POLITECNICO
MILANO 1863

A Deep Learning approach for Cystoscopy and Ureteroscopy tumors segmentation

Medical Robotics Laboratory

Prof: **De Momi Elena**
Tutor: **Jorge Francisco Lazo**

Carpi Andrea
Person Code: 10618258
Student Number: 964148
andrea1.carpi@mail.polimi.it

Fati Francesca
Person Code: 10790892
Student Number: 966344
francesca.fati@mail.polimi.it

Karim Kassem
Person Code: 10610471
Student Number: 971063
karim.kassem@mail.polimi.it

Abstract

Ureteroscopy and cystoscopy are considered the gold standards to identify tumors along the urinary tract, yet normal procedures miss up to 20% of lesions. Deep learning augmented cystoscopy/ureteroscopy procedures may improve tumor localization, intraoperative navigation, and surgical resection of tumors, helping early detection which could be translated in better chance of survival for patients [1]. In this work the implementation of ResUnet is studied to segment images from the urinary tract. A total of 2391 images from ureteroscopy and cystoscopy procedures have been manually segmented, 1201 have been excluded due to the poor quality. To overcome the low dimensionality and low variability of the dataset, different augmentation techniques have been exploited, which have demonstrated to increase the stability of the training, and to improve the robustness and the performance of the model. Our approach proposes a classification of the dataset into two classes based on tumors' colors and textural characteristics; two different ResUnet models have been trained specifically on each class and then combined together. In the case of Single ResUnet an Accuracy value of 0.7869 and a Dice coefficient value of 0.4010 have been obtained. Double ResUnet Ensemble, combination of two ResUnet, has achieved better results, with an Accuracy value and a Dice value respectively of 0.8946 and 0.5395. This work opens the opportunity to further investigate new strategies for overcoming issues regarding the datasets available in ureteroscopy and cystoscopy, and consequently for improving the detection of lesions in the urinary tract.

I. Introduction

Urinary tract tumors have become common and comprise different types of lesions ranging from small benign tumors to aggressive neoplasm with high mortality, among which the Bladder Cancer (BC) is the predominant one, and the ninth most common malignancy globally with an estimated 430'000 new diagnoses annually [1].

Standard diagnosis and surveillance of upper tract urothelial UTUC and BC rely on cystoscopy and ureteroscopy exams. During these procedures an endoscope is inserted into the ureter cavity and the imaging procedure is displayed on a monitor being often recorded for post-analysis. New endoscopic techniques, such as narrow-band imaging (NBI) and photodynamic diagnosis (PDD), have been developed to improve the visibility of the tumor, and they

proved to be useful in improving diagnosis and tumor identification [2][3]; however, white-light imaging WLI is still the primary method of observation.
The visual interpretation of clinical endoscopists can be hindered by artifacts mainly due to motion, low contrast, bubbles, bodily fluid, and blood, which are usually confused with lesions leading to unreliable detection of the tumor by the surgeon, which are estimated to be up to over 15% [4].
Identification and complete resection of the tumor reduce the high recurrence rate (75% for BC) and progression in patients who had incomplete initial resection (40%).
To address the above-mentioned limitations making the procedures more independent on the clinician experience, endoscopic imaging-based diagnosis using Convolutional Neural Network (CNN) have been clinically applied in the field of medicine, but this application to urology was initiated only recently [5][6]; this might be due, among the other reasons, to the lack of publicly available, annotated datasets.
As an initial step, in this work ureteroscopy and cystoscopy images have been manually segmented, and a training with ResUnet is performed. To overcome artifacts and poor quality images, which are also affected by a very low variability, a data augmentation procedure based on the library 'Albumentations' [7] is performed, which helps improving the overall performance of the detection. To further boost the algorithm detection capabilities, the dataset has been spilt into two subsets based on similar textural characteristics and colors and a specific ResUnet has been trained for each one. Then, the predictions of the two models have been combined and a 'common convolution ensemble' has been trained over the entire dataset.

