus on the visual features of the surgical tool itself [6][7]. However, the existing methods rely on the detection of hand-crafted features, specific to each type of surgical tool. In these methods, to detect a new tool, a new feature detector must be designed from scratch.

To tackle this problem, deep learning-based methods have been proposed that train end-to-end neural networks that take an image of an object as input and predict its pose directly without relying on hand-crafted features. These methods can be categorized into holistic methods [8][9][10][11][12] and intermediate representation methods [13][14][15]. Holistic methods estimate the location and orientation of an object of interest using a single shot. Recently, some high-performance holistic methods such as ROPE [11] and EfficientPose [16]

apply a two-stage pipeline. The first stage consists of 2D object detection, and the second stage regresses the 6DoF pose from object-level features, which are visual features extracted from the object as a whole. Although these methods achieve high accuracy on public datasets, they cannot perform well when under occlusion or when the object of interest is only partially visible, because they depend on detecting features from the object as a whole. In surgical scenarios, our object of interest (the surgical tool) is always partially observable since the endoscopic field of view is limited.

To achieve robustness to occlusion and partial object visibility, deep-learning methods using intermediate representation have been created [13][14][15]. Intermediate representations include keypoints, pixel-wise heat maps, or vectors. These representations differ from object-level features, to encourage the network to focus on local image information. However, these methods are more sensitive to variations in the scene such as changing lighting conditions as well as, specular reflections and shadows which affect the local feature extraction.

Therefore, the direct application of the above methods to RMIS has significant limitations due to the following challenges. a) **Partial object visibility**. Due to the confined operating space, the endoscopic camera will always be very close to surgical instruments, allowing only part of their body to be visible in the camera's FOV. This will affect the performance of holistic methods. b) **Surgical scene variations**. In RMIS, the surgical tools interact with soft tissue and organs. In this case, occlusion of the tool tip (e.g., due to blood), will make the pose estimation unstable. Furthermore, changes in the surgical scene such as lighting conditions and specular reflections will affect the accuracy of the pose estimation.c) **Lack of highly accurate ground truth data**. Many pose estimation datasets rely on RGBD cameras to generate ground truth data which introduces errors at the centimeter scale. However, the common diameter of surgical tools is about 5 millimeters and therefore, the accuracy requirements are of millimeter scale.

In this work, we propose a novel keypoint-graph-based network to estimate the pose of texture-less cylindrical objects of small diameter, focusing on MIS applications. Our proposed keypoint-graph-based network addresses the limitations of previous holistic methods by leveraging a keypoint object representation that allows for more efficient handling of partial object visibility For this purpose, a keypoint prediction module is introduced to detect 2D keypoints on the shaft of the object. The 2D keypoints are then connected in a graph format to extract relative geometric information. To make our method robust to object occlusions and changes in the surgical scene, temporal information is incorporated in our model. More specifically, graphs from consecutive frames are concatenated and a novel keypoint refinement module is designed using Graph Convolutional Networks (GCN). Finally, a stable and accurate object pose is computed using the PnP solver directly. The performance of the proposed method is evaluated on highly accurate ground truth data captured during autonomous surgical robot manipulation. Our method provides a convenient and effective markless solution for textless surgical tools in practical settings.

II. RELATED WORK

A. Marker-based methods

In RMIS, markers have been attached to surgical instruments [4][17][18] for pose estimation by detecting features such as corners and blobs on the marker's pattern. Although hand-crafted patterns can be designed for different types of tools, the generalization of marker-based methods is still very poor. This is because to apply a pattern to a new surgical instrument, complicated parameter modifications are required. In addition, occlusions and the small diameter of surgical instruments make pattern detection challenging, introducing errors in pose estimation.

