Kidney Stone Segmentation and Improved Generalization using SAM

Martin VILLAGRANA[1], Francisco LOPEZ-TIRO[1,2], Clement LAROSE[2,3], Gilberto OCHOA-RUIZ[1], Christian DAUL[2]

[1]Escuela de Ingenieria y Ciencias, Tecnologico de Monterrey
Av. Eugenio Garza Sada 2501 Sur, Tecnológico, 64849 Monterrey, N.L., Mexico

[2]Université de Lorraine, CNRS, CRAN (UMR 7039),
2 avenue de la Forêt de Haye, 54518 Vandœuvre-lès-Nancy, France

[3]CHRU Nancy-Brabois, service d'urologie, Vandœuvre-les-Nancy, France
gilberto.ochoa@tec.mx, christian.daul@univ-lorraine.fr

Résumé – La segmentation des calculs rénaux est un prétraitement facilitant l'identification du type des calculs urinaires à l'aide de classifieurs. La segmentation manuelle est fastidieuse et peu appropriée car la base de d'images urétéroscopiques est de grande taille. Cette étude examine le potentiel du modèle "segment anything model" (SAM, une approche d'apprentissage profond de référence) pour automatiser la segmentation des calculs rénaux. Les performances de SAM sont comparées à celles de modèles classiques tels que U-Net, Residual U-Net, et Attention U-Net qui sont efficaces, mais qui disposent d'une faible capacité de généralisation à des données "encore non vues". Cette contribution démontre l'adaptabilité et l'efficacité de SAM. Bien que SAM conduise à des résultats comparables à ceux d'U-Net pour des images dont la distribution est similaire à celle des données d'apprentissage (Accuracy : 97,68±3,04 ; Dice : 97,78±2,47 ; IoU : 95,76±4,18), SAM démontre une plus grande capacité de généralisation en surpassant de 23% les performances de toutes les variantes d'U-Net pour des distributions non apprises.

Abstract – Kidney stone segmentation is a preliminary step for facilitating the identification of urinary stone types using machine- or deep-learning methods. IPerforming a manual segmentation is tedious and impractical since the ureteroscopic image database is large. This study investigates the potential of the segment anything model (SAM, a state-of-the-art deep learning approach) to automate the kidney stone segmentation. The performance of SAM is compared to that of traditional models such as U-Net, Residual U-Net, and Attention U-Net which are efficient, but often struggle to generalize to unseen datasets. This contribution emphasize the adaptability on different data-distributions of SAM. While SAM shows a performance similar to that of U-Net on in-distribution data (Accuracy: 97.68±3.04, Dice: 97.78±2.47, IoU: 95.76±4.18), it demonstrates superior generalization capability on out-of-distribution data, outperforming all U-Net variants by margins of up to 23%.

1 Introduction

Kidney stone formation is an illness that affects millions worldwide. The lifetime prevalence in the United States ranges from 7.2% to 10.1%, whereas other in other world parts higher rates are reported, such as 21.1% in certain populations [1]. Early detection and accurate diagnosis behind the kidney stone formation are critical for prescribing an effective treatment to prevent complications such as renal damage, infections, and recurrent stone formation. With such timely intervention, patients can experience improved outcomes and reduced healthcare costs. Given its high incidence and the severe pain it causes, kidney stone detection plays a vital role in medical practice, particularly in patients with recurrent episodes or in those at high risk [2].

Recent advancements in deep learning (DL) offer promising potential for enhancing the classification of kidney stones [3, 4]. These methods have the potential of speeding up diagnosis by automating a highly operator dependent task when performed directly during the stone extraction process (ESR, endoscopic stone recognition) which is crucial for patients needing early intervention. An accurate segmentation of the detected kidney stones is a necessary preliminary step before classification, as it enables the model to isolate the stone itself from surrounding tissues, thereby improving diagnostic precision [5]. DL, particularly convolutional neural networks (CNNs) and transformer models, have demonstrated significant efficacy in segmentation tasks, as they can automatically discern relevant features within complex medical images, enhancing the reliability of image-based diagnosis [6].

