

# Supplementary material

## Multi-class motion-based semantic segmentation for ureteroscopy and laser lithotripsy

Soumya Gupta<sup>1,\*</sup>, Sharib Ali<sup>1,2</sup>, Louise Goldsmith<sup>3</sup>, Ben Turney<sup>3</sup>, and Jens Rittscher<sup>1,\*</sup>

<sup>1</sup>Institute of Biomedical Engineering, Big Data Institute, Department of Engineering Science, University of Oxford, Oxford, UK

<sup>2</sup>Oxford NIHR Biomedical Research Centre, Oxford, UK

<sup>3</sup>Department of Urology, The Churchill, Oxford University Hospitals NHS Trust, Oxford, UK

\*soumya.gupta@eng.ox.ac.uk & jens.rittscher@eng.ox.ac.uk

### 1 Overview of the data splits for the *in vitro* and *in vivo* experiments

Dataset	Split	Number of samples		Distribution ( $\mu \pm \sigma$ )	
		Single	Sequence ( $\times 5$ )	Stone	Laser
<i>In vitro</i>	Train	52	260	$0.3570 \pm 0.1702$	$0.0345 \pm 0.0107$
	Validation	18	90	$0.3517 \pm 0.1635$	$0.0295 \pm 0.0099$
	Test-I	18	–	$0.2858 \pm 0.1975$	$0.0369 \pm 0.0101$
<i>In vivo</i>	Train	92	460	$0.1084 \pm 0.0865$	$0.0222 \pm 0.0067$
	Validation	32	160	$0.0836 \pm 0.0337$	$0.0205 \pm 0.0069$
	Test-I	30	–	$0.0949 \pm 0.0628$	$0.0240 \pm 0.0033$
	Test-II	20	–	$0.1384 \pm 0.0566$	$0.0277 \pm 0.0076$

**Table S1.** Overview of the train, validation, and test splits for the *in vitro* and *in vivo* experiments. Mean ( $\mu$ ) and standard deviation ( $\sigma$ ) for each stone and laser annotations in each split are also provided. Here, Test-I refers to the split samples from the curated dataset as 20% split and Test-II refers to the unseen animal samples.

### 2 Optimal data augmentation strategy

The list of augmentation techniques with their corresponding settings used in this study is mentioned below in Table S2.

Augmentation type	Settings
Horizontal Flip	Probability of application ( $p$ ) = 0.5
Vertical Flip	Probability of application ( $p$ ) = 0.5
Shift scale rotate (SSR)	Shift limit=0.0625, Scale limit=0.1, Rotate limit=45, $p=0.5$
Sharpen	Range [0.2,0.5], $p=0.5$
Gaussian Blur (GB)	Gaussian filter with the size of the kernel in the range [3,7]
Random Brightness Contrast (RBC)	Range [-0.2, 0.2], $p=0.5$
Equalize	Probability of application ( $p$ ) = 0.5
Contrast Limited Adaptive Histogram Equalization (CLAHE)	Contrast clip limit = 4.0, $p=0.5$

**Table S2.** Augmentation settings used in this study for both *in vitro* and *in vivo* datasets

Table S3 below presents our ablation study associated with determining the optimal data augmentation strategy for our dataset. It can be seen that the *Random Brightness Contrast(RBC)* and *Equalize* transformation improve the segmentation accuracy in the case of *in vitro* dataset as compared to no augmentation scenario. On the other hand, *RBC* and *CLAHE* both provide a higher DSC compared to no augmentation in the case of the *in vivo* dataset.

Augmentation type	DSC					
	<i>In vitro</i>			<i>In vivo</i>		
	Stone	Laser	Average	Stone	Laser	Average
None	0.8732	0.8438	0.8585	0.8319	0.7936	0.81275
Horizontal Flip	0.8241	0.8309	0.8275	0.8203	0.7947	0.8075
Vertical Flip	0.8179	0.8054	0.81165	0.8102	0.7194	0.7648
Shift scale rotate (SSR)	0.8505	0.8217	0.8361	0.8261	0.7875	0.8068
Sharpen	0.849	0.8177	0.83335	0.8065	0.8001	0.8033
Gaussian Blur (GB)	0.8629	0.8467	0.8548	0.8126	0.8072	0.8099
<b>Random Brightness Contrast (RBC)</b>	0.8743	<b>0.8774</b>	<b>0.87585</b>	0.8265	0.8168	0.8217
<b>Equalize</b>	<b>0.8895</b>	0.8257	0.8576	0.775	0.8217	0.7983
<b>CLAHE</b>	0.8571	0.8251	0.8411	<b>0.839</b>	<b>0.832</b>	<b>0.8355</b>

**Table S3.** Comparison of augmentation techniques (applied on HybResUNet) for both *in vitro* and *in vivo* datasets

### 3 Ablation study of network design

Table S4 presents an ablation study for the integration of various combinations of dilations, ASPP and attention gate in different networks.