B. Holistic Methods

Recently, deep-learning methods are widely used in pose estimation tasks. Deep-learning models can learn geometric information automatically, without relying on markers. Simple end-to-end frameworks have been proposed to train pose estimation directly. For instance, PoseNet [10] directly regresses the camera pose from a single RGB image. PoseCNN [8] firstly localises the object and then estimates the depth to acquire the translation and rotation. However, due to the non-linearity of the rotation space, direct rotation estimation is very difficult which makes the method less generalisable. Several approaches use extra methods to solve the issue. EfficientPose [12] uses the backbone of EfficientNet [16] to detect the object. Then it regresses the bounding box and with a carefully designed iterative refinement module, it provides precise pose estimation. ROPE [11] combines the keypoint prediction into the holistic method pipeline. The model firstly predicts 2D detection using a Mask-RCNN [19] backbone. Based on the extracted ROI Align feature, the model further predicts corresponding keypoints on the object. Holistic methods achieve high performance in the LineMOD [20] benchmark. However, they are not yet suitable for RMIS applications because they rely on the detection of the whole object.

C. Intermediate Representations Methods

Due to the non-linearity of the rotation space, intermediate representations methods have been proposed. A prevalent representation is the keypoint, because keypoint prediction can easily be combined with pose estimation with the PnP algorithm [21]. A common approach is to represent keypoints as vector fields [14] and peaks of heat maps [22][11]. In addition to keypoints, another common intermediate representation is dense point clouds, which include the 3D coordinates of every image pixel. These methods [23][24] provide dense 2D-3D correspondence and achieve high robustness under occlusion. However, dense point cloud prediction is time-consuming and less accurate due to the lack of geometric constraints. Considering the inference time requirement in the real application, sparse keypoint prediction is more practical.

D. Spatial-Temporal Refinement

To refine pose estimation, both temporal and spatial information has been applied [25][26][27][28][29]. Various types of temporal models have been applied in human pose estimation tasks (e.g., Gated Recurrent Units (GRUs) [27], Temporal Convolutional Networks (TCNs) [29],and Transformers [28]. Recently, SmoothNet applies a simple series of fully connected layers to process a sequence of keypoints in human skeletons and achieve good performance. However, it processes each axis of each point separately. This method does not integrate spatial information together. ST-GCN[26] has illustrated Graph Convolution Network (GCN) could efficiently extract the spatial and temporal relationship between human skeleton points. Although GCN has been widely applied in non-rigid human pose estimation, only a few works have tried to apply it to rigid object pose estimation directly. Our work firstly tries to integrate the GCN to extract spatial-temporal information for robust object pose estimation.

III. METHODOLOGY

In this work, a two-stage pipeline is proposed to estimate the pose of cylindrical surgical instruments without requiring any markers to be attached to them. The first stage includes a keypoint prediction network, which outputs a set of N 2D keypoints $\{\mathbf{x}_i|i = 1, ..., N\}$, over the shaft of the tool on the endoscopic image. Here we set $N = 10$. The 2D keypoints correspond to a set of pre-defined 3D points on the 3D model of the tool, sampled using the Farthest Point Sampling (FPS) algorithm [30]. In the second stage, graph representation is used to describe the spatial and temporal relations of the 2D keypoints. Their locations are then refined by the proposed Graph Refinement Network. Finally, the tool pose is derived from the 2D-3D correspondences using the PnP algorithm.

A. Keypoint Prediction Module

The keypoint prediction module aims to identify the object-relevant pixels and predict the keypoints based on these pixels. It is built on the ResNet18 [31] pre-trained on ImageNet [32] which is used as the backbone network to extract feature maps from the endoscopic images. The module consists of two branches. One branch is for the object segmentation mask and consists of a Fully Convolutional layer. It aims to mask out all the irrelevant background information from the target object. This makes our method robust to changes in the surgical scene. For semantic segmentation, we use the binary cross-entropy loss for every image pixel as below:

$$L_{seg} = -v_{pred} \log(v_{gt}) - (1 - v_{pred}) log(1 - v_{gt}) \quad (1)$$

where, v_{pred} and v_{gt} are predicted and ground truth segmentation labels, respectively.

The other branch generates a vector-field representation of each image pixel employing the vector-field prediction model [14]. Compared to other keypoint representations like heatmap peaks [11][22], vector-field representation does not require the whole body of the object to be visible, so it is more generalisable to occlusion. In addition, vector-field representation enables the keypoint detection to be invariant to image transformations such as rotation, translation and scaling.

