International Journal of Computer Assisted Radiology and Surgery
https://doi.org/10.1007/s11548-021-02376-3

ORIGINAL ARTICLE

# Using spatial-temporal ensembles of convolutional neural networks for lumen segmentation in ureteroscopy

Jorge F. Lazo[1,2] · Aldo Marzullo[3] · Sara Moccia[4,5] · Michele Catellani[6] · Benoit Rosa[2] · Michel de Mathelin[2] · Elena De Momi[1]

Received: 11 January 2021 / Accepted: 9 April 2021

## Abstract

**Purpose** Ureteroscopy is an efficient endoscopic minimally invasive technique for the diagnosis and treatment of upper tract urothelial carcinoma. During ureteroscopy, the automatic segmentation of the hollow lumen is of primary importance, since it indicates the path that the endoscope should follow. In order to obtain an accurate segmentation of the hollow lumen, this paper presents an automatic method based on convolutional neural networks (CNNs).

**Methods** The proposed method is based on an ensemble of 4 parallel CNNs to simultaneously process single and multi-frame information. Of these, two architectures are taken as core-models, namely U-Net based in residual blocks ($m_1$) and Mask-RCNN ($m_2$), which are fed with single still-frames $I(t)$. The other two models ($M_1$, $M_2$) are modifications of the former ones consisting on the addition of a stage which makes use of 3D convolutions to process temporal information. $M_1$, $M_2$ are fed with triplets of frames ($I(t-1)$, $I(t)$, $I(t+1)$) to produce the segmentation for $I(t)$.

**Results** The proposed method was evaluated using a custom dataset of 11 videos (2673 frames) which were collected and manually annotated from 6 patients. We obtain a Dice similarity coefficient of 0.80, outperforming previous state-of-the-art methods.

**Conclusion** The obtained results show that spatial-temporal information can be effectively exploited by the ensemble model to improve hollow lumen segmentation in ureteroscopic images. The method is effective also in the presence of poor visibility, occasional bleeding, or specular reflections.

**Keywords** Deep learning · Ureteroscopy · Convolutional neural networks · Image segmentation · Upper tract urothelial carcinoma (UTUC)

## Introduction

Upper tract urothelial cancer (UTUC) is a sub-type of urothelial cancer which arises in the renal pelvis and the ureter. The disease has an estimated number of 3,970 patients affected in 2020 [1] in the USA. Flexible ureteroscopy (URS) is nowadays the gold standard for UTUC diagnosis and conservative

This work was supported by the ATLAS Project. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 813782.
B. Rosa was also partly supported by the Investissements d'Avenir program of the French Government, Labex CAMI ANR-11-LABX-0004 and IHU Strasbourg ANR-10-IAHU-02.

✉ Jorge F. Lazo
jorgefrancisco.lazo@polimi.it

[1] Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, Italy

[2] ICube, UMR 7357, CNRS-Université de Strasbourg, Strasbourg, France

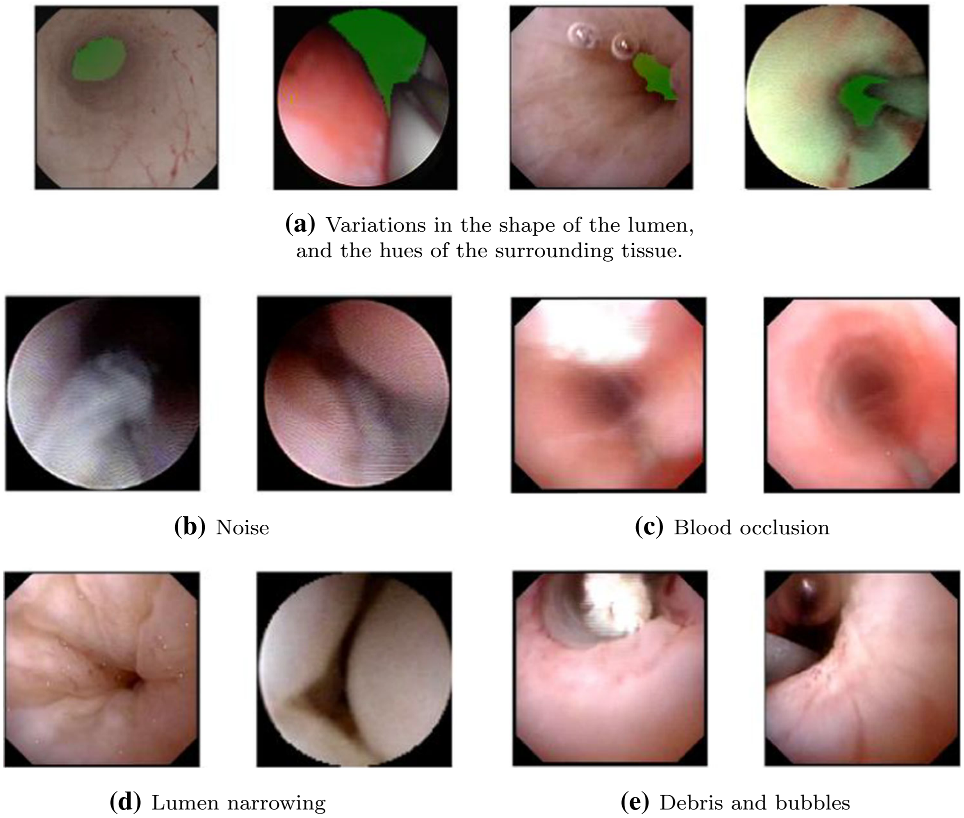
[3] Department of Mathematics and Computer Science, University of Calabria, Rende, CS, Italy

