MI-UNET: IMPROVED SEGMENTATION IN URETEROSCOPY

Soumya Gupta*, Sharib Ali*, Louise Goldsmith [†]*, Ben Turney*[†]   *Jens Rittscher**

*Institute of Biomedical Engineering, Department of Engineering Science,
University of Oxford, Oxford, UK
†Department of Urology, The Churchill, Oxford University Hospitals NHS Trust, Oxford, UK

ABSTRACT

Ureteroscopy has evolved into a routine technique for treatment of kidney stones. Laser lithotripsy is commonly used to fragment the kidney stones until they are small enough to be removed. Poor image quality, presence of floating debris and severe occlusions in the endoscopy video make it difficult to target stones during the ureteroscopy procedure. A potential solution is automated localization and segmentation of the stone fragments. However, the heterogeneity of stones in terms of shape, texture, as well as colour and the presence of moving debris make the task of stone segmentation challenging. Further, dynamic background, motion blur, local deformations, occlusions and varying illumination conditions need to be taken into account during segmentation. To address these issues, we compliment state-of-the-art U-Net based segmentation strategy with the learned motion information. This technique leverages difference in motion between the large stones and surrounding debris and additionally tackles problems due to illumination variability, occlusions and other factors that are present in the frame-of-interest. The proposed motion induced U-Net (MI-UNet) architecture consists of two main components: 1) U-Net and 2) DVFNet. The quantitative results show consistent performance and improvement over most evaluation metrics. The qualitative validation also illustrate that our complimentary DVFNet is able to effectively reduce the effect of surrounding debris in contrast to U-Net.

Index Terms— Ureteroscopy, U-Net, Optical flow, MI-UNet

1. INTRODUCTION

The past two decades have witnessed a significant rise in the incidence of kidney stone diseases [1]. Kidney stones are formed when minerals from urine separate and get aggregated inside the kidney or ureters. Most stones smaller than 5 mm will pass while stones larger than 5 mm may cause blockage of ureter and severe pain in the abdomen or lower back [2]. Abdominal X-ray, CT and Ultrasound are used for imaging kidney stones with CT being the most common as it provides the most accurate result [3]. Once the stone is located, ultrasound or laser energy is used to fragment the stones into smaller chunks that are either removed or further

broken down to be flushed out in urine. The available kidney stone treatment options include percutaneous nephrolithotomy (PCNL), ureteroscopy, and extra-corporeal shockwave lithotripsy (ESWL) [2]. The choice of treatment depends on the location, size and kind of kidney stone. PCNL is the most invasive technique recommended for very large stones or in case of failure from previous surgeries; ESWL is effective for small stones; and ureteroscopy techniques are recommended for cases when size of calculi falls in the range of 10-20 mm [4].

Ureteroscopy has evolved into a routine procedure for the treatment of kidney stones and the diagnosis of pathologies of the upper urinary tract [5]. It is performed using a long, thin and flexible tube that consists of a light source and a camera. The scope has a working channel through which tools like laser fiber can be inserted for stone fragmentation. A real-time video signal is produced by the ureteroscope and is available to the surgical team for guidance.

However, the endoscopy video is often of poor quality, corrupted by noise and motion blur, and has presence of stone debris obscuring the vision. The application of computer vision techniques to enhance the endoscopy video can provide assistance to the clinician and can potentially increase the efficiency of the procedure. Real-time estimation of stone size is the most important parameter during ureteroscopy as it determines if the stone requires further fragmentation or can be removed. Automatic segmentation is the primary step necessary for estimating the size of the stone in ureteroscopic images.

2. RELATED WORK

Previously, methods for the detection and segmentation of kidney stones in ultrasound (US) images have been proposed [6, 7]. To deal with noisy US images, an extensive pre-processing step was first applied. A level-set based segmentation was adopted to segment the kidney and kidney stones [6]. Watershed segmentation has also been used for the segmentation of renal calculi [7]. KNN and SVM classification techniques have been employed for the analysis of kidney stones from US images [8]. Unlike US, ureteroscopic image quality is compromised by breathing motion,

and the irrigation fluid used to clear out particles [9, 4]. A region growing algorithm for segmentation of renal calculi on ureteroscopic images was proposed [4]. However, such approaches are computationally expensive and require the user to define seed pixel, a similarity criterion and a stopping criterion. The similarity criterion depends on difference in colour values of the already found region and region to be examined. But, there exists a large variability in both texture, colour and position of kidney stones.

