A Novel Video-CTU Registration Method with Structural Point Similarity for FURS Navigation

Mingxian Yang[1](✉), Yinran Chen[1](✉), Bei Li[1], Zhiyuan Liu[1], Song Zheng[3](✉), Jianhui Chen[3], and Xiongbiao Luo[1,2](✉)

[1] Department of Computer Science and Technology,
Xiamen University, Xiamen, China
yangmingxian@stu.xmu.edu.cn, yinran_chen@xmu.edu.cn

[2] National Institute for Data Science in Health and Medicine,
Xiamen University, Xiamen, China
xiongbiao.luo@gmail.com

[3] Fujian Medical University Union Hospital, Fuzhou, China
zhengwu_99@aliyun.com

Abstract. Flexible ureteroscopy (FURS) navigation remains challenging since ureteroscopic images are poor quality with artifacts such as water and floating matters, leading to a difficulty in directly registering these images to preoperative images. This paper presents a novel 2D-3D registration method with structure point similarity for robust vision-based flexible ureteroscopic navigation without using any external positional sensors. Specifically, this new method first uses vision transformers to extract structural regions of the internal surface of the kidneys in real FURS video images and then generates virtual depth maps by the ray-casting algorithm from preoperative computed tomography urogram (CTU) images. After that, a novel similarity function without using pixel intensity is defined as an intersection of point sets from the extracted structural regions and virtual depth maps for the video-CTU registration optimization. We evaluate our video-CTU registration method on in-house ureteroscopic data acquired from the operating room, with the experimental results showing that our method attains higher accuracy than current methods. Particularly, it can reduce the position and orientation errors from (11.28 mm, 10.8°) to (5.39 mm, 8.13°).

Keywords: Surgical navigation · Flexible Ureteroscopy · Vision transformer · Computed tomography urogram · Ureteroscopic lithotripsy

1 Introduction

Flexible ureteroscopy (FURS) is a routinely performed surgical procedure for renal lithotripsy. This procedure inserts a flexible ureteroscope through the blad-

M .Yang and Y. Chen—Shows the equally contributed authors.

X. Luo and S. Zheng is the corresponding author.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14228, pp. 123–132, 2023.
https://doi.org/10.1007/978-3-031-43996-4_12

der and ureters to get inside the kidneys for diagnosis and treatment of stones and tumors. Unfortunately, such an examination and treatment depends on skills and experiences of surgeons. On the other hand, surgeons may miss stones and tumors and unsuccessfully orientate the ureteroscope inside the kidneys due to limited field of views, just 2D images without depth information, and the complex anatomical structure of the kidneys. To this end, ureteroscope tracking and navigation is increasingly developed as a promising tool to solve these issues.

Many researchers have developed various methods to boost endoscopic navigation. These methods generally consist of vision- and sensor-based tracking. Han et al. [3] utilized the porous structures in renal video images to develop a vision-based navigation method for ureteroscopic holmium laser lithotripsy. Zhao et al. [15] designed a master-slave robotic system to navigate the flexible ureteroscope. Luo et al. [7] reported a discriminative structural similarity measure driven 2D-3D registration for vision-based bronchoscope tracking. More recently, Huang et al. [4] developed an image-matching navigation system using shape context for robotic ureteroscopy. Additionally, sensor-based methods are widely sued in surgical navigation [1,6]. Zhang et al. [14] employed electromagnetic sensors to estimate the ureteroscope shape for navigation.

Although these methods mentioned above work well, ureteroscopic navigation is still a challenging problem. Compared to other endoscopes such as colonoscope and bronchoscope, the diameter of the ureteroscope is smaller, resulting in more limited lighting source and field of view. Particularly, ureteroscopy involves much solids (impurities) and fluids (liquids), making ureteroscopic video images low-quality, as well as these solids and fluids inside the kidneys cannot be regularly observed in computed tomography (CT) images. On the other hand, the complex internal structures such as calyx, papilla, and pyramids of the kidneys are difficult to be observed in CT images. These issues introduce a difficulty in directly aligning ureteroscopic video sequences to CT images, leading to a challenge of image-based continuous ureteroscopic navigation.