DL-segmentation models such as U-Net and its variations (Attention U-Net and Residual U-Net) have shown good performance in segmenting images. U-Net, known for its limited generalization efficiency, excels at preserving spatial information through its encoder-decoder structure, although it may have difficulties with complex and noisy images [7]. Residual U-Net, which integrates residual connections, addresses problems like vanishing gradients, improving performance in deeper networks, but its deeper architecture can increase training time [8]. Attention U-Net improves U-Net by incorporating care mechanisms, allowing the model to focus on relevant areas, but this added complexity can lead to increased computational costs [9]. More recently, models such as the "segment anything model" (SAM) have emerged. These transformer-based

architectures offers increased flexibility across a variety of image types without requiring extensive task-specific tuning. Although SAM is very promising for handling various data sets, its complexity and resource demand remains a challenge [10].

The segmentation of kidney stone images extracted from ureteroscopy videos is traditionally performed manually by an expert (see the top row in Fig. 1). While this method can be effective for small datasets, it becomes impractical when applied to large-scale databases due to its time-intensive nature and reliance on operator expertise. Advanced DL models like SAM offer a transformative solution, enabling the automatic segmentation of extensive video datasets with high accuracy and efficiency. The ability of SAM to generalize across diverse datasets presents a significant advancement, making it a promising alternative to traditional manual methods for large-scale segmentation tasks (see the bottom row in Fig. 1).

The main contributions of this work include the assessment of both traditional and modern deep learning segmentation techniques. By examining the performance of models such as U-Net and comparing them with SAM, this research seeks to identify the most appropriate approaches for kidney stone segmentation in two classes, namely kidney stones and surrounding tissues.

The paper is structured as follows. Section 2 provides details about the dataset and methodology used in this study, focusing on traditional deep learning-based segmentation techniques, contrasting them with SAM. Section 3 discusses the findings and presents the results. Finally, future directions and conclusions are discussed in Section 4.

2 Materials and Methods

2.1 Clinical image datasets

Four kidney stone datasets were utilized in our experiments. The images were obtained from two sources : standard CCD cameras and endoscopic images that were captured using an ureteroscope. The main characteristics of the used datasets are described below.

Database A. In-vivo endoscopic images. It contains a total of 156 in-vivo (endoscopic) images. In the scene, three classes are observed : surrounding tissue, kidney stone, and laser used for stone fragmentation. The dataset can be used to segment two or three classes. The dimensions of the images in this dataset are of different sizes around 1008×1042 pixels [11].

Database B. Ex-vivo endoscopic images. It includes 409 ex-vivo images acquired using a realistically simulated environment (i.e., a phantom that mimics the characteristics of the ureters in which stones may be blocked). The dataset consists of two classes : kidney stone and tissue. The dimensions of the all images on this dataset are 1920×1080 pixels [12].

Database C. In-vivo endoscopic images. It is comprised of 56 in-vivo images. The images were acquired from videos where several classes can be observed : living tissue, kidney stones, and instruments such as the laser for fragmentation. In this dataset, there are two resolutions of the images : 400×400 and 720×720 pixels. This dataset can be used to evaluate the performance of segmentation mode for either two or three classes.

Database D. Ex-vivo CCD-camera. It consists of 356 ex-vivo charge-coupled device (CCD) images, showing exclusively two classes : the extracted kidney stone fragments and their environment. The dimensions of the images in this dataset are of different sizes around 4288×2848 pixels [13].

All images from these databases were resized to square dimensions of 512×512 pixels for analysis. Additionally, each database is accompanied by ground-truth segmentation masks of the original images, manually created by a specialist.

2.2 Kidney Stone Segmentation

In this work, the SAM model is employed to perform kidney stone segmentation in ureteroscopy. Segmented masks from the input dataset (Distribution 1) are required to accomplish this task. These masks are obtained through the "traditional method for segmenting kidney stones" (see the top row in Fig. 1), performed by an expert. Subsequently, the "automatic kidney stone segmentation and generalization" method utilizing SAM (see the bottom row in Fig. 1), is applied to generate new automatic segmentation masks,either from the same distribution or other distributions.

Traditional method for segmenting kidney stones. Frames displaying both kidney stones and the surrounding tissues are first selected in ureteroscopy videos to generate binary (two class) segmentation masks. These images comprise the dataset from Distribution 1 thatis manually annotated at the pixel level by an urologist. The resulting masks are then paired with their corresponding RGB-images, forming the input dataset (X_1, Y_1), where X_1 represents the original images and Y_1 their segmentation labels.