The output of the vector-field representation model is multiplied with the tool segmentation mask to consider for further processing only pixels that belong to the tool shaft. Then, keypoints are localised by applying a voting scheme on the extracted vector-fields. Since the object is rigid, the relative position of every tool pixel is fixed. Hence, even if part of the tool is occluded, keypoints can still be localised by the visible part. The keypoint positions are predicted using a smooth L1 loss as below:

$$L_{keypoints} = \frac{1}{N} \sum_i SmoothL_1(\mathbf{x}_i^{pred} - \mathbf{x}_i^{gt}) \quad (2)$$

So the total loss is defined as:

$$L_1 = L_{keypoints} + \mu L_{seg} \quad (3)$$

B. Spatial-Temporal Keypoint Refinement Module

Given a set of keypoints detected by the keypoint prediction module, a graph is built to connect every point together. In order to improve the generalisation to the image size, the coordinates of every point are normalized by the image size. To achieve robustness to the challenges introduced due to the changes in the surgical scene, in our framework, temporal information is combined for the first time with spatial information in the graph representation, to refine the initially detected keypoints. For this purpose, we design a graph refinement network to predict the final 2D keypoint locations.

Graph Refinement Network: The input of this network are the detected keypoints of *L* consecutive frames denoted as $T \in \mathbb{R}^{L \times N \times C}$, where C represents the dimension channel for each keypoint. Since we only focus on 2D keypoints, $C = 2$. A keypoint set can be represented as $X = \{\mathbf{x}_{ti}|t = 1, ..., L, i = 1, ..., N\}$. The edges of the graph connect the keypoints as $E = \{x_{ti}x_{tj} \in N\}$. The adjacency matrix $\mathbf{A}$ is built from the edge set, where $A_{ij} = 1$ when x_i and x_j are connected. In this work, we assume all nodes are equally connected together with the same weight. A normalized adjacency matrix can be represented as

$$\hat{\mathbf{A}} = \mathbf{\Lambda}^{-1/2}(\mathbf{A} + \mathbf{I})\mathbf{\Lambda}^{-1/2} \quad (4)$$

where, $\mathbf{\Lambda}^{ii} = \sum_j(\mathbf{A}^{ij} + \mathbf{I}^{ij})$ aims to normalize the adjacency matrix.

The proposed Graph Refinement Network is composed of a block with one layer of TCN, followed by a block with 9 layers of GCN which apply convolutions along the temporal L and spatial C dimensions, respectively. The GCN layer consists of the normalized adjacency matrix $\hat{\mathbf{A}}$ and the convolution kernel $\mathbf{W}$ and it formulated as:

$$\mathbf{f}_{out} = \hat{\mathbf{A}}\mathbf{f}_{in}\mathbf{W}. \quad (5)$$

The first three GCN layers have an output of 64 channels, the next three GCN layers have an output of 128 channels