[4] The BioRobotics Institute, Scuola Superiore Sant'Anna, Pisa, Italy

[5] Department of Excellence in Robotics and AI, Scuola Superiore Sant'Anna, Pisa, Italy

[6] Istituto Europeo di Oncologia, Milan, Italy

Published online: 28 April 2021



**(a)** Variations in the shape of the lumen, and the hues of the surrounding tissue.

**(b)** Noise

**(c)** Blood occlusion

**(d)** Lumen narrowing

**(e)** Debris and bubbles

**Fig. 1** Sample images in our dataset showing: **a** the hue variability of the surrounding tissue as well as the shape and location of the lumen (the hollow lumen is highlighted in green to show clearly the variety of shapes in which it could appear). **b**–**e** Samples of artifacts (the lumen was not highlighted to have a clear view of the image artifacts)

treatment. URS is used to inspect the tissue in the urinary system, determine the presence and size of tumor [2] as well as for biopsy of suspicious lesions [3]. The procedure is carried out under the visual guidance of an endoscopic camera [4].

Navigation and diagnosis through the urinary tract are highly dependent upon the operator expertise [5]. For this reason, the current development of methods in computer-assisted interventions (CAI) intends to support surgeons by providing them with relevant information during the procedure [6]. Additionally, within the endeavors of developing new tools for robotic ureteroscopy, a navigation system which relies on image information from the endoscopic camera is also needed [7].

In this study, we focus on the segmentation of the ureter's lumen. In ureter-endoscopic images, the lumen appears most likely as a tunnel or hole in the images with its center being the region with the lowest illuminance inside the field of view (FOV). Lumen segmentation presents some particular challenges such as the difficulty of defining the concrete boundary of it, the narrowing of the ureter around the ureteropelvic junction [4], and the appearance of image artifacts such as blur, occlusions due to the appearance of floating debris or bleeding. Some examples of these, present in our data, are shown in Fig. 1.

In the CAI domain, deep learning (DL)-based methods represent the state of the art for many image processing tasks, including segmentation. In [8], an eight-layer fully convolutional network (FCN) is presented for semantic segmentation of colonoscopy images for different classes, including lumen in the colon, polyps and tools. In [9], a U-Net-like architecture based on residual blocks for lumen segmentation in ureteroscopy images is proposed. However, these DL-based approaches in the field of CAI only use single frames, which dismisses the chance of obtaining extra information from temporal features.

The exploitation of spatial-temporal information has shown to obtain better performance than approaches that only process single frames. In [10], a model based on 3D convolutions is proposed for the task of tool detection and articulation estimation, and in [11], a method for infants limb-pose estimation in intensive care uses 3D convolutions to encode the connectivity in the temporal direction.

Additionally, recent results in different biomedical image segmentation challenges have shown the effectiveness of DL ensemble models, such as in [12] where an ensemble consisting of 4 UNet-like models and one Deeplabv3+ network was proposed obtaining the second place in the 2019 SIIM-ACR pneumothorax challenge, and in [13] where an ensemble which analyzed single-slices data 3D volumetric data separately was presented, obtaining top performance in the HVSMR 3D Cardiovascular MRI in Congenital Heart Disease 2016 challenge dataset.