- (I) **Base network with ASPP:** Here, we compare the effect of dilated convolutions and ASPP module connected to each base network presented in (I) in Table S4. It can be seen that in the case of ASPP-HybResUNet, ASPP but does not provide any overall improvement as compared to only HybResUNet. It can also be seen that ASPP increases the sensitivity for both stone and laser class but decreases the PPV of stone class in most networks for the *in vitro* case.
- (II) **Base network with Attention(Att):** In the second set of experiments in Table S4, we can see that using attention increases the overall sensitivity of most networks.
- (III) **Base network with Att-ASPP:** In our third set of experiments, we can see that the combination of both Attention and ASPP does not lead to any improvement in the segmentation of stone or laser in both *in vitro* and *in vivo* cases.

### 4 Quantitative and qualitative results - *in vitro*

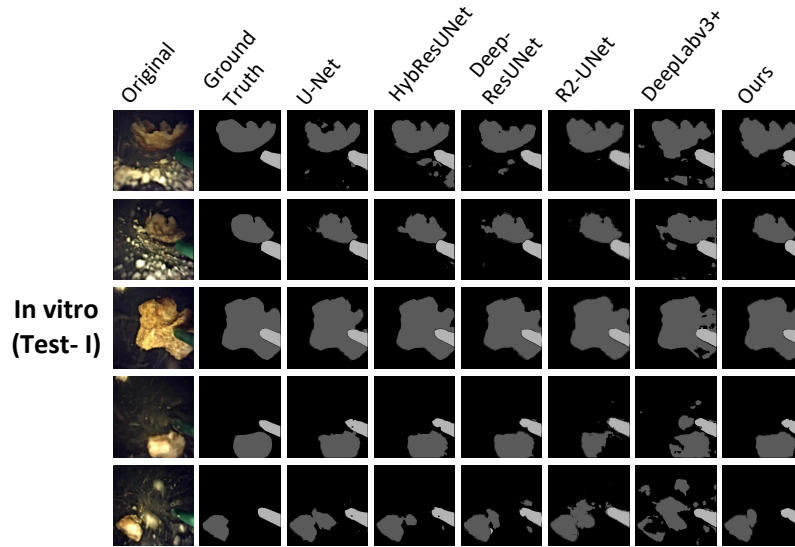
In case of the *in vitro* dataset, the mean of the two DVFs ( $DVF_{i \leftarrow (i+2)}$  and  $DVF_{(i+2) \leftarrow (i+4)}$ ) provides the input to the semantic segmentation network HybResUNet to obtain the first semantic map,  $p_i^1$  as opposed to the *in vivo* where the mean of the warped images ( $I_{warp_{i \leftarrow (i+2)}}$  and  $I_{warp_{(i+2) \leftarrow (i+4)}}$ ) is fed to the HybResUNet network to obtain the first semantic map,  $p_i^1$ . Table S5 below presents a quantitative comparison of proposed network architectures against existing approaches on our ureteroscopy and laser lithotripsy *in vitro* test set (Test-I).

- (I) **Base network:** It can be seen that HybResUNet provided a significantly higher DSC with a value of 0.8781. A higher JI, lower HD, higher overall PPV, and sensitivity were also seen as compared to other baseline networks included in the experiment (U-Net, DeepResUNet, R2UNet, DeepLabv3+, and Joint model).
- (II) **Base network with DVFNet (with DVF):** In the case of *in vitro*, DVFNet (with DVF) significantly improved the DSC, JI, and HD of HybResUNet by nearly 1.2%, 1%, and 4%, respectively. It can also be seen that the DVFNet further improved the overall PPV and sensitivity of most networks in this set.
- (III) **Base network with DVFNet (with warped image):** In the case of the *in vitro*, DVFNet (with warped image) can be seen to improve the mean of DSC and JI of Att-HybResUNet(refer to Table S4) and Att-ASPP-HybResUNet by 1% and 1.63%, respectively. However, the overall performance is still lower than that obtained from the use of DVFNet (with DVF).