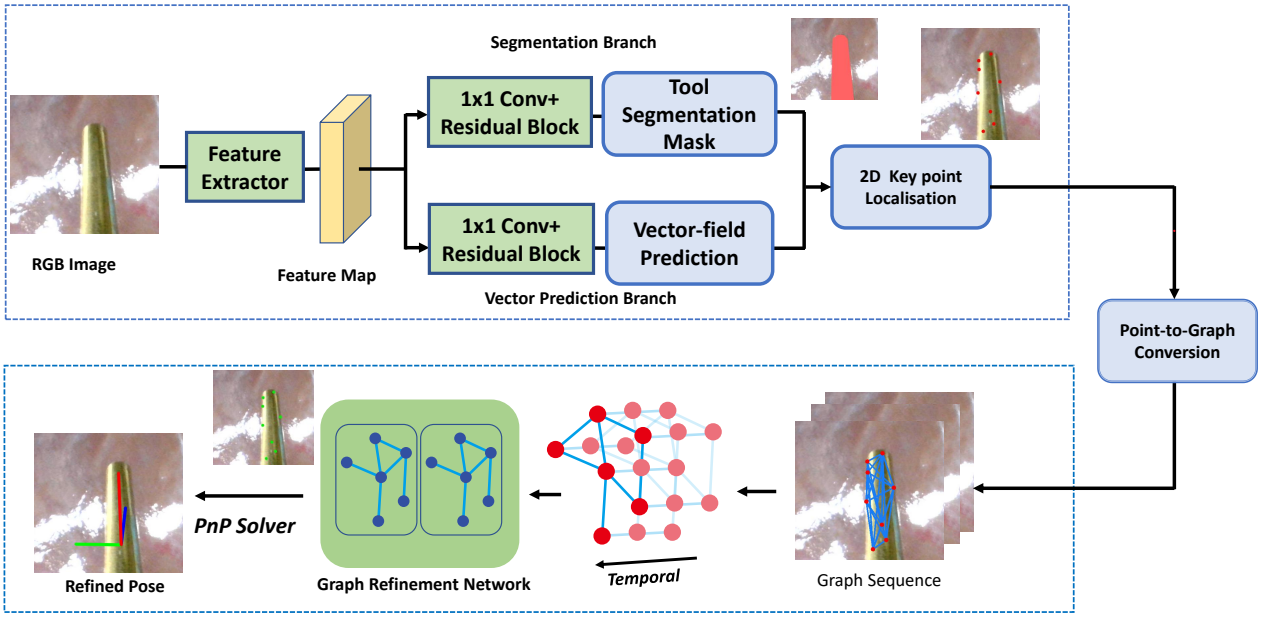

Fig. 2: The overview of our proposed pose estimation network. RGB images labeled with Semantic Mask and 2D keypoint locations are used to train the keypoint estimator in the first stage. Then, in the second stage, a sequence of keypoints will be converted into a sequence of the spatial graph.

and the final three GCN layers have 256 channels output. The output the GCN will be forwarded to two fully convolution layers for final keypoint prediction. The output of the Graph Refinement Network $Y \in \mathbb{R}^{1\times N\times C}$ is the set of refined keypoint locations for frame L, the last frame of the sequence.

The pose of the tool at frame L is then derived from the 2D-3D correspondences between the refined keypoints and the 3D model points using the PnP algorithm [21].

IV. EXPERIMENTS AND RESULTS

A. Dataset Collection

The proposed framework has been applied to estimate the pose of the cylindrical tool of a soft robot to enable autonomous tissue scanning with an imaging probe. Given that in our soft robot, the robotic shaft does not rotate along its main axis, in our captured datasets we have only translated the robotic shaft, to keep it consistent with the final real-world application. Additionally, since the diameter of the robotic shaft is five millimeters only, the shaft is not detected by off-the-shelf commercial depth cameras, such as the RealSense™RGB-D camera, which is mainly designed to capture humans or natural scenes. Therefore, to acquire ground truth, for the five millimeter shaft, an extra marker is used, that is rigidly attached to the robotic shaft and that does not occlude the shaft.

1) Dataset A: The goal of Dataset A was to train and test the keypoint prediction network, shown at the top of Figure 2. In this dataset, the robotic shaft is not installed on the soft robot to provide more freedom and diversity in the training data. The keypoint prediction module is trained purely on Dataset A and since segmentation masks of the robotic shaft are required for training the keypoint prediction network, they have been obtained using the ground truth pose by projecting a CAD model of the robotic shaft into the image plane. Dataset A is composed of a total of 13,244 pictures. 90% of the images are set as training data and 10% of the images are set as test data. These images capture the robotic shaft, without any occlusions, and next to it is a ChArUco marker, which is an extended version of the ArUco [33] marker. In these images, the ChArUco marker is masked out of the image, as illustrated in Figure 4, to guarantee that ChArUco does not influence the performance of the network. Additionally, we have also collected data with a different marker in Dataset B, to further test the keypoint prediction in different scenarios.

