A Comparative Study of Spatio-Temporal U-Nets for Tissue Segmentation in Surgical Robotics

Aleks Attanasio[1*], *Student Member IEEE,* Chiara Alberti[2*], Bruno Scaglioni[1], *Member IEEE,* Nils Marahrens[1], *Student Member IEEE,* Alejandro F. Frangi[3], *Fellow IEEE,* Matteo Leonetti[3], *Member IEEE,* Chandra Shekhar Biyani[4], Elena De Momi[2], *Member IEEE* and Pietro Valdastri[1], *Senior Member IEEE*

***Abstract*—In surgical robotics, the ability to achieve high levels of autonomy is often limited by the complexity of the surgical scene. Autonomous interaction with soft tissues requires machines able to examine and understand the endoscopic video streams in real-time and identify the features of interest. In this work, we show the first example of spatio-temporal neural networks, based on the U-Net, aimed at segmenting soft tissues in endoscopic images. The networks, equipped with Long Short-Term Memory and Attention Gate cells, can extract the correlation between consecutive frames in an endoscopic video stream, thus enhancing the segmentation's accuracy with respect to the standard U-Net. Initially, three configurations of the spatio-temporal layers are compared to select the best architecture. Afterwards, the parameters of the network are optimised and finally the results are compared with the standard U-Net. An accuracy of $83.77\% \pm 2.18\%$ and a precision of $78.42\% \pm 7.38\%$ are achieved by implementing both Long Short Term Memory (LSTM) convolutional layers and Attention Gate blocks. The results, although originated in the context of surgical tissue retraction, could benefit many autonomous tasks such as ablation, suturing and debridement.**

***Index Terms*—Medical Robotics, Computer Assisted Interventions, Minimally Invasive Surgery, Surgical Vision**

I. Introduction

Compared to open surgery, Robotic Minimally Invasive Surgery (rMIS) provides substantial benefits to the patient, such as reduced blood loss, decreased tissue trauma and shortened post-operative recovery. Although manual laparoscopy offers similar advantages, the skills required to perform complex procedures with manually manipulated instruments demand expensive and time-consuming training for surgeons. The use of such instruments significantly increases the cognitive load, with potential negative effects on the procedure outcomes. For these reasons, rMIS became popular in surgical disciplines with limited anatomical access, such as urology, gynaecology and thoracic surgery and is gaining momentum in other practices like Ear-Nose-Throat (ENT) and gastric surgery. Significant portions of rMIS procedures consist of dissecting and mobilising healthy tissues to reach the diseased area. During this phase, the surgeon heavily relies on the assistant to clear the surgical field from obstructing tissues, facilitating the surgeon's navigation in the anatomy.

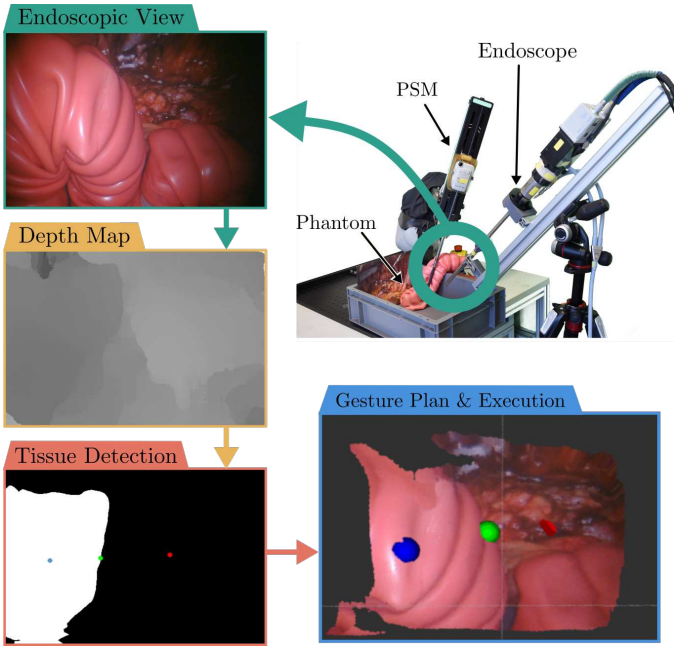

Fig. 1. Tissue flap segmentation workflow. The stereo images acquired by the endoscope are combined to evaluate Depth Maps fed into a neural network to detect the shape and boundaries of the tissue flap. The tissue flap profile is used to define three waypoints which are used to plan the retraction gesture.

The coordination between surgeon and assistant can be difficult and requires highly specialised personnel. Immersive consoles, such as the one in the Intuitive Surgical DaVinci robot, limit the communication between members of the clinical staff. In particular scenarios such as newly formed teams or lack of adeguate training on emergency situations, the limited communication could increase the risk of adverse events. Some robotic systems, (e.g. the DaVinci robot), allow the clinician to operate three arms, thus reducing the need for external assistance, but the switching process could increase the cognitive load on the clinician [1], particularly for less experienced surgeons.

A semi-autonomous assistance system, capable of operating one arm of the surgical robot and supporting the clinician during the manipulation of soft tissues would solve many issues and open the way for a shared control paradigm, in which the clinician can rely on the robot to perform minor repetitive

* these two authors contibuted equally
[1] A. Attanasio, B. Scaglioni, N. Marahrens and P. Valdastri are with with the Storm Lab UK, School of Electronic and Electrical Engineering, University of Leeds, Leeds, UK,{elaat, b.scaglioni, elnma, p.valdastri}[at]leeds.ac.uk .
[2] C. Alberti and E. De Momi are with Near Lab, Politecnico di Milano, Milan, Italy chiara1.alberti[at]mail.polimi.it, elena.demomi[at]polimi.it
[3] M. Leonetti and A. F. Frangi are with School of Computing, University of Leeds, Leeds, UK, {m.leonetti,a.frangi}[at]leeds.ac.uk
[4] C.S. Biyani is with Department of Urology, St James University Hospital, Leeds, UK, shekhar.biyani[at]nhs.net

tasks and focus on the clinical aspects of the procedure. The first step towards the autonomous execution of surgical tasks is the analysis of the scene. The autonomous system must segment the endoscopic scene and isolate the tissue flaps that can be manipulated to plan and execute the gesture. This is a crucial step in the accomplishment of many tasks, as any lack of accuracy at this stage could negatively affect the execution of the gesture and possibly lead to hazardous situations. For this reason, it is extremely important to provide an accurate segmentation system, capable of offering the best possible performance.

In previous work [2], we proposed a feasibility study on autonomous tissue retraction, developed on a DaVinci Research Kit (DVRK). To detect a candidate flap of tissue for the retraction, a single endoscopic Depth Map was segmented with deep-learning techniques, and the system was autonomously executing the retraction, based on the analysis of the image. The experimental setup is shown in Figure 3: the images captured by the endoscopic stereo-camera were segmented by means of a deep neural network (i.e. the U-Net [3]), the result of the segmentation was subsequently used to define starting and end point of the retraction. Although the images processed by the system were part of a video stream, the segmentation stage was performed on a single image, thus discarding the obvious relation between consecutive images in the stream. This approach neglects the information provided by the relation between consecutive images and therefore is sub-optimal, with negative consequences on the performance of the segmentation and of the whole task.