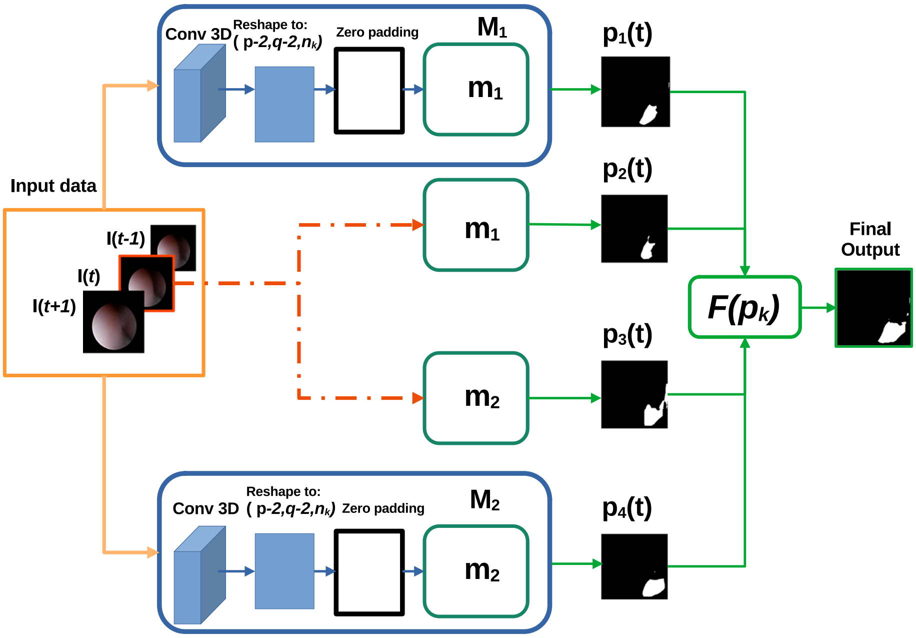


**Fig. 2** Workflow of the proposed ensemble for lumen segmentation in ureteroscopic videos. Blocks of 3 consecutive frames $I(t-1), I(t), I(t+1)$ of size $p \times q \times n_c$ (where $p$ and $q$ refer to the spatial dimensions and $n_c$ to the number of channels of each individual frame) are fed into the ensemble. Models $M_1$ and $M_2$ (orange line) take these blocks as input, whereas models $m_1$ and $m_2$ only take the central frame (red line). Each of the $p_i(t)$ predictions made by each model is ensembled with the function $F(p_k)$ defined in Eq. 1 to perform the final output

Inspired by both paradigms, our research hypothesis is that the use of ensembles which use both single-frame and consecutive-frames information could achieve a better generalization in data than models which uses only one of them. For this purpose, we propose an ensemble model which uses in parallel 4 convolutional neural networks which can exploit the information contained in single-frame and continue-frames, of ureteroscopy videos.

## Proposed method

As introduced in [12,14], we considered the use of ensembles to reach a better generalization of the model when testing it on unseen data. The proposed ensemble of CNNs for ureter's lumen segmentation is depicted in Fig. 2.

Our ensemble is fed with three consecutive frames $[I(t-1), I(t), I(t+1)]$ and produces the segmentation for the frame $I_t$. The ensemble is made of two pairs of branches. One pair (the red one in Fig. 2) consists of U-Net with residual blocks ($m_1$) and Mask-RCNN ($m_2$), which process the central frame $I_t$. The other pair (orange path in Fig. 2) processes the three frames with $M_1$ and $M_2$, which extend $m_1$ and $m_2$ as explained in "Proposed method" Section.

It is important to notice that frames constituting the input for any $M$ are expected to have the minimal possible changes, but still significant to provide extra information which could not be obtained by other means. Some specific examples in our case study include the appearance of debris crossing rapidly the FOV, the sudden appearance or disappearance of some image specularity, a slightly change in the illumination or the position of the element we are interested to segment. For this reason, we consider only three consecutive frames $I_{t-1}, I_t, I_{t+1}$ as input for the model.

The core models $m_1, m_2$ on which our method is based are two state-of-the-art architectures, for instance segmentation:

1. ($m_1$): The *U-Net* implementation used in this work is based on residual units as used in [9], instead of using the classical convolutional blocks, and this is meant to address the degradation as proposed in [15].
2. ($m_2$): Is an implementation of *Mask-RCNN* [16] using ResNet50 as backbone. Mask-RCNN is composed of different stages. The first stage is composed of two networks: a "backbone", which performs the initial classification of the input given a pretrained network, and a region proposal network. The second stage of the model consists of different modules which include a network that predicts the bounding boxes, an object classification network and a FCN which generate the masks for each RoI.

Since our implementation is made of different sets of models, the final output is determined using an ensemble function $F(p_i(t))$ defined as:

$$F(p_i(t)) = \frac{1}{k}\sum_{i}^{k} p_i(t) \tag{1}$$

where $p_i(t)$ corresponds to the prediction of each of the $k = 4$ models for a frame $I(t)$.

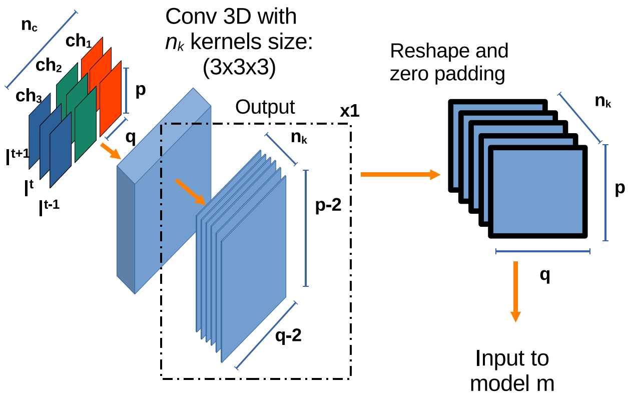


**Fig. 3** The initial stage of the models **M**. The blocks of consecutive frames $I(t-1)$, $I(t)$, $I(t+1)$ of size $p \times q \times n_c$ (where $p$ and $q$ refer to the spatial dimensions and $n_c$ to the number of channels ($ch$) of each individual frame) pass through an initial 3D convolution with $n_k$ number of kernels. The output of this step has a shape of size $(1,\ p-2,\ q-2,\ n_k)$ which is padding with zeros in the second and third dimensions to latter, and then reshaped to fit as input for the **m** core-models

### Extending the core models for handling multi-frame information

For each core model $m$, an extension $M$ is obtained by adapting the architecture for processing multi-frame information.

