Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

journal homepage: www.elsevier.com/locate/compmedimag

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

Soumya Gupta [a,b,*], Sharib Ali [a,b,c,f], Louise Goldsmith [e], Ben Turney [e], Jens Rittscher [a,b,c,d,*]

[a] Institute of Biomedical Engineering (IBME), Department of Engineering Science, University of Oxford, Oxford, UK
[b] Big Data Institute, University of Oxford, Li Ka Shing Centre for Health Information and Discovery, Oxford, UK
[c] Oxford NIHR Biomedical Research Centre, University of Oxford, Oxford, UK
[d] Ludwig Institute for Cancer Research, Nuffield Department of Clinical Medicine, University of Oxford, Oxford, UK
[e] Department of Urology, The Churchill, Oxford University Hospitals NHS Trust, Oxford, UK
[f] School of Computing, University of Leeds, Leeds, UK

ARTICLE INFO

Keywords:
Ureteroscopy
Laser lithotripsy
Kidney stone
Semantic segmentation
U-net
DVFNet
Deep learning

ABSTRACT

Ureteroscopy with laser lithotripsy has evolved as the most commonly used technique for the treatment of kidney stones. Automated segmentation of kidney stones and the laser fiber is an essential initial step to performing any automated quantitative analysis, particularly stone-size estimation, that can be used by the surgeon to decide if the stone requires further fragmentation. However, factors such as turbid fluid inside the cavity, specularities, motion blur due to kidney movements and camera motion, bleeding, and stone debris impact the quality of vision within the kidney, leading to extended operative times. To the best of our knowledge, this is the first attempt made towards multi-class segmentation in ureteroscopy and laser lithotripsy data. We propose an end-to-end convolution neural network (CNN) based learning framework for the segmentation of stones and laser fiber. The proposed approach utilizes two sub-networks: (I) HybResUNet, a hybrid version of residual U-Net, that uses residual connections in the encoder path of the U-Net to improve semantic predictions, and (II) a DVFNet that generates deformation vector field (DVF) predictions by leveraging motion differences between the adjacent video frames which is then used to prune the prediction maps. We also present ablation studies that combine different dilated convolutions, recurrent and residual connections, atrous spatial pyramid pooling, and attention gate models. Further, we propose a compound loss function that significantly boosts the segmentation performance in our data. We have also provided an ablation study to determine the optimal data augmentation strategy for our dataset. Our qualitative and quantitative results illustrate that our proposed method outperforms state-of-the-art methods such as UNet and DeepLabv3+ showing a DSC improvement of 4.15% and 13.34%, respectively, in our *in vivo* test dataset. We further show that our proposed model outperforms state-of-the-art methods on an unseen out-of-sample clinical dataset with a DSC improvement of 9.61%, 11%, and 5.24% over UNet, HybResUNet, and DeepLabv3+, respectively in the case of the stone class and an improvement of 31.79%, 22.15%, and 10.42% over UNet, HybResUNet, and DeepLabv3+, respectively, in case of the laser class.

1. Introduction

Kidney stones present a considerable burden for public healthcare systems, with the total healthcare expenditure for kidney stones exceeding 2 billion USD annually in the USA alone. It has a recurrence rate of 10% after one year, 50% over a period of 5–10 years, and 75% over 20 years in most patients (Alelign and Petros, 2018). Kidney stones, also known as renal calculi, are formed when crystal-forming substances separate from the urine and accumulate inside the upper urinary tract, kidney, ureter, or bladder (Alelign and Petros, 2018). Typically, stones larger than 5 mm can result in a blockage in the ureter, inducing severe pain in the abdomen and the lower back (Miller and Lingeman, 2007). Ureteroscopy has evolved into a minimally invasive routine technique for the treatment of a number of urological conditions such as urolithiasis, strictures, and hematuria (Reddy and DeFoor, 2010). Technological advancements have led to the development of low-cost single-use endoscopes with improved flexibility and image quality. The data used in this study has been acquired using single-use LithoVue™

* Corresponding authors at: Institute of Biomedical Engineering (IBME), Department of Engineering Science, University of Oxford, Oxford, UK.
E-mail addresses: soumya.gupta@eng.ox.ac.uk (S. Gupta), jens.rittscher@eng.ox.ac.uk (J. Rittscher).

https://doi.org/10.1016/j.compmedimag.2022.102112
Received 11 February 2022; Received in revised form 9 June 2022; Accepted 28 July 2022
Available online 8 August 2022

scope and Lithovue Elite scope (Boston Scientific). The procedure involves inserting a long flexible ureteroscope into the urinary tract through the urethra passing through the bladder and then into the ureter and kidney to access the kidney stones. The scope has a working channel through which tools like laser fiber can be inserted to perform laser lithotripsy, i.e., stone fragmentation using laser energy. Based on the size, location, and composition of the stone, the surgeon decides if the stone requires dusting or fragmentation and sets the laser settings accordingly (Aldoukhi et al., 2017). Fragmented stones are either left in place to clear out by themselves over time or extracted using a special wire basket. The surgeon tries to carefully target the stone centrally, rather than peripherally, in order to limit the excess heat generated in the confined spaces of the kidney or ureter (Aldoukhi et al., 2017). A ureteral stent is usually inserted to allow for the easy passage of the residual stone debris and fragments. Stones that are larger than the diameter of the ureter can require additional surgery. In order to avoid such discomfort to patients and assist clinicians in performing targeted laser treatment, estimating the size and location of the kidney stones is important. Automated segmentation is the primary step to performing any analysis of the stone fragments and laser fiber. However, compared to standard radiology, very little work has been published to address the problem of automated segmentation in ureteroscopy videos. The ureteroscopy and laser lithotripsy data is significantly different and challenging when compared to the other endoscopy datasets. The ureteroscopy video has a small-field-of-view and the signal quality is affected by stone debris originating from stone fragmentation that obscures the vision in the kidney, making it difficult for surgeons to perform the stone-treatment procedure efficiently, thereby adding to the treatment time. Intra-operative bleeding can also occur during ureteroscopy due to the continuous application of laser energy, intra-renal pressure, and the trauma caused to the walls of the ureter (De Coninck et al., 2019). In addition to the aforementioned challenges, the segmentation task becomes even more complex due to motion blur arising from unavoidable kidney movements and camera motion, specular highlights, dynamic background, varying illumination conditions, artifacts from the turbid fluid inside the target cavity (Rosa et al., 2011), and high variation in the size, shape, and composition of the stone. Sample images from our *in vitro* and *in vivo* datasets are shown in Fig. 1 wherein the stone debris and blood are highlighted in yellow and red rectangles, respectively.

To the best of our knowledge, multi-class segmentation that delineates both stones and the laser fiber in ureteroscopy videos has not been addressed before. Our previous work (Gupta et al., 2020a) proposed a combination of UNet and DVFNet (Deformation Vector Fields Net) framework that used cross-entropy and cross-correlation loss for stone segmentation only. However, segmentation of laser fiber is also important considering the fact that inaccurate laser targeting can result in undesirable excess heat and extended operative times. The study differs from the previously published work in three major ways: Improved semantic segmentation framework using residual connections; novel loss function designed as a combination of smoothness constraint, cross-correlation loss, focal loss, and boundary loss; and multi-class segmentation of stone and laser class. Here, we have experimented with various combinations of residual and recurrent connections, dilated convolutions, ASPP layers, and attention modules to identify the semantic segmentation network with competitive performance. To further improve the network performance, we make use of the temporal information in sub-sequences by incorporating a sub-network called DVFNet that leverages motion between the adjacent frames to compute end-to-end deformation vector field (DVF) predictions. This motion information is then used to prune the segmentation map obtained from the semantic segmentation network, resulting in a context-aware edge enhanced multi-class segmentation. The main contributions made in this work can be summarized as follows:

- A novel end-to-end CNN-based learning framework with residual connections that leverages motion between image pairs to overcome inevitable challenges of motion blur, stone debris, and other artifacts, and provide real-time multi-class segmentation of both stone and laser fiber in ureteroscopy and laser lithotripsy dataset.
- A novel compound loss function is proposed that outperforms traditional loss functions on ureteroscopy and laser lithotripsy dataset.
- Experimental validations on diverse and challenging *in vitro* and *in vivo* ureteroscopy datasets demonstrate the effectiveness of our proposed multi-class segmentation approach.

We compare our proposed method with different state-of-the-art (SOTA) methods to exhibit the competitiveness of the approach on both *in vitro* and *in vivo* datasets. The experiments on the *in vitro* dataset have been presented in the **Supplementary material**. We conduct an extended out-of-sample test of our network on unseen animal data and patient data to measure the robustness compared to the SOTA methods. Finally, we provide an extensive ablation study that validates our network choices and data augmentation strategies (see **Supplementary material**). The type of dataset used in this study has not been well-explored in literature and is significantly different and challenging as compared to the other endoscopy datasets (example images are provided in Fig. 1).

2. Related work

This section builds on recent advances in semantic segmentation and image registration. Of particular relevance are those segmentation and registration methods that have been developed for endoscopy imaging.

2.1. Segmentation in ureteroscopy and other endoscopy

Here, we first outline different methods that have been proposed for the segmentation of kidney stones. This is then followed by some deep learning methods used for segmentation of various abnormalities in endoscopy imaging.

Segmentation of kidney stones. Several methods such as Region indicator with Contour segmentation (RICS) (Tamilselvi and Thangaraj, 2012b), modified watershed segmentation (Tamilselvi and Thangaraj, 2012a), and squared euclidean distance method (Tamiselvi, 2013) have been implemented for the detection and segmentation of kidney stones in ultrasound (US) images. Some studies have also explored techniques such as intensity, location, and size-based thresholds (Thein et al., 2018), Fuzzy C-means clustering followed by level set (Akkasaligar et al., 2017), and CNN (Längkvist et al., 2018) for detection and segmentation of kidney stones in CT images. Rosa et al. (2011) proposed a region growing algorithm for renal calculi segmentation on ureteroscopy images. However, such approaches require user intervention to define seed pixel, similarity criterion, and a stopping criterion which is very challenging to determine due to the nature of variability in kidney stones. Gupta et al. (2020b) proposed an optical flow-based segmentation technique for binary segmentation of stone fragments in ureteroscopy. All of these methods either use traditional segmentation approaches (Rosa et al., 2011; Tamilselvi and Thangaraj, 2012b; Thein et al., 2018; Akkasaligar et al., 2017), or use unsupervised machine learning techniques combined with CNN feature extraction (Längkvist et al., 2018) resulting in low performance and very large computational time. Previously we presented an end-to-end convolutional network (Gupta et al., 2020a) that leveraged motion differences between adjacent frames to further improve the segmentation of stones in ureteroscopy and laser lithotripsy data. However, all of these methods are limited to stone segmentation and a limited set of stone types.