Automatic kidney stone segmentation and generalization. As mentioned in Section 1, kidney stone segmentation is effective for small datasets but impractical for large ones. For this reason, we evaluate SAM (Segment Anything Model) to learn from an expert-labeled dataset and then use it to automatically generate segmentation masks—both for the original data distribution and for others.

SAM involves two stages : training and testing. Concerning the training, an initial step is first carried out to prepare the data. This image preprocessing step involves a normalization and resizing techniques which "standardizes" the images (512×512) to ensures uniform input image sizes for all models. In the next step, the training of SAM is split into two processes : first, a feature extraction transformer block condenses the input image in a feature matrix. These extracted features and the model's prompts, such as segmentation masks (Y_2), are then fed into a decoder head. After training, the learned embeddings enable the testing of SAM, either on the same data distribution $(Y_1 \rightarrow Y_1)$ or on a different distribution $(Y_1 \rightarrow Y_2)$.

Dataset A (in-vivo endoscopic images) was used to train SAM. Inference was performed on its original distribution (in-

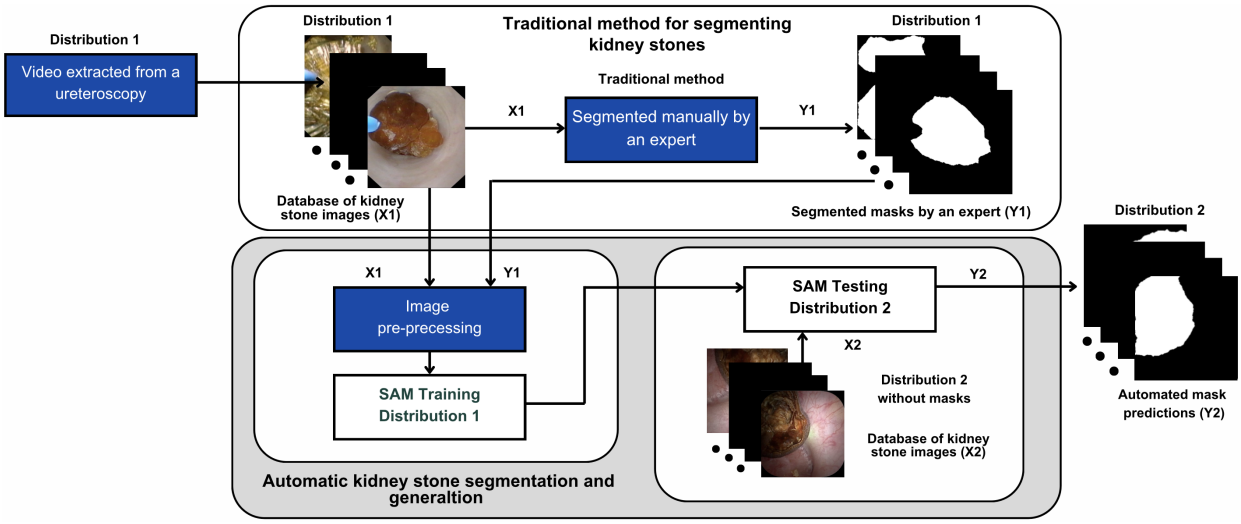

FIGURE 1 – Overview of the principle of the segmentation method compzrison. Traditional segmentation methods extract frames from ureteroscopy videos to create a dataset (distribution 1), which is manually labeled by an expert. When SAM for zn automatic kidney stone segmentation, the model is trained on distribution 1 and its labels. It then performs inference on a different dataset (distribution 2, without labels), producing the labels for this set. It has to be noticed that distributions 1 and 2 are distinct.

distribution) and on three additional datasets (B, C, and D), as detailed in Section 2.1. Three UNet variants were trained (standard UNet, Residual UNet, and Attention-UNet) to benchmark SAM's performance against classical architectures. All models were trained on their respective data distributions and subsequently evaluated on datasets A, B, C, and D. The primary distinction is that SAM was trained solely on Dataset A but tested across all datasets (A–D).