Let $\mathcal{I}$ be an ordered set of $n$ elements $I \in \mathbb{N}^{p,q,n_c}$ corresponding to frames of a video, where $p$ and $q$ represent spatial dimensions and $n_c$ the number of color channels (Fig. 3). Starting from any core model $(m)$, which takes as input elements from $\mathcal{I}$, we can define another segmentation model $(M)$ which receives multi-frame information from $\mathcal{I}$. Specifically, it receives inputs of the form $I \in \mathbb{N}^{r,p,q,n_c}$, where $r = 3$ represent the temporal dimension (number of frames). To this aim, the core model $m$ is extended by prepending an additional 3D convolution layer with $n_k$ kernels of size $(r \times 3 \times 3)$. The new layer produces an output $H \in \mathbb{N}^{1,p-2,q-2,n_k}$, so that feeding it into $m$ is straightforward. The issue of having $p-2$ and $q-2$ instead of $p$ and $q$ after the 3D convolution is fixed by padding the output with zeros in the two spatial dimensions. A graphical representation of the process is shown in Fig. 3.

## Evaluation

### Dataset

For this study, 11 videos from 6 patients undergoing ureteroscopy procedures were collected. Videos from five patients were used for training the model and tuning hyperparameters. Videos from the remaining patient, randomly chosen, were kept aside and only used for evaluating the performance. The videos were acquired from the European Institute of Oncology (IEO) at Milan, Italy, following the ethical protocol approved by the IEO and in accordance with the Helsinki Declaration.

The number of frames extracted and manually segmented by video is shown in Table 1. Data augmentation was implemented before starting the trainings. The operations used for this purpose were rotations in intervals of 90°, horizontal and vertical flipping and zooming in and out in a range of ± 2% the size of the original image.

### Training setting

All the models were trained, once at time, at minimizing the loss function based on the Dice similarity coefficient ($L_{\text{DSC}}$) defined as:

$$L_{\text{DSC}} = 1 - \frac{2\text{TP}}{2\text{TP} + \text{FN} + \text{FP}} \tag{2}$$

where true positive (TP) is the number of pixels that belong to the lumen, which are correctly segmented, false positive (FP) is the number of pixels miss-classified as lumen, and false negative (FN) is the number of pixels which are classified as part of lumen but actually they are not.

For the case of $(m1)$, the hyperparameters learning rate (lr) and mini batch size (bs) were determined using a fivefold cross-validation strategy with the data from patients 1, 2, 3, 4 and 6 in a grid search. The ranges in which this search was performed were $lr = \{1e{-}3, 1e{-}4, 1e{-}5, 1e{-}6\}$ and $bs = \{4, 8, 16\}$. The $DSC$ was set as the evaluation metric to determine the best model for each of the experiments. Concerning the extensions $M$, the same strategy was used to determine the number of kernels of the input 3D convolutional layer. The remaining hyperparameters were set the same as for $m_1$.

In case of $m_2$, the same fivefold cross-validation strategy was used. The hyperparameters tuned were: the backbone (from the options ResNet50 and ResNet101 [15]) and the value of minimal detection confidence in a range of 0.5–0.9 with differences of 0.1. To cover the range of different sizes of masks in the training and validation dataset, the anchor scales were set to the values of 32, 64, 128 and 160. In this case, the number of filters in the initial 3D convolutional layer was set to a value of 3 which is the only one that could match the predefined input-size, after reshaping, of ResNet backbone.

For each core models and their respective extensions, once the hyperparameters values were chosen, an additional training process was carried out using these values in order to obtain the final model. The training was performed using all the annotated frames obtained from the previously mentioned 5 patients, 60% of the frames were used for training and 40% for validation. The results obtained in this step were the ones

**Table 1** Information about the dataset collected

| Patient no. | Video no. | No. of annotated frames | Image size (pixels) |
|---|---|---|---|
| 1 | Video 1 | 21 | 356 × 256 |
| 1 | Video 2 | 240 | 256 × 266 |
| 2 | Video 3 | 462 | 296 × 277 |
| 2 | Video 4 | 234 | 296 × 277 |
| 3 | Video 5 | 51 | 296 × 277 |
| 4 | Video 6 | 201 | 296 × 277 |
| **5** | **Video 7** | **366** | **256 × 262** |
| 6 | Video 8 | 387 | 256 × 262 |
| 6 | Video 9 | 234 | 256 × 262 |
| 6 | Video 10 | 117 | 256 × 262 |
| 6 | Video 11 | 360 | 256 × 262 |
| Total | – | 2673 | – |

The video marked in bold indicates the patient-case that was used for testing

used to calculate the ensemble results the function defined in Eq. 1.