2) Dataset B: The goal of Dataset B was to train and test the graph refinement network, shown at the bottom of Figure 2. The keypoint prediction network is still used in these images, to infer the position of the keypoints, which is then used as input to the graph refinement network. We freeze the keypoint prediction network for the ablation study on the occlusion robustness test. In Dataset B we have captured different ground truth markers, specifically one ChArUco marker and another Keydot marker, to test the keypoint network in different scenarios. The goal of the

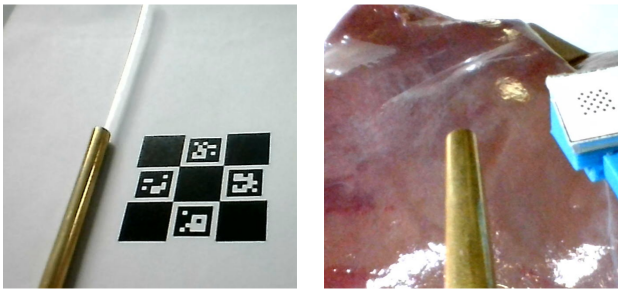

Fig. 3: Illustration of two generated datasets. (Left) A sample of Dataset A. (Right) A sample of Dataset B.

graph refinement network is to refine the predicted keypoints locations, therefore, 75 % of Dataset B consists of images of the robotic shaft under occlusion. Part of Dataset B, captures the robotic shaft over both a phantom and an ex vivo liver, under a real endoscopic light. This dataset is composed of a total of 22,684 pictures.

B. Hardware Setup

The image data was captured using an endoscopic camera, with 720×1080 pixels resolution. To hold the extra markers that are used to collect ground truth, we 3D printed a holder, so that the marker is rigidly attached to the robotic shaft. To accurately estimate the transformation from the markers to the robotic shaft, a temporary cylindrical marker [4] was attached to the robotic shaft, and then removed once the transformation was calibrated. We computed this relative translation by solving the hand-eye calibration $AX = XB$ [34]. The proposed vision system was applied to a parallel robot employing soft hydraulic actuators as in [35], as shown in Figure 1.

C. Evaluation

To assess the estimated camera to robotic shaft transformation, multiple metrics have been applied, namely, (a) the 5mm 5°metric, (b) the Average Distance-Symmetric (ADD-S) metric [8], and (c) a 2D project 5 metric. The (a) 5mm 5°metric, is the percentage of predictions with a rotation error smaller than 5 degrees and a translation error smaller than 5 millimeters. The (b) ADD-S metric, is the percentage of the cases where the mean distance between the points of the 3D tool model transformed with the predicted and ground truth pose is less than 10% of the model diameter. The (c) 2D projection metric is the percentage of images in which the 2D distance, between the predicted and ground truth robotic shaft, is smaller than 5 pixels. This distance is calculated by projecting the origin of the coordinate frames of the 3D model into the image plane.

To evaluate the advantages of the graph-keypoint-based method in estimating the pose of partially-visible objects, we compared the best performing pose estimation models on the LineMOD benchmark [20]. Among them, PVNet [14] is the most accurate keypoint-based method. ROPE [11] and EfficientPose [12] are both holistic methods. ROPE [11] uses a Mask RCNN [19] to capture the bounding box of the object and the corresponding features. It predicts keypoints based on the corresponding features. EfficientPose [12] builds on the architecture of EfficientNet [16] to extract features and predict the bounding box, rotation, and translation based on the extracted feature maps.

	5mm 5°	ADD-S	2D Projection
PVNet[14]	98.24	97.58	94.44
ROPE[11]	25.54	28.45	30.51
EfficientPose[12]	22.45	20.54	25.88
Ours	**99.45**	**98.71**	**96.07**

TABLE I: Evaluation on Dataset A, all values are percentages (%).

1) Evaluation Without Occlusion: We first compared the pose estimation accuracy on test images collected in Dataset A, which consists of images without occlusion of the robotic shaft, as previously mentioned. As shown in Table I, both holistic methods, ROPE [11] and EfficientPose [12] perform significantly worse in this dataset. This is because the robotic shaft is partially visible in the images and these methods expect images that capture the whole object. Therefore, the object detector of these methods fails to acquire accurate object features for pose estimation. Meanwhile, the keypoint-based methods perform significantly better since these methods use keypoint representations, which encourage the network to focus on the local features of the object.