The quantitative evaluation is based on standard metrics : Accuracy, Dice Coefficient, and Intersection over Union (IoU). These metrics can be used for an objective segmentation quality assessment. Testing across four datasets (with distinct distributions) demonstrated the models' generalization capability and robustness for kidney stone segmentation across diverse clinical scenarios.

The training of the models was performed on a 16 GB Nvidia DGX GPU, using the AdamW Optimizer with a dynamic learning rate adjustment and a warmup phase. A combination of Dice Loss and CE Loss was used to minimize discrepancies between predicted segmentation and ground truth masks. U-Net, Residual U-Net, and Attention U-Net were trained for 80 epochs, while SAM was trained for 200 epochs.

3 Results and Discussion

The performance of the segmentation models (SAM and UNet-based models), introduced in Section 2.2, was evaluated using the datasets described in Section 2.1. The experimental results are summarized in Table 1 and discussed below.

In distribution. The SAM model, trained on Dataset A and evaluated on its native distribution, achieves consistent performance across all evaluation metrics : 97.68 ± 3.04 Accuracy, 97.78 ± 2.47 Dice, 95.76 ± 4.18 IoU, and an Error Rate of 4.24. These results demonstrate SAM's ability to deliver high performance despite limited training data (156 images). While SAM does not achieve the absolute highest performance when trained and tested on the same distribution (Dataset A), it maintains comparable results. For comparison, all results in Table 1 are reported in terms of IoU.

Out-of-distribution. The SAM model, trained on Dataset A and evaluated on Datasets B, C, and D, consistently outperforms U-Net-based models : it achieves 93.74 ± 9.55 IoU on Dataset B (vs. U-Net's 60.75 ± 22.20 IoU), 86.64 ± 10.90 IoU on Dataset C (vs. 63.37 ± 23.70 IoU), and 93.81 ± 9.47 IoU on Dataset D (vs. 70.58 ± 27.59 IoU), demonstrating its ability to generate segmentation masks with out-of-distribution data.

4 Conclusions

The results from this study demonstrate that SAM outperforms traditional segmentation models such as U-Net, Residual U-Net, and Attention U-Net for kidney stone segmentation. While the traditional models performed well on the dataset they were trained on, their performance significantly declined when tested on new, unseen datasets. SAM, trained on the same dataset, not only excelled in the original dataset but also generalized effectively across different datasets, maintaining high segmentation accuracy.

SAM's ability to generate accurate, continuous segmentations without the artifacts commonly seen in other models, combined with its superior generalization across datasets, makes it the most reliable model for kidney stone segmentation. This reinforces the importance of using models that can adapt to new data, ensuring consistent performance in clinical applications where data variability is expected.

TABLE 1 – Comparison of the performance of four segmentation models (U-Net, Residual U-Net, Attention U-Net, and SAM) across four datasets (A to D). Segmentation is performed for two classes (surrounding tissue or background, and kidney stone). The presented results were measured using segmentation metrics (Accuracy, Dice Score, Intersection over Union, and Error Rate).

Dataset	Model	Accuracy ↑	Dice ↑	IoU ↑	Error rate ↓
A : In-vivo Endoscopic	SAM Model	97.68±03.04	97.78±02.47	95.76±04.18	4.24%
	U-Net	97.68±01.38	97.72±01.28	95.77±02.37	4.23%
	ResU-Net	**97.92±00.96**	**97.93±01.01**	**95.97±01.91**	**4.03%**
	AttentionU-Net	96.37±02.71	96.24±03.77	92.98±06.02	7.02%
B : Ex-vivo Endoscopic	**SAM Model**	**98.71±02.87**	**96.39±07.70**	**93.74±09.55**	**6.26%**
	U-Net	88.8±07.83	73.05±18.58	60.75±22.20	39.25%
	ResU-Net	87.9±07.03	70.85±18.49	57.95±21.82	42.05%
	AttentionU-Net	83.83±07.75	65.47±19.37	51.77±21.76	48.23%
C : In-vivo Endoscopic	**SAM Model**	**95.33±04.70**	**92.42±07.19**	**86.64±10.90**	**13.36%**
	U-Net	86.45±11.38	74.31±22.89	63.37±23.70	36.63%
	ResU-Net	86.09±11.64	71.69±25.58	60.83±25.27	39.17%
	AttentionU-Net	85.16±09.95	71.67±20.30	59.08±20.69	40.92%
D : Ex-vivo CCD Camera	**SAM Model**	**96.48±07.91**	**96.50±06.20**	**93.81±09.47**	**6.19%**
	U-Net	87.59±12.96	78.73±25.35	70.58±27.59	29.42%
	ResU-Net	89.51±11.02	84.13±18.92	76.19±22.41	23.81%
	AttentionU-Net	89.31±11.78	82.85±21.42	75.11±24.42	24.89%