Segmentation of various abnormalities in endoscopy. Hand-crafted features have been applied for the segmentation and detection of various abnormalities such as bleeding (Tuba et al., 2017),

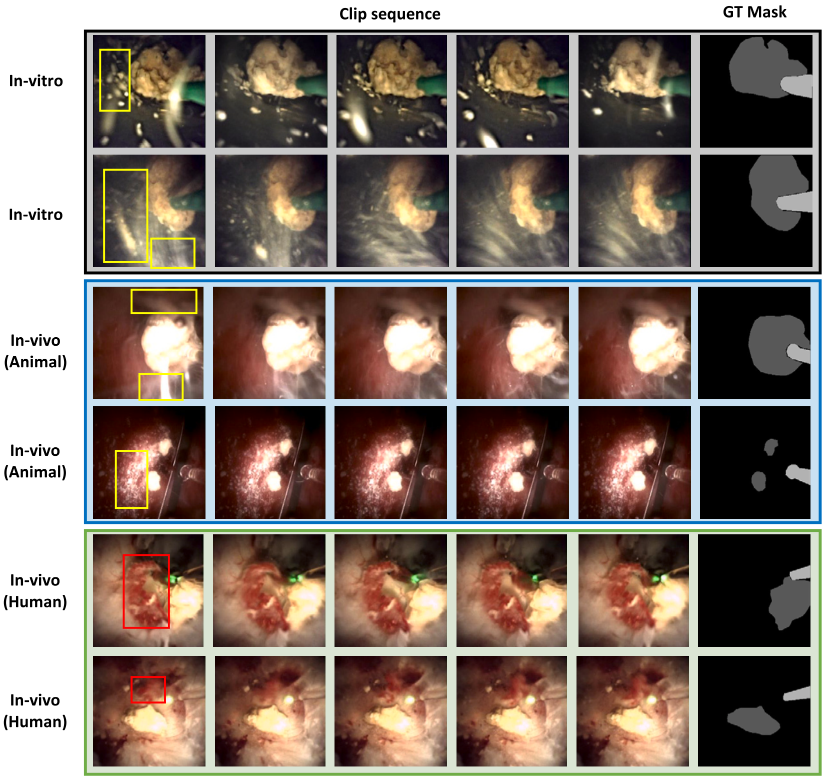

Fig. 1. Exemplary images for *in vitro*, *in vivo* (animal) and *in vivo* (human) clip sequences with corresponding ground truth masks showing stone and laser fiber. Stone debris and blood are highlighted in yellow and red rectangles (on left), respectively.

polyps (Prasath, 2017), ulcers (Yuan et al., 2015) and tumor regions (Alizadeh et al., 2014) in endoscopy videos. Various deep-learning-based strategies on automated segmentation of bleeding zones in wireless capsule endoscopy (WCE) have been proposed (Jia and Meng, 2017; Ghosh et al., 2018). Ali et al. (2021) presented a comprehensive analysis of various approaches that were submitted to EAD2020 challenge for artifact detection and segmentation and EDD2020 challenge for disease detection and segmentation. A multi-scale context-guided deep network based on FCN was proposed (Wang et al., 2020) for lesion segmentation in endoscopy images of the Gastrointestinal (GI) tract. Jha et al. (2021) presented a deep learning-based approach for real-time detection, localization, and segmentation of polyps in colonoscopy. Several groups have investigated the segmentation of various abnormalities such as bleeding (Jia and Meng, 2017; Ghosh et al., 2018) and other lesions (Wang et al., 2020) in endoscopy. However, as far as we know this work has not yet been applied to segmentation in ureteroscopy. Previously, we only considered the segmentation of kidney stones (Gupta et al., 2020a,b) ignoring the need to segment the laser fiber. It is to be noted that the laser fibers can be of different colors, sizes, and orientations (see Fig. 1). The presented study is an attempt towards multi-class segmentation of stone fragments and laser fiber in ureteroscopy and laser lithotripsy.

2.2. Advances in semantic segmentation

This section presents deep learning architectures used for segmentation with a particular focus on advancements in encoder–decoder networks applied to different medical image segmentation problems that are relevant to our work. Zhang et al. (2018) improved the performance of U-Nets by adapting a deep residual U-Net architecture (DeepResUNet) that combined the strengths of deep residual learning (He et al., 2016) and a U-Net architecture (Ronneberger et al., 2015). Peretz and Amar (2019) suggested a hybrid version of U-Nets called HybResUNet for brain tumor segmentation wherein residual blocks are only used in the encoding path of the U-Net. Alom et al. (2018) proposed a recurrent residual U-Net (R2-UNet) that uses recurrent convolutional layers with residual connectivity for improved medical image segmentation. Several studies have proposed that replacing conventional convolutions in CNN models with dilated convolutions significantly improves the network performance (Yu and Koltun, 2016; Hamaguchi et al., 2018; Piao and Liu, 2019). Hamaguchi et al. (2018) claimed that increasing dilation factors tend to increase the sparsity of the kernel and fail to aggregate local features. They then proposed a novel architecture for the segmentation of small object instances in remote satellite imagery by first increasing the dilation factors and then decreasing them. Recently, a variant of the residual UNet called ResUNet-a (Diakogiannis et al., 2020) was proposed where atrous convolutions and pyramid scene parsing pooling was incorporated in the network to improve the segmentation accuracy. Attention mechanisms have proven to be effective in highlighting only the relevant activations during training and are computationally efficient. An attention gated model (Oktay et al., 2018) was proposed and integrated into the standard U-Net for improved pancreas segmentation. This was also integrated with R2-UNet (Alom et al., 2018) for improved segmentation (LeeJunHyun, 2019). Jha et al. (2019) proposed ResUNet++ that took advantage of residual units, ASPP, and attention units to provide improved segmentation of colorectal polyps.

Methods for detecting bleeding (Jia and Meng, 2017; Ghosh et al., 2018), segmentation of various lesions such as ulcer, cancer, etc in the GI tract (Wang et al., 2020) and polyp segmentation (Jha et al., 2021) can also be investigated in the context of ureteroscopy. The ureteroscopy and laser lithotripsy data is significantly different and challenging as compared to the other endoscopy datasets in numerous aspects that include: a significant amount of stone, blood, and other debris that obscure the target; dynamic background; high variation in the appearance, size, shape, and composition of stone; specular highlights; high motion blur; and additional image artifacts from turbid fluid inside the target cavity. In this study, we have tried to overcome some of these challenges by experimenting with various combinations of residual and recurrent connections (Peretz and Amar, 2019; Alom et al., 2018; Diakogiannis et al., 2020), ASPP (Chen et al., 2018), dilated convolutions (Hamaguchi et al., 2018) and attention gate model (Oktay et al., 2018) that has been integrated into the base network U-Net to obtain an improved multi-class semantic segmentation in ureteroscopy and laser lithotripsy data.

2.3. Image registration

Several studies have shown the complementarity of image registration and segmentation tasks, meaning the features learned in image registration can be also used to improve the segmentation result (Qin et al., 2018; Mahapatra et al., 2018). To tackle limitations of simple rigid transformations, deformable image registration (DIR) methods are used for most works in medical image analysis (Cao et al., 2018; Ali and Rittscher, 2019). Some non-learning based approaches for DIR such as diffeomorphic Demons (Vercauteren et al., 2009), HAMMER (Shen and Davatzikos, 2002) and FNIRT (Andersson et al., 2008) have gained tremendous popularity. However, such traditional methods of image registration are iterative, time-consuming, and can fail when there is large variation in appearance between the source and the target images.

Supervised learning-based DIR methods. A CNN-based regression model was developed for brain MR images to directly learn the mapping between source and target images to their corresponding DVFs (Cao et al., 2018). Yang et al. (2017) also proposed a network to predict deformable registration followed by its refinement using a correction network for brain MR images. However, such neural network methods rely on strong supervision for training. The use of supervised methods are majorly limited by the fact that they require ground-truth deformation vector fields (DVFs) for model training which is difficult to obtain, especially in the case of medical datasets.

Unsupervised learning-based DIR methods. In order to handle large non-linear deformations, de Vos et al. (2019) used B-spline for transformation and interpolation for the predicted deformation fields and presented results on much complex datasets. Here, B-spline does not pass through all data points and can often lead to large interpolation errors (Ali and Rittscher, 2019). Ali and Rittscher (2019) presented an unsupervised end-to-end CNN framework for image registration that used a bicubic Catmul-Rom spline resampler to reduce the errors in the resampling of deformation fields. They also added a series of deformable convolutional filters to better capture complex deformations.

Building on the work of Ali and Rittscher (2019), our proposed framework involves a sub-network called DVFNet that leverages motion between the adjacent frames to compute end-to-end deformation vector field (DVF) predictions which are then used to prune the results of our semantic segmentation network.

3. Materials and method

This section presents a description of the dataset which is used in our study and details our proposed framework for multi-class segmentation in ureteroscopy and laser lithotripsy.

3.1. Materials

The data used in this study has been acquired using single-use LithoVue^{TM} scope and Lithovue Elite scope (Boston Scientific). The *in vitro* dataset was acquired under controlled settings wherein laser lithotripsy of four different human kidney stones was individually performed inside a container with irrigation fluid flowing through it. The *in vivo* dataset was provided by the Oxford University Hospitals and Boston Scientific. Sub-sequences containing intense lasering and stone movement were extracted and clip sequences containing 5 adjacent frames of these sub-sequences were used. One frame from each clip sequence (1 out of 5) was manually labeled using the following three class labels: stone fragments, laser fiber, and background. In Fig. 1, *in vitro*, *in vivo* (animal), and *in vivo* (human) samples have been highlighted in gray, blue, and red boxes, respectively. The VGG Image Annotator (VIA) tool (Dutta and Zisserman, 2019) was used to obtain a ground truth mask for each clip sequence as shown in Fig. 1. The annotations used in this study were performed by a Ph.D. student working on computer vision in ureteroscopy and independently verified by two experts, a senior research associate with over 5 years of experience in endoscopy and a senior urologist. A few samples did have disagreements but the senior urologist's decision was considered to be final in case of any disagreement.

The division of both *in vitro* and *in vivo* datasets into training, validation and test sets has been illustrated in Fig. 3. The *in vitro* dataset consists of 52, 18 and 18 clip sequences in train, validation and test splits, respectively while the *in vivo* consists of data from 4 subjects with 92, 32, and 30 clip sequences in train, validation, and test-I splits, respectively. As shown in Fig. 3 and mentioned in **Table S8 Supplementary material**, 13% of the combined train, validation, and test-I set consists of ureteroscopy images obtained from animal studies performed at Boston Scientific, and 87% consists of clinical ureteroscopy images collected at the Oxford University Hospitals. In the case of *in vivo*, we have also included an extra dataset (Test-II) that consists of 20 unseen new samples and is used in the final part of this study to validate our model as compared to existing SOTA approaches. Test-II consists of unseen ureteroscopy images obtained from another set of animal studies performed at Boston Scientific. An overview of the dataset configuration has also been shown in **Table S1 Supplementary material**. For a better understanding of the datasets, box-plots showing the relative size distribution of stone and laser class across training, validation, and test sets in the *in vitro* and *in vivo* datasets, respectively, has been shown in Fig. 2. Here, the size distribution, refers to the ratio of the object size in pixels to the size of the image (256 × 256). As evident from **Table S1 Supplementary material** and Fig. 2, the high standard deviation of the stone indicates high variability of kidney stones in each dataset. It can also be seen that the mean of stone is different for train, validation, and test sets, indicating that they come from different videos.

We understand that the patient-specific features can be very important and unique in certain image modalities. However, in the 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 with respect to the camera. These factors together create a dynamic scene in every frame. **Figure 4 in Supplementary material** illustrates random image frames from three patients with observable intra and inter patient variability across data. Therefore, our dataset was randomly split on sample level into train (60%), validation (20%), and test (20%) for both *in vivo* and *in vitro* data. To justify our experiments and demonstrate that no data leak has occurred in our test samples, we have: (i) evaluated model performance on 18 samples obtained from a separate unseen test patient data (see patient 3 in **Supplementary Figure 4**) that was not included in our training and held-out test dataset, and (ii) performed patient wise 4-fold cross validation on the *in vivo* data (**Table S9 Supplementary material)**.