Figure 1 shows that our model outperforms the existing approaches by overcoming the challenges and providing a more accurate delineation of stone and laser fiber. As evident in Figure 1, it can be clearly seen that laser fiber is nearly segmented well by all models. It can also be observed from Figure 1 that the existing approaches are not able to provide a clear segmentation of stone in most frames and hence some debris is segmented as part of the stone, resulting in either underestimation or overestimation of the stone size.

Dataset	Method	DSC		JI		HD		PPV		Sensitivity		
		Stone	Laser	Stone	Laser	Stone	Laser	Stone	Laser	Stone	Laser	
In vitro (Test-I)	I	ASPP-UNet <sup>†</sup>	0.8504	0.8656	0.7580	0.7958	6.563	3.5086	0.8845	0.8612	0.8553	0.8780
		ASPP-HybResUNet <sup>†</sup>	0.8777	0.8794	0.8015	0.8075	6.0554	3.6527	0.8917	0.8735	0.8997	0.9141
		ASPP-DeepResUNet <sup>†</sup>	0.7934	0.8647	0.6825	0.7891	7.5642	3.7578	0.7611	0.8537	0.9094	0.9006
		ASPP-R2-UNet <sup>†</sup>	0.8506	0.8832	0.7597	0.8106	6.7954	3.6081	0.8785	0.8717	0.8654	0.9192
	II	Att-UNet(Oktay et al., 2018)	0.8730	0.8488	0.7925	0.7777	6.3598	3.49	0.9021	0.8426	0.8796	0.8657
		Att-HybResUNet <sup>†</sup>	0.8844	0.8264	0.8192	0.7592	5.6916	4.0531	0.9402	0.8402	0.8730	0.8168
		Att-DeepResUNet <sup>†</sup>	0.8960	0.8491	0.8244	0.7729	5.8547	3.5731	0.9392	0.8404	0.8776	0.8746
		Att-R2-UNet(LeeJunHyun, 2019)	0.8788	0.8481	0.8007	0.7731	5.8442	3.6752	0.9118	0.8291	0.8779	0.8814
	III	Att-ASPP-UNet <sup>†</sup>	0.8301	0.8643	0.7357	0.7852	6.8324	3.8761	0.8340	0.8615	0.8733	0.8807
		Att-ASPP-HybResUNet <sup>†</sup>	0.8641	0.8693	0.7790	0.8038	6.4245	3.4374	0.8339	0.8794	0.9350	0.8788
		Att-ASPP-DeepResUNet <sup>†</sup>	0.8571	0.8578	0.7648	0.7775	6.457	3.7789	0.8788	0.8430	0.8635	0.9025
		Att-ASPP-R2-UNet <sup>†</sup>	0.8338	0.8782	0.7370	0.8056	6.6628	3.6545	0.8292	0.8577	0.8832	0.9340
In vivo (Test-I)	I	ASPP-UNet <sup>†</sup>	0.7963	0.7865	0.6854	0.6928	5.5427	3.9079	0.7991	0.7834	0.8416	0.801
		ASPP-HybResUNet <sup>†</sup>	0.8035	0.8334	0.6939	0.7453	5.7695	3.7815	0.814	0.8249	0.833	0.8558
		ASPP-DeepResUNet <sup>†</sup>	0.8234	0.7968	0.7152	0.699	5.2131	3.5706	0.8282	0.816	0.861	0.7855
		ASPP-R2-UNet <sup>†</sup>	0.7939	0.7409	0.6805	0.6453	5.5599	4.0746	0.7986	0.7261	0.8334	0.7638
	II	Att-UNet(Oktay et al., 2018)	0.8255	0.8094	0.7157	0.7292	5.3276	4.2002	0.8526	0.8325	0.8358	0.8316
		Att-HybResUNet <sup>†</sup>	0.819	0.8288	0.7388	0.7761	5.3323	4.096	0.8033	0.8763	0.8725	0.8231
		Att-DeepResUNet <sup>†</sup>	0.8119	0.7726	0.6994	0.6832	5.3048	4.5833	0.8334	0.7736	0.8141	0.7825
		Att-R2-UNet(LeeJunHyun, 2019)	0.7985	0.6966	0.6794	0.5858	5.745	4.5984	0.7902	0.7343	0.8293	0.6893
	III	Att-ASPP-UNet <sup>†</sup>	0.8079	0.8364	0.7019	0.7478	5.4895	3.6628	0.7923	0.8222	0.8712	0.8606
		Att-ASPP-HybResUNet <sup>†</sup>	0.803	0.8229	0.6842	0.7282	5.5082	3.9781	0.8391	0.8119	0.7982	0.8495
		Att-ASPP-DeepResUNet <sup>†</sup>	0.8185	0.8087	0.712	0.7071	5.3754	4.1094	0.8025	0.7999	0.87	0.8332
		Att-ASPP-R2-UNet <sup>†</sup>	0.7814	0.7913	0.6617	0.6794	5.5792	4.1598	0.7837	0.7943	0.8278	0.8032