Acknowledgements. The authors acknowledge the support of the "Secretaría de Ciencia, Humanidades, Tecnología e Innovación" (SECIHTI), the French Embassy in Mexico, and Campus France through postgraduate scholarships, as well as the Data Science Hub at Tecnológico de Monterrey. This work was also funded by Azure Sponsorship credits from Microsoft's AI for Good Research Lab under the AI for Health program and the French-Mexican Ecos Nord grant (MX 322537).

Références

[1] Leila Moftakhar, Fatemeh Jafari, Masoumeh Ghoddusi Johari, Ramin Rezaeianzadeh, Seyed Vahid Hosseini, and Abbas Rezaianzadeh. Prevalence and risk factors of kidney stone disease in population aged 40–70 years old in kharameh cohort study : a cross-sectional population-based study in southern iran. *BMC urology*, 22(1) :205, 2022.

[2] Ben H Chew, Larry E Miller, Brian Eisner, Samir Bhattacharyya, and Naeem Bhojani. Prevalence, incidence, and determinants of kidney stones in a nationally representative sample of us adults. *JU Open Plus*, 2(1) :e00006, 2024.

[3] Francisco Lopez-Tiro, Vincent Estrade, Jacques Hubert, Daniel Flores-Araiza, Miguel Gonzalez-Mendoza, Gilberto Ochoa, and Christian Daul. On the in vivo recognition of kidney stones using machine learning. *IEEE Access*, 12 :10736–10759, 2024.

[4] Jorge Gonzalez-Zapata, Francisco Lopez-Tiro, Elias Villalvazo-Avila, Daniel Flores-Araiza, Jacques Hubert, Gilberto Ochoa-Ruiz, Christian Daul, and Andres Mendez-Vazquez. A metric learning approach for endoscopic kidney stone identification. *Expert Systems with Applications*, 255, Part D, December 2024.

[5] Tonmoy Ghosh, Linfeng Li, and Jacob Chakareski. Effective deep learning for semantic segmentation based bleeding zone detection in capsule endoscopy images. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 3034–3038. IEEE, 2018.

[6] Soumya Gupta, Sharib Ali, Louise Goldsmith, Ben Turney, and Jens Rittscher. Multi-class motion-based semantic segmentation for ureteroscopy and laser lithotripsy. *Computerized Medical Imaging and Graphics*, 101 :102112, 2022.

[7] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net : Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015 : 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[8] Md Zahangir Alom, Chris Yakopcic, Mahmudul Hasan, Tarek M Taha, and Vijayan K Asari. Recurrent residual u-net for medical image segmentation. *Journal of medical imaging*, 6(1) :014006–014006, 2019.

[9] Zan Li, Hong Zhang, Zhengzhen Li, and Zuyue Ren. Residual-attention unet++ : a nested residual-attention u-net for medical image segmentation. *Applied Sciences*, 12(14) :7149, 2022.

[10] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[11] Vincent Estrade, Michel Daudon, Emmanuel Richard, Jean-Christophe Bernhard, Franck Bladou, Gregoire Robert, Laurent Facq, and Baudouin Denis de Senneville. Deep morphological recognition of kidney stones using intra-operative endoscopic digital videos. *Physics in Medicine & Biology*, 67(16), 2022.

[12] Jonathan El Beze, Charles Mazeaud, Christian Daul, Gilberto Ochoa-Ruiz, Michel Daudon, Pascal Eschwège, and Jacques Hubert. Evaluation and understanding of automated urinary stone recognition methods. *BJU Int.*, 130(6) :786–798, 2022.

[13] Mariela Corrales, Steeve Doizi, Yazeed Barghouthy, Olivier Traxer, and Michel Daudon. Classification of stones according to michel daudon : a narrative review. *European Urology Focus*, 7(1) :13–21, 2021.