II. Methods

A. Dataset Generation

The dataset consists of 24 videos (11 cystoscopy videos and 13 ureteroscopy videos) obtained during the surgical intervention of 24 patients with tumors in the urinary tract. From these videos, 2972 video fragments have been extracted and used for preprocessing. A total of 2391 images have been manually segmented, 1201 images of the original dataset have been excluded due to the poor quality of the image.
The segmentation procedure has been manually performed using an open-source tool developed by Oxford University [8]. All segmented images have been peer-reviewed to ensure the correctness of the segmentation procedure. The segmentation output of the software consists of a series of coordinated polygons. These have been rendered in a mask of the same size as the original image from the dataset and have been converted into a series of black and white pixels which represent the mask of the tumor.
On average, ureteroscopy images show a poor quality with respect to cystoscopy ones, so 6 ureteroscopy and 2 cystoscopy patients video data have been removed from the dataset. The main reason for the poor-quality image is the camera framing, which is usually very close to the tumor because of the tumor-removing procedure, and because of the blurring of the images.

B. Proposed Networks Architecture

The model selected for the study is the ResUnet, a semantic segmentation model inspired by the ResNet and U-Net. This architecture (initially developed for the extraction of road areas) exploits the residual blocks concept to facilitate the propagation of information and reduce the vanishing gradient phenomenon[9]. With ResUNet it is possible to obtain very deep U-Net with fewer parameters. [9]

C. Metrics and Optimization

- The metric used for performance evaluation is the Dice coefficient.

$$Dice = \frac{2 \cdot TP}{FN + (2 \cdot TP) + FP} \tag{1}$$

  - TP = True Positive
  - FP = False Positive
  - FN = False Negative

$$Dice = \frac{2 \cdot AreaOfOverlap}{TotalNumberOfPixels} \tag{2}$$

- The Optimization method used is the Adaptive Moment Estimation (Adam) with fixed $\beta 1$, $\beta 2$ parameters while the ϵ and starting learning rate (η) are optimized for each training.
  The performances of the network are highly related with size and heterogeneity of the dataset, comprising tumors in different configurations and shapes.
- The learning rate is reduced automatically during training whenever the model stop improving in the validation performance.
- The reduction factor is optimized for each model training.

D. Pre-processing

The original images dimensions were 300x300 with 3 channels (RGB). Before introducing the samples into the segmentation network 2 pre-processing steps are applied:

- *Cropping*: to obtain images with dimensions 256x256x3 (width, height, channels)
- *Normalization*: the intensity each pixel is divided by 255 to obtain a value between 0 and 1.

E. Data Augmentation

The performances of the network are highly related with size and heterogeneity of the dataset, comprising tumors in different configurations and shapes. The dataset contains only a small number of patients, and the images are obtained as frames extracted by single ureteroscopy and cystoscopy. For this reason, there is a lot of redundancy

in the images (consecutive frames of the same tumor).
The solution implemented to increase variability and improve the model performances is data augmentation by the use of the library: "Albumentations" [7]. Thanks to a tool written *ad hoc* for the dataset used, it is possible not only to select specific transformations for different classes of tumors, but also to determine the number of augmented images generated for each patient to obtain a balanced dataset.

1) Image samples of the geometrical transformations used to augment the dataset: These transformations are used for increasing the variability of samples associated (same lesion but different spatial positions and orientations)

- Random Rotation
- Random Horizontal/Vertical Flip and Transposition
- Random Crop of the image

Distortions
The goal in this case is to produce (starting from a single lesion) plausible images of lesions of the same type but with different geometries.

- Elastic Transform
- Grid Distortion
- Perspective

Noise introduction

- Static Blur
- Motion Blur: used to simulate rapid movement of the camera in the endoscope

Each transformation is performed with a random weighted probability and multiple operations were applied to the images to obtain greater variability. The distorted transformations are rarely combined together to avoid excessive distortion of the original image.