In this work, we propose a new approach to the segmentation of soft tissues in surgical endoscopic video streams. We take advantage of the correlation between consecutive images and demonstrate that, by considering sequences as an alternative to single images, the segmentation system outperforms our previous architecture. The main goal of the work is to provide a robust framework to segment soft tissues in abdominal surgery. The approach, based on deep-learning, could be applied to a wide range of surgical tasks and is suitable for real-time tracking of the tissue motion. The main technical contribution of this work is the development of three novel deep-learning network models for video stream segmentation. Starting from a standard network architecture such as the U-Net, we combine the use of Long Short-Term Memory (LSTM) [4] and Attention Gate blocks [5], [6] to develop three network variants. The performances of these networks are compared to our previous work, the process of parameters optimisation is discussed in detail and the effectiveness of a pre-training stage is evaluated. Additionally, a dataset, based on the FlapNet [2], is developed to train and verify the performances of the networks. The dataset, comprising labels and training images, and the code are publicly available for the research community at https://github.com/Stormlabuk/dvrk_ULSTM. Although the techniques described in this work originate in the context of retraction, robust segmentation of soft tissues could be used in developing many autonomous surgical tasks such as ablation [7], resection [8] and suturing [9]. The paper is organised as follows: in Section II the dataset processing and organisation (Section II-A), the model architecture (II-B) and the training setup (Section II-C) are desribed. Then, in Section III the performances of the three architectures are discussed. Additionally, a comparison with a pre-trained model [10] and our previous work [2] is carried out, to demonstrate the benefits in adopting LSTM layers and Attention Gate blocks in video segmentation. Section IV concludes the paper, summarising the contribution and discussing future developments.

A. Technical Contribution

Despite the great interest on autonomy in surgical robotics, demonstrated by the amount of literature [11], research on soft tissues manipulation is limited. The vast majority of the literature focuses on the automation of tasks [12] involving extraneous elements such as suturing [13], [14] and interventional needle passing [8] [15]. On the other hand, automation of tasks that involve tissue manipulation are challenging due to the complex geometry and compliance of the soft tissues. Few examples of autonomous tissue manipulation are available [16], [17], mostly demonstrated in simplified scenarios with reduced complexity. The main barrier for development of realistic applications is the complexity of the scene, difficult to analyse autonomously. A significant contribution can be provided by machine learning. Techniques based on neural networks are widely adopted for medical and surgical image analysis [18]. Deep Learning models have been employed in medicine for the segmentation from MRI and CT scans [19] of either organs [20], [21] or compromised tissue such as polyps [22] and tumours [23]. The U-Net [3] is commonly used in segmentation of medical images such as the segmentation of blood vessels, brain and skin tumours [24], [25], [26]. This network consists of an encoder-decoder architecture which captures contextual information, simultaneously providing accurate detection of the image features. The main drawback of the standard U-Net is the incapacity to correlate frames in a video sequence, thus not taking advantage of the tissues motion and consequently offering limited performances in continuous tissue manipulation. To overcome this limitation, a simplistic approach could consist in linearly merging several independent U-Nets. However, literature has shown outstanding results with the adoption of recurrent neural network architectures such as the Long Short-Term Memory (LSTM) cells [4]. LSTM provide memory to the model, thus allowing a representation of the features' evolution in time. Adding LSTMs on top of fully convolutional network proved to significantly enhance the accuracy of video segmentation [27] of street scenes. In medical imaging, LSTMs have been used to predict the growth of tumours from 4D patient's data [28] with a simple encoder/decoder model. LSTM cells have been adopted on top of a U-Net model for cell segmentation, showing a remarkable ability in discriminating both the cell's body and its boundaries from the background [10]. An alternative recurrent structure used for video segmentation is the Gated Recurrent Unit (GRU) [29]. These units, significantly simpler than LSTMs, have been implemented by means of convolutional networks to enhance the precision in prostate [30] and brain [31] segmentation.

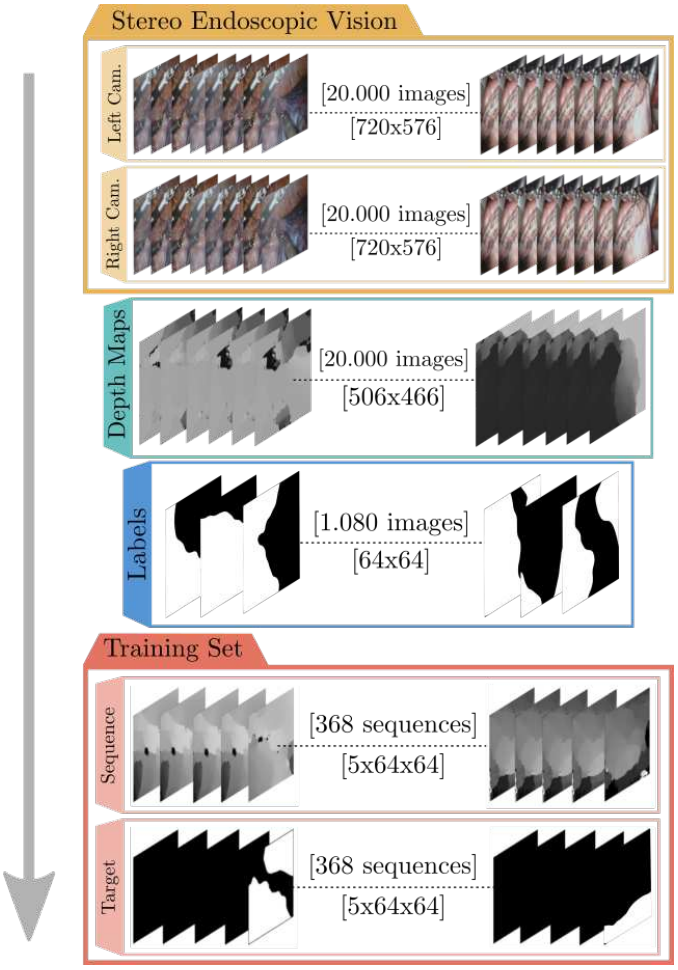

Fig. 2. The dataset is created from images collected with a stereoscopic endoscope. Depth Maps are evaluated from the stereo pairs and manually labelled. Subsequently, sequences are created by extracting the previous 4 frames from the whole operation video and batching them with the corresponding label of the 5th frame.

Additionally, an approach for video segmentation adopting 3D convolutional layers to extract the temporal information from image sequences was recently proposed in [32]. These blocks are particular structures that support the network's training and inference by identifying focus regions of the image where relevant information is contained. These blocks have shown effectiveness in medical image segmentation [5] for pancreas segmentation and classification [33].

II. Methods

A. Data Setup

The first step in the development of a tissue segmentation system is the conditioning of the input data. Since most surgical robots and advanced endoscopic systems are equipped with stereo-vision, we take advantage of the stereoscopic endoscope by considering pairs of stereo images as starting point. With a modified version of the Semi-Global Matching algorithm [34] implemented in the `stereo_img_proc` ROS package, each pair of stereo images generates a Depth Map (DM). Depth Maps are single-channel images in which pixel intensity represent the distance of each pixel from the camera frame. Distances are computed from the features' disparity in the left and right images. As DMs do not contain light and colour information, their use guarantees robustness against variations of lighting conditions and tissue colours. This aspect is particularly important in this work, as the instruments frequently cross the endoscope field of view during tissue manipulation, therefore, it is crucial to guarantee satisfactory performances in presence of the instruments. Additionally, as the color information is represented in images with three channels (RGB), DMs allow to work on single-channel images, thus speeding up the training phase.