2) Evaluation With Occlusion: Table II shows the performance of the compared methods on test images from Dataset B. In this evaluation, it is evident that the presence of occlusions makes the prediction of the pose of the robotic shaft challenging. Here, the holistic methods (ROPE and EfficientPose) cannot recognize the object correctly in most images. The performance of the intermediate representation methods is less affected by object occlusions, with our method significantly outperforming PVNet. The lower accuracy of the estimated poses in Dataset B is attributed to the different lighting conditions of the training and testing surgical scenes. The different lighting conditions cause a large domain shift in the data of texture-less object appearance. The results in Fig.4 (b) illustrate the initial keypoint and pose prediction of our model. Although most keypoints are predicted correctly, there are still a few wrong detections. These outliers affect the accuracy of the pose estimation with the PnP solver. Still, even with the presence of these outliers, our graph refinement network can rectify the keypoint positions based on spatial-temporal information. Another example is shown in the results on the second row of Fig.4, where the keypoint prediction network detected the keypoints over the wrong cylindrical object. Even in this challenging situation, our proposed graph refinement network is able to rectify the position of the keypoints. In this experiment, the graph refinement network improved our results by roughly 20% since it was able to recover accurate keypoint locations in those situations. The accuracy improvement demonstrates the

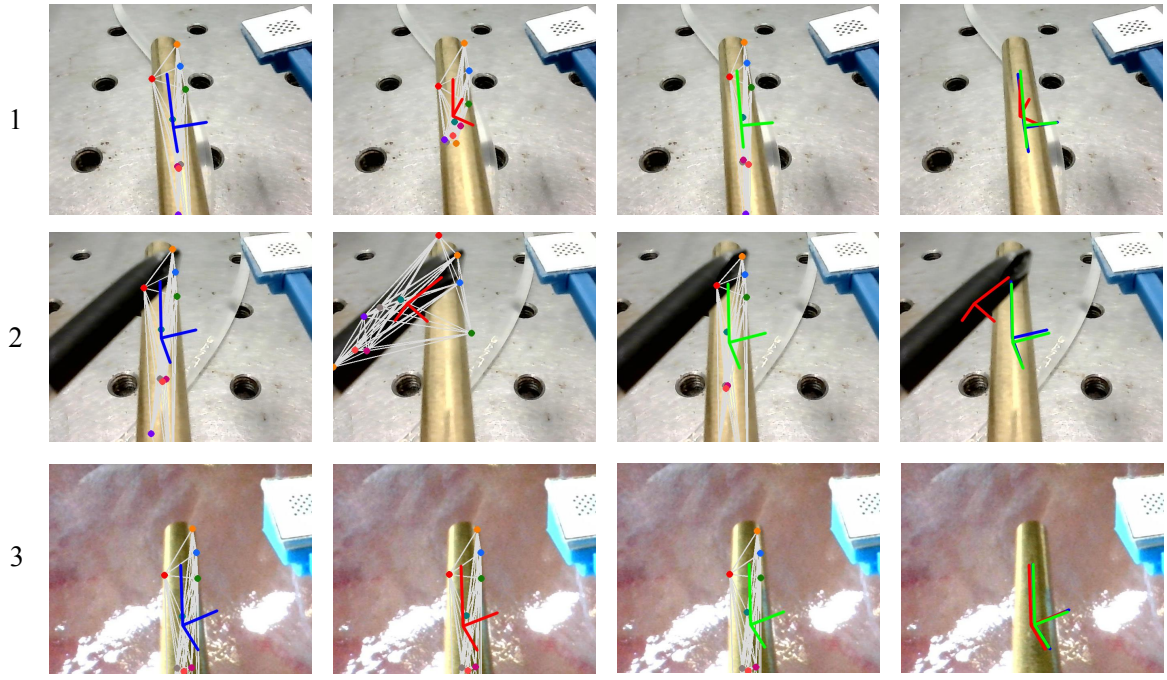

Fig. 4: The keypoint prediction network output and estimated tool poses. (a) Ground truth pose and keypoints (b) the initial prediction (c) refined pose and keypoints estimated with the Graph Refinement Network (d) comparison of the above tool poses.

	5mm 5°	ADD-S	2D Projection
PVNet [14]	9.57	22.38	**10.23**
ROPE [11]	2.37	3.27	4.23
EfficientPose [12]	0.47	0.23	0.64
Ours	12.33	27.87	7.08
Ours w/Graph Refiner	**33.39**	**55.69**	**37.49**

TABLE II: Evaluation on Dataset B with Occlusion. All values are percentages (%)

graph-based refinement module can effectively increase the generalisability and robustness of our proposed method under partial occlusion and variational lighting conditions.