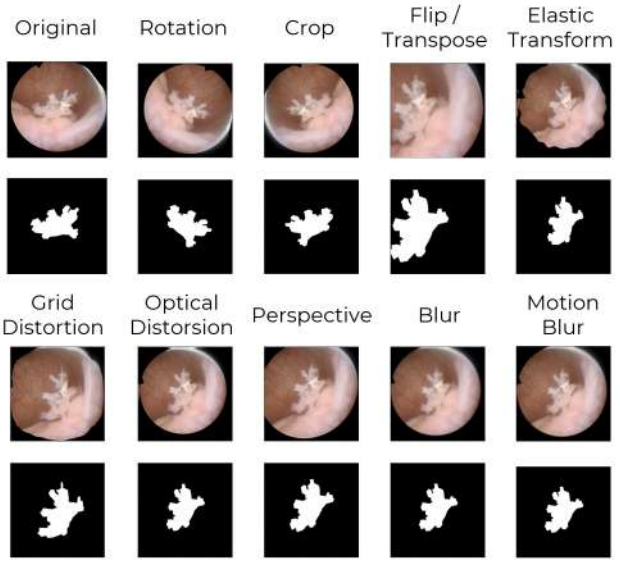

Fig. 1: Image samples of the geometrical transformations used to augment the dataset

2) Colour Transformations: Also, some transformations related to colours are used:

- Random Brightness
- Random Gamma
- Contrast enhancement: to improve the contrast in the images, the "contrast limited adaptive histogram equalization" technique (CLAHE) is used.
  CLAHE parameters: Tile-Size = 8, Clip-Value = 2.0

F. Narrow Band Imaging conversion

In the Dataset some images are acquired with the classical White Light Imaging (WLI) while others use Narrow Band Imaging (NBI). The number of samples belonging to the second class is very small with respect to the WLI class. To tackle this problem, "WLI2NBI" and "NBI2WLI" filters have been introduced to convert images from one style to another and vice versa. The filters are created using a genetic algorithm optimization technique written specifically to solve this problem which is discussed in Appendix A [A].

Fig. 2: On the left the different color filters applied in the augmentations. On the right the filters WLI2NBI and NBI2WLI applied on example images.

G. Model Training

Urs images are very poor in quality and in semantic information. This is mainly due to the presence of noise and close-ups extremely near to the tissue lesion that guarantees little context relative to the surrounding healthy tissue.

A possible solution is to focus only on the Cys cases in order to obtain a robust model as base on which train the Urs cases (transfer learning approach).

1) ***Single Model Approach:*** The initial dataset consisting of all segmented images (both cys and urs) prove to be particularly difficult to train with ResUnet.
As it is possible to observe in the Figure 4, the validation curve is highly unstable (even with small learning rate) and with poor performances.
A posteriori analysis shows that the network is not really able to segment tumor associated with different textures and colors.

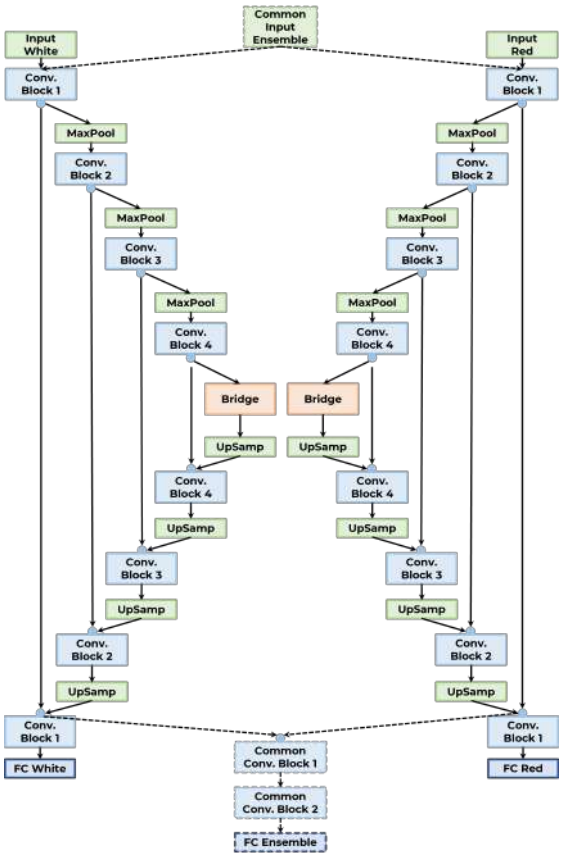

Fig. 3: Ensemble schema

2) ***Ensemble Approach****:* The new approach introduced is to use multiple networks for a tumor specific segmentation:

- The dataset is split into 2 subsets "*White*" and "*Red*" containing tumors with similar textural characteristics and colors.
- Two different ResUnet are trained specifically on each datasets.
- The predictions of the 2 models are then combined.
- The new "Common Convolution Ensemble" is trained again over the entire dataset.