In order to train the networks, DMs must be associated with labels highlighting the areas of the image covered by tissue flaps and by the surgical tools. In a previous work [2], our group developed FlapNet: a dataset of 1080 DMs extracted from images collected during a robotic surgery course, performed with a DaVinci Xi at the University of Leeds, on Thiel-embalmed cadavers [35] by experienced surgeons. Starting from the full stereo video stream of a lobectomy, the most relevant frames of the stream are extracted and labelled: for each DM, a binary mask is created, classifying each pixel as background (0) or tissue (1). The labelling process is carried out by researchers under the guidance of experienced urological and colorectal surgeons. Initially, the video sequences containing tissue flaps are identified and isolated. Subsequently, a set of single frames is manually selected. Depth Maps are generated for the identified images. The labelling process is carried out manually on the Depth Map. However, during the process, the user can visualize the RGB image to ease the label creation. Labels with Structural Similarity Index higher than 70% have been discarded to avoid similarity between the dataset entries, guaranteeing a significant variety of samples. To represent the tool's appearance in the endoscopic scene, regions of the DM containing surgical instruments are labelled, extracted from the original DM and superimposed over scenes where tools are not present. The instruments' labels are not available in the FlapNet, as the tissue flaps are the only targets for the segmentation.

The networks developed in this work require a sequence of images. To this end, the FlapNet dataset has been enriched by adding the four frames preceding every labelled image already available in the dataset. To account for this, entries of the original dataset are grouped with the four stereo-frames preceding every labelled image, thus obtaining a set of sequences, in which the last image of each set associated to a binary label (Figure 2). Since the majority of the samples (712 images) contained in the FlapNet are artificial images (i.e. created by the superposition of tools on the scene), no preceding frames are available for these entries, reducing the size of the dataset to 368 sequences. The images contained in the dataset are reduced to a size of 64x64. During preliminary tests this proved to be a satisfactory compromise between the amount of detail available in the image and time required to train the model. If required by a specific application, the output of the network can be up-sampled and linearised to the original size of the input image. Over the whole set of images, the pixels associated with the background are 70% of the total, leading to a slightly unbalanced dataset. Therefore, particular

attention is required during the training phase to limit the amount of predicted false negative and false positive. It is well known in the literature [36] that unbalanced datasets may create issues in the modeling of the less-represented response, leading to a degradation in performance. The original dataset contains only DMs where at least one area is classified as tissue. In order to represent the case in which no foreground tissue is present in the scene, 88 new sequences associated with black mask (only background) are added to the dataset, raising the number of the total sequences to 456.

Given the limited size of the dataset, data augmentation is required. Standard augmentation computer vision techniques are adopted to enlarge the dataset, including: contrast and brightness adjustment, horizontal and vertical flipping, image shifting and rotation. These transformations, randomly selected, are equally applied to every image and label of the sequence to maintain coherence between the input and the target. Moreover, elastic deformation [37] is applied to enhance the variety among the augmented entries by distorting the input image. This technique consists of convoluting two random displacement fields Δx and Δy with a Gaussian filter having standard deviation σ, which represents the elasticity coefficient. The resulting displacement fields are scaled by a factor α that defines the deformation intensity. An additional method for video augmentation comprising the inversion of the sequences' frames to obtain new sequences is herein adopted. This technique allows to create new sequences of images with a coherent time evolution of the scene, thus doubling the number of entries while maintain the correlation within subsequent frames. By means of this augmentation, the initial 456 sequences are doubled to 912, additionally every single sequence is distorted with the aforementioned computer vision techniques up to 3 times, thus increasing the number of entries to 2736 sequences.

B. Neural Networks Development

One of the most common neural network architectures utilised for the segmentation of medical images is the U-Net [3]. Satisfactory performances are reported in literature regarding image segmentation adopting this class of network even with limited amount of data and with high resolution images. As show in Figure 4, the network comprises two symmetric encoding and decoding branches, with parallel connections linking the encoders to the decoders. The standard U-Net architecture is suitable for segmenting single images in endoscopic scenarios, as demonstrated by our previous work [2], but cannot correlate consecutive frames (e.g. a video stream) and therefore has limited robustness. For this reason, we build upon the basic U-Net architecture by adding features that implement memory (i.e. recurrency) and take advantage of the relation between consecutive frames to enhance performances. We use recurrent structures such as LSTMs, proposing three network architectures. Additionally, in one of the network variants, the use of attention gates is explored. A summary of the features implemented is reported in Table I.

All the U-Net variants are developed in the TensorFlow [38] framework. The basic structure, identical for all the networks,

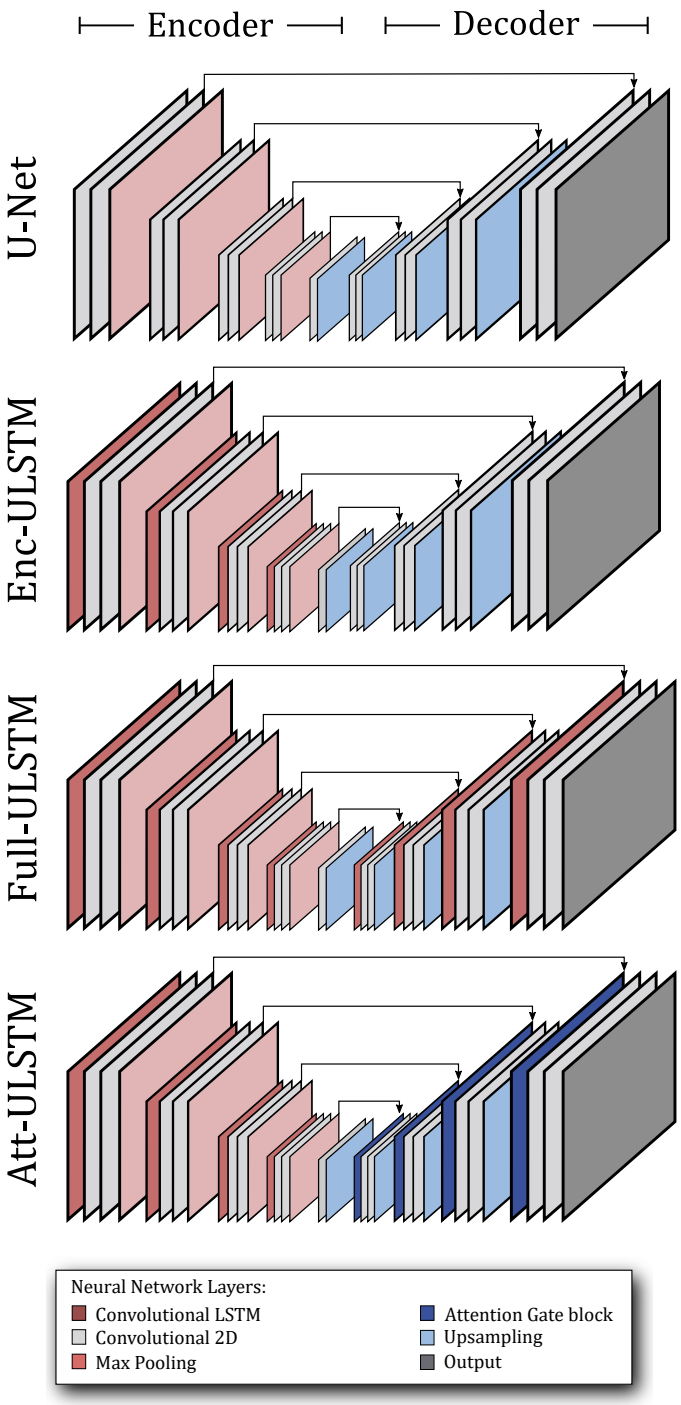

Fig. 3. Comparison between different extensions of the U-Net[3]. The Enc-ULSTM contains LSTM cells in the encoding branch, the Full-ULSTM model incorporates LSTMs in both branches. Finally, the Att-ULSTM includes Attention Gate blocks in the decoding block.

is composed of 4 encoding and 4 decoding blocks that constitute the contracting and expanding paths, respectively. The encoding blocks consist of 2 convolutional layers with batch normalisation and adopt the Rectifier Linear Unit (ReLU) activation function. Subsequently, a layer with pool size of 2 halves the output size, grouping the features detected by the previous layers to reduce over-fitting, while limiting the memory allocation required. In parallel, the decoding blocks

TABLE I
SUMMARY OF THE IMPLEMENTED FEATURE AND ARCHITECTURES IN THE THREE DIFFERENT MODELS PROPOSED.