The *in vitro* dataset supports preliminary investigation using different stone shapes, sizes, and compositions, and debris levels. This allowed us to understand the performance of segmentation methods and their ability to differentiate between background, stone, and instrument under controlled settings. Clinical human kidney stones were used for *in vitro* experiments to mimic more realistic imaging conditions. As can be seen from Fig. 2 and t-SNE plot (**Figure 3 Supplementary material**), there is a large variability between the *in vitro* and *in vivo* datasets. We, therefore, conducted independent experiments for *in vitro* and *in vivo* datasets in our work. To further justify our approach and understand the role of *in vitro data*, we trained HybResUNet model with the *in vitro* and *in vivo* datasets: trained separately and trained together (**Table S7 and Figure 2 Supplementary material**).

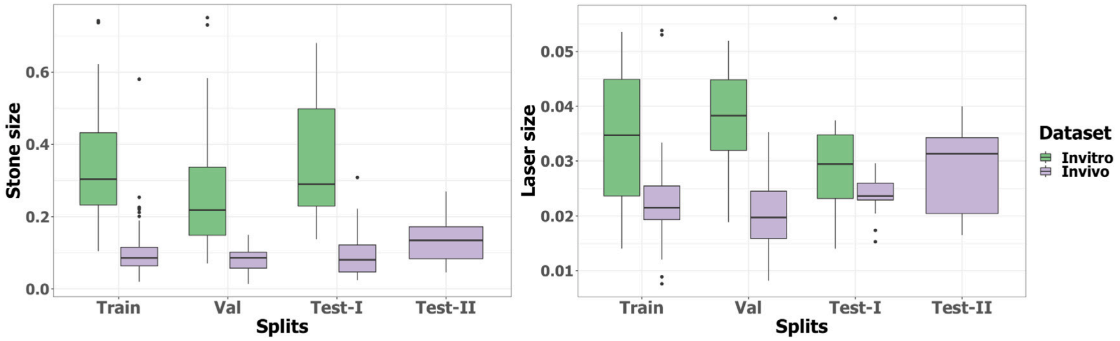

Fig. 2. Box-plots showing the size distribution of stone and laser class across training, validation, and test sets in *in vitro* and *in vivo* datasets. It can be seen that the size distribution of kidney stones is different across train, validation, and test sets, indicating that they come from different videos. The large variability between *in vitro* and *in vivo* datasets is also clearly evident, explaining why we ran experiments independently for the *in vitro* and *in vivo* datasets. All in-vitro experiments are provided in the Supplementary material.

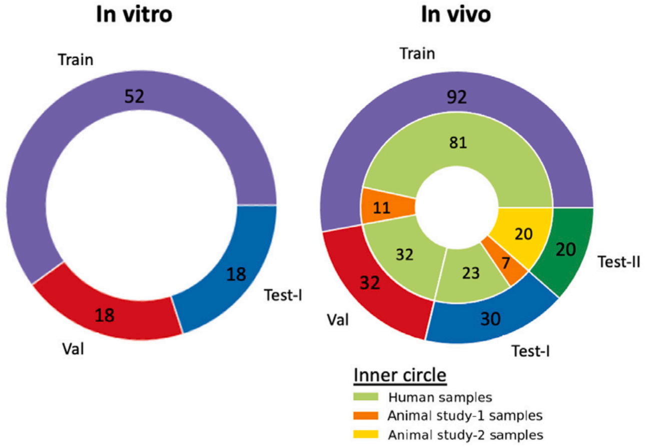

Fig. 3. Pie charts showing the division of *in vitro* and *in vivo* datasets. The proportion of both human and animal samples has also been illustrated for the *in vivo dataset*. All in-vitro experiments are provided in the Supplementary material.

3.2. Method

Our proposed framework utilizes two sub-networks: HybResUNet, which is a hybrid version of a residual U-Net that uses residual connections in the encoder path only; and a DVFNet that leverages motion between adjacent frames to compute end-to-end deformation vector fields. This motion information obtained from a DVFNet is used to prune the segmentation mask obtained from the semantic segmentation network resulting in an improved multi-class segmentation. The entire framework is designed as an end-to-end CNN model that optimizes our proposed compound loss function. In this section, the semantic segmentation network module, DVFNet, and the compound loss function are presented.

3.2.1. HybResUNet

Base network. The first sub-network of our segmentation framework is an encoder–decoder-based network called HybResUNet (Peretz and Amar, 2019). It is basically a 9-level deep U-Net architecture wherein residual blocks are used instead of traditional feed-forward units in the contracting (encoder) path as shown in Fig. 4. This is because the replacement of all feed-forward units in conventional encoder–decoder network with residual blocks increases the network complexity and tend to overfit the training data (Peretz and Amar, 2019). Each of the four residual units in the encoder consists of repeated application of two 3 × 3 convolutions, followed by a Batch Normalization (BN) and a Rectified Linear Unit (ReLU). These convolutions are followed by the addition of output to its initial input as residual units and a 2 × 2 max pooling operation with a stride of 2 for downsampling. The decoder in the HybResUNet uses transposed and regular convolutions to gradually increase the image size while reducing the number of features. The network also consists of skip connections that circumvent the information loss during downsampling by concatenating the output obtained from each residual block with the output of transposed convolution from the up-scaled features at the decoder layers. Each of these concatenations is further followed by the sequential application of two regular convolutions. For a fair comparison, other networks namely U-Net (Ronneberger et al., 2015), DeepResUNet (Zhang et al., 2018), and R2-UNet (Alom et al., 2018) are all implemented as 9 levels deep architecture to obtain the best performing base network for our dataset.

Dilations, ASPP, and Attention gate. To further improve our base network, we have tried incorporating dilated convolutions, Atrous Spatial Pyramid Pooling (ASPP), and attention gate mechanisms (Oktay et al., 2018). Replacing conventional convolutions in CNN models with dilated convolutions is known to improve the aggregation of multi-scale contextual information without losing resolution (Yu and Koltun, 2016; Hamaguchi et al., 2018; Piao and Liu, 2019). Hamaguchi et al. (2018) introduced a novel architecture for segmentation of small object instances in remote satellite imagery by first increasing the dilation factors and then decreasing them. Inspired by this idea, we have empirically obtained a series of dilation rates that work best for our data: $[1, 2, 3, 4, 3, 2, 1, 2, 1]$. These networks also consist of an Atrous Spatial Pyramid Pooling (ASPP) module at the end of the encoding path with an output stride (ratio of the input image size to the output feature map size) of 16 and six parallel convolutions with dilation rates $[1, 2, 4, 8, 16, 32]$. To sum up, the experiments labeled with ASPP in Table 3 (and **Supplementary material Table S4 and Table S5**) comprise a series of dilated convolutions and an ASPP module at the end of the encoding path. To leverage the attention gate mechanism (Oktay et al., 2018), we add attention gates to the skip connections just before the concatenation operation (refer to Att in Table 3 and **Supplementary material Table S4 and Table S5**). Such a mechanism suppresses the propagation of irrelevant and noisy responses in the network. We further tried to incorporate ASPP, series of dilated convolutions, and attention gate all together to observe if this improves the segmentation accuracy (Table 3 and **Supplementary material Table S4 and Table S5**).

3.2.2. DVFNet

Building on the work of Ali and Rittscher (2019), our DVFNet is also based on an encoder–decoder architecture where the parameters of the spline resampler are learned from training data. It consists of a total of 12 layers, that include three linear convolutional layers, two average pooling layers, and two deformable convolutional layers in the encoder. Each of these convolutional layers is combined with Batch Normalization (BN) and exponential linear unit (ELU) as shown in Fig. 4. The decoder layer consists of a Catmull–Rom spline resampler to rescale the obtained DVF from the encoder, which is then further resampled with two additional layers that consist of a convolution layer, an ELU activation, and the spline resampler. The final DVF obtained at 0th scale is then applied on the original image to obtain the corresponding warped image I_{warp}.

3.2.3. Loss function

Although commonly used, the cross-entropy loss does not differentiate between easy (correctly-classified) and hard (misclassified) samples causing the easily classified negatives in hard samples to compromise the majority of the loss and dominate the gradient (Lin et al., 2017).

Focal loss (FL), which is an improved version of CE loss was introduced by Lin et al. (2017). The focal Loss is defined as:

$$L_{FL} = -(1-\hat{p}_y)^{\gamma}\log(\hat{p}_y), \tag{1}$$

where $y \in \{0, \ldots,$ C-1$\}$ is an integer class label (C denotes the number of classes), $\hat{p} = \{(\hat{p}_0, \ldots, \hat{p}_{C-1})\} \in [0,1]^C$ is a vector representing an estimated probability distribution over the C classes and $\gamma \geq 0$ is the free focusing parameter (set to default value of 2) wherein higher the γ, the higher is the rate at which easy-to-classify examples are down-weighted. FL is a dynamically scaled CE loss where the scaling factor γ decays to zero as confidence in the correct class increases (Eq. (1)). Intuitively, the scaling factor γ automatically down-weights the easy examples and forces the model to focus on hard examples (Lin et al., 2017).

Bokhovkin and Burnaev (2019) proposed a novel loss function that is essentially a differentiable surrogate of a metric accounting accuracy of boundary detection. Let us say y_{pd} and y_{gt} represent the binary map predicted by a neural network and ground truth map, respectively for arbitrary class c for an image. The boundaries y^b_{gt} and y^b_{pd} can then be defined as:

$$y^b_{gt} = pool(1-y_{gt}, \theta_\circ) - (1-y_{gt}) \quad \text{and}$$
$$y^b_{pd} = pool(1-y_{pd}, \theta_\circ) - (1-y_{pd}), \tag{2}$$

where (1-$y_{gt,pd}$) refers to the inversion of any pixel of the map and *pool* $(\cdot,\ \cdot)$ denotes a pixel-wise max-pooling operation to the inverted binary map with a sliding window of size, $\theta_\circ$ set to 3. The euclidean distances between pixels to boundaries requires computation of a supporting map which is the map of extended boundary given by $y^{b,ext}_{gt} = pool(y^b_{gt},\theta)$ and $y^{b,ext}_{pd} = pool(y^b_{pd},\theta)$, where θ set to 5.