**Table S4.** Quantitative comparison of proposed network architectures against existing approaches on our ureteroscopy and laser lithotripsy test sets (Test-I). The published networks have been accordingly cited and our experiment networks have been indicated by a <sup>†</sup>superscript.



**Figure 1.** Qualitative analysis of our proposed method (Att-HybResUNet+DVFNet (with DVF) for the *in vitro* against existing SOTA methods on our ureteroscopy and laser lithotripsy test sets (Test-I). Each row shows a test image, followed by its ground truth segmentation mask (showing laser fiber and stone), followed by SOTA approaches: UNet, HybResUNet, Deep-ResUNet, R2-UNet, and DeepLabv3+, and finally our proposed model which is Att-HybResUNet+DVFNet (with DVF) for the *in vitro*

## 5 Single vs sequence sample in the test set

Table S6 below presents a Dice score-based comparison with their corresponding computation time. Here, one can observe that using sequence samples does not show significant improvement on Dice, however, inference time exceeds 6 times that when computed with a single frame. This indicates that the weights of HybResUNet updated during training using sequence samples

<i>In vitro (Test-I)</i>							
Class	Method	DSC	<i>p</i> -values	J1	HD	PPV	Sensitivity
Stone	UNet (Ronneberger et al., 2015)	0.8521±0.19	0.0790	0.7736±0.19	6.3096±1.83	0.9155±0.11	0.8479±0.20
	HybResUNet(Peretz and Amar, 2019)	0.8838±0.14	0.1063	0.8143±0.18	6.0125±2.37	0.9123±0.10	0.8914±0.17
	DeepResUNet(Zhang et al., 2018)	0.8828±0.12	0.1781	0.8054±0.15	6.3286±1.55	0.9239±0.10	0.8732±0.15
	R2-UNet(Alom et al., 2018)	0.8215±0.20	0.0258*	0.7344±0.22	6.7252±2.08	0.8855±0.13	0.8280±0.22
	DeepLabv3+(ResNet-50)(Chen et al., 2018)	0.8004±0.18	0.0153*	0.6982±0.21	7.2936±1.96	0.8411±0.21	0.8110±0.20
	Joint model(Qin et al., 2018)	0.7484±0.17	0.0042*	0.6268±0.21	7.5695±1.77	0.7470±0.23	0.8057±0.15
	MI-UNet(Gupta et al., 2020a)	0.8391±0.15	0.0139*	0.7485±0.20	6.8246±2.66	0.8383±0.18	0.8985±0.17
	<b>HybResUNet+DVFNet (with DVF)<sup>†</sup></b>	<b>0.9068±0.09</b>	—	<b>0.8413±0.14</b>	<b>5.7454±2.11</b>	0.9299±0.10	0.9040±0.13
	ASPP-HybResUNet+DVFNet (with DVF) <sup>†</sup>	0.8253±0.14	0.0201*	0.7250±0.19	7.0101±2.17	0.8244±0.20	0.8847±0.15
	Att-HybResUNet+DVFNet (with DVF) <sup>†</sup>	0.8872±0.14	0.1821	0.8180±0.17	6.2642±2.32	0.9136±0.09	0.8919±0.17
	Att-ASPP-HybResUNet+DVFNet (with DVF) <sup>†</sup>	0.8238±0.19	0.0417*	0.7380±0.23	6.3834±2.15	0.8620±0.19	0.8585±0.20
	HybResUNet+DVFNet (with warped image) <sup>†</sup>	0.8801±0.14	0.2451	0.7985±0.18	6.2620±1.93	0.8899±0.19	0.8931±0.15
	ASPP-HybResUNet+DVFNet (with warped image) <sup>†</sup>	0.8295±0.19	0.1426	0.7447±0.22	6.5654±1.92	0.8165±0.23	0.8768±0.17
	Att-HybResUNet+DVFNet (with warped image) <sup>†</sup>	0.8899±0.13	0.3757	0.8211±0.16	5.8026±1.97	<b>0.9534±0.05</b>	0.8626±0.18