This work aims to explore an accurate and robust vision-based navigation method for FURS procedures without using any external positional sensors. Based on ureteroscopic video images and preoperative computed tomography urogram (CTU) images, we propose a novel video-CTU registration method to precisely locate the flexible ureteroscope in the CTU space. Several highlights of this work are clarified as follows. To the best of our knowledge, this work shows the first study to continuously track the flexible ureteroscope in preoperative data using a vision-based method. Technically, we propose a novel 2D-3D (video-CTU) registration method that introduces a structural point similarity measure without using image pixel intensity information to characterize the difference between the structural regions in real video images and CTU-driven virtual image depth maps. Additionally, our proposed method can successfully deal with solid and fluid ureteroscopic video images and attains higher navigation accuracy than intensity-based 2D-3D registration methods.

2 Video-CTU Registration

Our proposed video-CTU registration method consists of several steps: (1) ureteroscopic structure extraction, (2) virtual depth map generation, and (3) structural point similarity and optimization. Figure 1 illustrates the flowchart of our method.

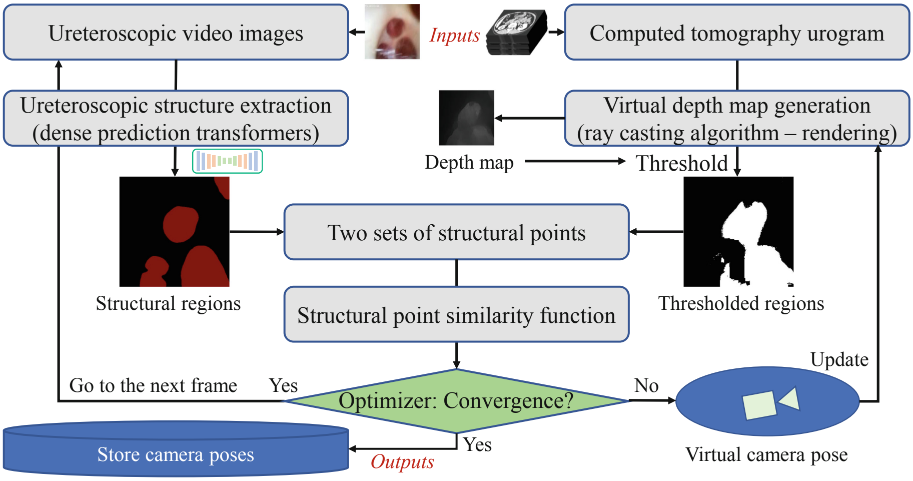

Ureteroscopic video images
Inputs
Computed tomography urogram
Ureteroscopic structure extraction (dense prediction transformers)
Virtual depth map generation (ray casting algorithm – rendering)
Depth map
Threshold
Two sets of structural points
Structural regions
Thresholded regions
Structural point similarity function
Update
Go to the next frame
Yes
Optimizer: Convergence?
No
Yes
Store camera poses
Outputs
Virtual camera pose

Fig. 1. Processing flowchart of our ureteroscopic navigation method.

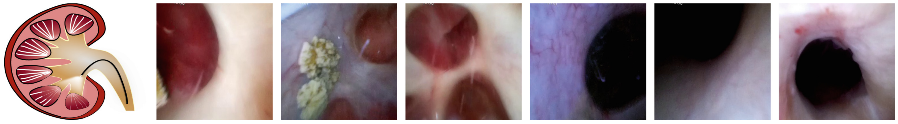
2.1 Ureteroscopic Image Characteristics

The internal kidneys consist of complex anatomical structures such as pelvis and calyx and also contain solid particles (e.g., stones and impurities) floating in fluids (e.g., water, urine and small blood), resulting in poor image quality during ureteroscopy. Therefore, it is a challenging task to extract meaningful features from these low-quality images for achieving accurate 2D-3D registration.