The precision P^c and recall R^c can then be defined as:

$$P^c = \frac{sum(y^b_{pd} \circ y^{b,ext}_{gt})}{sum(y^b_{pd})} \tag{3}$$

$$R^c = \frac{sum(y^b_{gt} \circ y^{b,ext}_{pd})}{sum(y^b_{gt})}, \tag{4}$$

where '∘' denotes the pixel-wise multiplication of two binary maps and *sum*(.) refers to the pixel-wise summation of a binary map. The reconstructed boundary metric, B^c is averaged over all classes and is then used to formulate the loss function, $L_{boundary}$ that can be defined as:

$$L_{boundary} = 1 - B^c \text{ with boundary metric } B^c = \frac{2P^cR^c}{P^c + R^c} \tag{5}$$

where P^c and R^c refer to the precision and recall. Bokhovkin and Burnaev (2019) performed a comparative analysis of their proposed boundary loss with various loss functions such as IOU loss, Dice loss, and Sensitivity–Specificity (SS) loss. In the first part of our study where we use non-sequence data, we aim to find the best performing baseline network by using a compound loss function that combines this boundary loss with SOTA Focal loss. For the second part of this study where we have integrated DVFNet for motion estimation, we propose to use our extended novel compound loss function which is a combination of Focal loss, Boundary loss, Cross-correlation loss between warped and target image, and smoothness constraint (Zhang, 2018) on the predicted deformation fields. The cross-correlation loss is computed between the warped images ($I_{warp_{i\leftarrow(i+2)}}$ and $I_{warp_{(i+2)\leftarrow(i+4)}}$) with their corresponding source images (I_i and $I_{(i+2)}$), respectively and is given by:

$$L_{sim} = L_{sim1}(I_i, I_{warp_{i\leftarrow(i+2)}}) + L_{sim2}(I_{(i+2)}, I_{warp_{(i+2)\leftarrow(i+4)}}) \tag{6}$$

i.e.,

$$L_{sim} = \frac{1}{2N}\sum\left(\frac{I_i(x)-\mu_1}{\sqrt{{\sigma_1}^2+\epsilon^2}} - \frac{I_{warp_{i\leftarrow(i+2)}}(x)-\mu_{warp_{i\leftarrow(i+2)}}}{\sqrt{{\sigma_{warp_{i\leftarrow(i+2)}}}^2+\epsilon^2}}\right)^2 +$$
$$\frac{1}{2N}\sum\left(\frac{I_{(i+2)}(x)-\mu_2}{\sqrt{{\sigma_2}^2+\epsilon^2}} - \frac{I_{warp_{(i+2)\leftarrow(i+4)}}(x)-\mu_{warp_{(i+2)\leftarrow(i+4)}}}{\sqrt{{\sigma_{warp_{(i+2)\leftarrow(i+4)}}}^2+\epsilon^2}}\right)^2, \tag{7}$$

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 estimated deformation vector fields (DVFs) can be locally smoothed using a smoothness constraint on its spatial gradients. Further, the smoothness constraint on the estimated deformation vector fields ($DVF_{i\leftarrow(i+2)}$ and $DVF_{(i+2)\leftarrow(i+4)}$) can be expressed as:

$$L_{smo} = \sum(\left\|\nabla\ DVF_{i\leftarrow(i+2)}\right\|_2^2 + \left\|\nabla\ DVF_{(i+2)\leftarrow(i+4)}\right\|_2^2), \tag{8}$$

where ∇ indicates the gradient of flow fields and $\|.\|$ indicates its L_2 norm.

Finally, our proposed compound loss function can be formulated as:

$$L = L_{FL} + \alpha L_{boundary} + \beta L_{sim} + \zeta L_{smo}, \tag{9}$$

where α, β, and ζ are the hyper-parameters used to balance the contribution from the boundary loss, similarity loss and smoothness loss, respectively with initial values set to 1, 0.5 and 1, respectively.

3.2.4. Proposed framework

Our proposed framework is shown in Fig. 4. Each clip sequence comprises of five images, I_i to $I_{(i+4)}$ wherein image pairs (I_i, $I_{(i+2)}$) and ($I_{(i+2)}$, $I_{(i+4)}$), respectively, are provided as an input in gray-scale format to different DVFNet networks during training. As shown in Fig. 4, we skip over $I_{(i+1)}$ and $I_{(i+3)}$ in order to have significant stone and debris movement across images fed into DVFNet. Each DVFNet computes a Deformation Vector Field (DVF) map ($DVF_{i\leftarrow(i+2)}$ and $DVF_{(i+2)\leftarrow(i+4)}$) and their corresponding warped image ($I_{warp_{i\leftarrow(i+2)}}$ and $I_{warp_{(i+2)\leftarrow(i+4)}}$). The obtained DVFs are locally smoothed via a smoothness constraint L_{smo} on its spatial gradients. Further, Normalized Cross Correlation (NCC) is used as a similarity metric to minimize shape differences between the obtained warped images ($I_{warp_{i\leftarrow(i+2)}}$ and $I_{warp_{(i+2)\leftarrow(i+4)}}$) and their corresponding source images I_i and $I_{(i+2)}$, respectively, as shown in Fig. 4 ($Lsim^1$ and $Lsim^2$ correspond to these losses).

As illustrated in Fig. 4, 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^1_i. A second semantic map, p^2_i is obtained by using the fifth input image $I_{(i+4)}$ in the RGB format to another HybResUNet. Finally, the two semantic maps are averaged to obtain a final map, p_i. The network then optimizes the final output semantic map by minimizing a combined loss function represented in Eq. (9). The DVFNet part of the framework is only used during network training (indicated by solid path in Fig. 4) while the learned weights of the HybResUNet are used during frame-wise inference (indicated by dotted path in Fig. 4).

Table 1 presents the number of trainable parameters for different networks discussed in this paper.

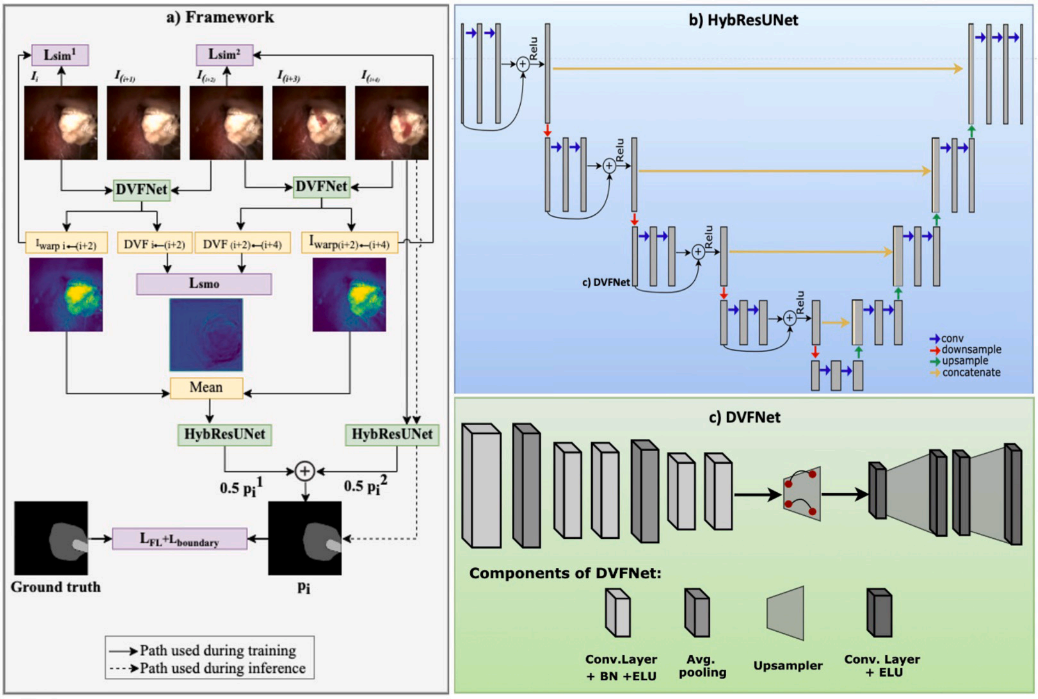

Fig. 4. Proposed framework: (a) Overall proposed semantic segmentation framework illustrating both training and inference paths presented in this study. Here, DVFNet utilizing image pairs is only used for training shown by solid path while the learned weights of HybResUNet are used during frame-wise inference indicated by the dotted line. (b) Network representation of the HybResUNet used in this work. It is a 9 level deep U-Net architecture with residual connections in the encoder path and (c) an encoder–decoder DVFNet architecture used to compute deformation vector fields (DVFs) between image pairs that are used to prune the segmentation results obtained from the HybResUNet.

4. Results

In this section, before comparing our proposed compound loss function with the SOTA focal loss (Lin et al., 2017), we outline our experimental set-up and data-augmentation strategy. Subsequently, quantitative and qualitative results of the proposed networks against existing SOTA networks on our laser lithotripsy and ureteroscopy datasets are presented. Finally, we provide an extended out-of-sample test of our proposed framework followed by a quantitative comparison of inference time of the networks. Additional details and experiments that support our algorithmic development are provided in the **Supplementary material**. This includes ablation studies for data augmentation strategies, network design and model performance on *in vitro* dataset. Also, Section 6 of the **Supplementary material** includes justification for our set experimental modes, i.e., rational behind not merging *in vitro* and *in vivo* datasets. Further, cross validation results of models trained on patient-wise split in *in vivo* data are also provided (**Supplementary material**, Section 8).

4.1. Experimental setup

All image samples were each resized to 256 × 256 pixels in a 3-channel RGB format to train the deep learning models (original image size depended on the acquisition settings of the source ranging from 394 × 392 to 1080 × 1080). Networks were trained with a batch size of 2 on NVIDIA Quadro RTX 6000 for 100 epochs using Adam optimizer with a learning rate of $1e^{-3}$, initial decay rates were set to default 0.9 and 0.999 for estimation of the first and second moments of gradient respectively, with validation performed after every epoch.

4.2. Data augmentation

Data augmentation techniques such as flips, random crops, rotate and color jittering have been commonly used for deep learning in medical imaging. In this study, we first intend to determine the optimal augmentation choices by studying the effect of different strategies on segmentation accuracy. Initially, we performed 8 training experiments for both *in vitro* and *in vivo* datasets, where each experiment involved training HybResUNet on one of the aforementioned augmentation types. Refer to the **Supplementary material Table S2** for the list of augmentation techniques with their corresponding settings used in our study. Dice similarity coefficient (DSC) was recorded for each experiment for the stone and laser-class (refer to the **Supplementary material Table S3**). It can be seen that the *Random Brightness Contrast(RBC)* and *Equalize* transformation improve the segmentation accuracy in the *in vitro* datasets as compared to no augmentation scenario(refer to the **Supplementary material Table S3**). On the other hand, *RBC* and *Contrast Limited Adaptive Histogram Equalization (CLAHE)* both provide a higher DSC compared to no augmentation in the *in vivo* datasets (refer to the **Supplementary material Table S3**). The difference in the results of the *in vitro* and *in vivo* can be majorly attributed to the difference in background appearance between the two datasets (refer to Fig. 1). Further experiments in this study have therefore used *RBC+Equalize* and *RBC+CLAHE* for augmentation of the *in vitro* and *in vivo* datasets, respectively.

Table 1
Number of trainable parameters for different CNN networks explored in this work.

Network	# Trainable params
UNet (Ronneberger et al., 2015)	31,390,851
HybResUNet (Peretz and Amar, 2019)	31,564,995
DeepResUNet (Zhang et al., 2018)	32,613,193
R2-UNet (Alom et al., 2018)	39,091,523
HybResUNet+DVFNet[a]	31,916,421
ASPP-HybResUNet+DVFNet[a]	59,194,757
Att-HybResUNet+DVFNet[a]	32,442,033
Att-ASPP-HybResUNet+DVFNet[a]	59,720,369

[a] The SOTA methods have been accordingly referenced and our experimental methods.

Table 2
Comparison of loss functions showing accuracy improvement with compound loss as opposed to state-of-the-art focal loss.

Network	Loss function	DSC					
		In vitro			*In vivo*		
		Stone	Laser	**Mean**	Stone	Laser	**Mean**
UNet	L_{FL}	0.8544	0.7643	0.8094	0.7257	0.6991	0.7124
	L_{FL}+$L_{boundary}$	0.8631	0.8401	0.8516	0.7948	0.7657	0.7803
	L_{FL}+ $L_{boundary}$+L_{sim}+L_{smo}	0.8892	0.8582	0.8737	0.7825	0.8144	0.7985
HybResUNet	L_{FL}	0.8698	0.8026	0.8362	0.7712	0.7110	0.7411
	L_{FL}+$L_{boundary}$	0.9037	**0.8854**	0.8945	0.8011	0.8100	0.8055
	L_{FL}+ $L_{boundary}$+L_{sim}+L_{smo}	**0.9055**	0.8842	**0.8948**	**0.8251**	**0.8413**	**0.8332**

4.3. *Loss function*

We introduce a novel compound loss function that is a combination of focal loss, boundary loss, cross-correlation loss, and a smoothness loss for improved segmentation performance as described in Section 3.2.3. Table 2 presents a quantitative comparison of the segmentation results obtained when a U-Net and HybResUNet was trained with the focal loss vs when trained with a combination of other losses in the case of both *in vitro* and *in vivo* test datasets.