White Tumor Segmentation

The first ResUnet is trained using two patients cases applying augmentation of 2 and 3 times to obtain a balanced set.
The performance has been evaluated also for the network without augmentation.
In the graph below is possible to observe a higher dice score in validation, less overfitting, and a more robust training.

Red Tumor Segmentation

The second ResUnet is trained using two patients which have been augmented 2 and 3 times respectively. Also in this case both the performances and the stability during training improve.

Common Convolution Ensemble

The previous 2 network (Red vs White) are combined to obtain the Ensemble.

- Both the network are cut at the level of the last concatenation (after the residual connection) and 2 common convolution blocks (64 and 32 filters respectively) have been added to combine the 2 predictions.
- The output mask is then computed by the use of a Fully convolutional layer.
- For the Fine Tuning of the ensemble model, the White/Red Backbones are frozen in order to train only the last layers over the entire augmented dataset.

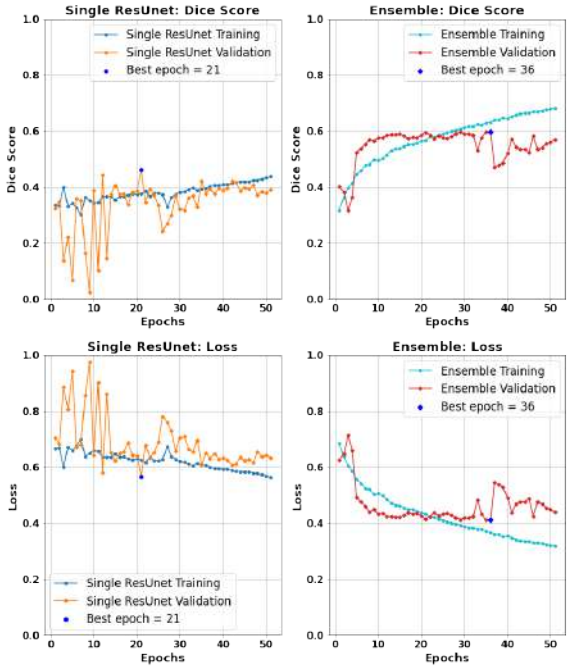

Fig. 4: Training evolution of the single ResUnet compared with the ensemble model on the same dataset.
Single ResUnet training shows an erratic behaviour, even though different learning rate have been proved; while the ensemble model shows an extremely stable behaviour and an increase of 13% in the Dice coefficient both in test and validation.

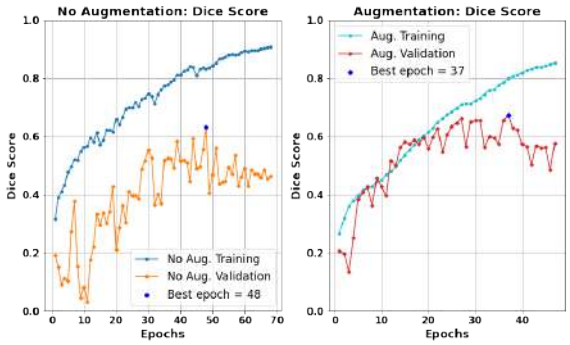

Fig. 5: Training evolution of the ResUnet for White Tumor Segmentation

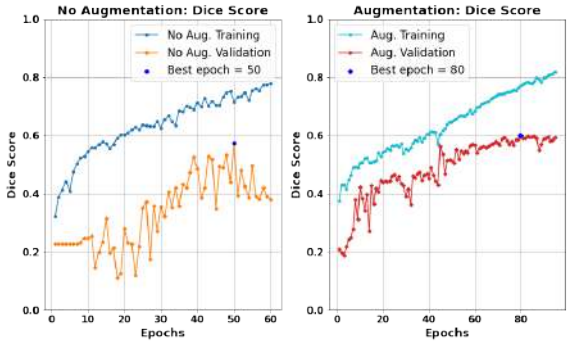

Fig. 6: Training evolution of the ResUnet for Red Tumor Segmentation

III. Results

A. Models Performances

TABLE I: White ResUnet Performances

	Accuracy	Dice
Without Augmentation	0.9197	0.6549
With Augmentation	**0.9457**	**0.6700**