Our idea is to introduce specific structures inside the kidneys to boost the video-CTU registration since these structural regions are meaningful features that can facilitate the similarity computation. During ureteroscopy, various anatomical structures observed in ureteroscopic video images indicate different poses of the ureteroscope inside the kidneys. While some structural features such as capillary texture and striations at the tip of the renal pyramids are observed ureteroscopic images, they are not discernible in CT or other preoperative data. Typical structural or texture regions (Columns 1∼3 in Fig. 2 (b)) observed both in ureteroscopic video and CTU images are the renal papilla when the ureteroscope gets into the kidneys through the ureter and renal pelvis to reach the major and minor calyxes. Additionally, we also find that the renal pelvis (dark or ultra-low light) regions (Columns 4∼6 in Fig. 2 (b)) are also useful to enhance the registration. Hence, this work employs these interior renal structural characteristics to calculate the similarity between ureteroscopic images and CTU.

2.2 Ureteroscopic Structure Extraction

Deep learning is widely used for medical image segmentation. Lazo et al. [5] used spatial-temporal ensembles to segment lumen structures in ureteroscopic images. More recently, vision transformers show the potential to precisely segment various medical images [8,12]. This work employs the dense prediction transformer (DPT) [11] to extract these structural regions from ureteroscopic images.



(a) Poses

(b) Specific structures inside the kidneys

Fig. 2. Ureteroscopic video images with various specific structure information

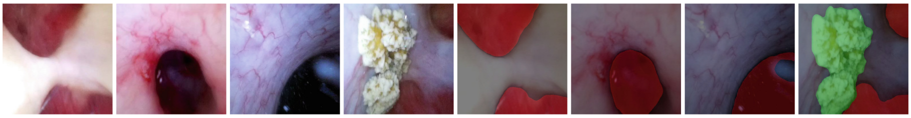


Fig. 3. Input images (the first four images) and their corresponding segmented results: Red indicates structural feature regions and green denotes stones (Color figure online)

DPT is a general deep learning framework for dense prediction tasks such as semantic segmentation and has three versions of DPT-Base, DPT-Large, and DPT-Hybrid. This work use DPT-Base since it only requires a small number of parameters but provides a high inference speed. DPT-Base consists of a transformer encoder and a convolutional decoder. Its backbone is vision transformers [2], where input images are transformed into tokens by non-overlapping patches extraction, followed by a linear projection of their flattened representation. The conventional decoder employs a reassemble operation [11] to assemble a set of tokens into image-like feature representations at various resolutions:

$$\text{Reassemble}_s^{\hat{D}}(t) = (\text{Resample}_s \circ \text{Concatenate} \circ \text{Read})\,(t) \tag{1}$$

where s is the output size ratio of the feature representation and $\hat{D}$ is the output feature dimension. Tokens from layers $\text{l} = \{6, 12, 18, 24\}$ are reassembled in DPT-Base. These feature representations are subsequently fused into the final dense prediction. In the structure extraction, we define three classes: Non-structural regions (background), structural regions, and stones. We manually select and annotate ureteroscopic video images for training and testing. Vision transformers require large datasets for training, so we initialize the encoder with weights pretrained on ImageNet and further train it on our in-house database. Figure 3 displays some segmentation results of ureteroscopic video images.

2.3 Virtual CTU Depth Map Generation and Thresholding

This step is to compute depth maps of 2D virtual images generated from CTU images by volume rendering [13]. Depth maps can represent structural information of virtual images. Note that this work uses CTU to create virtual images since non-contrast CT images cannot capture certain internal structures of the kidneys, although these structures can be observed in ureteroscopic video images.