	Att-ASPP-HybResUNet+DVFNet (with warped image) <sup>†</sup>	0.8860±0.10	0.4739	0.8089±0.15	6.2607±2.10	0.8898±0.14	<b>0.9113±0.12</b>
Laser	UNet(Ronneberger et al., 2015)	0.8388±0.25	0.2771	0.7787±0.27	3.3187±1.05	0.8393±0.26	0.8408±0.24
	HybResUNet(Peretz and Amar, 2019)	0.8724±0.17	0.8292	0.8031±0.20	3.6028±1.30	0.8792±0.20	0.8780±0.15
	DeepResUNet(Zhang et al., 2018)	0.8276±0.23	0.1048	0.7541±0.25	3.6314±1.15	0.8609±0.24	0.8065±0.21
	R2-UNet(Alom et al., 2018)	0.8061±0.27	0.1080	0.7368±0.28	4.1269±1.52	0.8380±0.22	0.8013±0.28
	DeepLabv3+(ResNet-50)(Chen et al., 2018)	0.8446±0.20	0.2520	0.7717±0.23	3.8366±1.43	0.8626±0.18	0.8452±0.22
	Joint model(Qin et al., 2018)	0.7122±0.23	0.0001*	0.5943±0.23	4.6308±1.11	0.7399±0.25	0.6991±0.24
	MI-UNet(Gupta et al., 2020a)	0.8094±0.23	0.0914	0.7307±0.26	4.0275±1.68	0.7821±0.23	0.8561±0.24
	HybResUNet+DVFNet (with DVF) <sup>†</sup>	0.8702±0.14	—	0.7923±0.17	3.4938±1.16	0.8510±0.18	0.9069±0.10
	ASPP-HybResUNet+DVFNet (with DVF) <sup>†</sup>	0.8278±0.24	0.1349	0.7572±0.25	3.8447±1.27	0.8320±0.26	0.8320±0.22
	Att-HybResUNet+DVFNet (with DVF) <sup>†</sup>	<b>0.9004±0.08</b>	0.1069	<b>0.8263±0.11</b>	<b>3.2977±0.95</b>	0.8891±0.12	0.9227±0.04
	Att-ASPP-HybResUNet+DVFNet (with DVF) <sup>†</sup>	0.8475±0.22	0.3302	0.7809±0.24	3.5967±1.08	0.8326±0.24	0.8740±0.19
	HybResUNet+DVFNet (with warped image) <sup>†</sup>	0.8708±0.19	0.4718	0.7898±0.17	3.6420±1.42	<b>0.8955±0.15</b>	0.8697±0.13
	ASPP-HybResUNet+DVFNet (with warped image) <sup>†</sup>	0.8758±0.16	0.6223	0.8066±0.19	3.5898±1.27	0.8556±0.20	0.9118±0.11
	Att-HybResUNet+DVFNet (with warped image) <sup>†</sup>	0.8393±0.23	0.2714	0.7711±0.24	3.7712±1.73	0.8312±0.24	0.8536±0.22
	Att-ASPP-HybResUNet+DVFNet (with warped image) <sup>†</sup>	0.8769±0.12	0.5193	0.7984±0.16	3.6418±0.98	0.8548±0.17	<b>0.9233±0.05</b>

**Table S5.** Quantitative comparison of proposed network architectures against existing approaches on our ureteroscopy and laser lithotripsy *in vitro* test set (Test-I) where listed values represent average performances. Here, ‘I’ in the table represents performance of various baseline models, ‘II’ represents combination of the best performing baseline model (HybResUNet) and DVFNet (with DVF) under different configurations, and ‘III’ represents combination of the best performing baseline model (HybResUNet) and DVFNet(with warped image) under different configurations. The SOTA methods have been accordingly referenced and our experimental methods have been indicated by a †superscript. *p*-values that represent statistical significance between proposed method and other implementations with *p*-value < 0.05 are computed using paired t-test (highlighted by a \* superscript)

are sufficient to provide good performance in the test set. We have therefore not used sequence samples in the test dataset.