It can be observed for UNet that the combination of focal loss and boundary loss improved the mean DSC by nearly 5.2% and 9.5% in the case of *in vitro* and *in vivo* datasets, respectively. While for the HybResUNet, this improvement was recorded to be 6.5% and 7.9% for *in vitro* and *in vivo* datasets, respectively. It can also be seen that incorporating similarity loss and smoothness loss to this combination of focal and boundary loss further boosted the mean DSC by 2.6% and 2.3% for the UNet and in case of HybResUNet, it provided a DSC boost of nearly 0.03% and 3.3% in case of *in vitro* and *in vivo* datasets, respectively. When combined losses are used instead of focal loss, the DSC for the laser class is boosted by a higher margin as compared to the boost in the stone class.

4.4. *Quantitative results*

We have evaluated our proposed method and compared it with other existing SOTA approaches using standard computer vision metrics such as Dice similarity coefficient (DSC), Jaccard index (JI), Hausdorff distance (HD), positive predictive value (PPV) and sensitivity (defined per pixel) in Table 3. We have established a quantitative comparison of our proposed framework against SOTA network architectures for the *in vivo* dataset in Table 3 (and for *in vitro* dataset in **Table S5 Supplementary material**). Similarly, Our **Supplementary material Table S4** presents an ablation study for integration of dilations, ASPP, and attention gate in our network.

(I) **Base network**: SOTA methods are compared as the base network of choice for our proposed assembled network. For this case, it can be observed that the HybResUNet provided a significantly higher DSC with a value of 0.829 for the *in vivo* dataset. 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 this set of experiments, we propose to incorporate DVFNet together with the best base network in (I), i.e., HybResUNet. For this network, the mean of $DVF_{i\leftarrow(i+2)}$ and $DVF_{(i+2)\leftarrow(i+4)}$ is fed as input to the first HybResUnet as shown in Fig. 4. Although DVFNet (with DVF) can be seen to improve the stone segmentation of HybResUNet in *the in vivo*, it showed no overall improvement in the segmentation results.

(III) **Base network with DVFNet (with warped images)**: This set of experiments involved the incorporation of DVFNet (with warped image) together with the best base network HybResUNet and its derivatives. Here, warped image corresponds to the case when mean of $I_{warp_{i\leftarrow(i+2)}}$ and $I_{warp_{(i+2)\leftarrow(i+4)}}$ is fed to the input of the first HybResUNet as shown in Fig. 4. For the *in vivo* data, it can be observed that DVFNet (with warped image) improved the performance of all networks: HybResUNet, ASPP-HybResUNet, Att-HybResUNet, Att-ASPP-HybResUNet, particularly for HybResUNet where DSC, JI, and HD were improved by nearly 1.15%, 2.2%, and 2.83%, respectively. In addition to this, DVFNet (with warped image) can be seen to increase the sensitivity of all networks in this set.

4.5. *Qualitative results*

In this section, we have presented the qualitative results of segmentation obtained from our proposed framework as opposed to the ground truth and other SOTA approaches in the *in vivo* dataset (qualitative results on the *in vitro* dataset presented in **Figure 1 Supplementary material**). Fig. 5 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 Fig. 5, it can be clearly seen that laser fiber is nearly segmented well by all models except for some difficult frames like the first image wherein it is only our model that is able to clearly segment the laser. It can also be observed from Fig. 5 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.

4.6. *Extended out-of-sample assessment*

We evaluate our proposed *in vivo* framework on a separate out-of-sample dataset (Test-II) which is acquired from a second site at Boston Scientific. We established our proposed method comparison with existing SOTA approaches. As shown in Table 4, the proposed model HybResUNet+DVFNet (with warped image) provided a significant DSC improvement of 9.61%, 11%, 8.31%, 9.86% and 5.24% over UNet, HybResUNet, DeepResUNet, R2-UNet and DeepLabv3+, respectively in case of the stone class. While for the laser class, DSC saw an improvement of 31.79%, 22.15%, 30.13%, 14.52% and 10.42% over UNet, HybResUNet, DeepResUNet, R2-UNet and DeepLabv3+, respectively. Further, we can also see from Table 4 that the ASPP module further improved the performance of HybResUNet+DVFNet by 1.62% in dice score and 1.74% in Jaccard index for the laser class. Qualitative results in Fig. 6 also demonstrate that our proposed model is able to provide the most accurate segmentation of all frames as compared to the other existing approaches.

To justify our experiments, we also evaluated our trained models on 18 samples obtained from a separate unseen test patient data (exemplary samples of patient 3 are shown in **Supplementary Figure 4**). It can be seen in Table 5 and Fig. 7 that our proposed approach on out-of-sample patient data outperforms SOTA approaches. Additionally, in order to demonstrate that no data leak has occurred in our test samples, we also performed a patient wise 4-fold cross validation on the *in vivo* data and our proposed approach provided an average (across folds) DSC improvement of 9.98% and 11.5% over SOTA UNet for the stone and laser classes, respectively (**Table S9 Supplementary material**).

Table 3
Quantitative comparison of proposed network architectures against existing approaches on our ureteroscopy and laser lithotripsy *in vivo* 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.

In vivo (Test-I)

Class		Method	DSC	*p*-values	JI	HD	PPV	Sensitivity
Stone	I	UNet (Ronneberger et al., 2015)	0.8129 ± 0.13	0.5556	0.7025 ± 0.16	5.4626 ± 1.40	0.8312 ± 0.19	0.8374 ± 0.12
		HybResUNet (Peretz and Amar, 2019)	0.8339 ± 0.15	0.2862	**0.7381** ± 0.17	5.3870 ± 1.75	**0.8462** ± 0.21	0.8525 ± 0.09
		DeepResUNet (Zhang et al., 2018)	0.8214 ± 0.13	0.8819	0.7140 ± 0.15	5.4958 ± 1.33	0.8167 ± 0.18	0.8543 ± 0.11
		R2-UNet (Alom et al., 2018)	0.7734 ± 0.16	0.0162*	0.6536 ± 0.18	6.0680 ± 1.35	0.7580 ± 0.21	0.8328 ± 0.11
		DeepLabv3+(ResNet-50) (Chen et al., 2018)	0.7653 ± 0.16	0.0008*	0.6438 ± 0.19	5.7150 ± 1.32	0.7573 ± 0.24	0.8375 ± 0.13
		Joint model (Qin et al., 2018)	0.6900 ± 0.22	0.0007*	0.5660 ± 0.23	5.8173 ± 1.44	0.7218 ± 0.26	0.7003 ± 0.23
		MI-UNet (Gupta et al., 2020a)	0.7126 ± 0.20	0.0017*	0.5852 ± 0.21	6.0840 ± 1.43	0.6875 ± 0.24	0.7938 ± 0.19
	II	HybResUNet+DVFNet (with DVF)[a]	**0.8347** ± 0.14	0.3036	0.7357 ± 0.17	5.3560 ± 1.47	0.8429 ± 0.19	0.8473 ± 0.09
		ASPP-HybResUNet+DVFNet (with DVF)[a]	0.8115 ± 0.18	0.4706	0.7127 ± 0.20	**5.3123** ± 1.73	0.8148 ± 0.23	0.8430 ± 0.15
		Att-HybResNet+DVFNet (with DVF)[a]	0.7996 ± 0.16	0.2450	0.6914 ± 0.19	5.6316 ± 1.47	0.7506 ± 0.22	**0.9026** ± 0.09
		Att-ASPP-HybResUNet+DVFNet (with DVF)[a]	0.8072 ± 0.17	0.3592	0.7036 ± 0.19	5.4765 ± 1.60	0.7881 ± 0.22	0.8668 ± 0.13
	III	**HybResUNet+DVFNet (with warped image)**[a]	0.8203 ± 0.14	–	0.7158 ± 0.17	5.4264 ± 1.43	0.8226 ± 0.20	0.8562 ± 0.12
		ASPP-HybResUNet+DVFNet (with warped image)[a]	0.7992 ± 0.16	0.1227	0.6911 ± 0.19	5.5953 ± 1.55	0.7817 ± 0.23	0.8682 ± 0.11
		Att-HybResUNet+DVFNet (with warped image)[a]	0.8183 ± 0.16	0.8483	0.7156 ± 0.18	5.5431 ± 1.65	0.8032 ± 0.21	0.8713 ± 0.09
		Att-ASPP-HybResUNet+DVFNet (with warped image)[a]	0.8016 ± 0.15	0.0975	0.6918 ± 0.18	5.3543 ± 1.59	0.8062 ± 0.21	0.8382 ± 0.12
Laser	I	UNet (Ronneberger et al., 2015)	0.7974 ± 0.21	0.0548*	0.7043 ± 0.24	4.1557 ± 1.48	0.8050 ± 0.24	0.8137 ± 0.20
		HybResUNet (Peretz and Amar, 2019)	0.8241 ± 0.18	0.0331*	0.7328 ± 0.21	3.9540 ± 1.35	0.8640 ± 0.17	0.8214 ± 0.20
		DeepResUNet (Zhang et al., 2018)	0.7851 ± 0.27	0.0277*	0.7023 ± 0.26	4.2869 ± 1.40	0.8084 ± 0.27	0.7783 ± 0.27
		R2-UNet (Alom et al., 2018)	0.7678 ± 0.22	0.0315*	0.6636 ± 0.23	4.4229 ± 1.14	0.7875 ± 0.22	0.7979 ± 0.23
		DeepLabv3+(ResNet-50) (Chen et al., 2018)	0.7144 ± 0.29	0.0148*	0.6200 ± 0.29	4.4581 ± 1.06	0.7287 ± 0.29	0.7111 ± 0.29
		Joint model (Qin et al., 2018)	0.6348 ± 0.30	0.0001*	0.5271 ± 0.29	4.7468 ± 0.98	0.6991 ± 0.28	0.6111 ± 0.31
		MI-UNet (Gupta et al., 2020a)	0.7249 ± 0.22	0.0009*	0.6082 ± 0.24	4.8526 ± 1.31	0.7445 ± 0.23	0.7635 ± 0.19
	II	HybResUNet+DVFNet (with DVF)[a]	0.7834 ± 0.26	0.0245*	0.6965 ± 0.25	4.3907 ± 1.41	0.7748 ± 0.23	0.8409 ± 0.27
		ASPP-HybResUNet+DVFNet (with DVF)[a]	0.8048 ± 0.18	0.0999	0.7034 ± 0.21	4.4334 ± 1.04	0.7826 ± 0.17	0.8385 ± 0.20
		Att-HybResUNet+DVFNet (with DVF)[a]	0.8017 ± 0.24	0.0234*	0.7144 ± 0.23	4.3347 ± 1.20	0.8352 ± 0.19	0.8300 ± 0.26
		Att-ASPP-HybResUNet+DVFNet (with DVF)[a]	0.8374 ± 0.13	0.5780	0.7397 ± 0.17	4.2352 ± 0.81	0.7979 ± 0.17	**0.8975** ± 0.12
	III	HybResUNet+DVFNet (with warped image)[a]	0.8568 ± 0.21	–	**0.7878** ± 0.21	3.6501 ± 0.99	**0.8894** ± 0.18	0.8487 ± 0.22
		ASPP-HybResUNet+DVFNet (with warped image)[a]	**0.8658** ± 0.15	0.7439	0.7852 ± 0.17	**3.6361** ± 0.64	0.8634 ± 0.16	0.8770 ± 0.15
		Att-HybResUNet+DVFNet (with warped image)[a]	0.8389 ± 0.19	0.0701	0.7558 ± 0.20	3.9573 ± 0.95	0.8798 ± 0.17	0.8370 ± 0.21
		Att-ASPP-HybResUNet+DVFNet (with warped image)[a]	0.8478 ± 0.18	0.7321	0.7662 ± 0.20	3.9358 ± 0.90	0.8276 ± 0.20	0.8784 ± 0.18

**p*-values that represent statistical significance between proposed method and other implementations with *p*-value < 0.05 are computed using paired t-test.