D. Discussion

Based on our experiment, we found that although some holistic methods achieve top performance on public datasets such as LineMOD [20], and YCB [8], these methods cannot generalize to the presented surgical robot scenario since their pose regressor can not deal with partial object visibility. Meanwhile, our keypoint prediction network shows higher accuracy on partially visible objects like surgical instruments. By incorporating spatial-temporal information with the proposed graph refinement network, our pose estimation framework outperforms the above methods and successfully deals with the challenges of the MIS data.

V. CONCLUSIONS

In this paper, we propose a novel keypoint-graph-based pose estimation method tailored for texture-less cylindrical objects of small size. The proposed framework has been applied for the estimation of the pose of the cylindrical tool of a soft robot to enable autonomous tissue scanning with an imaging probe. The method's performance has been validated on two datasets captured from phantom and ex-vivo scenes. The proposed method is able to accurately predict the pose of a textureless robotic shaft with an ADD-S score of over 98%. The method outperforms state-of-the-art pose estimation models under challenging conditions such as object occlusion and changes in the lighting of the scene. In the future, the deployment of this method on a robot in real-time application still needs to be explored. Further experiments can be conducted to integrate this method into a robotic control loop to provide visual feedback for surgical soft robots.

ACKNOWLEDGEMENTS

This work was supported by the Royal Society [URF\R\2 01014] , EPSRC [EP/W004798/1] and the NIHR Imperial Biomedical Research Centre.

REFERENCES

[1] M. Runciman, A. Darzi, and G. P. Mylonas, "Soft Robotics in Minimally Invasive Surgery," *Soft Robotics*, vol. 6, pp. 423–443, 3 2019.

[2] "Soft robotics: Challenges and perspectives," *Procedia Computer Science*, vol. 7, pp. 99–102, 2011.

[3] F. Schmitt, O. Piccin, L. Barbé, and B. Bayle, "Soft robots manufacturing : A review," vol. 5, 2018.

[4] J. Cartucho, C. Wang, B. Huang, D. S. Elson, A. Darzi, and S. Giannarou, "An enhanced marker pattern that achieves improved accuracy in surgical tool tracking," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 10, no. 4, pp. 400–408, 2022.

[5] B. Huang, Y.-Y. Tsai, J. Cartucho, K. Vyas, D. Tuch, S. Giannarou, and D. Elson, "Tracking and visualization of the sensing area for a tethered laparoscopic gamma probe," *International Journal of Computer Assisted Radiology and Surgery*, vol. 15, 06 2020.

[6] M. Ye, L. Zhang, S. Giannarou, and G.-Z. Yang, "Real-time 3d tracking of articulated tools for robotic surgery," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016* (S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells, eds.), (Cham), pp. 386–394, Springer International Publishing, 2016.

[7] A. Reiter, P. K. Allen, and T. Zhao, "Feature classification for tracking articulated surgical tools," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012* (N. Ayache, H. Delingette, P. Golland, and K. Mori, eds.), (Berlin, Heidelberg), pp. 592–600, Springer Berlin Heidelberg, 2012.

[8] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes," 2018.

[9] W. Kehl, F. Manhardt, F. Tombari, S. Ilic, and N. Navab, "SSD-6D: making rgb-based 3d detection and 6d pose estimation great again," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 1530–1538, IEEE Computer Society, 2017.

[10] A. Kendall, M. Grimes, and R. Cipolla, "Posenet: A convolutional network for real-time 6-dof camera relocalization," in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 2938–2946, 2015.

[11] B. Chen, T.-J. Chin, and M. Klimavicius, "Occlusion-robust object pose estimation with holistic representation," in *WACV*, 2022.

[12] Y. Bukschat and M. Vetter, "Efficientpose: An efficient, accurate and scalable end-to-end 6d multi object pose estimation approach," 2020.

[13] M. Rad and V. Lepetit, "Bb8: A scalable, accurate, robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth," pp. 3848–3856, 10 2017.

[14] S. Peng, Y. Liu, Q. Huang, X. Zhou, and H. Bao, "Pvnet: Pixel-wise voting network for 6dof pose estimation," in *CVPR*, 2019.