We introduce a ray casting algorithm in volume rendering to generate depth maps of virtual rending images [13]. The ray casting algorithm is to trace a ray that starts from the viewpoint and passes through a pixel on the screen. When the tracing ray intersects with a voxel, the properties of that voxel will affect the value of corresponding pixel in the final image. For a 3D point (x_0, y_0, z_0) and a normalized direction vector (r_x, r_y, r_z) of its casting ray, a corresponding point (x, y, z) at any distance d on the tracing ray is:

$$(x, y, z) = (x_0 + dr_x,\ y_0 + dr_y,\ z_0 + dr_z). \tag{2}$$

The tracing ray $\mathbf{R}(x, y, z, r_x, r_y, r_z)$ will stop when it encounters opaque voxels. For 3D point (x,y,z), its depth value V can be calculated by projecting the ray $\mathbf{R}(x, y, z, r_x, r_y, r_z)$ onto the normal vector $\mathbf{N}(x, y, z)$ of the image plane:

$$V(x, y, z) = \mathbf{R}(x, y, z, r_x, r_y, r_z) \bullet \mathbf{N}(x, y, z), \tag{3}$$

where symbol • denotes the dot product.

To obtain the depth map with structural regions, we define two thresholds t_u and t_v. Only CT intensity values within $[t_u, t_v]$ are opaque voxels that the casting rays cannot pass through in the ray-casting algorithm. According to CTU characteristics [10], this work uses our excretory-phase data to generate virtual images and set $[t_u, t_v]$ to [–1000, 120], where –1000 represents air and 120 was determined by the physician's experience and characteristics of contrast agents.

Unfortunately, the accuracy of thresholded structural regions suffers from inaccurate depth maps caused by renal stones and contrast agents. Stones and agents are usually high intensity in CTU images, which result in incorrect depth information of structural regions (e.g., renal papilla). To deal with these issues, we use the segmented stones as a mask to remove these regions with wrong depth. On the other hand, the structural regions usually have larger depth values than the agent-contrasted regions. Therefore, we sort the depth values outside the mask and only use the thresholded structural regions with the largest depth values for the structural point similarity computation.

2.4 Structural Point Similarity and Optimization

We define a point similarity measure between DPT-base segmented structural regions in ureteroscopic images and thresholded structural regions in virtual depth maps generated by the volume rendering ray casting algorithm.

The structural point similarity function (cost function) is defined as an intersection of point sets from the extracted real and virtual structural regions:

$$\mathcal{F}(\mathbf{I}_i, \mathbf{D}_v) = \frac{2\,|\mathcal{P}_i \bigcap \mathcal{P}_v|}{|\mathcal{P}_i| + |\mathcal{P}_v|}, \mathcal{P}_i \in E_i(x, y), \mathcal{P}_v \in E_v(x, y), \tag{4}$$