	U-Net	Enc-ULSTM	Full-ULSTM	Att-ULSTM
Conv. Layers	✔	✔	✔	✔
Encoder LSTM	✗	✔	✔	✔
Decoder LSTM	✗	✗	✔	✗
Attention Gate	✗	✗	✗	✔

are composed of 2 convolutional layers. The output of each block is up-sampled by a factor 2 with bilinear interpolation, to restore the original image size. The up-sampled outputs are subsequently combined with the feature maps from the encoding branch by means of parallel skip connections. The number of kernels, set to 64 for the encoding block, is doubled for every contraction step in the encoding branch and halved for every expansion step in the decoding branch, resulting in a symmetric structure. To save memory in the training phase, 128 kernels are maintained between the second and third encoder and decoder. The two branches of the network are connected by a single convolutional layer with 512 kernels. The output layer comprises a convolutional layer with a sigmoid activation function.

Starting from the basic structure, we propose three variations implementing LSTM and attention gates:

- Enc-ULSTM: the U-Net model contains convolutional LSTM layers at the beginning of each encoding block.
- Full-ULSTM: the U-Net model contains convolutional LSTM layers in the encoding and decoding branch.
- Att-ULSTM: using the Enc-USLTM as base model, attention gate blocks are added before each decoder block.

In the Enc-ULSTM and Full-ULSTM, convolutional LSTMs are used. The detailed structure of an LSTM is described in Figure 4. LSTMs are composed of three gates (forget, input and output) which, combined with the previous cell state c_{t-1}, the previous hidden state h_{t-1} and the input x_t, allow to extract the correlation between subsequent frames, thus rejecting lower-level responses. By means of the forgetting gate contained in the LSTM cells, the non-relevant information at time t is discarded, enhancing the accuracy of the response at time $t+1$. In this particular application, LSTM cells support the network in detecting relevant information such as the position and geometry of a tissue while ignoring and forgetting the appearance of tools. This contributes to the robustness of the network against instruments crossing the endoscopic scene.

In the Att-ULSTM, each decoding block includes a first layer composed of attention gates. In these blocks, capitalising on a gating signal g, the lower activations are discarded, thus allowing the network to autonomously find the relevant areas of the image to focus on, hence resulting in a precise segmentation. The Attention Gate unit takes x_t as input. The gating signal g is applied to every pixel in order to define the focus regions. Three linear transformation W_g, W_x and Σ define the set of parameters of the single unit and are evaluated with channel-wise [1x1x1] convolutions. These blocks contribute to the extraction of focus regions, thus helping identifying the candidate areas of the image where a flap could be found.

C. Models Training

The adoption of convolutional LSTM layers allows the networks to rely on both temporal and spatial features. For this recurrent architecture, a modified version of the Back Propagation Through Time algorithm has been adopted, namely the Truncated Back Propagation Through Time [39]. This algorithm, commonly adopted for recurrent networks, periodically updates the gradient a fixed amount of times over the batch. In this work, this parameter was set to $\tau = 5$. Hence, the gradient is weighted on the previous input and hidden state, yielding a simultaneous evaluation of the temporal and spatial features in the convolutional layers. The networks are trained for 10.000 iterations over 650 epochs using the Adam [40] optimiser, capable of managing sparse gradients and preventing noise, as well as vanishing of weak gradients.

A step profile is scheduled for the value of the learning rate, decreasing from an initial value of 10^{-3} to 10^{-5} to speed-up the initial phase of the training. The kernel's weights are randomly initialised with the He uniform distribution [41] which allows to regulate the initial values depending on the preceding layers' dimension, thus reducing the time required for training. Dropout is applied in the LSTM layers and in the central block to limit over-fitting. While standard dropout is implemented for convolutional layers, the same approach is not suitable for long-term memory. As standard dropout applies a mask to the layer to randomly deactivate the neurons, if applied to LSTM cells it would resets the forget gate at each iteration, thus ereasing the cell's memory. For this reason, a recurrent of dropout [42] is applied to LSTM layers to maintain the dropout mask fixed, preventing the loss of memory of the cells. The dataset is split into 75% training set, 15% validation and 10% test set. The models are trained on a Linux (Ubuntu 18.04) machine equipped with an Intel Xeon Gold 6140 (2.30GHz) CPU, an Nvidia Quadro 5000 RTX GPU and 128 GB DDR4 2666MHz RAM.

Two loss functions are compared in this work. The Combo Loss (CL) [43] is the weighted Dice Loss $DL = \frac{2 \cdot P \cdot G}{P+G}$, where G is the ground truth and P the sigmoid output, [44] and the Weighted Cross-Entropy (WCEL) defined as:

$$WCEL = p \cdot -log(\hat{p}) \cdot \beta + (1-p) \cdot -log(1-\hat{p}) \qquad (1)$$

where p is the ground truth label, $\hat{p}$ is the sigmoid activation of the logits and β is a trade-off factor to foster either false negatives or false positives. The CL is finally defined as:

$$CL = \alpha \cdot WCEL + (1-\alpha)DL \qquad (2)$$

where α controls the contribution of the single DL and WCEL. Given the unbalanced dataset and considering that for the surgical application false positives must be minimised, we defined $\beta = 0.8$ and $\alpha = 0.6$ to favour the contribution of the WCEL over the DL.

The other function considered here is the Tversky Loss (TL) [45], widely used in medical image segmentation for its ability