[a] The SOTA methods have been accordingly referenced and our experimental methods.

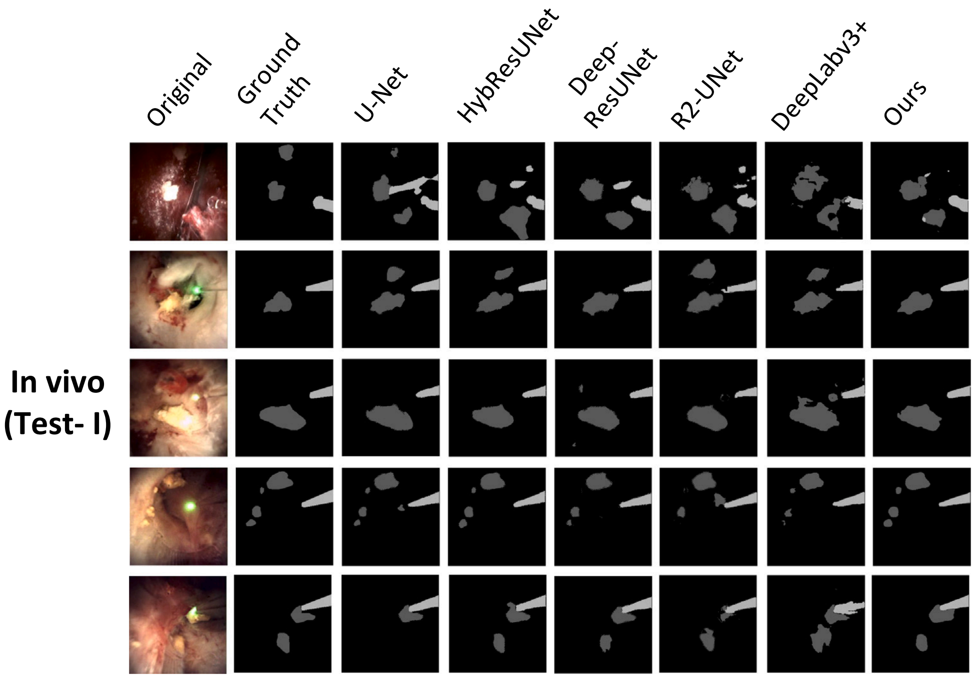

Fig. 5. Qualitative analysis of our proposed method HybResUNet+DVFNet (with warped image) for the *in vivo* against existing SOTA methods on our ureteroscopy and laser lithotripsy test sets (Test-I). Each row shows a test image, followed by its ground truth segmentation mask (showing laser fiber and stone), followed by SOTA approaches: UNet, HybResUNet, Deep-ResUnet, R2-UNet, and DeepLabv3+, and finally our proposed model which is HybResUNet+DVFNet (with warped image).

Table 4
Quantitative comparison of proposed network architecture for *in vivo* data against existing approaches on an unseen *in vivo* test dataset (Test-II) where listed values represent average performances.

In vivo (Test-II)							
Class	Method	DSC	*p*-values	JI	HD	PPV	Sensitivity
Stone	UNet (Ronneberger et al., 2015)	0.3516 ± 0.22	0.3185	0.2382 ± 0.19	8.8137 ± 1.69	0.4192 ± 0.30	0.3426 ± 0.22
	HybResUNet (Peretz and Amar, 2019)	0.3471 ± 0.25	0.0400*	0.2413 ± 0.21	9.0627 ± 1.95	0.3977 ± 0.31	0.3451 ± 0.24
	DeepResUNet (Zhang et al., 2018)	0.3558 ± 0.24	0.1138	0.2471 ± 0.21	8.5908 ± 1.63	0.4071 ± 0.31	0.3518 ± 0.23
	R2-UNet (Alom et al., 2018)	0.3508 ± 0.23	0.0954	0.2406 ± 0.20	8.9529 ± 1.73	0.4221 ± 0.30	0.3513 ± 0.26
	DeepLabv3+(ResNet-50) (Chen et al., 2018)	0.3662 ± 0.24	0.0062*	0.2520 ± 0.20	8.6843 ± 1.73	0.4288 ± 0.30	0.3480 ± 0.21
	HybResUNet+DVFNet (with warped image)†	**0.3854** ± 0.25	0.1227	0.2728 ± 0.22	8.6852 ± 1.90	0.4262 ± 0.29	**0.3873** ± 0.25
	ASPP-HybResUNet+DVFNet (with warped image)†	0.3852 ± 0.25	–	**0.2733** ± 0.22	**8.5904** ± 1.80	**0.4383** ± 0.31	0.3775 ± 0.24
	Att-HybResUNet+DVFNet (with warped image)†	0.3757 ± 0.25	0.1599	0.2656 ± 0.22	8.9544 ± 2.03	0.4228 ± 0.30	0.3794 ± 0.24
	Att-ASPP-HybResUNet+DVFNet (with warped image)†	0.3607 ± 0.25	0.8447	0.2538 ± 0.22	8.7278 ± 1.96	0.4157 ± 0.32	0.3432 ± 0.23
Laser	UNet (Ronneberger et al., 2015)	0.5564 ± 0.29	0.0284*	0.4385 ± 0.27	5.7087 ± 0.90	0.5824 ± 0.31	0.5746 ± 0.31
	HybResUNet (Peretz and Amar, 2019)	0.6003 ± 0.24	0.0988	0.4673 ± 0.22	6.0062 ± 1.11	0.6443 ± 0.29	0.5969 ± 0.25
	DeepResUNet (Zhang et al., 2018)	0.5635 ± 0.30	0.0459*	0.4465 ± 0.27	5.9400 ± 0.85	0.6255 ± 0.31	0.5452 ± 0.31
	R2-UNet (Alom et al., 2018)	0.6403 ± 0.21	0.0086*	0.5050 ± 0.22	5.4860 ± 1.10	0.6860 ± 0.25	0.6417 ± 0.23
	DeepLabv3+(ResNet-50) (Chen et al., 2018)	0.6641 ± 0.11	0.0020*	0.5071 ± 0.12	5.4833 ± 0.91	0.8078 ± 0.14	0.5854 ± 0.14
	HybResUNet+DVFNet (with warped image)†	0.7333 ± 0.13	0.7438	0.5956 ± 0.16	**5.2917** ± 1.09	0.8131 ± 0.17	**0.7057** ± 0.18
	ASPP-HybResUNet+DVFNet (with warped image)†	**0.7452** ± 0.11	–	**0.6060** ± 0.14	5.4093 ± 1.14	**0.8417** ± 0.15	0.7016 ± 0.15
	Att-HybResUNet+DVFNet (with warped image)†	0.6742 ± 0.17	0.2658	0.5332 ± 0.19	5.8803 ± 1.01	0.7300 ± 0.22	0.6732 ± 0.20
	Att-ASPP-HybResUNet+DVFNet (with warped image)†	0.7164 ± 0.14	0.1468	0.5765 ± 0.16	5.4492 ± 0.95	0.7712 ± 0.17	0.7005 ± 0.17

*p-values represent statistical significance between proposed method and other implementations with p-value < 0.05.

Table 5
Quantitative comparison of proposed network architecture for *in vivo* data against baseline methods on 18 samples from an unseen patient data where listed values represent average performances.

In vivo (Test on unseen patient data)					
Class	Method	DSC	JI	PPV	Sensitivity
Stone	UNet	0.8883 ±0.10	0.8121 ±0.15	0.8820 ±0.14	0.9103 ±0.09
	HybResUNet	0.9179 ±0.03	0.8500 ±0.06	0.9328 ±0.06	0.9091 ±0.06
	HybResUNet+DVFNet(with warped image)	0.9205 ±0.03	0.8545 ±0.06	0.9267 ±0.04	0.9180 ±0.06
Laser	UNet	0.8697 ±0.08	0.7775 ±0.11	0.8600 ±0.09	0.8850 ±0.10
	HybResUNet	0.8581 ±0.04	0.7539 ±0.06	0.8446 ±0.08	0.8800 ±0.07
	HybResUNet+DVFNet(with warped image)	0.8702 ±0.07	0.7769 ±0.11	0.8671 ±0.11	0.8843 ±0.07

Table 6
Inference time of base network for 10 images of the test set on NVIDIA Quadro RTX 6000.

Network	Computation time (s)
UNet (Ronneberger et al., 2015)	0.8427
HybResUNet (Peretz and Amar, 2019)	0.8623
DeepResUNet (Zhang et al., 2018)	0.8736
R2-UNet (Alom et al., 2018)	1.2164
DeepLabv3+(ResNet50) (Chen et al., 2018)	0.7440
HybResUNet+DVFNet	5.2272
ASPP-HybResUNet+DVFNet	5.1638
Att-HybResUNet+DVFNet	5.2366
Att-ASPP-HybResUNet+DVFNet	5.0530

4.7. Inference time analysis

In this section, we have shown a computation time analysis of different networks involved in this study, and the results are presented in Table 6. The inference time was calculated by running 10 test images on NVIDIA Quadro RTX 6000. As mentioned in Section 3.2.4, sequence samples and DVFNet are only employed during training and not used during test time. In **Supplementary material Table S6**, we have provided a comparison of DSC 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. Here, one can observe that using sequence samples does not show significant improvement with respect to Dice. However, inference time for sequence sample exceeds 6 times as compared to that when computed with a single frame. Therefore, the weights of the HybResUNet learned using sequence samples during training are used during frame-wise inference. This allows the network to perform inference in real-time.

5. Discussion

To our knowledge, this study is the first attempt to develop a multi-class segmentation approach for ureteroscopy and laser lithotripsy. In this work, we conducted experiments for *in vitro* and *in vivo* datasets independently. This is due to the large variability that exists between the two datasets, as evident from Fig. 2. **Table S7 and Figure 2** of the Supplementary material present quantitative and qualitative comparison, respectively, for both *in vitro* and *in vivo*, trained independently and jointly. The performance decrease in the case of the joint training can be attributed to the significant variability between the *in vitro* and *in vivo* datasets as illustrated by a t-SNE plot in **Supplementary material Figure 3** which shows a distinct disjoint between *in vitro* and *in vivo* dataset. The variability between the *in vitro* and *in vivo* datasets can also be seen from Fig. 1, which shows that the *in vivo* background has more textural information, higher heterogeneity, debris from tissue and blood, and other image artifacts that are almost absent in the *in vitro* dataset.