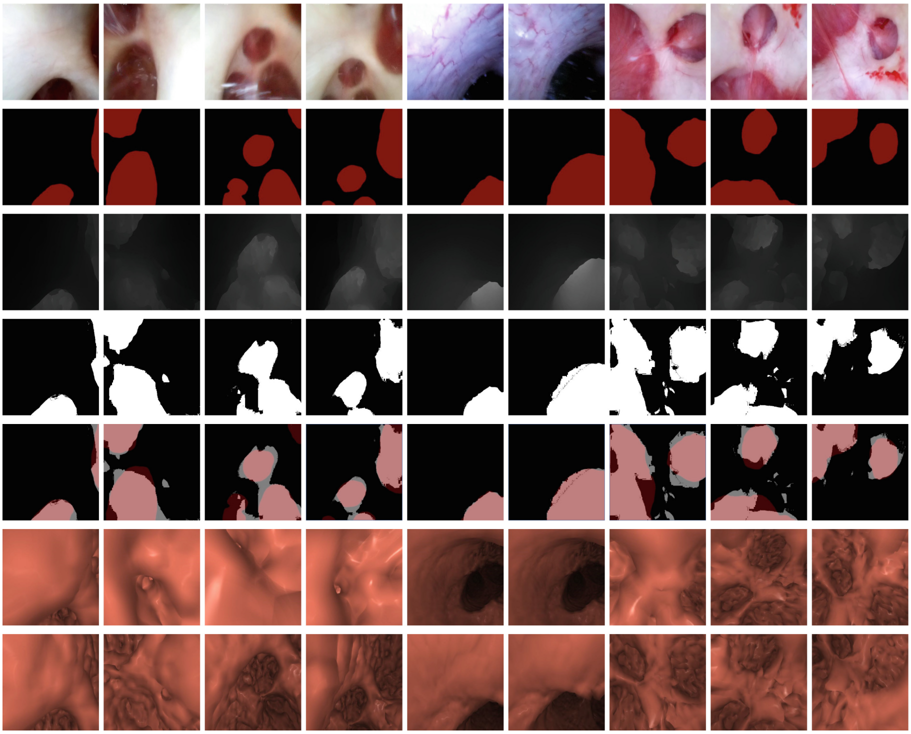


Fig. 4. Visual navigation results: Rows 1∼5 show the real video images, segmented structural regions, depth maps, thresholded structural regions, and overlapped regions for similarity computation, while Rows 6∼7 illustrate the tracking results (2D virtual rendering images) of using Luo et al. [7] and our method.

where $\mathbf{I}_i$ is the ureteroscopic video image at frame i, point sets $\mathcal{P}_i$ and $\mathcal{P}_v$ are from the ureteroscopic image extracted structural region $E_i(a,b)$ and the thresholded structural region $E_v(a,b)$ ((a,b) denotes a point)from the depth map $\mathbf{D}_v$ of the 2D virtual rendering image $\mathbf{I}_v(\mathbf{p}_i,\mathbf{q}_i)$, respectively:

$$\mathbf{D}_v \propto \mathbf{I}_v(\mathbf{p}_i,\mathbf{q}_i), \tag{5}$$

where $(\mathbf{p}_i,\mathbf{q}_i)$ is the endoscope position and orientation in the CTU space.

Eventually, the optimal pose $(\tilde{\mathbf{p}}_i,\tilde{\mathbf{q}}_i)$ of the ureteroscope in the CTU space can be estimated by maximizing the structural point similarity:

$$(\tilde{\mathbf{p}}_i,\tilde{\mathbf{q}}_i) = \arg\max_{\mathbf{p}_i,\mathbf{q}_i} \mathcal{F}(\mathbf{I}_i,\mathbf{D}_v), \qquad \mathbf{D}_v \in \mathbf{I}_v(\mathbf{p}_i,\mathbf{q}_i), \tag{6}$$

where Powell method [9] is used as an optimizer to run this procedure.

Table 1. DPT-base segmented results of intersection over union (IoU), accuracy (Acc), and dice similarity coefficient (DSC), and estimated ureteroscope (position, orientation) errors of using the two vision-based navigation methods

Classes	IoU%	Acc%	DSC%	Methods	Luo et al. [7]	Our method
Background	92.14	98.20	95.91	Case A	(18.13 mm, 11.78°)	(5.66 mm, 7.05°)
Structures	80.09	83.17	88.95	Case B	(11.04 mm, 4.87°)	(2.59 mm, 3.66°)
Renal stones	92.80	95.40	96.27	Case C	(8.60 mm, 12.37°)	(6.20 mm, 10.04°)
Average	88.34	92.26	93.71	Average	(11.28 mm, 10.82°)	(5.39 mm, 8.13°)

Table 2. Sensitivity analysis results of threshold in virtual CTU depth map generation. In this work, the threshold range was set to [–1000, 120].

Threshold	[–1000, 70]	[–1000, 95]	[–1000, 120]	[–1000, 145]	[–1000, 170]
Position Errors	11.08 mm	8.48 mm	5.39 mm	6.70 mm	5.59 mm
Orientation Errors	15.55°	10.52°	8.13°	8.92°	11.10°

3 Results and Discussion

We validate our method on clinical ureteroscopic lithotripsy data with video sequences and CTU volumes. Ureteroscopic video images were a size of 400 × 400 pixels, while the space parameters of CTU volumes were 512 × 512 pixels, 361∼665 slices, 0.625∼1.25 mm slice thickness. Three ureteroscopic videos more than 30000 frames were acquired from three ureteroscopic procedures for experiments. While we manually annotated ureteroscopic video images for DPT-base segmentation, three experts also manually generated ureteroscope pose ground-truth data by our developed software, which can manually adjust position and direction parameters of the virtual camera to visually align endoscopic real images to virtual images, evaluating the navigation accuracy of the different methods.