The networks were implemented using *Tensorflow* and *Keras* frameworks in Python 3.6 trained on a *NVIDIA GeForce RTX 280* GPU.

### Performance metrics

The performance metrics chosen were DSC, precision (Prec) and recall (Rec), defined as:

$$\mathrm{DSC} = 1 - L_{\mathrm{DSC}} \tag{3}$$

$$\mathrm{Prec} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{4}$$

$$\mathrm{Rec} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{5}$$

### Ablation study and comparison with state of the art

First, the performance of the proposed method was compared with the one presented in [9], where the same U-Net based on residual blocks architecture was used. Then, as ablation study, four versions of the ensemble model were tested:

1. ($m_1$,$m_2$): only single-frame information was considered in the ensemble;
2. ($M_1$,$M_2$): only multi-frame information was considered in the ensemble;
3. ($m_1$,$M_1$), ($m_2$,$M_2$): each of the core models and its respective extension were considered in the ensemble, separately.

In these cases, the ensemble function was computed using the values of the predictions of each of the models. The Kruskal–Wallis test on the $DSC$ was used to determine the statistical significance between the different single models tested.

## Results

The box plots of the Prec, Rec and the DSC are shown in Fig. 4. Results of the ablation study are shown in Table 2. The proposed method achieved a DSC value of 0.80 which is 8% better than $m_1$ using single frames ($p < 0.01$) and 3% than $m_2$ trained as well with single frames ($p < 0.05$). When using single-frame information, $m_2$ performs 5% better than $m_1$. However, the result is the opposite using multi-frame information. The ensembles of single-frame models ($m_1$, $m_2$) perform 7% better with respect to ensembles of models exploiting multi-frame information ($M_1$, $M_2$). In the case of spatiotemporal-based models, U-Net based on residual blocks ($M_1$) performs 3% better than the one based on Mask-RCNN ($M_2$). This might be due to the constraint of fitting the output of the 3D convolution into the layers of the backbone of Mask-RCNN. The same limitation might explain the similar behavior when it comes to the comparison of the ensembles composed only of U-Net based in residual blocks models and Mask-RCNN-based models, where the former one performs 4% better than the second one. The only model which achieves a better performance than the proposed one in any metric is U-Net based on residual blocks with the Rec, obtaining a value 0.04 better than the model we proposed. Visual examples of the achieved results are shown in Fig. 5 and in the video attached to this paper. Here, the first 2 rows show frames in which the lumen appears clearly and there is no presence of major image artifacts. As observable, each single model underestimates the ground-truth mask. However, their ensemble gives a better approximation. The next 2 rows show cases in which some kind of occlusions (such as

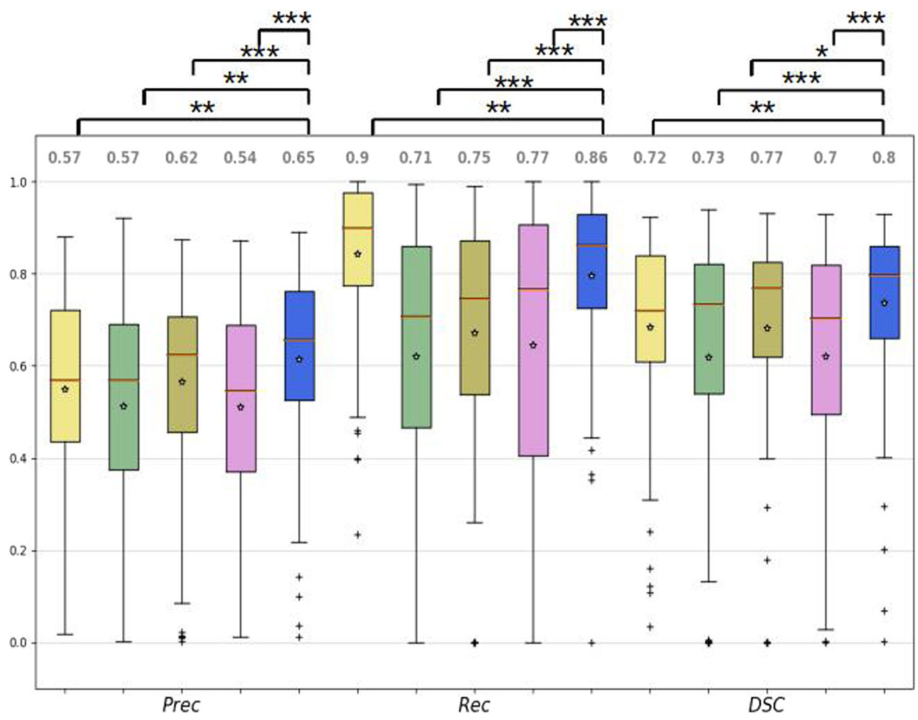


**Fig. 4** Box plots of the precision (Prec), recall (Rec) and the Dice similarity coefficient (DSC) for the models tested. $m_1$ (yellow): ResUNet with single image frames, $m_2$ (green): ResUNet using consecutive temporal frames, $M_1$ (brown): Mask-RCNN with single image frames, $M_2$ (pink): Mask-RCNN using consecutive temporal frames, and the proposed ensemble method (blue) formed by all the previous models. The asterisks represent the significant difference between the different architectures in terms of the Kruskal–Wallis sign rank test ($^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$)