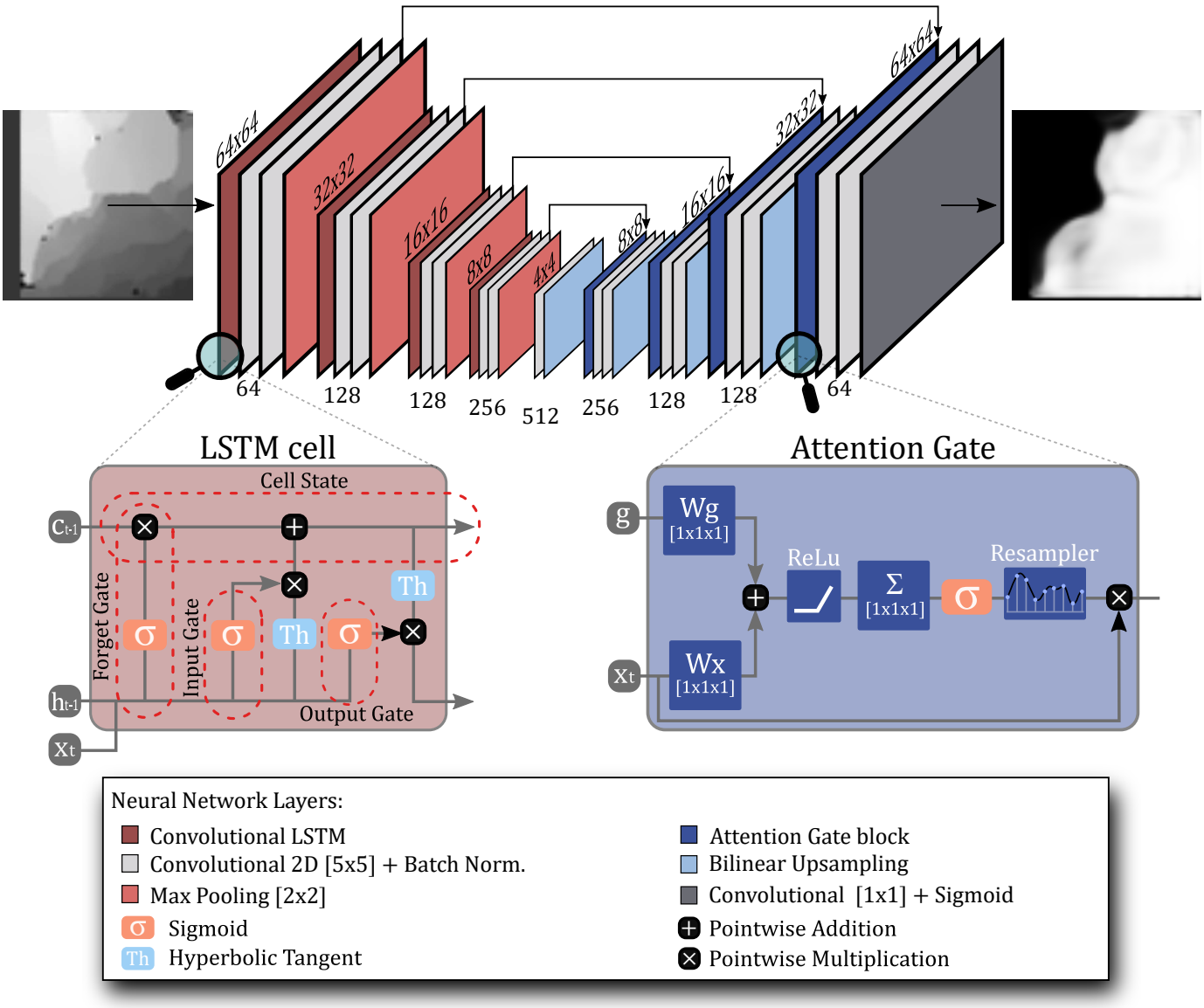

Fig. 4. The structure of the Att-ULSTM model comprises 4 encoders, 4 decoders and the central block connecting the two branches. Each encoding block is composed of a LSTM cell, two convolutional and one max pooling layers, while the decoding blocks present an Attention Gate block, two convolutional layers and a linear upsampling layer. The output layer is a convolutional layer with sigmoid activation function.

to train over highly unbalanced training sets. The TL formula is a generalisation of the DL:

$$TL = \frac{2 \cdot P \cdot G}{P + G + \gamma \cdot P \setminus G + \eta \cdot G \setminus P} \tag{3}$$

where G is the ground truth, P is the prediction, $P \setminus G = P \cdot (1 - G)$ is the relative complement and γ, η are weights to balance false positives or false negatives.

III. Results

In this section, the performance of the three networks models is evaluated. Four metrics, all aimed at evaluating the ratio between True Positive (TP), True Negative (TN) and False Positive (FP), False Negative (FN), are proposed:

- The Precision: $P = \frac{TP}{TP+FP}$, represents the capability of the algorithm to reject false postives.
- The Recall: $R = \frac{TP}{TP+FN}$ describes the sensitivity of the network in detecting TP and TN. Combined with Precision, it provides a reliable measure of the network robustness. The Recall is particularly meaningful with unbalanced datasets.
- The Accuracy: $A = \frac{TP+TN}{TP+TN+FP+FN}$, reports correct predictions over the full testing set.
- The Jaccard Index: $J = \frac{TP}{TP+FP+FN}$, estimates the similarity between the ground truth and the prediction, computing the ratio between intersection and union of the two. If used in conjunction with the accuracy, accurately predicts the quality of the segmentation.

The joint analysis of these metrics provide a comprehensive insight of the networks' performance in terms of rejection to disturbances and management of false positives/negatives. A K-fold cross-validation with $K = 10$ is adopted to validate

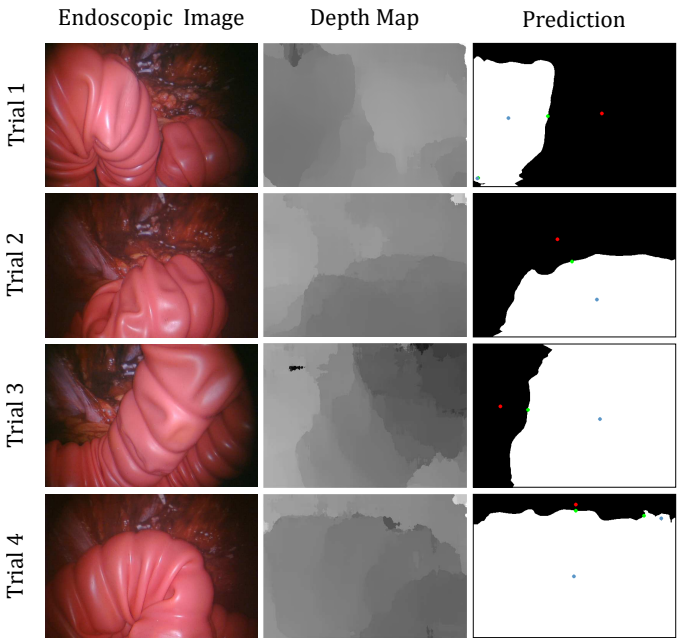

Fig. 5. Predictions examples of the Att-ULSTM model at the end of the training phase. The tissue is placed in different regions of the endoscopic scene to verify the robust inferring of the model, independently from the tissue position.

the network's robustness against data variability. Initially, the models are trained and tested using the CL, discussed in Section II-C. As in Figure 6, the Att-ULSTM model provides better performances in terms of accuracy, precision and Jaccard Index, while the best values of Recall is given by the Full-ULSTM network. The Att-ULSTM structure provides superior identification of the tissue flaps and a sufficient rejection to FP and FN as shown in Figure 5. Further analysis will be carried out only on Att-ULSTM model, comparing this architecture with the state of the art. To further improve the network performances, the Att-ULSTM is trained using the Tversky Loss instead of the Combo Loss. The results are reported in Table II and compared with the performance of the same network structure trained with the CL.

TABLE II
PERFORMANCE COMPARISON OF THE MODEL TRAINED WITH BOTH TVERSKY AND COMBO LOSS FUNCTIONS

	Tversky Loss	Combo Loss
Accuracy	82.25% ± 2.80%	83.77% ± 2.18%
Precision	74.89% ± 9.35%	78.42% ± 7.38%
Recall	70.60% ± 6.49%	74.32% ± 3.83%
Jaccard Index	72.53% ± 7.54%	75.83% ± 3.38%

The adoption of the Tversky Loss entails a slight loss of performance in the Att-ULSTM model with respect to the Combo Loss. For this reason, the combo Loss is selected. In Figure 7 and 3 the precision and accuracy during the training phase are reported for the worst ($K = 1$), the average ($K = 2$) and the best ($K = 3$) performing model over the K validations.