Sample type	Network (during testing)	DSC				Computation time (secs)
		<i>In vitro</i>		<i>In vivo</i>		
		Stone	Laser	Stone	Laser	
Single	HybResUNet	<b>0.9068</b>	0.8702	0.8203	0.8568	<b>0.86</b>
Sequence	HybResUNet+DVFNet	0.9067	<b>0.8731</b>	<b>0.8221</b>	<b>0.8603</b>	5.23

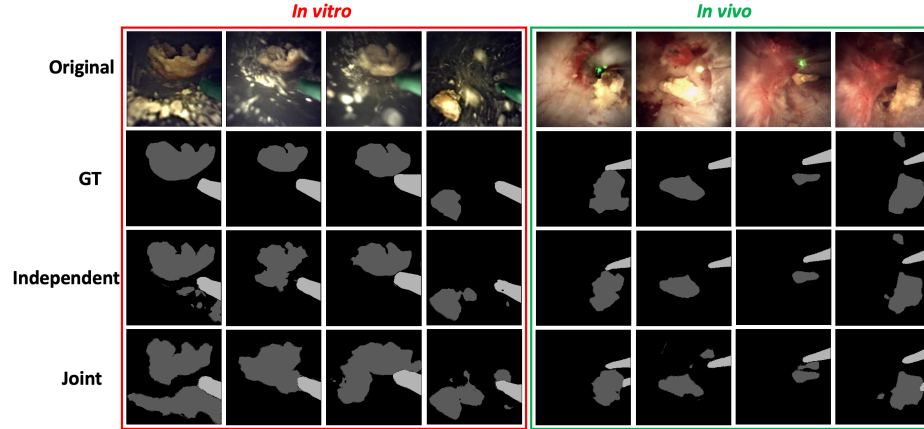
**Table S6.** Comparison of dice score and computation time (for 10 samples on NVIDIA Quadro RTX 6000 ) for HybResUNet+DVFNet applied on sequence data vs single sample data in the test dataset

## 6 Why independent runs for *In vitro* and *In vivo*

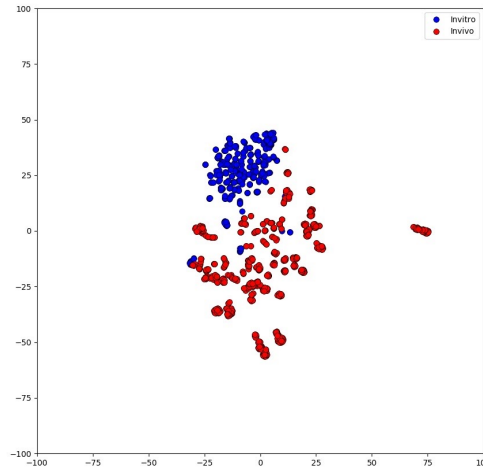
The *in vitro* dataset supports preliminary investigation using different stone shapes, sizes and composition. This allowed us to understand how the segmentation method performs under controlled settings. Clinical human kidney stones were used for *in vitro* experiments to mimic more realistic imaging conditions. Initially, we trained our model on a combination of both *in vitro* and *in vivo* datasets. Table S7 and Figure 2 presents a quantitative and qualitative comparison between two scenarios where *in vitro* and *in vivo*: i) were trained separately and ii) were trained together with HybResUNet. The performance decrease in the case of joint training can be attributed to the significant variability between the *in vitro* and *in vitro* datasets. Figure 3 shows a t-SNE plot for better visualization of the variability between *in vitro* and *in vivo* datasets. We then proceeded our study with independent runs for each of them.

Training strategy	DSC			
	<i>In vitro</i>		<i>In vivo</i>	
	Stone	Laser	Stone	Laser
Independent	<b>0.8838</b>	0.8724	<b>0.8339</b>	<b>0.8241</b>
Joint	0.8192	<b>0.9222</b>	0.8027	0.8095