TABLE II: Red ResUnet Performances

	Accuracy	Dice
Without Augmentation	0.8959	0.4587
With Augmentation	**0.9194**	**0.6168**

TABLE III: ResUnet Performances

	Accuracy	Dice
Single ResUnet	0.7869	0.4010
Double ResUnet Ensable	**0.8946**	**0.5395**

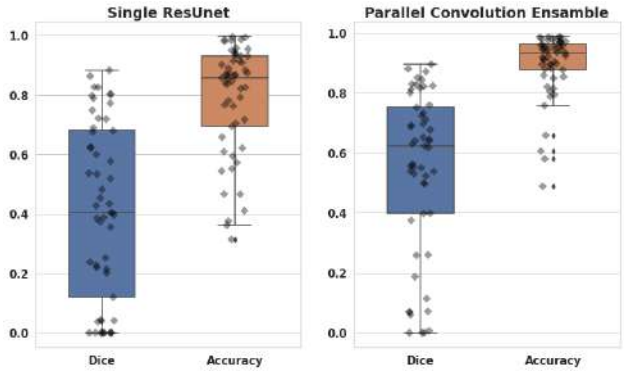

Fig. 7: Comparison between the test results obtained with the single ResUnet vs the Ensemble model.

B. Ensemble Test Predictions

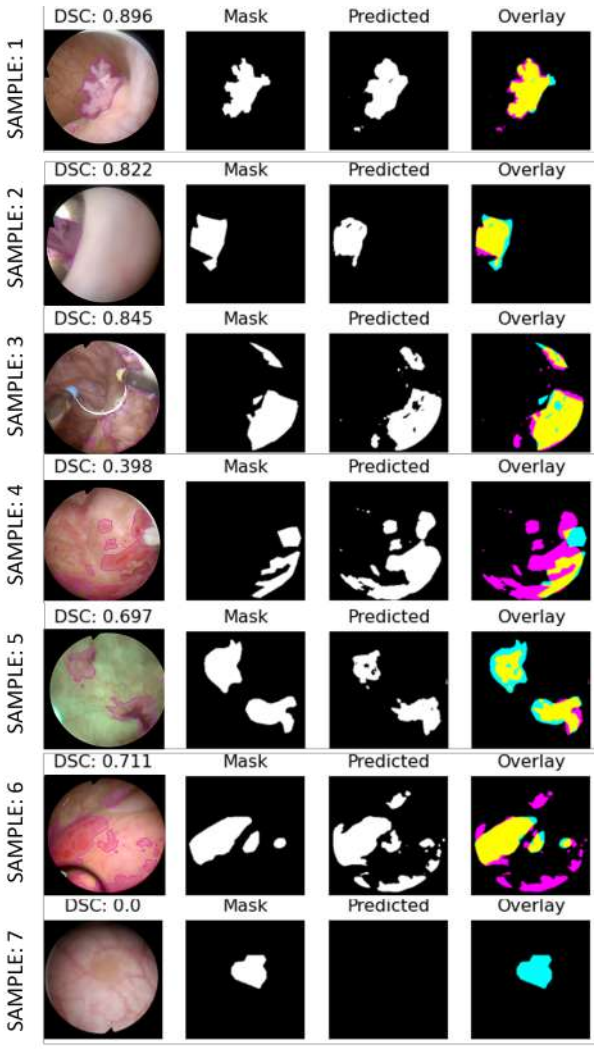

Fig. 8: Test segmented images computed using the Ensemble Model

1) The network manages to precisely identify pailary tumor even in harsh conditions such as the presence of instruments.
2) Network is also able to avoid the erroneous segmentation of surgical instrumentation.
3) In the sample 4 case the prediction is far from the ground-truth (DSC: 0.398) because the model is capable to put in evidence suspicious regions of tissue that are not present in the mask.
4) Nevertheless the model is not sufficiently sensitive to segment lesion with texture similar to the healthy tissue.