Given the restricted data available for this particular application, pre-training is evaluated, with the aim of limiting the over-fitting during training. The neural network model proposed in [10] is considered, due to its similarity with the Enc-ULSTM structure. Despite the similar structure, the pre-trained convolutional layers are characterised by an higher number of filters, thus increasing the model complexity. As shown in Table III, the pre-trained model offers no performances improvement. This is motivated by the higher amount of kernels in the convolutional layers of the pre-trained model which increases the complexity of the model. Moreover, the model is pre-trained with microscopic images of cells, requiring a smaller amount of data augmentation with respect to endoscopic images, in which the geometrical constraints of the anatomy limit the image augmentation. Moreover, the amount of images contained in the pre-training dataset is limited, thus preventing the model to generalise the predictions.

TABLE III
PERFORMANCE COMPARISON BETWEEN THE PRE-TRAINED MODEL [10] AND THE MODEL TRAINED FROM SCRATCH.

	P-ConvULSTM	Att-ULSTM
Accuracy	77.59% ± 2.30%	83.77% ± 2.18%
Precision	73.31% ± 5.64%	78.42% ± 7.38%
Recall	58.76% ± 5.59%	74.32% ± 3.83%
Jaccard Index	64.65% ± 4.83%	75.83% ± 3.38%

Finally, to demonstrate the increased performances provided by the approach in this work regarding the segmentation of single images, the Att-ULSTM model and the standard U-Net presented in [2] are compared. In Section II-A, the U-Net implemented in our previous work is fed with single images from the video stream and produces a single prediction for each frame. By comparing these two networks it is possible to assert if the adoption of LSTM layers and attention gates is beneficial for tissue flap segmentation in video. The networks performances are evaluated in terms of accuracy and precision, as defined in Section II-C.

TABLE IV
PERFORMANCE COMPARISON BETWEEN THE ORIGINAL FLAPNET AND THE PROPOSED ATT-ULSTM

	Accuracy	Precision
U-Net [2]	80.90% ± 1.32%	72.63% ± 1.94%
Att-ULSTM	83.77% ± 2.18%	78.42% ± 7.38%
p-value	0.0173	0.0376

With the adoption of spatio-temporal layers and Attention Gates blocks in the Att-ULSTM, the model outperforms a standard feed-forward U-Net model, as shown in Table IV. In particular, the adoption of LSTM provides the ability to extract temporal information from subsequent frames, thus guaranteeing a more robust prediction. It is worth to mention that both the Att-ULSTM and U-Net models are trained over the same DMs, thus no evaluation bias is introduced in the comparison of the two models. The standard deviation of the precision is slightly higher for the Att-ULSTM. This is related to a better characterisation of the tools' presence in the augmented entries of the FlapNet, which are omitted in the training of the Att-ULSTM, as explained in Section II-A. This enhances the robustness of the the U-Net with respect to the presence of tools, compared to the Att-ULSTM model. However, as shown by the other metrics, the segmentation of the Att-ULSTM is more reliable. Using the computer

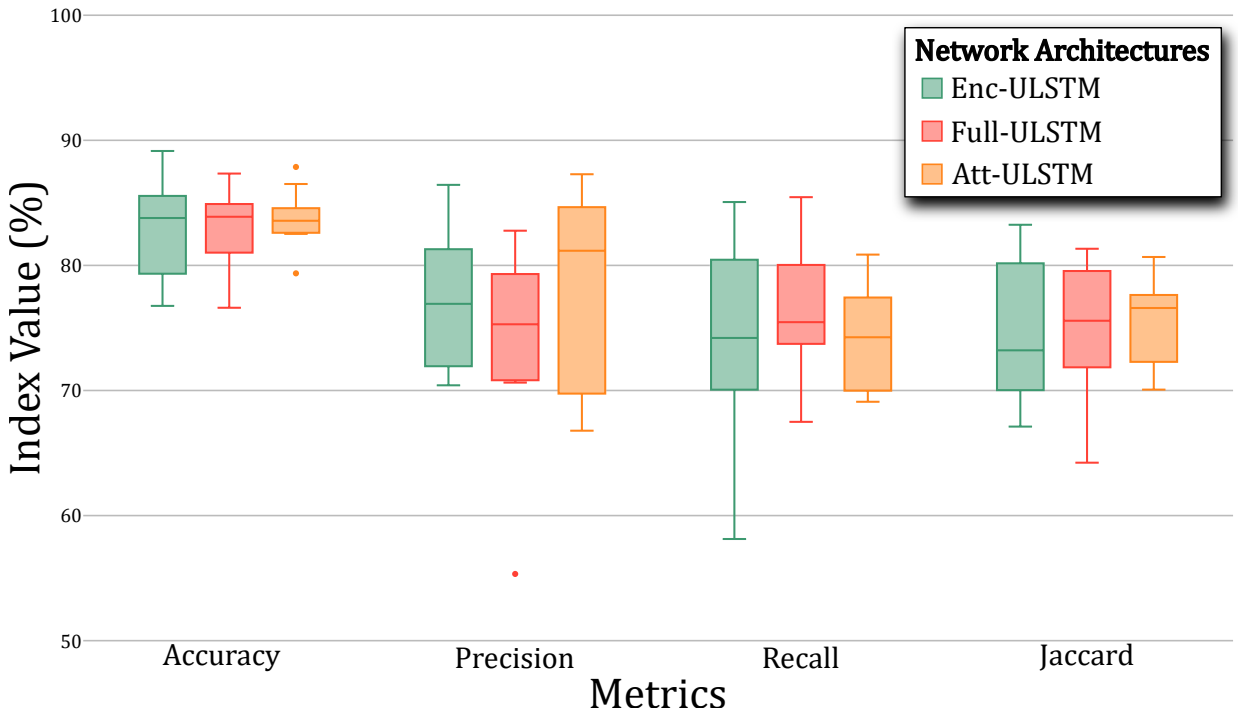

Fig. 6. Performance comparison among the three proposed models. The metrics considered for this comparison are Accuracy, Precision, Recall and Jaccard Index. Results show that the best performance is achieved by the Att-ULSTM model in terms of accuracy, precision and Jaccard index while the Full-ULSTM show the best performance in terms of Recall.