Today, convolutional neural networks have been established for building effective and robust segmentation methods. U-Net is one of the most popular CNN architectures used for semantic segmentation in biomedical datasets. This model employs an encoder-decoder architecture with multi-scale features that can be trained on less training images and still yield promising segmentation results [10]. U-Net and its variants have shown to achieve state-of-art results in a variety of biomedical datasets [11, 12, 13]. An improved U-Net where the convolutional layers were replaced by residual-inception blocks was proposed for nuclei segmentation on histology data [11]. Classical 2D U-Net can be extended to 3D U-Net to obtain volumetric segmentation [12]. A graph search based UNet-D was proposed for analysis of endoscopy images [14]. U-Net has also been used for semantic segmentation on laryngeal endoscopic images [13] and polyp detection in wireless capsule endoscopy [15].

To the best of our knowledge, not much work has been reported on segmentation of kidney stones in ureteroscopy images. This is due to the challenges and complexities involved in the ureteroscopy data that include stone heterogeneity, poor image quality, motion blur, presence of stone debris obscuring the vision. The video quality is further affected by the kidney movement induced by breathing and the irrigation fluid that is continuously passed through the procedure [4]. Varying illumination conditions and large refractory errors from the irrigation fluid further increase the level of difficulty for the segmentation tasks. In this paper, we present a motion induced U-Net (MI-UNet) that leverages the motion vector fields between the frame-of-interest and the previous frame to improve the segmentation results. Our approach involves the estimation of the deformation vector field (DVF) between the two frames which is then fed into the training network to assist U-Net based segmentation module. Our experiments demonstrate the effectiveness of our proposed method compared to the performance of U-Net alone.

3. MATERIALS AND METHOD

3.1. Materials

Four human kidney stones of different shapes were individually targeted with laser in an in-vitro environment and imaged using a Boston Scientific Lithovue scope. Parts of videos with intense lasering and stone movements were selected and broken down into frames. Consecutive frames were then orga-
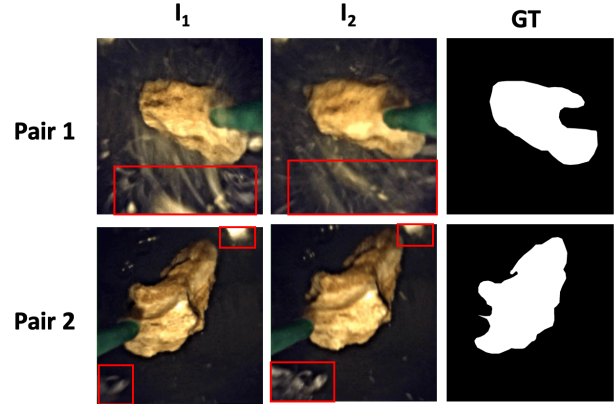
Fig. 1: Paired images $\{I_1, I_2\}$ showing different viewpoints with variable surrounding debris shown in red rectangles and corresponding ground truth masks (GT).

nized in pairs (I_1 and I_2) and first image of each pair was manually annotated using the VGG Image Annotator (VIA) tool [16] to obtain the ground truth masks as shown in Fig. 1. 248 sets of image pairs (I_1, I_2) and corresponding ground truth binary segmentation mask of I_1 were resized to 256×256 pixels and used to train the network. 28 independent test samples were used to evaluate the segmentation quality. This study also establishes a performance comparison with Vanilla U-Net which was trained using set of images I_1 and its corresponding ground truth masks.

3.2. Method

Our proposed segmentation approach (see Fig. 2) compliments the U-Net based segmentation through DVFNet. The entire learning process is designed as an end-to-end convolutional neural network that optimises a combination of two different losses. Each module and its combination is detailed below:

3.2.1. U-Net

The first sub-network is U-Net consisting of an encoder path (left side) and a decoder path (right side). The encoder path, also known as the contracting path, consists of repeated application of two 3×3 convolutions, each followed by a rectified linear unit (ReLU) and a 2×2 max pooling operation with a stride of 2 for downsampling. In the contracting path, the image size gradually reduces and the depth gradually increases as each downsampling step results in double the number of feature channels. The decoder path, also known as the expansive path, involves application of transposed convolutions along with regular convolutions to gradually increase the size of image and reduce the depth of image. To improve localization information, every step of the decoder path uses skip connections by concatenating the output of the transposed con-
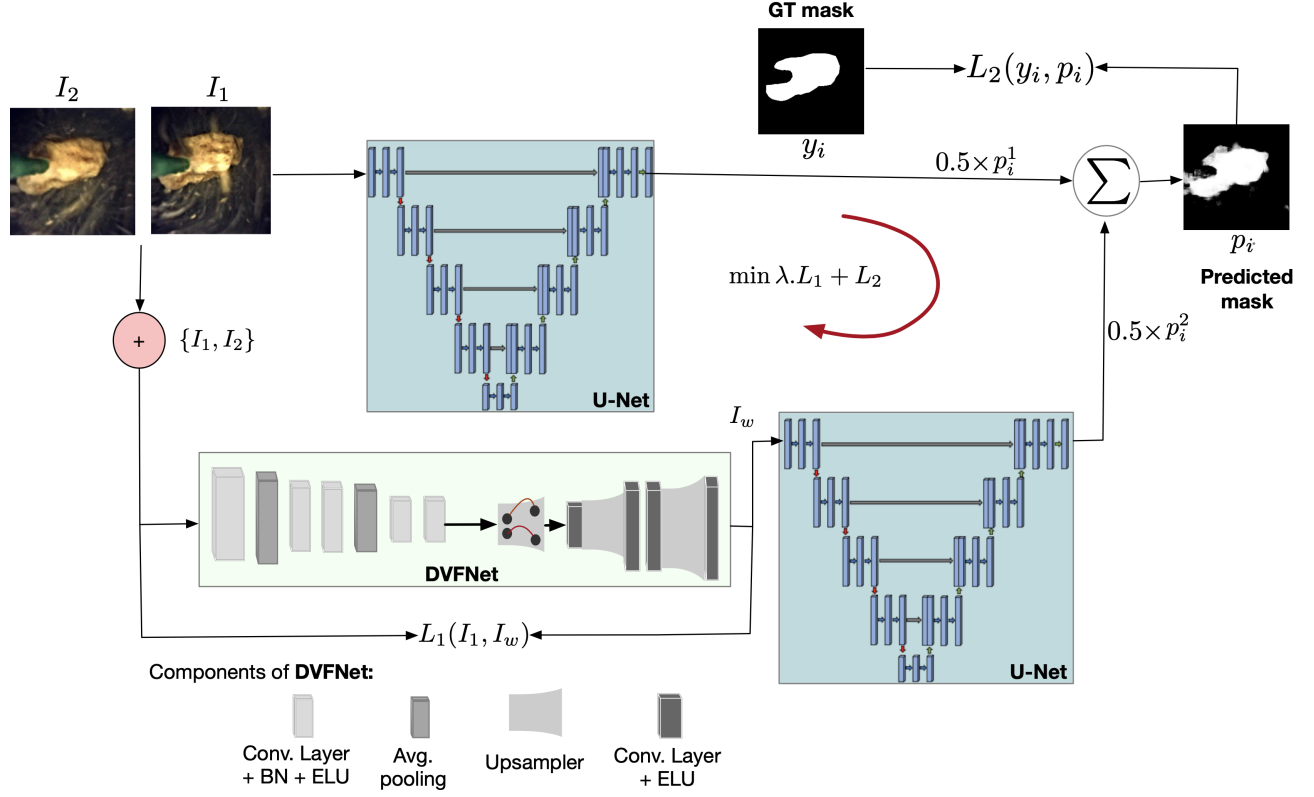
Fig. 2: Proposed MI-UNet network architecture

volution layers with the feature maps from the encoder at the same level where each concatenation is followed by application of two consecutive regular convolutions [10].

3.2.2. DVFNet

The second sub-network (DVFNet) is also based on the encoder-decoder architecture but with a learnable spline re-sampler [17]. It consists of 12 layers that is a combination of convolution layers, exponential linear unit (ELU) layers, and an average pooling for encoder layer, while for decoder layer, first a resampler is used to rescale the encoded DVF which is then further resampled with 2 additional layers consisting of convolution layer, an ELU activation function and a cubic spline based resampler. The final DVF prediction at 0th scale is then applied on the original image I_2 to obtain a warped image I_w.

3.2.3. Loss function

The loss function used in this study is a combination of two losses: $L_1(I_1, I_w)$ is the cross-correlation loss calculated between source image I_1 and computed warped image and L_2 is the binary cross-entropy loss calculated between the predicted mask for I_1 and its ground truth mask. The cross-correlation

loss is given by:

$$L_1(I_1, I_w) = \frac{1}{2N} \sum \left(\frac{I_1(x) - \mu_1}{\sqrt{\sigma_1^2 + \epsilon^2}} - \frac{I_w(x) - \mu_w}{\sqrt{\sigma_w^2 + \epsilon^2}} \right)^2 \quad (1)$$

where, μ and σ are the mean and standard deviation, N is the total number of pixels and $\epsilon = 10^{-3}$ to avoid division by zero. The binary cross-entropy loss is computed based on average prediction (p_i) of U-Net prediction mask of I_1 (p_i^1) and warped image I_w (p_i^2) and is given by:

$$L_2 = \frac{1}{N} \sum_{i=1}^{N} (y_i(log(p_i)) + (1 - y_i)(log(1 - p_i))) \quad (2)$$

where y is the label (1 or 0) and p is the predicted probability.