IV. Conclusions

The above-described approach proves to be effective in identifying and segmenting images coming from cystoscopy examination procedures, obtaining good results even in photograms affected by artifacts and occlusions, such as blood, air bubbles or instruments.
The algorithm performs very well with white polyp tumors, while the dice score drops with other kinds of lesions, such as the flat reddish ones. This behaviour might be related to the labelling: since the edge of a tumor is not always sharp, especially for red tumors (due to the presence of tissue thickening and redness) it is hard to define a gold standard for generating medical-grade masks. It is worth to notice that the model is able to find new suspicious regions which were not initially considered during manual segmentation, which is a very good clue of the network generalization capability (probably obtained thanks to the data augmentation). Unfortunately in some cases this predictions are penalized in dice coefficient (Sample 4 in Figure 8). In a future application this approach could be extended also to ureteroscopy images, which have been excluded during preprocessing because of the dataset poor quality. The conversion of images from one type of light to another (NBI to WLI and vice versa) likely to be a promising strategy and one possible solution to investigate is the use of a Cycle Coherence GAN (Generative Adversarial Network) for style transfer.[10] The use of GAN for synthetic data generation can be also exploited in case of data scarcity.[11]

Appendix

To convert an image from the WLI to the NBI modality a combination of RBG shift transformation (3 parameters for the colors component tuning), HSV shift (3 parameters for the tone, saturation, luminosity) and CLAHE (2 parameters) for contrast enhancement are adopted.
For fitting whole the parameters, the problem is codified as a list (genome) with length 8.
Two pairs of images with the same semantic content but different styles (NBI vs WLI) are used to fitting the algorithm.
For the identification of the optimal solution the fitness score used is obtained by maximizing the mean correlation of the RGB and HSV colour histograms computed for 4 different sectors of the transformed image and the target one.
Algorithm execution:

1) A population of 100 possible solutions (genomes) is initialized randomly.
2) Each solution is evaluated using the fitness score defined above.
3) New solutions are generated from the original population using a 1-point crossover function with random genomes generation to avoid falling in local minima.
4) The lowest solutions of the population are discarded and the new generation is finished.
5) This process is repeated 20 times (generations) to reach convergence.

After the application of the filter the noise introduced by the CLAHE is removed using a Non-Local Means denoising method.[12]

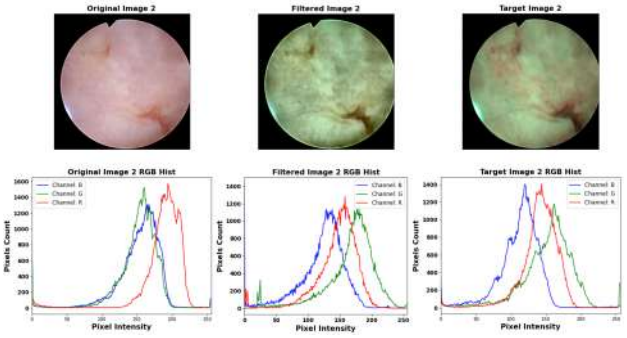

Fig. 9: Example of application of the WLI2NBI filter.

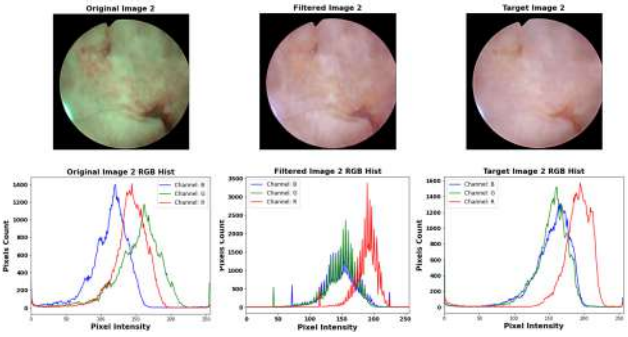

Fig. 10: Example of application of the NBI2WLI filter.

References

[1] Giorgio Santoni, Maria B. Morelli, Consuelo Amantini, and Nicola Battelli. Urinary Markers in Bladder Cancer: An Update. *Frontiers in Oncology*, 8 (SEP):362, 9 2018. ISSN 2234943X. doi: 10.3389/FONC.2018.00362. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6137202/.