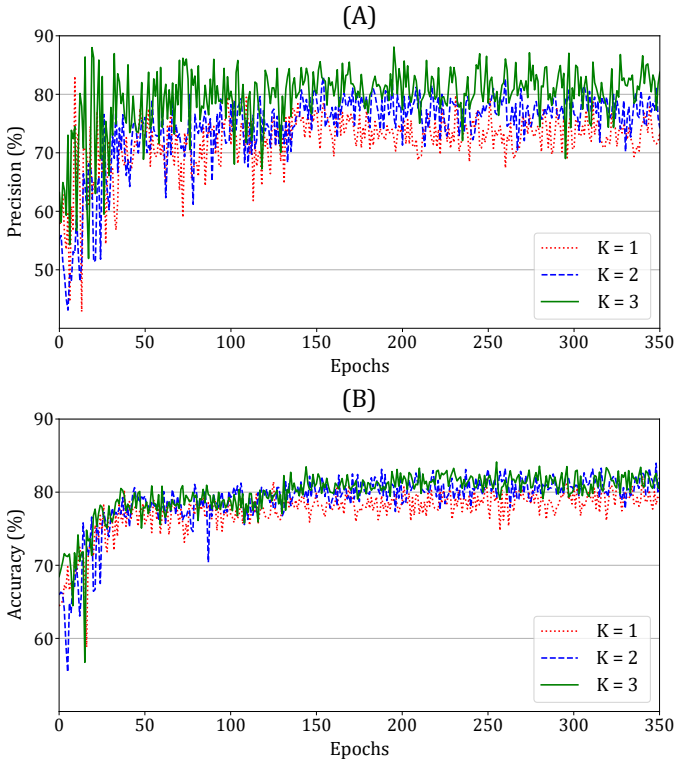

Fig. 7. Precision (A) and accuracy (B) reported during training of the K=10 models for K-fold cross-validation. Only the best (K=3), the average (K=2) and the worst (K=1) are represent on the plot to simplify its visualisation. As the plateau is reached within the first 200 epochs, only 350 epochs are shown.

mentioned in Section II-C, an inference time of $t_i = 0.5$ s was recorded, with a maximum speed of 2 FPS against the recorded $t_i < 42$ ms recorded for the standard feed-forward U-Net. This result is acceptable, considering that the surgeon motion are generally relatively slow to guarantee a safe interaction with the anatomy.

Given the limited training and testing data for the Att-ULSTM, a non-parametric test is required to prove the normal distribution of the two groups. A Wilcoxon rank sum test [46] is carried out for accuracy and precision to assess statistical significance of the two models' performances. This test assesses the null hypothesis that the two groups are continuous distribution with equal medians. In Table IV, the comparison between the models' accuracy and precision are shown. The p-value indicates a low probability for the two distribution to have equal median, thus there is a statistically significant improvement in the prediction performances using the Att-ULSTM model.

IV. Conclusion

A novel approach to the segmentation of tissue in endoscopic video streams is herein discussed. Three neural network architectures for tissue segmentation in endoscopic images are proposed. The tissue detection and segmentation are considered the initial step towards intelligent interaction with the anatomy. On top of this, an estimation of the physical interaction is needed to accomplish a particular task. This evaluation however varies depending on the specific objective task to reproduce. The adoption of attention gates and recurrent structures such as LSTMs enhance the accuracy of the tissue detection, compared to a standard feed-forward network. The

performances of the three variants are compared and the Att-ULSTM is selected for further investigation. For this network, different cost functions are compared, and the use of pre-training is evaluated. Experimental results show enhanced performances with respect to the state of the art for what concerns the network's precision (78.42 % $\pm$ 7.38 %) and prediction stability. The adoption of LSTM and attention gates to take advantage of the time-related features, embedded in the images sequence, can improve the performances and robustness of the detection in the context of endoscopic images for surgical robotics. To achieve this result, the FlapNet dataset is enhanced to meet the requirements of the recurrent network's structure, thus resulting in a new dataset, now available to the research community.

The approach discussed in this work, demonstrating an enhanced ability to segment soft tissues, can significantly improve the implementation of autonomous tasks involving the elaboration of endoscopic images. Examples range from laparoscopic procedures, to non-autonomous robotic and semi-autonomous robot-assisted surgical tasks such as ablation, retraction and suturing. Localising the target tissue flap is indeed a key step towards surgical gesture automation and, given the variety and complexity of the human anatomy, this task is extremely challenging.

The major limitation of this work is the limited availability of labelled medical images. As pre-training over different dataset did not show promising results [2], weak labelling and unsupervised learning could be beneficial in dealing with such limited amount of data. As discussed above, the pre-training does not provide improved results, due to the unique characteristics of the surgical images. In conclusion, the most promising approach to increase the networks performances would consist of an increased number of entries in the dataset. However, labelling endoscopic images is time-consuming and requires specialised medical knowledge, thus hindering the process. The adoption of generative adversarial networks (GANs) could be beneficial to improve the network's ability to reject the surgical instrument, thus guaranteeing a correct and precise segmentation of the tissue flaps. Future work could include the adoption of endoscopic RGB image along with DMs to enhance the performance of the proposed model.

ACKNOWLEDGEMENTS

Research reported in this article was supported by Intuitive Surgical Inc. under the Technology Research Grants program 2019, by the Royal Society, by the Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/R045291/1, and by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 818045). Any opinions, findings and conclusions, or recommendations expressed in this article are those of the authors and do not necessarily reflect the views of the Royal Society, EPSRC, or the ERC. The authors thank Intuitive Surgical, and in particular Simon Di Maio and Dale Bergman, for supportng the research and allowing us to work on the dVRK. Alejandro F. Frangi acknowledges support from the Royal Academy of Engineering Chair in Emerging Technologies Scheme (CiET1819/19).

REFERENCES

[1] M. Liu and M. Curet, "A Review of Training Research and Virtual Reality Simulators for the da Vinci Surgical System," *Teaching and Learning in Medicine*, vol. 27, no. 1, pp. 12–26, 2015.

[2] A. Attanasio, B. Scaglioni, M. Leonetti, A. Frangi, W. Cross, C. S. Biyani, and P. Valdastri, "Autonomous Tissue Retraction in Robotic Assisted Minimally Invasive Surgery - A Feasibility Study," *IEEE Robotics and Automation Letters*, 2020. [Online]. Available: Pre-printathttp://eprints.whiterose.ac.uk/163725/

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9351, pp. 234–241, 2015.

[4] F. Gers, "Learning to forget: continual prediction with LSTM," in *9th International Conference on Artificial Neural Networks: ICANN '99*, vol. 1999. IEE, 1999, pp. 850–855.

[5] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-Net: Learning Where to Look for the Pancreas," Apr 2018.

[6] M.-T. Luong, H. Pham, and C. D. Manning, "Effective Approaches to Attention-based Neural Machine Translation," Aug 2015.

[7] B. Su, J. Tang, and H. Liao, "Automatic laser ablation control algorithm for an novel endoscopic laser ablation end effector for precision neurosurgery," *IEEE International Conference on Intelligent Robots and Systems*, vol. 2015-December, pp. 4362–4367, 2015.

[8] N. Sarli, G. Del Giudice, S. De, M. S. Dietrich, S. D. Herrell, and N. Simaan, " Preliminary Porcine In Vivo Evaluation of a Telerobotic System for Transurethral Bladder Tumor Resection and Surveillance ," *Journal of Endourology*, vol. 32, no. 6, pp. 516–522, 2018.

[9] A. Shademan, R. S. Decker, J. Opfermann, S. Leonard, P. C. Kim, and A. Krieger, "Plenoptic cameras in surgical robotics: Calibration, registration, and evaluation," *Proceedings - IEEE International Conference on Robotics and Automation*, vol. 2016-June, pp. 708–714, 2016.

[10] A. Arbelle and T. R. Raviv, "Microscopy Cell Segmentation Via Convolutional LSTM Networks," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, Apr 2019, pp. 1008–1012.

[11] G.-Z. Yang, J. Cambias, K. Cleary, E. Daimler, J. Drake, P. E. Dupont, N. Hata, P. Kazanzides, S. Martel, R. V. Patel, V. J. Santos, and R. H. Taylor, "Medical robotics—Regulatory, ethical, and legal considerations for increasing levels of autonomy," *Science Robotics*, vol. 2, no. 4, p. eaam8638, 2017.