**Table 2** Average Dice similarity coefficient (DSC), precision (Prec) and recall (Rec) in the cases in which the ensembles were formed only by: 1. spatial models $(m_1, m_2)$; 2. spatial-temporal $(M_1, M_2)$, 3. ResUnet with both spatial and temporal inputs $(M_1, m_1)$ and 4. Mask-RCNN with the same setup $(M_2, m_2)$

| $F(*)$ | DSC | Prec | Rec |
|---|---|---|---|
| $(m_1, m_2)$ | 0.78 | 0.65 | 0.71 |
| $(M_1, M_2)$ | 0.71 | 0.55 | 0.57 |
| $(M_1, m_1)$ | 0.72 | 0.56 | 0.66 |
| $(M_2, m_2)$ | 0.68 | 0.51 | 0.63 |

$F(*)$ refers to the ensemble function used (Eq. 1), and the components used to form the ensemble are stated between the parenthesis

blood or debris) is covering most of the FOV. In those cases, single-frame models ($m$) give better results than its counterparts handling temporal information ($M$). Finally, the last 2 rows of the image contain samples showing minor occlusions (such as small pieces of debris crossing the FOV) and images where the lumen is not on focus.

The average inference time was also calculated. Results for $m_1$ and $M_1$ are 26.3±3.7 ms and 31.5±4.7 ms, respectively. In case of $m_2$ and $M_2$, the average inference times are 29.7±2.1 ms and 34.7±6.2 ms, respectively. In the case of the ensemble, the average inference time was 129.6±6.7 ms when running the models consecutively.

## Discussion

The proposed method achieved satisfactory results, outperforming existing approaches for lumen segmentation [9]. Quantitative evaluation, together with a visual inspection of the obtained segmentations, highlights the advantage of using ensembles, confirming our research hypotheses. This is particularly appreciable in the presence of occlusions such as blood or dust covering the FOV (Fig. 5 rows 5–6). In those cases, single-frame-based models tended to include non-lumen regions in the predicted segmentation. An opposite behavior was observed when using only multi-frame-based models, which tended to predict smaller regions with respect to the ground-truth and which is also noticeable in the general performances carried during the ablation studies (Table 2). The ensemble of all of them resulted, instead, in a predicted mask closer to the ground-truth and exemplifies why the use of it in general turns into better performances. It was also observed that the proposed ensemble method was able to correctly manage undesirable false positives appearing in single models. This is due the fact that those false positives did not appear in all the models at the same regions; therefore, the use of ensembles eliminates them from the final result. This is of great importance in the clinical practice, given that false positive classifications during endoluminal inspection might result in a range of complications of the surgical operation, including tools colliding with tissues [17], incorrect path planning [18], among others.

Despite the positive results achieved by the proposed approach, some limitations are worth to be mentioned. Computational time required for inference is one of those. In terms of inference time, the proposed model requires 4 times more than previous implementations. However, it is important to state that when it comes to applications of minimal inva-