We first trained a HybResUNet (Peretz and Amar, 2019) on our *in vitro* and *in vivo* datasets by using different augmentation techniques and recorded the DSC for each class to determine the optimized augmentation strategy that best captures our target data (refer to the **Supplementary material Table S3**). It can be observed from Fig. 1 that the laser fiber is always present in a certain orientation and in the right part of the image. This explains why networks trained with spatial-level transforms such as flips, shift and rotate are not able to perform well on the test dataset. It can also be seen from **Supplementary material Table S3** that the *random brightness contrast (RBC)* and the histogram equalization in case of *in vitro*, and *RBC* and *CLAHE* in the case of *in vivo* dataset seem to either improve the segmentation accuracy or provide a competitive performance as compared to the case with no augmentations. Therefore, *RBC+Equalize* and *RBC+CLAHE* were used

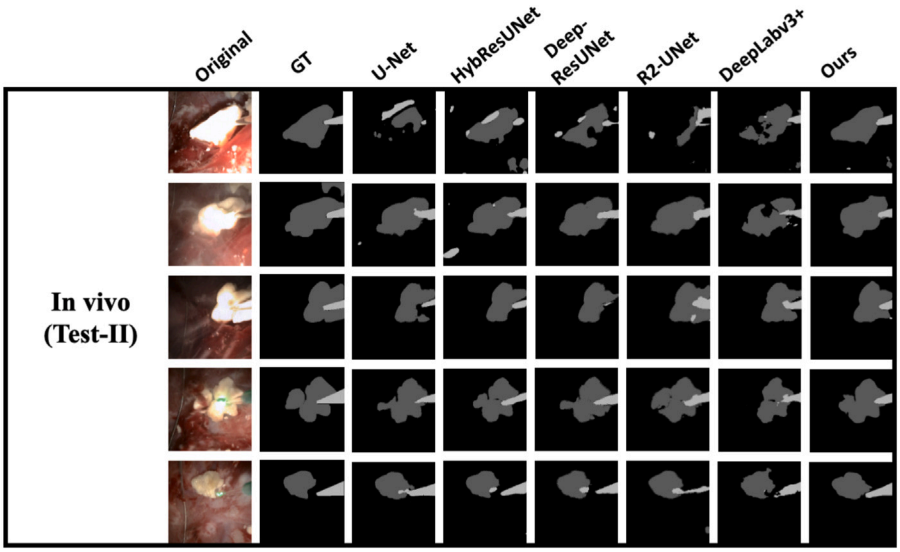

Fig. 6. Qualitative analysis of our proposed method (HybResUNet+DVFNet (with warped image)) for *in vivo* against existing SOTA methods on our unseen *in vivo* test dataset (Test-II).

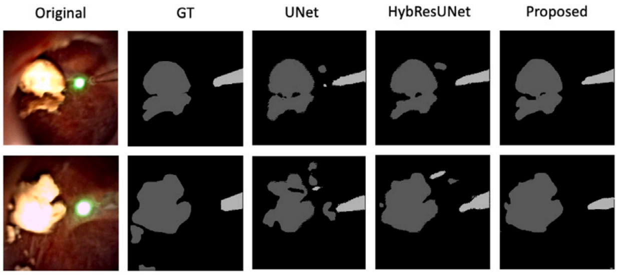

Fig. 7. Qualitative comparison of proposed network architecture for *in vivo* data against baseline methods on 18 samples from an unseen patient data (exemplary images shown in row 3 of **Figure 4 Supplementary material**).

for data augmentation in all experiments for *in vitro* and *in vivo* datasets, respectively.

A compound loss function has been used to enable better differentiation between easy or hard examples and improve the segmentation performance. In Table 2, we can see that the boundary loss when combined with focal loss leads to an improvement in the case of *in vitro* and *in vivo* datasets, respectively. The *in vitro* dataset has less scope for improvement as it has no textural information in the background and relatively more pronounced boundaries for stone and laser class as compared to the *in vivo* dataset. This explains the relatively higher improvement for the *in vivo* as compared to the *in vitro* dataset. Incorporation of similarity loss and smoothness constraint further boosted the DSC for both *in vitro* and *in vivo* datasets. It can also be noted that the accuracy of the laser class was boosted by an overall higher margin as compared to the stone class. We hypothesize that such discrepancy in the improvement percentage between stone class and laser class is possibly due to the stone class having more variability in terms of shape and texture as compared to the laser class.

HybResUNet has a relatively simpler model with residual blocks in the encoder path only as opposed to the DeepResUNet (Zhang et al., 2018) and R2-UNet (Alom et al., 2018). It can also be seen from Table 1 that the number of trainable parameters is significantly less for HybResUNet as compared to DeepResUNet and R2-UNet. We hypothesize that DeepResUNet and R2-UNet make the U-Net network more complicated and tend to overfit the training data, thereby leading to poor performance (refer to Table 3 and **Table S5 Supplementary material**) as compared to HybResUNet. Atrous Spatial Pyramid Pooling (ASPP) technique uses a series of different dilation rates in parallel to capture multi-scale contextual information. The use of dilated convolutions in an increasing order followed by decreasing order helps better aggregation of local features and improves detection of small objects (Hamaguchi et al., 2018). This explains the improvement in the Dice similarity coefficient and Jaccard index of the laser class for ASPP-HybResUNet in both *in vitro* and *in vivo* datasets (refer to the **Supplementary material Table S4**). It can also be observed that when the attention gate (AG) model is integrated into our encoder–decoder HybResUNet network, it improved the segmentation of the stone class and also improved the sensitivity by suppressing irrelevant responses in the network. However, when we tried integrating the attention gate mechanism, dilated convolutions, and ASPP, the networks tend to become overly complicated and hence we did not observe any improvement in the segmentation performance (refer to the **Supplementary material Table S4**).

To further improve the segmentation performance of our network, we integrated DVFNet into our framework. As can be observed in Table 3, DVFNet (with warped image) when integrated with HybResUNet improved its performance by 1.15%, 2.22%, and 2.83% in DSC, JI, and HD, respectively. In the case of *in vitro*, HybResUNet+DVFNet (with DVF) provided an improvement of 1.18%, 1%, and 3.9% in DSC, JI, and HD, respectively (**Supplementary material Table S5**). This network when combined with the attention gate model, Att-HybResUnet+DVFNet (with DVF), slightly improved the segmentation performance, thereby providing the best results on the *in vitro* dataset. It can also be seen in **Table S5 Supplementary material** that for the *in vitro* data, there is not much difference in the results for DVFNet (with DVF) and DVFNet (with warped image). On the other hand, it can be seen in Table 3 that for the *in vivo* samples, DVFNet (with warped image) provided significantly better results as compared to DVFNet (with DVF). Therefore, we hypothesize that as the *in vitro* lacks any background texture, it does not make a big difference as to which DVFNet approach should be adopted. But, for the *in vivo* that has the texture-rich background, DVFNet (warped image) performs better as it retains the high-level background information.

In Table 3 for the clinical data, one can observe that our proposed method provided a significant improvement (p-value < 0.05) over all SOTA baseline methods for the laser class and over most SOTA methods for the stone class. For the *in vitro* dataset, it can be seen in **Supplementary material Table S5** that in the case of the stone class, our proposed method “HybResUNet+DVFNet (with DVF)” has the least standard deviation (±0.09 in DSC) and p-value < 0.05 for most SOTA baseline methods. Similarly, for the laser class, our method achieved a standard deviation of only ±0.14 in DSC. Its variant (Att-HybResUNet+DVFNet (with DVF)) improved the performance further and showed the least standard deviation (±0.08 in DSC) as compared to the SOTA baseline methods. Our *in vitro* dataset has no textural information in the background, making it relatively easier for the networks to learn the segmentation of stone and laser class. This leaves less scope for improvement, especially for the laser class which has lesser variability as opposed to the stone class. We, therefore do not observe any statistically significant improvement (p-values > 0.05) in the segmentation of laser fiber for the *in vitro* dataset.

In Fig. 5, our proposed framework is able to outperform other existing networks by overcoming the effect of debris and other artifacts. The first image in Fig. 5 gets one of the worst-performing results across all networks. It can be observed that the small stone on the top part of the image is less visible due to relatively less light falling on it and is therefore not getting captured by any models. It can also be observed that the bright-red tissue protrusion in this image is being misclassified as a stone by all the networks. This shows that the networks are sensitive to any tissue protrusions in the images. In addition to this, it should also be observed in this image that our model is able to provide the most accurate segmentation of the laser fiber as opposed to the other networks. We hypothesize that the other models fail to segment the laser accurately due to the fact that it is transparent in appearance and the laser light is not activated making it difficult for the network to spot it. Further, the segmentation prediction on the second and third *in vivo* images show that our model is not only able to overcome the effect of stone debris but also blood and white tissue debris. Based on the results of the fourth *in vivo* image, we hypothesize that our proposed model is able to successfully pick up secondary small stones and perform well in dynamic illumination situations.

In addition, to further justify that sample-level split presented in the paper did not cause a data leak in our test results, we evaluated our trained models on unseen patient data (Table 5 and Fig. 7). It is evident that our proposed model outperforms the SOTA baseline models. Similarly, on a 4-fold patient wise cross validation using the *in vivo data*, our proposed approach showed consistent performance outperforming the SOTA methods (**Table S9 Supplementary material)**. We also performed an extended out-of-sample test study on only animal data referred to as “Test-II” of our proposed framework. These data were acquired from a new site (Boston Scientific) and the quantitative results and qualitative results are presented in Table 4 and Fig. 6, respectively. As shown in Table 4, our proposed framework outperformed the existing SOTA approaches on Test-II samples as well, thereby demonstrating that our proposed model not only outperforms other SOTA networks but also is robust to new ureteroscopy and lithotripsy datasets. Even though Table 4 shows a decrease in metric values on the Test-II as compared to Test-I dataset, it is important to note that the models were trained on mainly human ureteroscopy samples while Test-II consists of images of ureteroscopy performed on animal subjects only (**Table S8 Supplementary material**). Additionally, the hardware settings such as LED illumination and software configuration settings used at the time of surgery and image acquisition are different for some animal studies present in Test-I compared to Test-II dataset. 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. Increasing the number of animal samples in the training set and adding variability in training data, for example, in terms of kidney stone heterogeneity and background scene, can improve model robustness. Additionally, using learning-based data augmentation techniques such as generative adversarial networks can be further investigated to enhance model accuracy and robustness. Although we have tried to mimic real scenarios in our *in vitro* dataset by using real kidney stones for fragmentation, yet it has not been possible to capture every aspect of the clinical settings such as tissue background, proper flow of irrigation fluid, movement constraints on stone fragments, blood debris and other image artifacts. This possibly explains why our results and proposed framework is different for both *in vitro* and *in vivo* datasets.

Stone and laser localization and segmentation can assist clinical stone fragmenting procedures. While the video frame rate of most ureteroscopy is 30 FPS, our proposed framework achieves 12 FPS. As can be seen from Table 6, the DVFNet part of the framework requires over 6× more computation time than that for a single frame. Also, our **Supplementary material Table S6** shows DVFNet during test inference does not significantly improve the results but adversely affects the inference time. The DVFNet part of the framework, when used in conjunction with the HybResUNet during training, helps to prune the segmentation mask by adjusting/refining the weights of the network layers of HybResUNet. Therefore, even though we require a combination of HybResUNet and DVFNet for training, only HybResUNet is used during test inference minimizing the required time for segmentation of laser and stone.