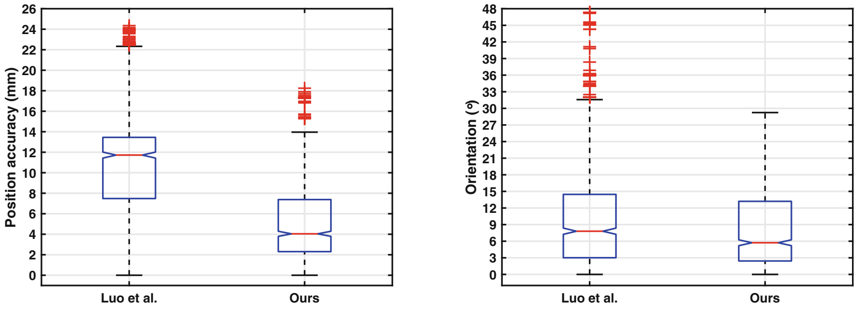


Fig. 5. Boxplotted position and orientation errors of using the two methods

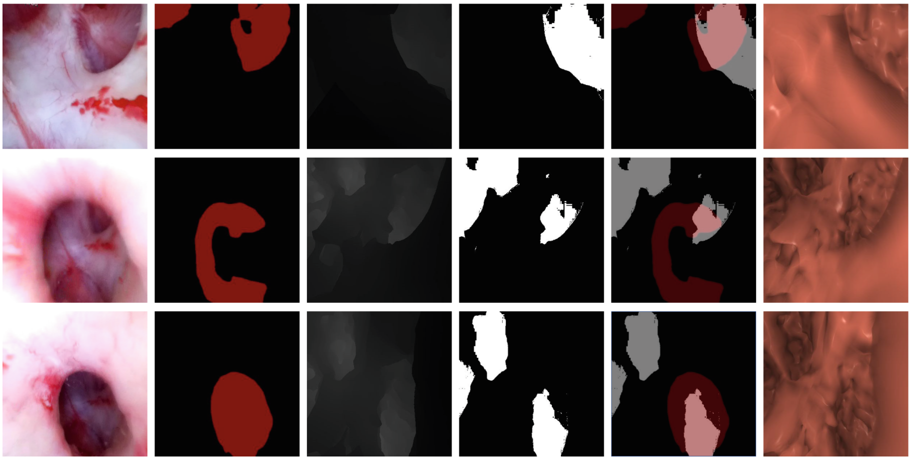


Fig. 6. Our method fails to track the ureteroscope in CTU: Columns 1∼6 correspond to input images, segmented structural regions, depth maps, thresholded structural regions, overlapped regions, and generated virtual images.

Figure 4 illustrates the navigation results of segmentation, depth maps, extracted structural regions for similarity calculation, and generated 2D virtual images corresponding to estimated ureteroscope poses. Structural regions can be extracted from ureteroscopic images and virtual depth maps. Particularly, we can see that our method generated virtual images (Row 7 in Fig. 4) resemble real video images (Row 1 in Fig. 4) much better than Luo et al. [7] generated ones (Row 6 in Fig. 4). This implies that our method can estimate the ureteroscope pose much more accurate than Luo et al. [7]. Table 1 summarizes quantitative segmentation results and position and orientation errors. DPT-Base can achieve average segmentation IoU 88.34%, accuracy 92.26%, and DSC 93.71%. The average position and orientation errors of our method were 5.39 mm and 8.14°, which much outperform the compared method. Table 2 shows the results of sensitivity analysis for the threshold values. It can be seen that inappropriate threshold selection can lead to an increase in errors. Figure 5 boxplots estimated position and orientation errors for a statistical analysis of our navigation accuracy.

The effectiveness of our proposed method lies in several aspects. First, renal interior structures are insensitive to solids and fluids inside the kidneys and can precisely characterize ureteroscopic images. Next, we define a structural point similarity measure as intersection of point sets between real and virtual structural regions. Such a measure does not use any point intensity information for the similarity calculation, leading to an accurate and robust similarity characterization under renal floating solids and fluids. Additionally, CTU images can capture more renal anatomical structures inside the kidneys compared to CT slices, still facilitating an accurate similarity computation.