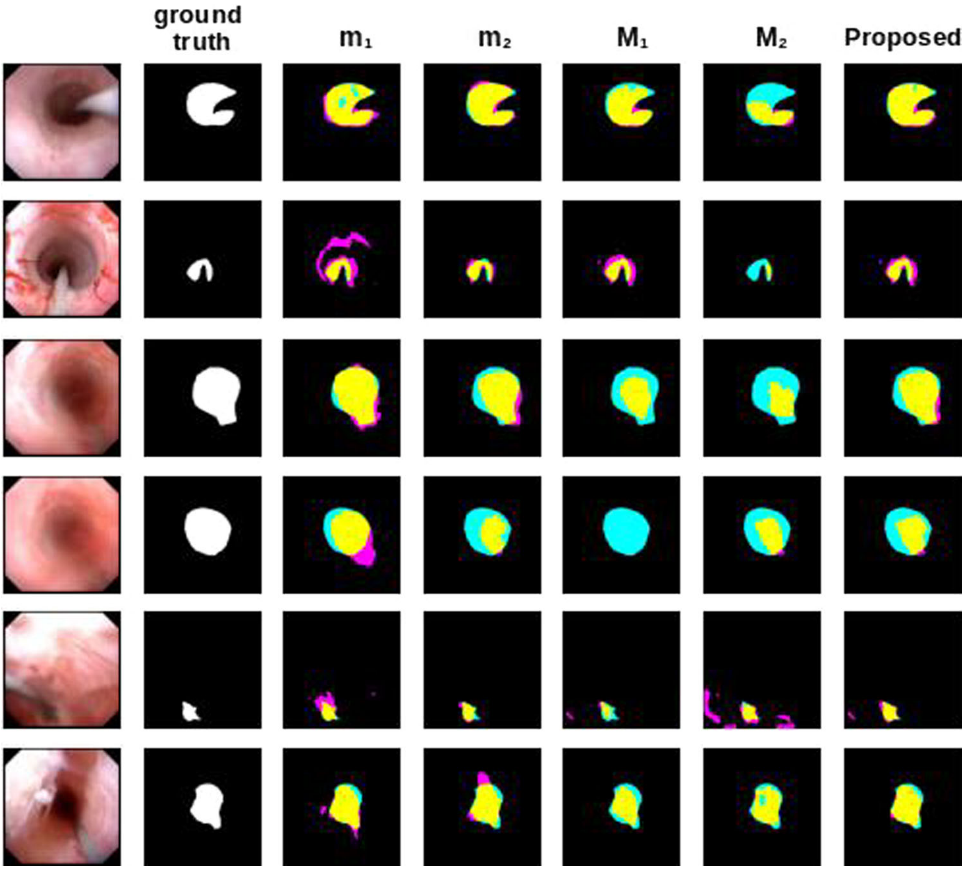


**Fig. 5** Samples of segmentation with the different models test. The colors in the overlay images represent the following for each pixel. True positive (TP): yellow, false positive (FP): pink, false negative (FN): blue, true negative (TN): black. The first two rows depict images where the lumen is clear with the respective segmentation from each model. Rows 3–4 show cases in which some kind of occlusion appears. Finally, the rows 5–6 depict cases in which the lumen is contracted, and/or there is debris crossing the FOV

sive surgery, accuracy may be preferred over speed to avoid any complication, such as perforations of the ureter [5]. Furthermore, such time could be improved by taking advantage of distributed parallel setups. A final issue is related to the scarcity of public available and annotated data, necessary to train and benchmark, which is a well-known problem in the literature. However, this can be overcome in future as new public repositories containing spatial-temporal data are released. Regarding the effectiveness, we consider it as the metric defined for DL systems proposed in [19] which takes into account the product of data quality, robustness and information gain, and we can assert the proposed model is more effective than previous implementations since: (1) the data quality produced with it is better in terms of the mean DSC, Prec and Rec values; (2) the method is more robust against the appearance of artifacts as shown in Fig. 5 and the additional videos attached; and 3) the information gain is higher since the lumen area is delineated better. The disclosed cost-effectiveness of this method for its clinical application such as the one presented in [20] for diabetic retinopathy screening is beyond the scope of this paper. However, a rough estimation should consider 1) the economical cost of the GPU model used to train the networks presented in this work (NVIDIA RTX 2080); 2) the current cost that requires to perform ureteroscopy procedures, according to national health system of each country; and 3) the rate in which this method could reduce complications and thus reduce hospitalization time or the requirement of further interventions.

## Conclusion

In this paper, we introduced a novel ensemble method for ureter's lumen segmentation. Two core models based on U-Net and Mask-RCNN were exploited and extended, in order to capture both single-frame and multi-frame information. Experiments showed that the proposed ensemble method outperforms previous approaches for the same tasks [9], by achieving an increment of 7% in terms of $DSC$. In the future, evident extensions of the present work will be investigated, including better methods to fit spatial-temporal data into models which were pre-trained in single image datasets (such as Mask-RCNN). Furthermore, we will investigate methods for decreasing the inference time, thus allowing real-time applications.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s11548-021-02376-3.

**Acknowledgements** This work was supported by the ATLAS Project. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 813782. B. Rosa was also partly supported by the Investissements d'Avenir program of the French Government, Labex CAMI ANR-11-LABX-0004 and IHU Strasbourg ANR-10-IAHU-02. J.F. Lazo would like to thank Dr. Armando Daniel Lovera Mandujano for all his advice and valuable suggestions to understand better intraluminal organs.

**Funding** Open access funding provided by Politecnico di Milano within the CRUI-CARE Agreement.

## Declarations

**Conflict of interest** The authors declare that they have no conflict of interest.

**Ethical approval** The collection of data was in accordance with the ethical standards of the Istituto Europeo di Oncologia and with the 1964 Helsinki Declaration, revised in 2000. All the subjects involved in this research were informed and agreed to data treatment before the intervention.