6. Conclusion

We believe to be the first to present a multi-class segmentation method for ureteroscopy and laser lithotripsy imaging that is supported by a comprehensive experimental analysis. The proposed framework effectively makes use of residual connections and motion information between adjacent frames to produce robust and reliable segmentation of renal stones and laser fiber in real-time. The qualitative and quantitative results demonstrate that our algorithm can efficiently tackle the challenging vision quality within the kidney, resulting in increased segmentation accuracy as compared to the existing state-of-the-art methods. Our approach makes effective use of the temporal information within five adjacent frames only to improve the segmentation results. Future research direction includes using different forms of recurrent neural networks in order to improve the temporal information and use it to further improve the segmentation results. In order to perform accurate size estimation, the present study serves as a preliminary work that needs to be extended to 3D. The segmentation method can also be used in conjunction with depth estimation techniques in order to estimate parameters like stone size. Future work also includes a quantitative assessment of the segmented stone fragments and laser fiber in order to help the clinician gain a better understanding of the target and improve patient outcomes. Through this study, we also highlight that the clinical workload of endoscopists can be tackled by the development of medical image analysis tools that can shorten procedure time whilst improving diagnosis and therapy.

CRediT authorship contribution statement

Soumya Gupta: Conceptualization, Methodology, Software, Data curation, Writing – original draft, Writing – review & editing, Visualization, Investigation. **Sharib Ali:** Conceptualization, Methodology, Validation, Writing – original draft, Writing – review & editing, Supervision. **Louise Goldsmith:** Resources. **Ben Turney:** Resources, Validation, Supervision. **Jens Rittscher:** Conceptualization, Validation, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We would like to thank Boston Scientific (BSc) for funding this project (Grant No: DFR04690). We would also like to thank Aditi Ray, Longquan Chen, and Niraj Rauniyar from BSc for facilitating the equipment, data, and guidance. SG, LG, and BT are funded by BSc,

SA and JR were supported by NIHR Oxford Biomedical Research center and JR is also supported by Ludwig Institute for Cancer Research. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health. The computational aspects of this research were funded from the NIHR Oxford BRC with additional support from the Wellcome Trust Core Award Grant Number 203141/Z/16/Z.

Ethics approval

The two datasets of animal studies acquired by Boston Scientific were approved by Institutional Animal Care and Use Committee (IACUC protocol number: TCK379-IS98 and I00323).

Additional information

Disclaimer

Some of the data was acquired by or on behalf of Boston Scientific. Data on file using LithoVue and/or prototype devices. Concept device or technology not available for sale.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.compmedimag.2022.102112.

References

Akkasaligar, P.T., Biradar, S., Kumbar, V., 2017. Kidney stone detection in computed tomography images. In: 2017 International Conference on Smart Technologies for Smart Nation (SmartTechCon). IEEE, pp. 353–356.

Aldoukhi, A.H., Roberts, W.W., Hall, T.L., Ghani, K.R., 2017. Holmium laser lithotripsy in the new stone age: dust or bust? Front. Surg. 4, 57.

Alelign, T., Petros, B., 2018. Kidney stone disease: an update on current concepts. Adv. Urol. 2018.

Ali, S., Dmitrieva, M., Ghatwary, N., Bano, S., Polat, G., Temizel, A., Krenzer, A., Hekalo, A., Guo, Y.B., Matuszewski, B., et al., 2021. Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy. Med. Image Anal. 70, 102002.

Ali, S., Rittscher, J., 2019. Conv2Warp: An unsupervised deformable image registration with continuous convolution and warping. In: International Workshop on Machine Learning in Medical Imaging. Springer, pp. 489–497.

Alizadeh, M., Zadeh, H.S., Maghsoudi, O.H., 2014. Segmentation of small bowel tumors in wireless capsule endoscopy using level set method. In: 2014 IEEE 27th International Symposium on Computer-Based Medical Systems. IEEE, pp. 562–563.

Alom, M.Z., Hasan, M., Yakopcic, C., Taha, T.M., Asari, V.K., 2018. Recurrent residual convolutional neural network based on U-Net (R2-UNet) for medical image segmentation. arXiv preprint arXiv:1802.06955.

Andersson, J., Smith, S., Jenkinson, M., 2008. FNIRT–FMRIB's non-linear image registration tool. Hum. Brain Mapp. 2008.

Bokhovkin, A., Burnaev, E., 2019. Boundary loss for remote sensing imagery semantic segmentation. In: International Symposium on Neural Networks. Springer, pp. 388–401.

Cao, X., Yang, J., Zhang, J., Wang, Q., Yap, P.-T., Shen, D., 2018. Deformable image registration using a cue-aware deep regression network. IEEE Trans. Biomed. Eng. 65 (9), 1900–1911.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV).

De Coninck, V., Keller, E.X., Somani, B., Giusti, G., Proietti, S., Rodriguez-Socarras, M., Rodríguez-Monsalve, M., Doizi, S., Ventimiglia, E., Traxer, O., 2019. Complications of ureteroscopy: a complete overview. World J. Urol. 1–20.

de Vos, B.D., Berendsen, F.F., Viergever, M.A., Sokooti, H., Staring, M., Išgum, I., 2019. A deep learning framework for unsupervised affine and deformable image registration. Med. Image Anal. 52, 128–143.

Diakogiannis, F.I., Waldner, F., Caccetta, P., Wu, C., 2020. Resunet-a: a deep learning framework for semantic segmentation of remotely sensed data. ISPRS J. Photogramm. Remote Sens. 162, 94–114.

Dutta, A., Zisserman, A., 2019. The VIA annotation software for images, audio and video. In: Proceedings of the 27th ACM International Conference on Multimedia. http://dx.doi.org/10.1145/3343031.3350535.

Ghosh, T., Li, L., Chakareski, J., 2018. Effective deep learning for semantic segmentation based bleeding zone detection in capsule endoscopy images. In: 2018 25th IEEE International Conference on Image Processing (ICIP). IEEE, pp. 3034–3038.

Gupta, S., Ali, S., Goldsmith, L., Turney, B., Rittscher, J., 2020a. MI-UNet: Improved segmentation in ureteroscopy. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). IEEE, pp. 212–216.

Gupta, S., Ali, S., Goldsmith, L., Turney, B., Rittscher, J., 2020b. Motion induced segmentation of stone fragments in ureteroscopy video. In: Medical Imaging 2020: Image-Guided Procedures, Robotic Interventions, and Modeling, Vol. 11315. International Society for Optics and Photonics, 1131514.

Hamaguchi, R., Fujita, A., Nemoto, K., Imaizumi, T., Hikosaka, S., 2018. Effective use of dilated convolutions for segmenting small object instances in remote sensing imagery. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, pp. 1442–1450.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Jha, D., Ali, S., Johansen, H.D., Johansen, D.D., Rittscher, J., Riegler, M.A., Halvorsen, P., 2021. Real-time polyp detection, localization and segmentation in colonoscopy using deep learning. IEEE Access 9, 40496–40510.

Jha, D., Smedsrud, P.H., Riegler, M.A., Johansen, D., De Lange, T., Halvorsen, P., Johansen, H.D., 2019. Resunet++: An advanced architecture for medical image segmentation. In: 2019 IEEE International Symposium on Multimedia (ISM). IEEE, pp. 225–2255.

Jia, X., Meng, M.Q.-H., 2017. A study on automated segmentation of blood regions in wireless capsule endoscopy images using fully convolutional networks. In: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017). IEEE, pp. 179–182.

Längkvist, M., Jendeberg, J., Thunberg, P., Loutfi, A., Lidén, M., 2018. Computer aided detection of ureteral stones in thin slice computed tomography volumes using convolutional neural networks. Comput. Biol. Med. 97, 153–160.

LeeJunHyun, 2019. Image-segmentation. https://github.com/LeeJunHyun/Image_Segmentation.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988.

Mahapatra, D., Ge, Z., Sedai, S., Chakravorty, R., 2018. Joint registration and segmentation of xray images using generative adversarial networks. In: International Workshop on Machine Learning in Medical Imaging. Springer, pp. 73–80.

Miller, N.L., Lingeman, J.E., 2007. Management of kidney stones. Bmj 334 (7591), 468–472.

Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al., 2018. Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999.

Peretz, G., Amar, E., 2019. Brain tumor segmentation. Available at https://github.com/galprz/brain-tumor-segmentation.

Piao, S., Liu, J., 2019. Accuracy improvement of unet based on dilated convolution. J. Phys. Conf. Ser. 052066, IOP Publishing.

Prasath, V., 2017. Polyp detection and segmentation from video capsule endoscopy: A review. J. Imaging 3 (1), 1.

Qin, C., Bai, W., Schlemper, J., Petersen, S.E., Piechnik, S.K., Neubauer, S., Rueckert, D., 2018. Joint learning of motion estimation and segmentation for cardiac MR image sequences. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 472–480.

Reddy, P.P., DeFoor, W.R., 2010. Ureteroscopy: The standard of care in the management of upper tract urolithiasis in children. Indian J. Urol. IJU: J. Urol. Soc. India 26 (4), 555.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.

Rosa, B., Mozer, P., Szewczyk, J., 2011. An algorithm for calculi segmentation on ureteroscopic images. Int. J. Comput. Assist. Radiol. Surg. 6 (2), 237–246.

Shen, D., Davatzikos, C., 2002. HAMMER: hierarchical attribute matching mechanism for elastic registration. IEEE Trans. Med. Imaging 21 (11), 1421–1439.

Tamilselvi, P., Thangaraj, P., 2012a. A modified watershed segmentation method to segment renal calculi in ultrasound kidney images. Int. J. Intell. Inf. Technol. 8 (1), 46–61.

Tamilselvi, P., Thangaraj, P., 2012b. Segmentation of calculi from ultrasound kidney images by region indicator with contour segmentation method. Glob. J. Comput. Sci. Technol..

Tamiselvi, P., 2013. Segmentation of renal calculi using squared euclidean distance method. Int. J. Sci. Eng. Technol. 2 (7), 651–655.

Thein, N., Nugroho, H.A., Adji, T.B., Hamamoto, K., 2018. An image preprocessing method for kidney stone segmentation in CT scan images. In: 2018 International Conference on Computer Engineering, Network and Intelligent Multimedia (CENIM). IEEE, pp. 147–150.

Tuba, E., Tuba, M., Jovanovic, R., 2017. An algorithm for automated segmentation for bleeding detection in endoscopic images. In: 2017 International Joint Conference on Neural Networks (IJCNN). IEEE, pp. 4579–4586.

Vercauteren, T., Pennec, X., Perchant, A., Ayache, N., 2009. Diffeomorphic demons: Efficient non-parametric image registration. NeuroImage 45 (1), S61–S72.

Wang, S., Cong, Y., Zhu, H., Chen, X., Qu, L., Fan, H., Zhang, Q., Liu, M., 2020. Multi-scale Context-guided Deep Network for Automated Lesion segmentation with endoscopy images of gastrointestinal tract. IEEE J. Biomed. Health Inf..

Yang, X., Kwitt, R., Styner, M., Niethammer, M., 2017. Quicksilver: Fast predictive image registration–a deep learning approach. NeuroImage 158, 378–396.

Yu, F., Koltun, V., 2016. Multi-scale context aggregation by dilated convolutions. CoRR abs/1511.07122.

Yuan, Y., Wang, J., Li, B., Meng, M.Q.-H., 2015. Saliency based ulcer detection for wireless capsule endoscopy diagnosis. IEEE Trans. Med. Imaging 34 (10), 2046–2057.

Zhang, J., 2018. Inverse-consistent deep networks for unsupervised deformable image registration. arXiv preprint arXiv:1809.03443.

Zhang, Z., Liu, Q., Wang, Y., 2018. Road extraction by deep residual u-net. IEEE Geosci. Remote Sens. Lett. 15 (5), 749–753.