3.2.4. Proposed network

The source image I_1 is first fed to the U-Net to obtain first prediction map p_i^1. Then, images I_1 and I_2 are concatenated and provided as input to the DVFNet architecture that computes the deformation field and subsequently a warped image, I_w. The obtained I_w is then provided to the U-Net to obtain a second prediction map p_i^2. The two mask predictions are then averaged together to obtain a final prediction map p_i.

Table 1: Quantitative evaluation of U-Net and proposed MI-UNet

Method	Jaccard Index	Dice Coef.	F2-score	PPV
U-Net [10]	0.751 ± 0.053	0.833 ± 0.048	$\mathbf{0.888} \pm 0.073$	0.785 ± 0.067
MI-UNet (**proposed**)	$\mathbf{0.765 \pm 0.046}$	$\mathbf{0.850 \pm 0.040}$	$0.878 \pm \mathbf{0.062}$	$\mathbf{0.822 \pm 0.055}$

The network then optimizes the prediction map by minimizing a combined loss function ($\lambda L_1 + L_2$): cross-correlation loss L_1 between I_1 and warped image that enables the network to learn DVF prediction and a binary cross entropy loss L_2 between final predicted mask and the ground truth. The proposed network architecture is shown in Fig. 2.

4. RESULTS AND DISCUSSION

We randomly selected the image pairs for train-validation-test: 228 train samples and 20 validation samples. The training was done on a single 12 GB NVIDIA 2080Ti GPU.

Fig. 3 presents the Jaccard index (JI) and the validation loss for validation set during training time for 300 epochs with batch size of 2. It can be observed that MI-UNet was able to achieve higher boost in JI on the validation set of 0.92 while U-Net achieved only 0.89. Additionally, MI-UNet achieved a lower loss compared to U-Net.

14 independent image pairs with abundance of debris were used to test the efficacy of the method using standard metrics (Jaccard index (JI), Dice coefficient (DSC), F2-score and positive predictive value (PPV)) for the quantitative evaluation of the image segmentation. Tab. 1 represents mean values for each metric over 14 images. It can be observed that the proposed MI-UNet has JI, DSC, and PPV values 0.765, 0.850 and 0.822, respectively, while the U-Net based architecture has lower values of 0.751, 0.833, and 0.785, respectively. Also, the standard deviation for all evaluation metrics is lower for our proposed MI-UNet showing better consistency in all test data.

Fig. 4 shows an example of a pair of test set images with significant stone displacement and the corresponding DVF obtained from DVFNet network. As evident from the Fig. 4, our proposed network is able to efficiently learn the DVF predictions thereby providing improved segmentation of stone in the target images. Fig. 5 shows a qualitative comparison between the ground truth, U-Net prediction and MI-UNet prediction for three samples. It can be observed that the MI-UNet is able to provide a more accurate segmentation of the stone fragments when compared to the U-Net. Fig. 5 also demonstrates that MI-UNet is able to overcome the effect of motion blur and obstruction due to stone debris in the scene.

5. CONCLUSION

To the best of our knowledge, this is the first work introducing motion integrated U-Net algorithm for improved segmentation of kidney stones in ureteroscopy images. Although we
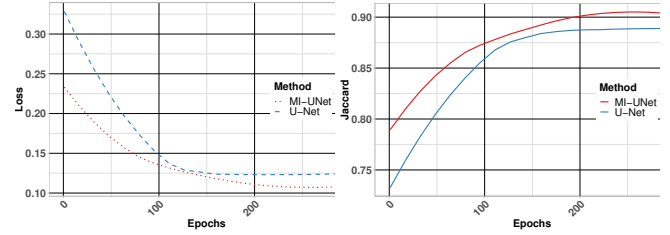
Fig. 3: Validation loss and Jaccard index during training
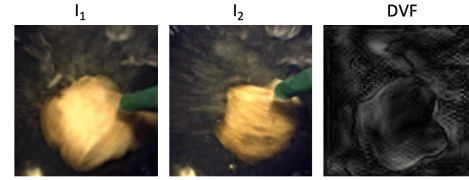
Fig. 4: A sample of paired images $\{I_1, I_2\}$ with corresponding DVF obtained from DVFNet
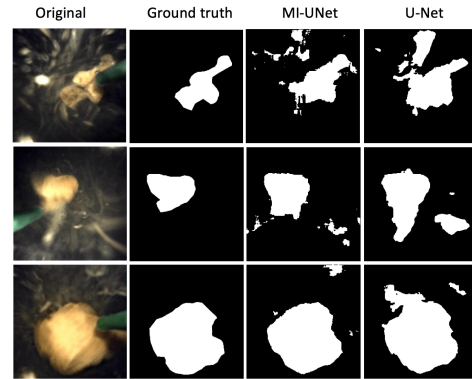
Fig. 5: Qualitative evaluation of segmentation using U-Net and MI-UNet