**Informed consent** Written informed consent was obtained from all patients included in the study.

## References

1. Siegel RL, Miller KD, Jemal A (2020) Cancer statistics, 2020. CA A Cancer J Clin 70(1):7–30. https://doi.org/10.3322/caac.21601
2. Cosentino M, Palou J, Gaya JM, Breda A, Rodriguez-Faba O, Villavicencio-Mavrich H (2013) Upper urinary tract urothelial cell carcinoma: location as a predictive factor for concomitant bladder carcinoma. World J Urol 31(1):141–145. https://doi.org/10.1007/s00345-012-0877-2
3. Rojas CP, Castle SM, Llanos CA, Cortes JAS, Bird V, Rodriguez S,Reis IM, Zhao W, Gomez-Fernandez C, Leveillee RJL, and Jorda M (2013) Low biopsy volume in ureteroscopy does not affect tumor biopsy grading in upper tract urothelial carcinoma. In: Urologic oncology: seminars and original investigations, vol. 31, Elsevier, pp 1696–1700. https://doi.org/10.1016/j.urolonc.2012.05.010
4. Wason SE, Leslie SW (2020) Ureteroscopy StatPearls. https://pubmed.ncbi.nlm.nih.gov/32809391/. Accessed 29 Nov 2020
5. de la Rosette JJ, Skrekas T, Segura JW (2006) Handling and prevention of complications in stone basketing. Eur Urol 50(5):991–999. https://doi.org/10.1016/j.eururo.2006.02.033
6. Münzer B, Schoeffmann K, Böszörmenyi L (2018) Content-based processing and analysis of endoscopic images and videos: a survey. Multim Tools Appl 77(1):1323–1362. https://doi.org/10.1007/s11042-016-4219-z
7. Borghesan G, Trauzettel F, Ansar MHD, Barata BF, Wu D, Li Z, Lazo JF, Finocchiaro M, Xuan TH, Lai C-F, Ramesh S, Sahu SK, Sestini L, Guiqiu L, Pore A (2020) ATLAS: autonomous intraluminal surgery—system specifications for targeted intraluminal interventions. https://atlas-itn.eu/d102_main/. Accessed 10 Dec 2020
8. Vázquez D, Bernal J, Sánchez FJ, Fernández-Esparrach G, López AM, Romero A, Drozdzal M, Courville A (2017) A benchmark for endoluminal scene segmentation of colonoscopy images. J Healthc Eng. https://doi.org/10.1155/2017/4037190
9. Lazo JF, Marzullo A, Moccia S, Cattellani M, Rosa B, Calimeri F, de Mathelin M, De Momi E(2020) A lumen segmentation method in ureteroscopy images based on a deep residual u-net architecture. In: International conference on pattern recognition (ICPR)
10. Colleoni E, Moccia S, Du X, De Momi E, Stoyanov D (2019) Deep learning based robotic tool detection and articulation estimation with spatio-temporal layers. IEEE Robot Autom Lett 4(3):2714–2721. https://doi.org/10.1109/LRA.2019.2917163
11. Moccia S, Migliorelli L, Carnielli V, Frontoni E (2019) Preterm infants' pose estimation with spatio-temporal features. IEEE Trans Biomed Eng. https://doi.org/10.1109/TBME.2019.2961448
12. Wang X, Yang S, Lan J, Fang Y, He J, Wang M, Zhang J, Han X (2020) Automatic segmentation of pneumothorax in chest radiographs based on a two-stage deep learning method. IEEE Trans Cognit Dev Syst. https://doi.org/10.1109/TCDS.2020.3035572
13. Zheng H, Zhang Y, Yang L, Liang P, Zhao Z, Wang C, Chen DZ (2019) A new ensemble learning framework for 3d biomedical image segmentation. Proc AAAI Conf Artif Intell 33:5909–5916. https://doi.org/10.1609/aaai.v33i01.33015909
14. Vuola AO, Akram SU, Kannala J (2019) Mask-RCNN and U-net ensembled for nuclei segmentation. In: 2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019). IEEE, pp 208–212
15. 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. https://doi.org/10.1109/CVPR.2016.90
16. He K, Gkioxari G, Dollár P, Girshick R (2017) Mask R-CNN. In: Proceedings of the IEEE international conference on computer vision, pp. 2980–2988. https://doi.org/10.1109/ICCV.2017.322
17. He Y, Zhang P, Qi X, Zhao B, Li S, Hu Y (2020) Endoscopic path planning in robot-assisted endoscopic nasal surgery. IEEE Access 8:17039–17048. https://doi.org/10.1109/ACCESS.2020.2967474
18. Alsunaydih FN, Arefin MS, Redoute J-M, Yuce MR (2020) A navigation and pressure monitoring system toward autonomous wireless capsule endoscopy. IEEE Sens J 20(14):8098–8107. https://doi.org/10.1109/JSEN.2020.2979513
19. Blasch E, Liu S, Liu Z, Zheng Y (2018) Deep learning measures of effectiveness. In: NAECON 2018-IEEE national aerospace and electronics conference. IEEE, pp. 254–261
20. Xie Y, Nguyen Q, Bellemo V, Yip MY, Lee XQ, Hamzah H, Lim G, Hsu W, Lee ML, Wang JJ et al (2019) Cost-effectiveness analysis of an artificial intelligence-assisted deep learning system implemented in the national tele-medicine diabetic retinopathy screening in singapore. Invest Ophthalmol Vis Sci 60(9):5471