[15] C. Song, J. Song, and Q. Huang, "Hybridpose: 6d object pose estimation under hybrid representations," 2020.

[16] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning* (K. Chaudhuri and R. Salakhutdinov, eds.), vol. 97 of *Proceedings of Machine Learning Research*, pp. 6105–6114, PMLR, 09–15 Jun 2019.

[17] L. Zhang, M. Ye, P.-L. Chan, and G.-Z. Yang, "Real-time surgical tool tracking and pose estimation using a hybrid cylindrical marker," *International Journal of Computer Assisted Radiology and Surgery*, vol. 12, 03 2017.

[18] U. Jayarathne, E. Chen, J. Moore, and T. Peters, "Robust, intrinsic tracking of a laparoscopic ultrasound probe for ultrasound-augmented laparoscopy," *IEEE Transactions on Medical Imaging*, vol. PP, pp. 1–1, 08 2018.

[19] W. Abdulla, "Mask r-cnn for object detection and instance segmentation on keras and tensorflow." https://github.com/matterport/Mask$_R CNN$, 2017.

[20] E. Brachmann, A. Krull, F. Michel, S. Gumhold, J. Shotton, and C. Rother, "Learning 6d object pose estimation using 3d object coordinates," vol. 8690, pp. 536–551, 09 2014.

[21] V. Lepetit, F. Moreno-Noguer, and P. Fua, "Epnp: An accurate o(n) solution to the pnp problem," *International Journal of Computer Vision*, vol. 81, 02 2009.

[22] J. Wu, T. Xue, J. J. Lim, Y. Tian, J. B. Tenenbaum, A. Torralba, and W. T. Freeman, "Single image 3d interpreter network," in *Computer Vision – ECCV 2016* (B. Leibe, J. Matas, N. Sebe, and M. Welling, eds.), (Cham), pp. 365–382, Springer International Publishing, 2016.

[23] Z. Li, G. Wang, and X. Ji, "Cdpn: Coordinates-based disentangled pose network for real-time rgb-based 6-dof object pose estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[24] K. Park, T. Patten, and M. Vincze, "Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation," pp. 7667–7676, 10 2019.

[25] A. Zeng, L. Yang, X. Ju, J. Li, J. Wang, and Q. Xu, "Smoothnet: A plug-and-play network for refining human poses in videos," in *European Conference on Computer Vision*, Springer, 2022.

[26] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *AAAI*, 2018.

[27] H. Choi, G. Moon, J. Y. Chang, and K. M. Lee, "Beyond static features for temporally consistent 3d human pose and shape from a video," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[28] Z. Wan, Z. Li, M. Tian, J. Liu, S. Yi, and H. Li, "Encoder-decoder with multi-level attention for 3d human shape and pose estimation," in *The IEEE International Conference on Computer Vision (ICCV)*, 2021.

[29] A. Zeng, X. Sun, F. Huang, M. Liu, Q. Xu, and S. C.-F. Lin, "Srnet: Improving generalization in 3d human pose estimation with a split-and-recombine approach," in *ECCV*, 2020.

[30] Y. Eldar, M. Lindenbaum, M. Porat, and Y. Zeevi, "The farthest point strategy for progressive image sampling," *IEEE Transactions on Image Processing*, vol. 6, no. 9, pp. 1305–1315, 1997.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.

[32] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

[33] S. Garrido-Jurado, R. Muñoz-Salinas, F. Madrid-Cuevas, and M. Marín-Jiménez, "Automatic generation and detection of highly reliable fiducial markers under occlusion," *Pattern Recognition*, vol. 47, no. 6, pp. 2280–2292, 2014.

[34] R. Tsai and R. Lenz, "A new technique for fully autonomous and efficient 3d robotics hand/eye calibration," *IEEE Transactions on Robotics and Automation*, vol. 5, no. 3, pp. 345–358, 1989.

[35] M. Runciman, J. Avery, A. Darzi, and G. Mylonas, "Open Loop Position Control of Soft Hydraulic Actuators for Minimally Invasive Surgery," *Applied Sciences*, vol. 11, p. 7391, 8 2021.