[2] Joo Yong Lee, Kang Su Cho, Dong Hyuk Kang, Hae Do Jung, Jong Kyou Kwon, Cheol Kyu Oh, Won Sik Ham, and Young Deuk Choi. A network meta-analysis of therapeutic outcomes after new image technology-assisted transurethral resection for non-muscle invasive bladder cancer: 5-aminolaevulinic acid fluorescence vs hexylaminolevulinate fluorescence vs narrow band imaging. *BMC cancer*, 15(1), 8 2015. ISSN 1471-2407. doi: 10.1186/S12885-015-1571-8. URL https://pubmed.ncbi.nlm.nih.gov/26232037/.

[3] Weiting Kang, Zilian Cui, Qianqian Chen, Dong Zhang, Haiyang Zhang, and Xunbo Jin. Narrow band imaging-assisted transurethral resection reduces the recurrence risk of non-muscle invasive bladder cancer: A systematic review and meta-analysis. *Oncotarget*, 8(14):23880–23890, 2017. ISSN 1949-2553. doi: 10.18632/ONCOTARGET.13054. URL https://pubmed.ncbi.nlm.nih.gov/27823975/.

[4] Jeonghun Lee, Sung Won Park, You Sun Kim, Kyung Jin Lee, Hyun Sung, Pil Hun Song, Won Jae Yoon, and Jeong Seop Moon. Risk factors of missed colorectal lesions after colonoscopy. *Medicine*, 96(27), 7 2017. ISSN 15365964. doi: 10.1097/MD.0000000000007468. URL /pmc/articles/PMC5502189//pmc/articles/PMC5502189/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC5502189/.

[5] Toshiaki Hirasawa, Kazuharu Aoyama, Tetsuya Tanimoto, Soichiro Ishihara, Satoki Shichijo, Tsuyoshi Ozawa, Tatsuya Ohnishi, Mitsuhiro Fujishiro, Keigo Matsuo, Junko Fujisaki, and Tomohiro Tada. Application of artificial intelligence using a convolutional neural network for detecting gastric cancer in endoscopic images. *Gastric cancer : official journal of the International Gastric Cancer Association and the Japanese Gastric Cancer Association*, 21(4): 653–660, 7 2018. ISSN 1436-3305. doi: 10.1007/S10120-018-0793-2. URL https://pubmed.ncbi.nlm.nih.gov/29335825/.

[6] Yoriaki Komeda, Hisashi Handa, Tomohiro Watanabe, Takanobu Nomura, Misaki Kitahashi, Toshiharu Sakurai, Ayana Okamoto, Tomohiro Minami, Masashi Kono, Tadaaki Arizumi, Mamoru Takenaka, Satoru Hagiwara, Shigenaga Matsui, Naoshi Nishida, Hiroshi Kashida, and Masatoshi Kudo. Computer-Aided Diagnosis Based on Convolutional Neural Network System for Colorectal Polyp Classification: Preliminary Experience. *Oncology*, 93 Suppl 1(1):30–34, 12 2017. ISSN 1423-0232. doi: 10.1159/000481227. URL https://pubmed.ncbi.nlm.nih.gov/29258081/.

[7] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. Albumentations: Fast and flexible image augmentations. *Information (Switzerland)*, 11(2), 2 2020. ISSN 20782489. doi: 10.3390/INFO11020125.

[8] VGG Image Annotator. URL https://www.robots.ox.ac.uk/~vgg/software/via/via-2.0.0.html.

[9] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. Road Extraction by Deep Residual U-Net. *IEEE Geoscience and Remote Sensing Letters*, 15(5):749–753, 11 2017. ISSN 15580571. doi: 10.1109/lgrs.2018.2802944. URL https://arxiv.org/abs/1711.10684v1.

[10] Jun Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. *Proceedings of the IEEE International Conference on Computer Vision*, 2017-October:2242–2251, 3 2017. ISSN 15505499. doi: 10.1109/ICCV.2017.244. URL https://arxiv.org/abs/1703.10593v7.

[11] Veit Sandfort, Ke Yan, Perry J. Pickhardt, and Ronald M. Summers. Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks. *Scientific Reports 2019 9:1*, 9(1):1–9, 11 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-52737-x. URL https://www.nature.com/articles/s41598-019-52737-x.

[12] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel. Non-Local Means Denoising. *Image Processing On Line*, 1:208–212, 9 2011. ISSN 2105-1232. doi: 10.5201/IPOL.2011.BCM{_}NLM. URL https://www.ipol.im/pub/art/2011/bcm_nlm/.