Our method still suffers from certain limitations. Figure 6 displays some ureteroscopic video images our method fails to track. This is because that the segmentation method cannot successfully extract structural regions, while the ray casting algorithm cannot correctly generate virtual depth maps with structural regions. Both unsuccessfully extracted real and virtual structural regions collapse the similarity characterization. We will improve the segmentation of ureteroscopic video images, while generating more ground-truth data for training and testing.

4 Conclusion

This paper proposes a new 2D-3D registration approach for vision-based FURS navigation. Specifically, such an approach can align 2D ureteroscopic video sequences to 3D CTU volumes and successfully locate an ureteroscope into CTU space. Different from intensity-based cost function, a novel structural point similarity measure is proposed to effectively and robustly characterize ureteroscopic video images. The experimental results demonstrate that our proposed method can reduce the navigation errors from (11.28 mm, 10.8°) to (5.39 mm, 8.13°).

Acknowledgements. This work was supported in part by the National Natural Science Foundation of China under Grants 61971367, 82272133, and 62001403, in part by the Natural Science Foundation of Fujian Province of China under Grants 2020J01004 and 2020J05003, and in part by the Fujian Provincial Technology Innovation Joint Funds under Grant 2019Y9091.

References

1. Attivissimo, F., Lanzolla, A.M.L., Carlone, S., Larizza, P., Brunetti, G.: A novel electromagnetic tracking system for surgery navigation. Comput. Assist. Surg. **23**(1), 42–52 (2018)
2. Dosovitskiy, A., et al.: An image is worth 16 × 16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
3. Han, M., Dai, Y., Zhang, J.: Endoscopic navigation based on three-dimensional structure registration. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 2900–2905. IEEE (2020)
4. Huang, Z.: Image-matching based navigation system for robotic ureteroscopy in kidney exploration. Master's thesis, Delft University of Technology, Netherlands (2022)
5. Lazo, J.F., et al.: Using spatial-temporal ensembles of convolutional neural networks for lumen segmentation in ureteroscopy. Int. J. Comput. Assist. Radiol. Surg. **16**(6), 915–922 (2021). https://doi.org/10.1007/s11548-021-02376-3
6. Luo, X.: A new electromagnetic-video endoscope tracking method via anatomical constraints and historically observed differential evolution. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12263, pp. 96–104. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59716-0_10
7. Luo, X.: Accurate multiscale selective fusion of CT and video images for real-time endoscopic camera 3D tracking in robotic surgery. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 33, pp. 1386–1390. IEEE (2022)

8. Matsoukas, C., Haslum, J.F., Söderberg, M., Smith, K.: Is it time to replace CNNs with transformers for medical images? arXiv preprint arXiv:2108.09038 (2021)
9. Nocedal, J., Wright, S.J.: Numerical Optimization, 2nd ed. Springer, New York (2006). https://doi.org/10.1007/978-0-387-40065-5
10. Noorbakhsh, A., Aganovic, L., Vahdat, N., Fazeli, S., Chung, R., Cassidy, F.: What a difference a delay makes! CT urogram: a pictorial essay. Abdom. Radiol. **44**(12), 3919–3934 (2019)
11. 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)
12. Shamshad, F., et al.: Transformers in medical imaging: a survey. arXiv preprint arXiv:2201.09873 (2022)
13. Wrenninge, M.: Production Volume Rendering: Design and Implementation, vol. 5031 (2020)
14. Zhang, C., et al.: Shape estimation of the anterior part of a flexible ureteroscope for intraoperative navigation. Int. J. Comput. Assist. Radiol. Surg. **17**(10), 1787–1799 (2022)
15. Zhao, J., Li, J., Cui, L., Shi, C., Wei, G., et al.: Design and performance investigation of a robot-assisted flexible ureteroscopy system. Appl. Bionics Biomech. **2021** (2021)