[12] A. Attanasio, B. Scaglioni, E. De Momi, P. Fiorini, and P. Valdastri, "Autonomy in surgical robotics," *Annual Review of Control Robotics and Autonomous Systems*, 2020. [Online]. Available: Pre-printathttps://www.stormlabuk.com/wp-content/uploads/2020/07/ARCAS2020_PREPRINT.pdf

[13] C. D'Ettorre, G. Dwyer, X. Du, F. Chadebecq, F. Vasconcelos, E. De Momi, and D. Stoyanov, "Automated Pick-Up of Suturing Needles for Robotic Surgical Assistance," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, vol. 1, no. c, May 2018, pp. 1370–1377.

[14] D. L. Chow and W. Newman, "Improved knot-tying methods for autonomous robot surgery," *IEEE International Conference on Automation Science and Engineering*, pp. 461–465, 2013.

[15] S. Sen, A. Garg, D. V. Gealy, S. McKinley, Y. Jen, and K. Goldberg, "Automating multi-throw multilateral surgical suturing with a mechanical needle guide and sequential convex optimization," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, May 2016, pp. 4178–4185.

[16] R. Elek, T. D. Nagy, D. Nagy, T. Garamvölgyi, B. Takács, P. Galambos, J. K. Tar, I. J. Rudas, and T. Haidegger, "Towards surgical subtask automation-blunt dissection," in *INES 2017 - IEEE 21st International Conference on Intelligent Engineering Systems, Proceedings*, vol. 2017-Janua, 2017, pp. 253–257.

[17] T. D. Nagy, M. Takacs, I. J. Rudas, and T. Haidegger, "Surgical subtask automation — Soft tissue retraction," in *2018 IEEE 16th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*. IEEE, Feb 2018, pp. 000 055–000 060.

[18] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, Dec 2017.

[19] A. Norouzi, M. S. M. Rahim, A. Altameem, T. Saba, A. E. Rad, A. Rehman, and M. Uddin, "Medical Image Segmentation Methods,

Algorithms, and Applications," *IETE Technical Review*, vol. 31, no. 3, pp. 199–213, May 2014.

[20] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P. M. Jodoin, and H. Larochelle, "Brain tumor segmentation with Deep Neural Networks," *Medical Image Analysis*, vol. 35, pp. 18–31, 2017.

[21] P. Hu, F. Wu, J. Peng, P. Liang, and D. Kong, "Automatic 3D liver segmentation based on deep learning and globally optimized surface evolution," *Physics in Medicine and Biology*, vol. 61, no. 24, pp. 8676–8698, Dec 2016.

[22] J. Bernal, N. Tajkbaksh, F. J. Sanchez, B. J. Matuszewski, H. Chen, L. Yu, Q. Angermann, O. Romain, B. Rustad, I. Balasingham, K. Pogorelov, S. Choi, Q. Debard, L. Maier-Hein, S. Speidel, D. Stoyanov, P. Brandao, H. Cordova, C. Sanchez-Montes, S. R. Gurudu, G. Fernandez-Esparrach, X. Dray, J. Liang, and A. Histace, "Comparative Validation of Polyp Detection Methods in Video Colonoscopy: Results From the MICCAI 2015 Endoscopic Vision Challenge," *IEEE Transactions on Medical Imaging*, vol. 36, no. 6, pp. 1231–1249, Jun 2017.

[23] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-DenseUNet: Hybrid Densely Connected UNet for Liver and Tumor Segmentation From CT Volumes," *IEEE Transactions on Medical Imaging*, vol. 37, no. 12, pp. 2663–2674, Dec 2018.

[24] M. Z. Alom, C. Yakopcic, M. Hasan, T. M. Taha, and V. K. Asari, "Recurrent residual U-Net for medical image segmentation," *Journal of Medical Imaging*, vol. 6, no. 01, p. 1, Mar 2019.

[25] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation," *IEEE Transactions on Medical Imaging*, pp. 1–1, 2019.

[26] W. Chen, B. Liu, S. Peng, J. Sun, and X. Qiao, "S3D-UNet: Separable 3D U-Net for Brain Tumor Segmentation," 2019, pp. 358–368. [Online]. Available: http://link.springer.com/10.1007/978-3-030-11726-9{_}32

[27] M. Fayyaz, M. H. Saffar, M. Sabokrou, M. Fathy, F. Huang, and R. Klette, "STFCN: Spatio-Temporal Fully Convolutional Neural Network for Semantic Segmentation of Street Scenes," 2017, pp. 493–509.

[28] L. Zhang, L. Lu, X. Wang, R. M. Zhu, M. Bagheri, R. M. Summers, and J. Yao, "Spatio-Temporal Convolutional LSTMs for Tumor Growth Prediction by Learning 4D Longitudinal Patient Data," *IEEE Transactions on Medical Imaging*, vol. 39, no. 4, pp. 1114–1126, Apr 2020.

[29] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," Dec 2014.

[30] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, "On the Properties of Neural Machine Translation: Encoder-Decoder Approaches," Sep 2014.

[31] S. Andermatt, S. Pezold, and P. Cattin, "Multi-dimensional Gated Recurrent Units for the Segmentation of Biomedical 3D-Data," 2016, pp. 142–151.

[32] E. Colleoni, S. Moccia, X. Du, E. De Momi, and D. Stoyanov, "Deep Learning Based Robotic Tool Detection and Articulation Estimation With Spatio-Temporal Layers," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2714–2721, Jul 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8715379/

[33] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Medical Image Analysis*, vol. 53, pp. 197–207, Apr 2019.

[34] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 30, no. 2, pp. 328–341, 2007.

[35] M. Benkhadra *et al.*, "Flexibility of thiel's embalmed cadavers: the explanation is probably in the muscles," *Surgical and Radiologic Anatomy*, vol. 33, no. 4, pp. 365–368, May 2011.

[36] A. More, "Survey of resampling techniques for improving classification performance in unbalanced datasets," Aug 2016.

[37] P. Simard, D. Steinkraus, and J. Platt, "Best practices for convolutional neural networks applied to visual document analysis," in *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.*, vol. 1. IEEE Comput. Soc, 2003, pp. 958–963.

[38] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org.

[39] R. J. Williams and J. Peng, "An Efficient Gradient-Based Algorithm for On-Line Training of Recurrent Network Trajectories," *Neural Computation*, vol. 2, no. 4, pp. 490–501, Dec 1990.

[40] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Proceedings of the 3rd International Conference on Learning representations*, Dec 2015, pp. 1–15.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," Feb 2015. [Online]. Available: http://arxiv.org/abs/1502.01852

[42] S. Semeniuta, A. Severyn, and E. Barth, "Recurrent Dropout without Memory Loss," Mar 2016.

[43] S. A. Taghanaki, Y. Zheng, S. Kevin Zhou, B. Georgescu, P. Sharma, D. Xu, D. Comaniciu, and G. Hamarneh, "Combo loss: Handling input and output imbalance in multi-organ segmentation," *Computerized Medical Imaging and Graphics*, vol. 75, pp. 24–33, Jul 2019.

[44] F. Milletari *et al.*, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Fourth International Conference on 3D Vision (3DV)*, 2016, pp. 565–571.

[45] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Tversky Loss Function for Image Segmentation Using 3D Fully Convolutional Deep Networks," 2017, pp. 379–387.

[46] F. Wilcoxon, S. Katti, and R. A. Wilcox, "Critical values and probability levels for the wilcoxon rank sum test and the wilcoxon signed rank test," *Selected tables in mathematical statistics*, vol. 1, pp. 171–259, 1970.