have developed this method for ureteroscopy video, it can also be applied to other video datasets. The proposed algorithm effectively leverages motion information between two images to produce more robust and reliable segmentation. The qualitative and quantitative results demonstrate that our algorithm can efficiently tackle the obstruction due to stone debris and improve the segmentation accuracy.

6. REFERENCES

[1] Salvatore Butticè, Tarik Emre Sener, Christopher Netsch, Esteban Emiliani, Rosa Pappalardo, and Carlo Magno, "LithoVue™: A new single-use digital flexible ureteroscope," *Central European Journal of Urology*, 2016.

[2] Nicole L. Miller and James E. Lingeman, "Management of kidney stones," *British Medical Journal*, 2007.

[3] Wayne Brisbane, Michael R. Bailey, and Mathew D. Sorensen, "An overview of kidney stone imaging techniques," *Nature Reviews Urology*, 2016.

[4] Benoît Rosa, Pierre Mozer, and Jérôme Szewczyk, "An algorithm for calculi segmentation on ureteroscopic images," *International Journal of Computer Assisted Radiology and Surgery*, 2011.

[5] Petrisor Geavlete, Razvan Multescu, and Bogdan Geavlete, "Pushing the boundaries of ureteroscopy: Current status and future perspectives," *Nature Reviews Urology*, 2014.

[6] K. Viswanath and R. Gunasundari, "Design and analysis performance of kidney stone detection from ultrasound image by level set segmentation and ANN classification," in *Proceedings of the 2014 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2014*, 2014.

[7] P. Thangaraj and P. R. Tamilselvi, "A Modified Watershed Segmentation Method to Segment Renal Calculi in Ultrasound Kidney Images," *Int. J. Intell. Inf. Technol.*, vol. 8, 2012.

[8] Jyoti Verma, Madhwendra Nath, Priyanshu Tripathi, and K. K. Saini, "Analysis and identification of kidney stone using Kth nearest neighbour (KNN) and support vector machine (SVM) classification techniques," *Pattern Recognition and Image Analysis*, vol. 27, 2017.

[9] John R. Van Sörnsen De Koste, Suresh Senan, Catharina E. Kleynen, Ben J. Slotman, and Frank J. Lagerwaard, "Renal mobility during uncoached quiet respiration: An analysis of 4DCT scans," *International Journal of Radiation Oncology Biology Physics*, 2006.

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015.

[11] Z. Zeng, W. Xie, Y. Zhang, and Y. Lu, "Ric-unet: An improved neural network based on unet for nuclei segmentation in histology images," *IEEE Access*, 2019.

[12] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger, "3d u-net: Learning dense volumetric segmentation from sparse annotation," *CoRR*, vol. abs/1606.06650, 2016.

[13] Max-Heinrich Laves, Jens Bicker, Lüder A. Kahrs, and Tobias Ortmaier, "A dataset of laryngeal endoscopic images with comparative study on convolution neural network based semantic segmentation," *CoRR*, vol. abs/1807.06081, 2018.

[14] Shufan Yang and Sandy Cochran, "Graph-search based unet-d for the analysis of endoscopic images," 2019.

[15] Liansheng Wang, Rongzhen Chen, and Yanxing Hu, "Iddf2018-abs-0261 polyp detection using an unet based model," *Gut*, vol. 67, 2018.

[16] Abhishek Dutta and Andrew Zisserman, "The VIA annotation software for images, audio and video," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019.

[17] Sharib Ali and Jens Rittscher, "Conv2warp: An unsupervised deformable image registration with continuous convolution and warping," in *Machine Learning in Medical Imaging*, 2019, pp. 489–497.