**Table S7.** Quantitative comparison of independent runs against joint training of *in vitro* and *in vivo* datasets



**Figure 2.** Qualitative comparison of independent runs against joint training of *in vitro* and *in vivo* datasets



**Figure 3.** t-SNE plot showing variability between the *in vitro* and *in vivo* datasets

## 7 Test-I vs Test-II *in vivo* dataset

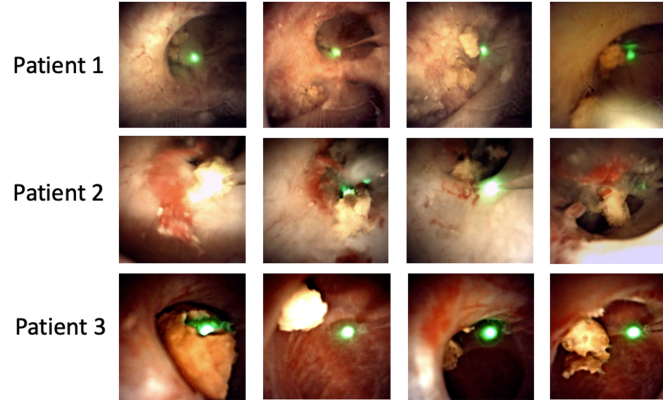
As shown in Table S8, 13 % of the combined Training, Validation, and Test-I set consists of ureteroscopy images obtained from animal studies (Setting 1) performed at Boston Scientific, and 87% consists of clinical ureteroscopy images collected at the Oxford University Hospitals. Test-II consists of unseen ureteroscopy images obtained from a new set of animal studies (Setting 2) performed at Boston Scientific. As can be seen from Table S8, the models were trained on mainly human ureteroscopy samples while the Test-II set consists of images of ureteroscopy performed on animal subjects. In addition, hardware settings such as LED illumination and software configuration settings used at the time of surgery and image acquisition are different for the animal studies in Test-I and Test-II datasets. This indicates that the trained models do not cover all the possible variance scenarios arising from different image acquisition settings, resulting in a performance decrease in the Test-II dataset.

Dataset split ( <i>In vivo</i> )	Number of sequence samples	
	From Human	From Animal
Train	81	11 (Setting 1)
Validation	32	-
Test-I	23	7 (Setting 1)
Test-II	-	20 (Setting 2)

**Table S8.** Number of human and animal samples in various splits of the *in vivo* dataset where Setting 1 and Setting 2 refers to the different hardware and software configurations used at the time of surgery and image acquisition

## 8 Inter and intra patient variability

In case of ureteroscopy, images from the same patient can have a lot of variability in terms of tissue appearance, illumination conditions, stone variability, and different viewpoints w.r.t the camera. These factors play together to create a dynamic scene in every frame. Figure 4 below illustrates random image frames from some patients and one can observe the large variability across data from the same patient. To further demonstrate that sample-level split did not cause a data leak in our test results, we have also presented quantitative results (Table S9) of patient-wise 4-fold cross validation performed on *in vivo* data. It can be seen that our proposed model outperforms state-of-the-art baseline models.



**Figure 4.** Random frames from different patients illustrating the intra and inter-patient variability across data. Here, patient 3 represents the out-of-sample test dataset used for generalisability assessment.

	Network	Fold1		Fold2		Fold3		Fold4		Average	
		DSC	JI	DSC	JI	DSC	JI	DSC	JI	DSC	JI
Stone	UNet	0.668	0.5314	0.6261	0.4825	0.5788	0.4452	0.6632	0.5321	0.634±0.04	0.4978±0.04
	HybResUNet	0.6966	0.5574	0.6152	0.4705	0.6492	0.5115	0.6572	0.5205	0.6546±0.03	0.515±0.04
	HybResUNet +DVFNet (warped img)	0.6777	0.5381	0.6223	0.4837	0.7514	0.633	0.7376	0.6034	<b>0.6973±0.06</b>	<b>0.5646±0.07</b>
Laser	UNet	0.6854	0.5649	0.5439	0.4274	0.0579	0.0358	0.8366	0.7242	0.531±0.34	0.4381±0.29
	HybResUNet	0.6922	0.5675	0.5895	0.4729	0.1223	0.0734	0.8599	0.7559	0.566±0.32	0.4674±0.29
	HybResUNet +DVFNet (warped img)	0.6682	0.543	0.6724	0.552	0.2405	0.1503	0.7874	0.6532	<b>0.5921±0.24</b>	<b>0.4746±0.22</b>

**Table S9.** Performance comparison of 4-fold patient-wise cross fold validation performed on *in vivo* data where each fold represents validation results on different subjects (Fold1, Fold2 and Fold4 are human patient data while Fold3 represents animal data).